RICE UNIVERSITY

Nonnormality in Lyapunov Equations

by

Jonathan Baker

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Arts

Approved, Thesis Committee:

Danny Sorensen, Chairman
Noah Harding Professor of
Computational and Applied Mathematics,
Rice University

Matthias Heinkenschloss
Professor of Computational and
Applied Mathematics,
Rice University

Adrianna Gillman
Assistant Professor of
Computational and Applied Mathematics,
Rice University

Mark Embree, Director
Professor of Mathematics,
Virginia Polytechnic Institute and State
University

Houston, Texas

December, 2015

ABSTRACT

Nonnormality in Lyapunov Equations

by

Jonathan Baker

The singular values of the solution to a Lyapunov equation determine the potential accuracy of the low-rank approximations constructed by iterative methods. Low-rank solutions are more accurate if most of the singular values are small, so a-priori bounds that describe coefficient matrix properties that correspond to rapid singular value decay are valuable. Previous bounds take similar forms, all of which weaken (quadratically) as the coefficient matrix departs from normality. Such bounds suggest that the farther from normal the coefficient matrix is, the slower the singular values of the solution will decay. This predicted slowing of decay is manifest in the ADI algorithm for solving Lyapunov equations, which converges more slowly when the coefficient is far from normal. However, simple examples typically exhibit an eventual acceleration of decay if the coefficient departs sufficiently from normality. This thesis shows that the decay acceleration principle is universal: decay *always* improves as departure from normality increases beyond a given threshold, specifically, as the numerical range of the coefficient matrix extends farther into the right half-plane. The concluding chapter gives examples showing that similar behavior can occur for general Sylvester equations, though the right-hand side plays a more important role.

Contents

Illustrations

Chapter 1

Introduction

Many physical phenomena can be modelled by linear time-invariant (LTI) control systems, which can be expressed in the state space control form

$$\begin{aligned} \dot{x}(t) &= \mathbf{A}x(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}x(t) + \mathbf{D}u(t). \end{aligned} \tag{1.1}$$

Here $x(t) \in \mathbb{R}^n$ is the system state, $u(t) \in \mathbb{R}^p$ is the control input vector, and $y(t) \in \mathbb{R}^q$ is the output at time t. Single input, single output (SISO) systems can be represented in this form with $p = q = 1$, u and y scalar-valued functions, and $\mathbf{B}$ and $\mathbf{C}$ column and row vectors. Controlling the system (1.1) consists of choosing u so that the solution x and input u satisfy some condition, such as driving x to a desired end state with minimal energy input $\|u\|_{L_2}$.

The system (1.1) is said to be (BIBO) stable if $y(t)$ is bounded for all choices of bounded u (bounded input, bounded output). The system is BIBO stable if $\mathbf{A}$ is stable, i.e., the eigenvalues of $\mathbf{A}$ have strictly negative real parts, $\sigma(\mathbf{A}) \subset \mathbb{C}_-$. The system (1.1) or the pair $(\mathbf{A}, \mathbf{B})$ is said to be (state) controllable if for every pair of states $(x_0,\ x_1)$ and time interval $[0, T]$ there is a choice of control function u such that $x(T) = x_1$ when control u is applied starting at $x(0) = x_0$. This condition is equivalent to the controllability matrix $[\mathbf{B} \quad \mathbf{AB} \quad \mathbf{A}^2\mathbf{B} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{B}]$ having full row rank [1, Section 4.2.1].

Simulation (solving the system to determine x for a given u) and controller design (choosing u to satisfy the conditions of interest) are computationally expensive for

large systems, such as those that come from spatially discretized differential equations. Model reduction methods construct systems with the same form as (1.1) such that the output of the reduced system approximates the output of the original system accurately, but with a much lower-dimensional state space.

"Balanced truncation" is an important family of methods of model reduction that relies on computing a pair of matrices called the system Gramians. The (infinite time) controllability (or reachability) Gramian is the matrix

$$\mathcal{P} = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^* t} dt, \tag{1.2}$$

which also solves the continuous time Lyapunov equation

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* = -\mathbf{B}\mathbf{B}^*. \tag{1.3}$$

Similarly, the observability Gramian of (1.1) is

$$\mathcal{Q} = \int_0^\infty e^{\mathbf{A}^* t}\mathbf{C}^*\mathbf{C} e^{\mathbf{A}t} dt, \tag{1.4}$$

which satisfies

$$\mathcal{Q}\mathbf{A} + \mathbf{A}^*\mathcal{Q} = -\mathbf{C}^*\mathbf{C}. \tag{1.5}$$

The Lyapunov equation in the form (1.3) is of primary interest for this work, but it will be useful to lay out a few properties of the more general case,

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}^* = \mathbf{G}. \tag{1.6}$$

In (1.6), if $\mathbf{A}$ is stable, then the solution $\mathbf{X}$ exists and is unique. By adding the assumption that $\mathbf{G}$ is Hermitian and negative *semi*definite, $\mathbf{X}$ is guaranteed to be Hermitian and positive *semi*definite. In the Gramian Lyapunov equation (1.3), if $\mathbf{A}$ is stable, and $(\mathbf{A}, \mathbf{B})$ is controllable, then the solution is (strictly) positive definite.

From here on, it is assumed that $\mathbf{A}$ is stable and $(\mathbf{A}, \mathbf{B})$ is controllable because of these useful results and because this is the most important case in control theory applications.

In their simplest form, balanced truncation methods need to compute $n \times n$ Gramian matrices as a first step. However, $\mathcal{P}$ and $\mathcal{Q}$ are typically dense, even if $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are sparse. For large n, computing or just storing such $n \times n$ dense matrices may be impossible. Fortunately, Gramians can often be accurately represented by a low rank factor (say $\mathcal{P} \approx \mathbf{Z}\mathbf{Z}^*$, where $\mathbf{Z}$ is $n \times k$ with $k \ll n$). When balanced truncation is performed using an approximate factored Gramian ($\mathbf{Z}\mathbf{Z}^*$ in place of $\mathcal{P}$), the result is "approximate balanced truncation." The singular values of $\mathcal{P}$ and $\mathcal{Q}$ determine how closely they may be approximated by such factorizations. Specifically,

$$\min_{\mathbf{Z} \in \mathbb{C}^{n \times k}} \frac{\|\mathbf{Z}\mathbf{Z}^* - \mathcal{P}\|_2}{\|\mathcal{P}\|_2} = \varsigma_{k+1}(\mathcal{P}),$$

where $\varsigma_k(\mathcal{P})$ is the kth largest singular value of $\mathcal{P}$. Thus, estimating the computational complexity of approximate balanced truncation requires a-priori estimates of the singular values of the system Gramians.

A secondary interest in the singular values of Gramians arises from the fact that some systems are more reducible than others. The "Hankel singular values" of (1.1) measure the error between the original and reduced systems. To be more precise, the Hankel singular values $\sigma_1 \geq \cdots \geq \sigma_n$ of (1.1) are the singular values of the input-output map or "Hankel operator"

$$y(t) = \mathcal{H}(u)(t) := \int_{-\infty}^{0} \mathbf{C} e^{\mathbf{A}(t-\tau)} \mathbf{B} u(\tau) d\tau. \tag{1.7}$$

The Hankel singular values reveal the best possible accuracy of low-rank approximations by

$$\min_{\operatorname{rank}(\hat{\mathcal{H}}) \leq k} \frac{\|\hat{\mathcal{H}} - \mathcal{H}\|_{L_2}}{\|\mathcal{H}\|_{L_2}} = \sigma_{k+1},$$

see also [1, Thm. 7.9]. It happens that the Hankel singular values are also the square roots of the singular values of the product of the system Gramians, $\mathcal{P}\mathcal{Q}$. If the singular values of $\mathcal{P}$, $\mathcal{Q}$, or both are small, then the Hankel singular values may be small (depending on the alignment of the eigenspaces of $\mathcal{P}$ and $\mathcal{Q}$), and the system may be reducible. Even in the worst case, the Hankel singular values are bounded by

$$\sigma_k = \sqrt{\varsigma_k(\mathcal{P}\mathcal{Q})} \leq \min\Big\{\sqrt{\|\mathcal{P}\|_2 \varsigma_k(\mathcal{Q})}, \sqrt{\|\mathcal{Q}\|_2 \varsigma_k(\mathcal{P})}\Big\}.$$

Most of the rest of this work will not need any properties of the system Gramians $\mathcal{P}$, $\mathcal{Q}$ other than satisfying (1.3), so the generic solution variable $\mathbf{X}$ will be used. A matrix $\mathbf{G}$ will denote a general right-hand side as in (1.6), while $\mathbf{B}$ will be used when only the factored case (1.3) is being considered. The singular values of a matrix $\mathbf{M}$ will be written as $\varsigma_1(\mathbf{M}) \geq \cdots \geq \varsigma_n(\mathbf{M})$. If the eigenvalues of $\mathbf{M}$ are real (in particular, if $\mathbf{M}$ is Hermitian), its eigenvalues will be written $\lambda_1(\mathbf{M}) \geq \cdots \geq \lambda_n(\mathbf{M})$. Because the singular values of $\mathbf{X}$ appear so often, they will be written $s_k := \varsigma_k(\mathbf{X})$.

The expressions "$\mathbf{X}$ exhibits fast singular value decay" and "the singular values of $\mathbf{X}$ decay quickly" mean most of the singular values of $\mathbf{X}$ are small compared to $s_1 = \|\mathbf{X}\|$. Fast singular value decay is necessary for fast convergence of low-rank algorithms, and bounds for singular values (or "decay bounds") are estimates of the best possible convergence rate. Section 1.2 describes several methods of obtaining upper bounds on the singular values of $\mathbf{X}$, but it will be shown that these bounds can be very pessimistic, particularly when $\mathbf{A}$ is far from normal. Developing an alternative bound that exploits properties of nonnormal matrices is the main goal of this work.

A few authors have previously noted that singular value decay bounds tend to be especially pessimistic when $\mathbf{A}$ is not normal, e.g. [20, Sec. 4.1]. Sabino is one of the few authors to provide more than experimental results in this area [19], but Lyapunov equations in general have received considerable attention. The rest of this chapter

explains this work's relationship to some of the most important ideas in the Lyapunov equation literature. The first section of this chapter is an introduction to techniques for solving (1.6), but it is not a complete summary of the field; the overviews of [19, 20] are more exhaustive. The second section discusses earlier theoretical bounds—as functions of $\mathbf{A}$ and $\mathbf{G}$—for the singular values of the solution $\mathbf{X}$. This chapter concludes with a summary of Sabino's investigation of the role of nonnormality.

1.1 Solution Methods

For very small n, it may be reasonable to solve (1.6) via the equivalent Kronecker product form

$$(\mathbf{I} \otimes \mathbf{A} + \overline{\mathbf{A}} \otimes \mathbf{I})\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{G}), \tag{1.8}$$

where vec stacks the columns of a matrix into a single vector: $\mathrm{vec}([\, y_1 \quad y_2 \quad \cdots \quad y_n \,]) = [\, y_1^T \quad y_2^T \quad \cdots \quad y_n^T \,]^T$. However, (1.8) is a system of n^2 equations. For large n, this is much too difficult to solve with general purpose linear solvers, notwithstanding the system's sparsity.

The Bartels–Stewart algorithm [3] transforms (1.6) by Schur factorization to a basis in which $\mathbf{A}$ is triangular. It is a simple matter to back-solve the transformed equation column-wise and then return the answer to the original basis. Unfortunately, because columns of the transformed solution are found independently from each other, this method cannot promise a numerically symmetric solution. The variant algorithm derived by Hammarling [11] has the advantage of enforcing both the symmetry and positive definiteness of $\mathbf{X}$ by constructing it from its Cholesky factorization. However, this method still requires the Schur factorization of $\mathbf{A}$, and even this step may be prohibitively expensive for large problems. For large problems, even explicit storage of the dense $n \times n$ solution (or its upper triangular Cholesky factor) is not feasible.

These are problems common among direct methods, which have a fixed, size-based cost.

Alternatively, iterative methods offer sequences of increasingly accurate solutions, and the cost may be controlled based on required accuracy and the user's resources. The iterative method Smith provided in [21] is the basis for many other algorithms. Given a scalar parameter $q \in \mathbb{C}$ with $\mathrm{Re}(q) < 0$, Smith's method computes a portion of an infinite series expression for $\mathbf{X}$. Let

$$\mathbf{A}_q := (\mathbf{A} + q\mathbf{I})^{-1}(\mathbf{A} - \overline{q}\mathbf{I}) \tag{1.9a}$$

$$\text{and} \quad \mathbf{G}_q := 2\mathrm{Re}(q)(\mathbf{A} + q\mathbf{I})^{-1}\mathbf{G}(\mathbf{A}^* + \overline{q}\mathbf{I})^{-1}. \tag{1.9b}$$

Then

$$\mathbf{X} = \sum_{k=0}^{\infty} (\mathbf{A}_q)^k \mathbf{G}_q ((\mathbf{A}_q)^k)^*. \tag{1.10}$$

Smith shows that (1.10) must converge, so truncating (1.10) provides a viable approximation to $\mathbf{X}$. Successive approximate solutions can then be expressed with the simple recurrence

$$\mathbf{X}_k = \mathbf{A}_q \mathbf{X}_{k-1} \mathbf{A}_q^* + \mathbf{G}_q. \tag{1.11}$$

(To speed up convergence, Smith also suggested an iterative matrix-squaring procedure that is less important for this discussion.)

The more versatile alternating direction implicit (ADI) iteration, first applied to (1.6) by Wachspress in [25], can be considered a generalization of Smith's method where a different parameter q_k, $\mathrm{Re}\,(q_k) < 0$, may be chosen at each step. This produces a modified recurrence

$$\mathbf{X}_k = \mathbf{A}_{q_k}^* \mathbf{X}_{k-1} \mathbf{A}_{q_k} + \mathbf{G}_{q_k}. \tag{1.12}$$

In the forms just given, (1.11) and (1.12) are impractical because they are dense updates. However, Penzl in [16] showed that the Smith and ADI methods can be

modified to produce a sequence of approximate solutions of increasing rank,

$$\mathbf{X}_k = \mathbf{Z}_k \mathbf{Z}_k^*, \tag{1.13}$$

where $\mathbf{Z}_k$ is a $n \times kp$ matrix ($\mathbf{G} = -\mathbf{B}\mathbf{B}^*$ has rank p) with the recursive formula comparable to (1.12):

$$\begin{aligned} \mathbf{Z}_1 &= \sqrt{-2\mathrm{Re}\,(q_1)}(\mathbf{A} + q_1\mathbf{I})^{-1}\mathbf{B} \\ \mathbf{Z}_k &= \left[\begin{array}{c|c} (\mathbf{A} - q_k\mathbf{I})(\mathbf{A} + q_k\mathbf{I})\mathbf{Z}_{k-1} & \sqrt{-2\mathrm{Re}\,(q_k)}(\mathbf{A} + q_k\mathbf{I})^{-1}\mathbf{B} \end{array} \right]. \end{aligned} \tag{1.14}$$

This is an attractive method since it does not require storing and updating a large $\mathbf{X}$ explicitly.

The accuracy of such low-rank approximations is related to the relative sizes of the singular values of $\mathbf{X}$. When most of the singular values of $\mathbf{X}$ are relatively small, an accurate low-rank approximation exists. The Eckart-Young theorem shows that

$$\min_{\mathrm{rank}\,\mathbf{Y} \le kp} \frac{\|\mathbf{X} - \mathbf{Y}\|_2}{\|\mathbf{X}\|_2} = \frac{s_{kp+1}}{s_1}. \tag{1.15}$$

Specifically, if the singular value decomposition of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_n \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^* \tag{1.16}$$

with $\mathbf{U}$, $\mathbf{V}$ unitary, then the optimal $\mathbf{Y}$ in (1.15) is

$$\mathbf{Y} = \begin{bmatrix} u_1 & \cdots & u_{kp} \end{bmatrix} \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_{kp} \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_{kp} \end{bmatrix}^*. \tag{1.17}$$

In the present context, $\mathbf{X}$ is symmetric, so $\mathbf{U} = \mathbf{V}$.

The goal of this work is finding a-priori bounds for the singular value decay (1.15). Such bounds give insight into the best performance that is achieveable by *any* iterative method that constructs low-rank approximations to $\mathbf{X}$.

1.2 Singular Value Decay Bounds

This section highlights the methods of other authors for bounding s_k. Existing singular value decay bounds depend monotonically on the departure of $\mathbf{A}$ from normality, allowing slower decay the farther $\mathbf{A}$ is from normal.

One of the most interesting results about decay bounds in general comes from Penzl [17] and is derived differently by Sabino [19]. They found that the spectra of $\mathbf{A}$ and $\mathbf{X}$ are independent, in that any spectrum of $\mathbf{A}$ can correspond to any decay of the singular values of $\mathbf{X}$. Consequently no bound for the singular values of $\mathbf{X}$ can simply be a function of the spectrum of $\mathbf{A}$ alone.

The bounds in this section and the new bound in Chapter 3 depend on the departure of $\mathbf{A}$ from normality in various ways, but the latter has the advantage of not depending directly on $\sigma(\mathbf{A})$, the spectrum of $\mathbf{A}$. So to make a fair comparison, $\sigma(\mathbf{A})$ must remain constant when comparing these bounds for varying $\mathbf{A}$ and $\mathbf{X}$ values.

1.2.1 A Diagonalization Bound

Most of the existing singular value decay bounds are derived from the low-rank ADI method that constructs an approximate solution $\mathbf{X}_k = \mathbf{Z}_k\mathbf{Z}_k^*$ with rank at most kp. These iterates satisfy

$$\mathbf{X} - \mathbf{X}_k = \phi_k(\mathbf{A})\mathbf{X}\phi_k(\mathbf{A}^*) \tag{1.18}$$

where

$$\phi_k(z) := \prod_{j=1}^{k} \frac{q_j - z}{\overline{q}_j + z}, \tag{1.19}$$

and $\{q_k\}$ are the shift parameters from (1.12). For any particular choice of $\{q_k\}$,

$$\frac{s_{kp+1}}{s_1} = \min_{\operatorname{rank} \mathbf{Y} \leq kp} \frac{\|\mathbf{X} - \mathbf{Y}\|_2}{\|\mathbf{X}\|_2} \tag{1.20a}$$

$$\leq \frac{\|\mathbf{X} - \mathbf{X}_k\|_2}{\|\mathbf{X}\|_2} \tag{1.20b}$$

$$\leq \|\phi_k(\mathbf{A})\| \|\phi_k(\mathbf{A}^*)\| \tag{1.20c}$$

$$= \|\phi_k(\mathbf{A})\|^2, \tag{1.20d}$$

so bounding $\|\phi_k(\mathbf{A})\|$ is a useful way of bounding both ADI approximation error and singular value decay.

For diagonalizable $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, Sorensen and Zhou [22] first found that

$$\frac{s_{kp+1}}{s_1} \leq \frac{\|\mathbf{X} - \mathbf{X}_k\|_2}{\|\mathbf{X}\|_2} \leq \|\phi_k(\mathbf{A})\|^2 \leq \kappa(\mathbf{V})^2 \max_{z \in \sigma(\mathbf{A})} |\phi_k(z)|^2 \tag{1.21}$$

where $\kappa(\mathbf{V}) = \|\mathbf{V}\|_2 \|\mathbf{V}^{-1}\|_2$ is the condition number of $\mathbf{V}$. The bound (1.21) generalizes Penzl's result in [17], which applied only to the Hermitian case $\mathbf{A} = \mathbf{A}^*$ (for which $\kappa(\mathbf{V}) = 1$).

Because $\kappa(\mathbf{V})$ is a measurement of the departure of $\mathbf{A}$ from normality, the above bound suggests monotonically slower singular value decay for $\mathbf{X}$ as $\mathbf{A}$ departs from normality.

1.2.2 A Numerical Range Bound

Now consider another method of bounding $\|\phi_k(\mathbf{A})\|$, which also bounds the singular values of $\mathbf{X}$ and the convergence rate of ADI via (1.20). For some choices of $\{q_k\}$, it is possible to use Crouzeix's bound for analytic functions [7] on the numerical range of $\mathbf{A}$ (or field of values),

$$W(\mathbf{A}) := \left\{ \frac{x^* \mathbf{A} x}{x^* x} : x \in \mathbb{C}^n \right\}. \tag{1.22}$$

If the shifts $\{q_k\}$ lie outside $W(-\mathbf{A})$ (in particular, if $W(\mathbf{A}) \subset \mathbb{C}_-$), then ϕ_k is analytic on $W(\mathbf{A})$ and

$$\|\phi_k(\mathbf{A})\| \leq \gamma \sup_{z \in W(\mathbf{A})} |\phi_k(z)| \tag{1.23}$$

where γ is defined as the smallest constant such that (1.23) holds for all $\mathbf{A}$ and appropriate $\{q_k\}$. The exact value of γ is unknown, but Crouzeix proved that

$$2 \leq \gamma \leq 11.08.$$

In particular, no example has been found for which (1.23) does not hold with $\gamma = 2$.

Departure from normality typically causes $W(\mathbf{A})$ to expand, so (1.23) is another bound that is weakened by nonnormality. In fact, if $\mathbf{A}$ is sufficiently far from normal, $W(\mathbf{A})$ may extend into the open right half-plane $\mathbb{C}_+$ (even though $\sigma(\mathbf{A}) \subset \mathbb{C}_-$). This problematic case $W(\mathbf{A}) \nsubseteq \mathbb{C}_-$ can also be characterized by $\omega(\mathbf{A}) \geq 0$, where $\omega(\mathbf{A})$ is the numerical abscissa of $\mathbf{A}$ or the rightmost extend of $W(\mathbf{A})$:

$$\omega(\mathbf{A}) := \max \mathrm{Re}\,(W(\mathbf{A})). \tag{1.24}$$

Note that the numerical abscissa is also the largest eigenvalue of the Hermitian part of $\mathbf{A}$, $H(\mathbf{A}) := (\mathbf{A} + \mathbf{A}^*)/2$. That is,

$$\omega(\mathbf{A}) = \lambda_1(H(\mathbf{A})). \tag{1.25}$$

Remark 1.1. For any $z \in \mathbb{C}_+$ and $q \in \mathbb{C}_-$, it follows that $|q - z|/|\overline{q} + z| > 1$. Thus, if $\omega(\mathbf{A}) > 0$, then there exists a point $z \in W(\mathbf{A}) \cap \mathbb{C}_+$ that makes each term in the product (1.19) greater than 1 for any choice of shifts, so $\sup_{z \in W(\mathbf{A})} |\phi_k(z)| > 1$. Consequently, when $\omega(\mathbf{A}) > 0$, (1.23) provides only the vacuous bound on the singular values of $\mathbf{X}$ and the convergence of ADI

$$\frac{s_{kp+1}}{s_1} \leq \frac{\|\mathbf{X} - \mathbf{X}_k\|_2}{\|\mathbf{X}\|_2} \leq 1 < \|\phi_k(\mathbf{A})\|^2 \leq \left(\gamma \sup_{z \in W(\mathbf{A})} |\phi_k(z)| \right)^2. \tag{1.26}$$

Because of the assumption $\sigma(\mathbf{A}) \subset \mathbb{C}_-$, the diagonalization bound (1.21) will not become vacuous via the maximization term, although it can be greater than 1 for sufficiently large $\kappa(\mathbf{V})$.

1.2.3 A Pseudospectral Bound

Sabino [19] derived another bound for $\|\phi_k(\mathbf{A})\|$ based on the pseudospectra of $\mathbf{A}$. The ε-pseudospectrum of $\mathbf{A}$ is defined as

$$\sigma_\varepsilon(\mathbf{A}) := \left\{ z \in \mathbb{C} : z \in \sigma(\mathbf{A}) \text{ or } \left\| (\mathbf{A} - z\mathbf{I})^{-1} \right\| > 1/\varepsilon \right\}. \tag{1.27}$$

If Γ_ε is the boundary of $\sigma_\varepsilon(\mathbf{A})$, then

$$\phi_k(\mathbf{A}) = \frac{1}{2\pi i} \int_{\Gamma_\varepsilon} \phi_k(z)(\mathbf{A} - z\mathbf{I})^{-1} dz \tag{1.28}$$

and

$$\|\phi_k(\mathbf{A})\| \leq \frac{|\Gamma_\varepsilon|}{2\pi\varepsilon} \sup_{z \in \sigma_\varepsilon(\mathbf{A})} |\phi_k(z)|, \tag{1.29}$$

where $|\Gamma_\varepsilon|$ denotes the contour length of Γ_ε. Varying $\varepsilon > 0$ in (1.29) produces a continuum of bounds, and the infimum of (1.29) over all ε is also a valid bound on $\|\phi_k(\mathbf{A})\|$. As $\mathbf{A}$ departs from normality, the set $\sigma_\varepsilon(\mathbf{A})$ typically grows, as does the scalar $|\Gamma_\varepsilon|/(2\pi\varepsilon) \geq 1$ [23, Ch. 48], and (1.29) also allows slow singular value decay when $\mathbf{A}$ is far from normal.

As observed in Remark 1.1, if $\mathbf{A}$ is so far from normal that $\sigma_\varepsilon(\mathbf{A})$ intersects $\mathbb{C}_+$, then $\sup_{z \in \sigma_\varepsilon(\mathbf{A})} |\phi_k(z)| > 1$, and (1.29) does not give a useful bound on ADI convergence or the singular values of $\mathbf{X}$. However, pseudospectra are more flexible than the parameterless numerical range; for any fixed stable $\mathbf{A}$, one can choose $\varepsilon > 0$ small enough that $\sigma_\varepsilon(\mathbf{A}) \subset \mathbb{C}_-$.

1.2.4 Choosing Shift Parameters

The bounds in the previous sections come from analyzing (1.12), and the strength of such bounds depends heavily on the choice of shift parameters $\{q_k\}$. Although the iteration (1.12) will converge for any choice of shifts $\{q_k\} \subset \mathbb{C}_-$, the convergence may be very slow if the shifts are not chosen carefully.

The previous section explained that the convergence of ADI is bounded below by the singular values of $\mathbf{X}$ and above by $\|\phi_k(\mathbf{A})\|$,

$$\frac{s_{kp+1}}{s_1} \leq \frac{\|\mathbf{X} - \mathbf{X}_k\|_2}{\|\mathbf{X}\|_2} \leq \|\phi_k(\mathbf{A})\|^2, \tag{1.30}$$

so *if* one can find a small number of shifts that make $\|\phi_k(\mathbf{A})\|$ small, then convergence will be fast. Rather than attempting to minimize $\|\phi_k(\mathbf{A})\|$ itself, one may attempt to minimize one of the bounds from the previous section: (1.21), (1.23), and (1.29), which are collected here

$$\|\phi_k(\mathbf{A})\| \leq \kappa(\mathbf{V}) \max_{z \in \sigma(\mathbf{A})} |\phi_k(z)| \tag{1.31a}$$

$$\|\phi_k(\mathbf{A})\| \leq \gamma \sup_{z \in W(\mathbf{A})} |\phi_k(z)| \tag{1.31b}$$

$$\|\phi_k(\mathbf{A})\| \leq \frac{|\Gamma_\varepsilon|}{2\pi\varepsilon} \sup_{z \in \sigma_\varepsilon(\mathbf{A})} |\phi_k(z)|. \tag{1.31c}$$

These are several ways to bound $\|\phi_k(\mathbf{A})\|$ with the maximum value of $|\phi_k|$ over sets of scalars, so one may find approximately optimal shifts by solving the "ADI minimax problem,"

$$\{\hat{q}_k\} = \arg\min_{q_1,\ldots,q_k} \max_{z \in \Omega} \prod_{j=1}^{k} \left| \frac{q_j - z}{\overline{q}_j + z} \right|, \tag{1.32}$$

where Ω is $\sigma(\mathbf{A})$, $W(\mathbf{A})$, or $\sigma_\varepsilon(\mathbf{A})$. In fact, one might reasonably expect to find good shifts by solving (1.32) with Ω as any set containing the spectrum of $\mathbf{A}$. The exact solution to (1.32) is known when Ω is a sublevel set of certain rational functions [26],

and in other cases, high quality shifts can be obtained heuristically [16]. However, even after (at least approximately) solving (1.32) for $\{\hat{q}_k\}$,combining (1.31) with (1.30) only gives a meaningful bound on ADI convergence if $\max_{z\in\Omega} |\phi_k(z)|$ is small enough that (1.31) can be used to show $\|\phi_k(\mathbf{A})\| < 1$.

Rather than further exploring shift parameters and the upper bound on convergence in (1.30), the rest of this work concentrates on the lower bound, i.e., the fastest possible convergence rate as revealed by the singular values of $\mathbf{X}$.

1.3 Nonnormality and Decay

Sorensen and Zhou [22] observed that the bounds in Section 1.2 are pessimistic, particularly when $\mathbf{A}$ is far from normal. Because these bounds for the singular values of $\mathbf{X}$ are based on specific low-rank approximations (the ADI iterates $\mathbf{X}_k$), pessimistic bounds correspond to approximations that are far from optimal. In other words, when $\mathbf{A}$ is far from normal, ADI may converge slowly compared to the optimal rate s_{kp+1}/s_1.

For the special case below, Sabino showed that increasing the departure form normality of $\mathbf{A}$ slows the singular value decay of $\mathbf{X}$ up to a point, beyond which the singular value decay starts to accelerate [19]. The core of this thesis in Chapter 3 shows that this eventual decay acceleration occurs generally. Part of these results were published separately [2].

The solution $\mathbf{X}$ can be found algebraically for Lyapunov equations of the form

$$\mathbf{A} := \begin{bmatrix} -1 & \alpha \\ 0 & -1 \end{bmatrix} \tag{1.33a}$$

$$\mathbf{G} = -\mathbf{B}^T\mathbf{B} := -[t \quad 1]^T[t \quad 1] = \begin{bmatrix} -t^2 & -t \\ -t & -1 \end{bmatrix} \tag{1.33b}$$

$$\mathbf{X} = \frac{1}{4}\begin{bmatrix} 2t^2 + 2\alpha t + \alpha^2 & \alpha + 2t \\ \alpha + 2t & 2 \end{bmatrix}. \tag{1.33c}$$

Section 2.1 will introduce several measures of departure from nonnormality, and these show that $\mathbf{A}$ is farther from normal as α increases.

Rather than choosing a single $\mathbf{B}$ matrix for this problem, it will be illustrative to use (for each α) the $\mathbf{B}$ that gives the slowest singular value decay. The singular values of (1.33c) can be computed algebraically, and it is easy to show that singular value decay is slowest for the right-hand side with $t = -\alpha/2$, which corresponds to

$$\max_t \frac{s_2}{s_1} = \begin{cases} \alpha^2/4, & \text{if } 0 \le \alpha \le 2; \\ 4/\alpha^2, & \text{if } \alpha \ge 2. \end{cases} \tag{1.34}$$

In other words, for fixed α, the singular value decay of the solution (1.33c) is at most (1.34) for any right-hand side. Because (1.34) uses the worst-case right-hand side, this is the best possible bound for the singular value decay of $\mathbf{X}$ as a function of α (or $\mathbf{A}$). This provides a standard with which to evaluate the bounds that are not specific to this example. As bounds for s_2/s_1, (1.31b) and (1.31c) are tight to the extent that they match (1.34).

As α increases from 0, $\mathbf{A}$ departs from normality and the best bound (1.34) increases until $\alpha = 2$ but then decreases. This non-monotone behavior of the ideal bound could not have been predicted by examining the previous theoretical decay bounds, which only grow as α increases as shown in Figure 1.1.

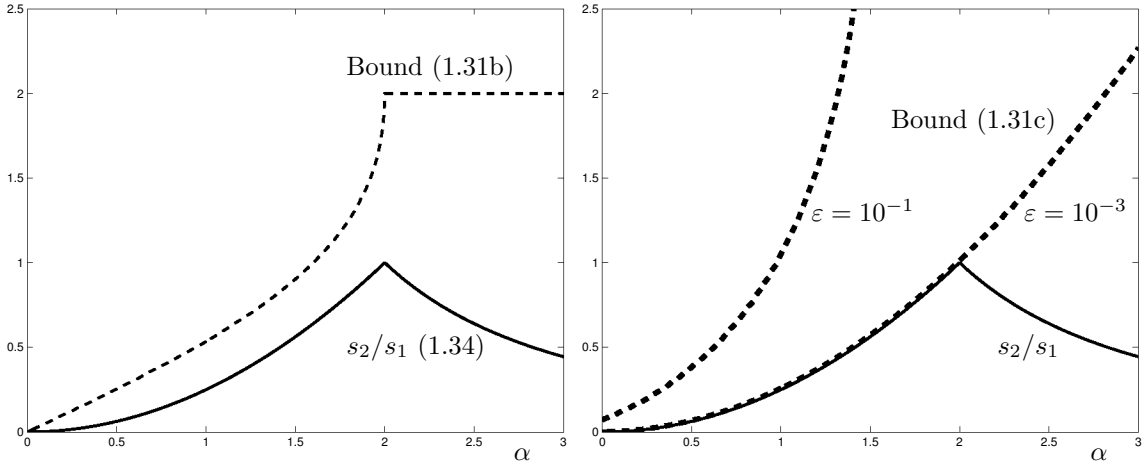

Figure 1.1 : The Jordan block (1.33a) is not diagonalizable, so (1.31a) does not apply, but (1.31b) and (1.31c) are not difficult to calculate. Both $W(\mathbf{A})$ and $\sigma_\varepsilon(\mathbf{A})$ are discs centered at -1, and a single optimal ADI shift was chosen as in [26]. The numerical range bound (1.31b) (left) assumes $\gamma = 2$. Remark 1.1 showed that (1.31b) cannot be helpful (i.e., less than 1) when $W(\mathbf{A})$ intersects $\mathbb{C}_+$, which occurs for $\alpha \geq 2$. The pseudospectral bound (1.31c) (right) is very descriptive of s_2/s_1 for $\alpha \leq 2$, but it fails to match the acceleration of decay that occurs for larger α, even though $\sigma_\varepsilon(\mathbf{A})$ remains within the left half-plane for the range of α shown.

Other than the recent paper [2], the literature does not appear to contain more thorough investigations of the effect of the nonnormality of $\mathbf{A}$ on singular value decay of $\mathbf{X}$.

Chapter 2

Motivating Observations

Although Lyapunov equations in general have been intensely researched, the particular issue of nonnormal coefficients is largely unexplored. The example of Sabino shown in Section 1.3 illustrates that the bounds in Section 1.2 can have misleading qualitative behavior. This chapter corroborates this observation with further examples, which also provide some evidence that the numerical abscissa of the coefficient $\mathbf{A}$ could be used to predict the decay acceleration.

2.1 Measuring Nonnormality

Fundamentally, normality is a binary property: $\mathbf{A}$ is normal if it is unitarily diagonalizable (i.e., $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^*$ for some unitary $\mathbf{V}$ and diagonal $\mathbf{\Lambda}$); otherwise, $\mathbf{A}$ is nonnormal. But for a specific application, some nonnormal matrices will behave more like normal matrices than others. In these cases, it may be useful to compare the "level" of nonnormality of matrices. The property of normal matrices that is important to an application may determine how nonnormality should be measured. Trefethen and Embree discuss several possibly useful scalar measures of nonnormality in [23, Ch. 48]. Seventy equivalent conditions for normality are listed in [10], many of which can be said to be more closely satisfied for some matrices than others, thereby providing many potential measures of nonnormality. A few such measures are listed below. For matrices with the same eigenvalues, the measures in this section (and

others) are shown to be essentially equivalent in [8]. Therefore it is not ambiguous to speak of one matrix having greater "departure from normality" than another without referring to a specific measure.

- **Distance from normality**: Perhaps the most obvious way to measure nonnormality is the minimum distance to a normal matrix:

$$\inf_{\mathbf{M} \text{ normal}} \|\mathbf{A} - \mathbf{M}\|_F. \tag{2.1}$$

  This unassuming quantity was first shown to be obtainable in [18] and is surprisingly expensive to compute. While it has interesting theoretical qualities, the distance to normality does not arise naturally in applications and will not be used in this work.

- **Distance from commutativity**: Normality of $\mathbf{A}$ is equivalent to (and often defined as) commutativity with $\mathbf{A}^*$, that is $\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$. The norm of the difference is therefore an obvious measure of nonnormality:

$$\|\mathbf{A}^*\mathbf{A} - \mathbf{A}\mathbf{A}^*\|. \tag{2.2}$$

  However, the commutativity of a normal matrix with its adjoint is not typically the most important consequence of nonnormality. Rather, the relevant properties of a normal matrix are often seen by considering that it has a complete set of eigenspaces that are orthogonal to each other. The next two measures of nonnormality relate to this fact.

- **Condition number of eigenvector matrix**: A matrix is normal if and only if it is unitarily diagonalizable (i.e., it is not defective and its eigenspaces are orthogonal). Thus, in the case that $\mathbf{A}$ is diagonalizable, $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$, $\mathbf{A}$ is

normal if and only if $\mathbf{V}$ can be taken to be a unitary matrix. So the degree of nonnormality of $\mathbf{A}$ can be measured by the condition number (or departure from orthogonality) of $\mathbf{V}$

$$\kappa(\mathbf{V}) := \|\mathbf{V}\|_2 \|\mathbf{V}^{-1}\|_2 \geq 1, \tag{2.3}$$

with $\kappa(\mathbf{V}) = 1$ if and only if $\mathbf{V}$ is unitary and $\mathbf{A}$ is normal.

Although $\mathbf{V}$ is not unique, it seems reasonable to choose the eigenvector matrix with the lowest condition number, which corresponds to choosing an orthogonal basis for each eigenspace of $\mathbf{A}$.

The scale of the columns of $\mathbf{V}$ is also arbitrary since $\mathbf{VD}$ is also an eigenvector matrix for any nonsingular diagonal matrix $\mathbf{D}$. It is not true that $\kappa(\mathbf{V})$ is always minimized when the columns of $\mathbf{V}$ have equal norm [24], but the ratio is no more than $\sqrt{n}$. In other words

$$\min_{\mathbf{D} \text{ diagonal}} \kappa([v_1 \quad \cdots \quad v_n]\mathbf{D}) \geq \tfrac{1}{\sqrt{n}} \kappa\Big([v_1 \quad \cdots \quad v_n]\tilde{\mathbf{D}}\Big) \tag{2.4}$$

with

$$\tilde{\mathbf{D}} = \begin{bmatrix} \|v_1\| & & \\ & \ddots & \\ & & \|v_n\| \end{bmatrix}^{-1}.$$

- **Henrici's ν-departure from normality**: Henrici suggests another measure of nonnormality based on the degree to which a matrix fails to have a full set of orthogonal eigenspaces [12] . Specifically, take the norm of the off-diagonal

portion of a unitary triangularization (Schur factorization)

$$\begin{aligned}\Delta(\mathbf{A}) := \min\{\|\mathbf{T}\| : \mathbf{A} = \mathbf{V}(\mathbf{\Lambda} + \mathbf{T})\mathbf{V}^*, \\ \mathbf{V}\,\text{unitary},\ \mathbf{\Lambda}\,\text{diagonal}, \\ \mathbf{T}\,\text{strictly upper triangular}\}.\end{aligned} \tag{2.5}$$

If the Frobenius norm is used in (2.5), then $\Delta(\mathbf{A})$ is independent of the choice of factorization, and

$$\|\mathbf{A}\|_F^2 = \|\mathbf{V}(\mathbf{\Lambda} + \mathbf{T})\mathbf{V}^*\|_F^2 = \|\mathbf{\Lambda}\|_F^2 + \|\mathbf{T}\|_F^2 \tag{2.6}$$

$$\Delta_F(\mathbf{A})^2 := \|\mathbf{T}\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{\Lambda}\|_F^2 \tag{2.7}$$

$$\Delta_F(\mathbf{A}) = \sqrt{\|\mathbf{A}\|_F^2 - \sum_{k=1}^{n} |\lambda_k|^2} \tag{2.8}$$

$$= \sqrt{\sum_{k=1}^{n} (\varsigma_k^2(\mathbf{A}) - |\lambda_k|^2)}. \tag{2.9}$$

Since the scale of a matrix does not affect its eigenspaces, one could also consider the scale-invariant measure of departure from normality $\Delta_F(\mathbf{A})/\|\mathbf{A}\|$ for any convenient norm.

Henrici also gave a bound for the numerical range that grows with nonnormality as measured by Δ_F

$$W(\mathbf{A}) \subseteq \mathrm{Co}(\sigma(\mathbf{A})) + B\left(\Delta_F(\mathbf{A})\sqrt{\frac{1 - 1/n}{2}}\right) \tag{2.10}$$

where $\mathrm{Co}(\cdot)$ denotes the convex hull of a set and $B(r) := \{z \in \mathbb{C} : |z| \leq r\}$. In particular, the numerical abscissa satisfies

$$\omega(\mathbf{A}) \leq \max \mathrm{Re}(\sigma(\mathbf{A})) + \Delta_F(\mathbf{A})\sqrt{\frac{1 - 1/n}{2}}. \tag{2.11}$$

After dividing both sides by $\|\mathbf{A}\|$, (2.11) gives one way of considering how the growth of $\omega(\mathbf{A})/\|\mathbf{A}\|$ requires increasing departure from normality.

- **Real eigenvalue displacement**: Let the eigenvalues of the Hermitian part

$$H(\mathbf{A}) := (\mathbf{A} + \mathbf{A}^*)/2$$

be $\omega_1 \geq \cdots \geq \omega_n$ and let the eigenvalues of $\mathbf{A}$ be $\lambda_1, \ldots, \lambda_n$ in any order. Define

$$\Omega(\mathbf{A}) := \min_{p \text{ permutes } \{1,\ldots,n\}} \sqrt{\sum_{j=1}^{n} (\mathrm{Re}(\lambda_j) - \omega_{p(j)})^2}. \tag{2.12}$$

It is shown in [10] that $\mathbf{A}$ is normal if and only if $\Omega(\mathbf{A}) = 0$. Equivalently, one could use the scale-invariant version $\Omega(\mathbf{A})/\|\mathbf{A}\|$. Notice that greater departure from normality is indicated—not only by the growth of $\omega_1/\|\mathbf{A}\|$ as discussed in Section 1.2.2—but by the growth of $\omega_k/\|\mathbf{A}\|$ for any k.

2.2 Lack of Singular Value Decay

To reveal the properties of $\mathbf{A}$ that determine the level of singular value decay of $\mathbf{X}$, consider the extreme case where $\mathbf{X}$ has no singular value decay at all, i.e., $s_k = \xi$ for all k. The fact that $\mathbf{X}$ is Hermitian requires $\mathbf{X} = \xi\mathbf{I}$, and (1.3) becomes

$$\mathbf{A} + \mathbf{A}^* = -\frac{1}{\xi}\mathbf{B}\mathbf{B}^*. \tag{2.13}$$

Thus the Hermitian part $(\mathbf{A} + \mathbf{A}^*)/2$ is negative semidefinite and $\omega(\mathbf{A}) \leq 0$. (Consequently, for any other $\mathbf{A}$ such that $\omega(\mathbf{A}) > 0$, $\mathbf{X}$ must exhibit at least some singular value decay.) Furthermore, if $p < n$ as in most control applications, then $\mathbf{B}\mathbf{B}^*$ is singular, so in the special case of (2.13),

$$0 \in \sigma(-\mathbf{B}\mathbf{B}^*) = \sigma(H(\mathbf{A}))$$

and $\omega(\mathbf{A}) = 0$. So, when $H(\mathbf{A})$ is a scalar multiple of $\mathbf{B}\mathbf{B}^*$, the slowest singular value decay occurs *exactly* when $\omega(\mathbf{A}) = 0$.

This result seems to recommend investigating the role of $\omega(\mathbf{A})$, but it does not directly apply to any of the examples that follow since $H(\mathbf{A})$ will not be required to be an exact multiple of $\mathbf{BB}^*$. Even so, it will be seen that in every case, decay is slowest at a point when $\omega(\mathbf{A}) \geq 0$. It is not known whether singular value decay can accelerate while $\omega(\mathbf{A}) < 0$.

2.3 Numerical Demonstrations

To isolate the effect of nonnormality on the decay of s_k, experiments need families of related matrices that differ in normality but are otherwise similar. Throughout this work, the chosen pencils of coefficients $\mathbf{A}_\alpha$ will have constant spectra because this is a condition of the equivalence of measures of nonnormality explained at the beginning of Section 2.1. The parameter α will be suppressed when there is no confusion.

For a synthetic example, consider the Jordan block $\mathbf{A} = \alpha\mathbf{S} - \mathbf{I}$, where $\mathbf{S}$ is the shift matrix

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}. \tag{2.14}$$

This is the same as the example in Section 1.3 but for general n. As α increases, the shift operator $\alpha\mathbf{S}$ dominates and $\mathbf{A}$ departs from normality: $\Delta_F(\mathbf{A}) = (n-1)\alpha$.

Figure 2.1 shows the singular values of $\mathbf{X}$ for a range of α values. As α increases (and $\mathbf{A}$ departs from normality), the singular values of $\mathbf{X}$ increase at first but then decrease. The dashed line marks the point at which $\omega(\mathbf{A}) = 0$ and the numerical range $W(\mathbf{A})$ first intersects the right half-plane. To the right of the dashed line, $\omega(\mathbf{A}) > 0$ and the singular value decay of $\mathbf{X}$ accelerates. The coincidence of decay acceleration

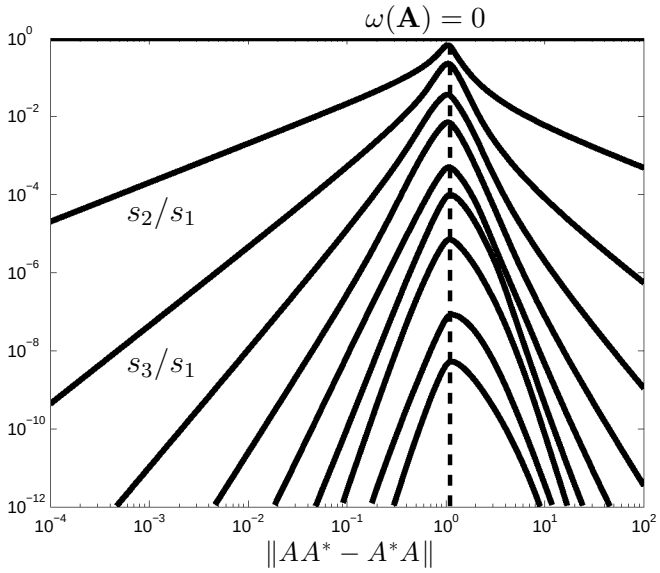

Figure 2.1 : For this example, $\mathbf{A} \in \mathbb{R}^{12\times 12}$ is a Jordan block and $\mathbf{B} \in \mathbb{R}^{12\times 1}$ is a fixed vector with random entries i.i.d. $\sim N(0,1)$. The nonnormality of $\mathbf{A}$ has a non-monotone relationship with the singular values of $\mathbf{X}$. Specifically, the normalized singular values of $\mathbf{X}$ begin to decrease after the numerical range of $\mathbf{A}$ crosses the imaginary axis. The dashed line marks this threshold.

and the crossing of the numerical range into $\mathbb{C}_+$ suggests that the sign-change of $\omega(\mathbf{A})$ might be used to predict decay acceleration.

In order to do the above study with matrices from an application, a method is needed to generate a family or pencil of matrices based on the matrix of interest. Henrici's ν-departure (2.5) suggests such a method. Start with a Schur factorization and rescale the off-diagonal portion of the triangular factor, as in

$$\mathbf{A} = \mathbf{URU}^* \tag{2.15a}$$

$$\mathbf{R} = \mathbf{\Lambda} + \mathbf{T} \tag{2.15b}$$

$$\mathbf{R}_\alpha = \mathbf{\Lambda} + \alpha\mathbf{T} \tag{2.15c}$$

where $\mathbf{\Lambda}$ and $\mathbf{T}$ are diagonal and strictly upper triangular. The parameter α directly controls the nonnormality of R_α as measured by (2.5). The original basis could be

restored by forming $\mathbf{A}_\alpha = \mathbf{U}\mathbf{R}_\alpha\mathbf{U}^*$, but since this orthogonal transformation does not affect the departure from normality or the spectrum of the Lyapunov solution $\mathbf{X}$, it is sufficient to use the family of Lyapunov equations

$$\mathbf{R}_\alpha\mathbf{X}_\alpha + \mathbf{X}_\alpha\mathbf{R}_\alpha^* = -\mathbf{U}^*\mathbf{B}\mathbf{B}^*\mathbf{U}. \tag{2.16}$$

Figure 2.2 was created by using the method (2.15) on data related to control of the International Space Station available from SLICOT*. The system feedback matrix $\mathbf{A}$ from the dataset was factored, and (2.16) was solved for various α. The singular value decay of the solutions $\mathbf{X}_\alpha$ did not accelerate significantly when using the right-hand side $-\mathbf{B}\mathbf{B}$ from the dataset. However, with a random $\mathbf{B}$, the behavior of the singular values of $\mathbf{X}_\alpha$ is similar to that of the previous example shown in Figure 2.1, except that the numerical range's threshold does not so closely coincide with the accelerating singular value decay. In this and other experiments drawn from applications, $W(\mathbf{A})$ crossed into the right half-plane at or below the level of nonnormality than was required to speed up the singular value decay of $\mathbf{X}$.

2.4 Symbolic Demonstration

This section develops a companion example to the one in Section 1.3, which confirms that singular value decay acceleration occurs for a family of diagonalizable $\mathbf{A}$. Using diagonalizable $\mathbf{A}$ allows a comparison of actual decay to the bound (1.31a). The right-hand side $-\mathbf{B}\mathbf{B}$ will again be chosen to make decay as slow as possible. The nonnormality of $\mathbf{A}$, the sign of $\omega(\mathbf{A})$, and the decay of singular values of $\mathbf{X}$ will then be compared

*Subroutine Library in Systems and Control Theory benchmark download page: http://slicot.org/20-site/126-benchmark-examples-for-model-reduction

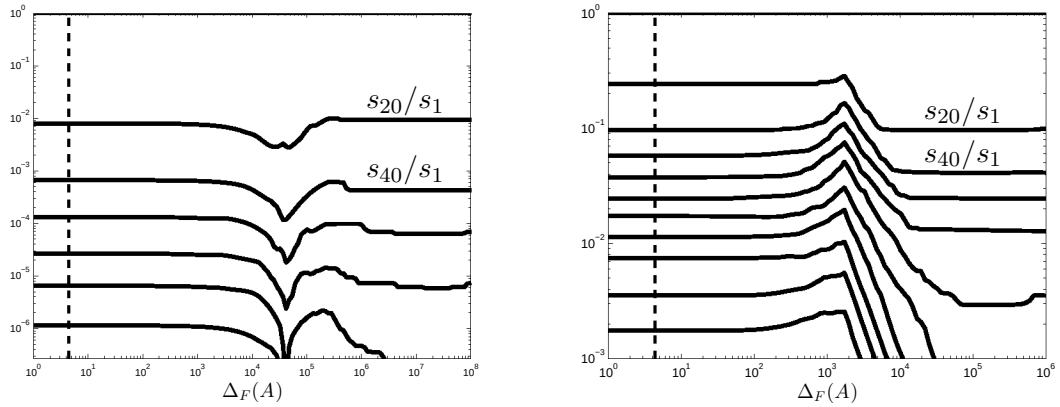

Figure 2.2 : For the original ISS problem (left), scaling up the off-diagonal part of the Schur factor as in (2.15) does not cause definite singular value decay acceleration. However, this is a highly structured, possibly exceptional equation. To remove any effect of the structure of the right-hand side of the problem, the same coefficient $\mathbf{A} \in \mathbb{R}^{270\times 270}$ may be paired with a random $\mathbf{B} \in \mathbb{R}^{270\times 3}$ with entries distributed i.i.d. $\sim N(0,1)$. For this setup (right), decay accelerates when $\Delta_F(\mathbf{A})$ grows large. This occurs after $\mathbf{A}$ departs from normality beyond the threshold $\omega(\mathbf{A}) > 0$, although the coincidence is not as striking as in Figure 2.1.

For fixed real r and M and parameters α and t, consider

$$\mathbf{A} = \begin{bmatrix} r+M & \alpha M \\ -2M/\alpha & r-M \end{bmatrix} \tag{2.17a}$$

$$\mathbf{B} = \begin{bmatrix} t & 1 \end{bmatrix}^T. \tag{2.17b}$$

The eigenvalues of $\mathbf{A}$ are $\sigma(\mathbf{A}) = \{r \pm iM\}$, so $\mathbf{A}$ is stable for any $r < 0$. By symmetry, attention can be limited to the case $\alpha > 0$. For simplicity, the calculations below also assume $|r| > M \geq 0$.

The solution to (1.3) is

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, \tag{2.18}$$

where

$$X_{11} = \frac{-2M^2 + 2(M^2 - Mr)\alpha t - (M^2 - Mr + r^2)\alpha^2 t^2}{2\alpha^2 r\,(M^2 + r^2)} \tag{2.19a}$$

$$X_{12} = X_{21} = \frac{-2M^2 - 2Mr + 2(M^2 - r^2)\alpha t - (M^2 - Mr)\alpha^2 t^2}{4\alpha r\,(M^2 + r^2)} \tag{2.19b}$$

$$X_{22} = \frac{-2M^2 - 2Mr - 2r^2 + 2(M^2 + Mr)\alpha t - M^2\alpha^2 t^2}{4r\,(M^2 + r^2)}. \tag{2.19c}$$

The singular value ratio of $\mathbf{X}$ is

$$\frac{s_2}{s_1} = \frac{\operatorname{tr}(\mathbf{X}) - \sqrt{\operatorname{tr}(\mathbf{X})^2 - 4\,\det(\mathbf{X})}}{\operatorname{tr}(\mathbf{X}) + \sqrt{\operatorname{tr}(\mathbf{X})^2 - 4\,\det(\mathbf{X})}}. \tag{2.20}$$

Choosing a single $\mathbf{B}$ matrix for this problem would be arbitrary and provide limited information. The behavior of this equation is better understood by choosing the $\mathbf{B}$ that gives the slowest singular value decay. It can be verified that s_2/s_1 achieves a maximum value of

$$\max_t \frac{s_2}{s_1} = \frac{1}{2}\Big(\beta - \sqrt{\beta^2 - 4}\Big) \tag{2.21}$$

with

$$\beta := \frac{1}{\alpha^2 M^2}\Big[r(4 + 2\alpha^2)\sqrt{(4 + \alpha^4)(M^2 + r^2)} + (4 + 2\alpha^2 + \alpha^4)(M^2 + 2r^2)\Big]$$

when

$$t = \frac{2M + \alpha^2 M - 2r + \alpha^2 r - \sqrt{(4 + \alpha^4)(M^2 + r^2)}}{\alpha^3 M - 2\alpha r}. \tag{2.22}$$

So for fixed α, the singular value decay of $\mathbf{X}$ is no faster than (2.21) for any right-hand side. Because (2.21) uses the worst-case right-hand side, this is the best possible bound for the singular value decay of $\mathbf{X}$ as a function of α (or $\mathbf{A}$). So, (1.31a) is tight as a bound for s_2/s_1 to the extent that it matches (2.21).

Earlier examples suggested that the sign of the numerical abscissa of $\mathbf{A}$ may control the rate of singular value decay of $\mathbf{X}$. To illustrate that hypothesis for this example,

it is useful to find the $\mathbf{A}$ that corresponds to the slowest decay. Basic calculus shows that (2.21) is maximized when

$$\alpha = \frac{\sqrt{2}}{|M|}\sqrt{r^2 \pm \sqrt{r^4 - M^4}}. \tag{2.23}$$

Also, observe that the numerical abscissa of $\mathbf{A}$ is

$$\omega(\mathbf{A}) = \lambda_1\left(\tfrac{1}{2}(\mathbf{A} + \mathbf{A}^*)\right) = r + |M|\sqrt{1/\alpha^2 + \alpha^2/4}, \tag{2.24}$$

and substituting (2.24) into (2.23) gives $\omega(\mathbf{A}) = 0$. In other words, the slowest singular value decay precisely coincides with the numerical range crossing into the right half-plane. This is exactly what occurred for the non-diagonalizable Jordan block examples in Sections 1.3 and 2.3.

In order to calculate the diagonalization bound (1.31a), the nonnormality $\kappa(\mathbf{V})$ as in (2.3) can also be found in closed form. The eigenvalues of $\mathbf{A}$ are $r \pm Mi$, and choosing equally scaled eigenvectors gives

$$\mathbf{V} = \begin{bmatrix} 1-i & 1+i \\ \alpha & \alpha \end{bmatrix} \tag{2.25}$$

$$\kappa(\mathbf{V}) = \frac{1}{2\alpha}\Big(2 + \alpha^2 + \sqrt{4 + \alpha^4}\Big). \tag{2.26}$$

As a function of α, $\kappa(\mathbf{V})$ is convex and minimized at $\alpha = \sqrt{2}$, so $\mathbf{A}$ departs from normality as α departs from $\sqrt{2}$. The slowest singular value decay occurs at the critical values (2.23), which lie on either side of $\sqrt{2}$. Therefore, moving $\mathbf{A}$ away from normality (whether by increasing or decreasing α away from $\sqrt{2}$) causes first a slowing and then acceleration of worst-case decay. Figure 2.3 illustrates this by plotting worst-case dacay (2.21) against a range of α values. The figure also shows how the bound (1.31a) (using $\sigma(\mathbf{A})$ as shifts) grows monotonically as α departs from $\sqrt{2}$, which qualitatively describes the behavior of s_2/s_1 only while $\omega(\mathbf{A}) < 0$.

Also note that, for smaller $\left|\frac{r}{M}\right|$, the diagonalization bound (1.31a) becomes even less tight. Using $\sigma(\mathbf{A})$ as shifts ($q_1 = r + iM$ and $q_2 = r - iM$) gives

$$\frac{s_2}{s_1} \leq \kappa(\mathbf{V})^2 \sup_{z \in \sigma(\mathbf{A})} |\phi_2(z)| = \kappa(\mathbf{V})^2 \left| \frac{1}{(r/M)^2 + 1} \right|. \tag{2.27}$$

From (2.26), the bound (2.27) is greater than 1—and therefore uninformative—for all α if $\left|\frac{r}{M}\right| < \sqrt{2 + 2\sqrt{2}}$.

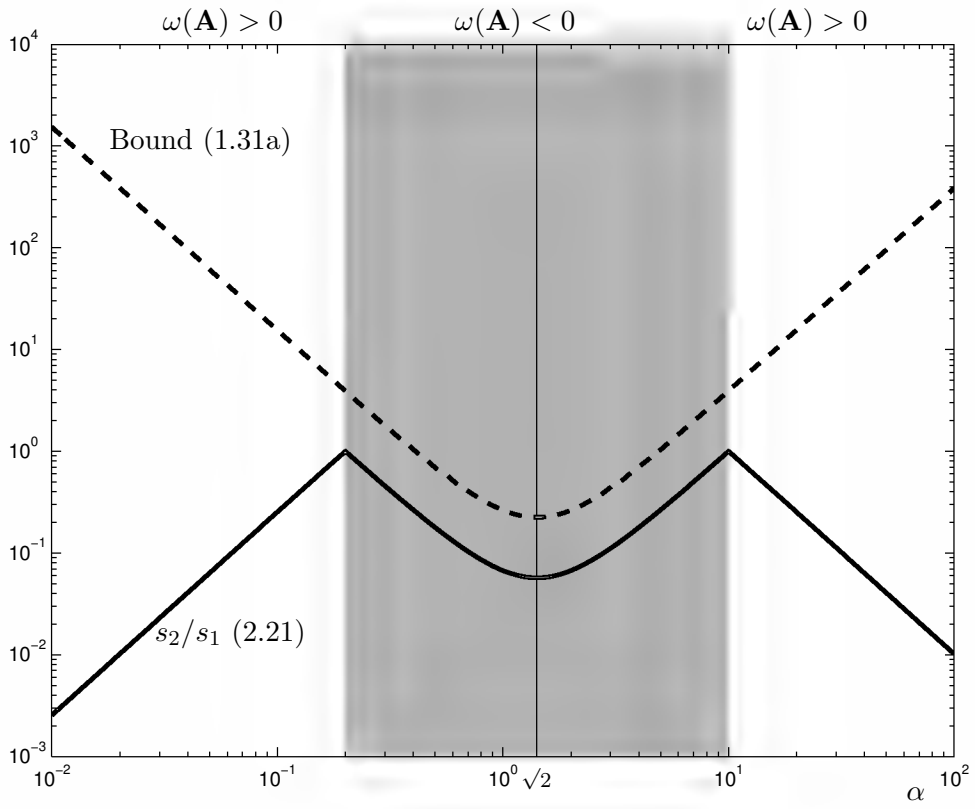

Figure 2.3 : The matrix $\mathbf{A}$ in (2.17) demonstrates a weakness of the bound (1.31a) (plot uses $r = -5$, $M = 1$). Decay is locally fastest at $\alpha = \sqrt{2}$ for which $\mathbf{A}$ is closest to normal. Near this point, the bound (1.31a) matches the actual behavior of the decay well. At first, decay slows as $\mathbf{A}$ departs from normality, i.e., as α departs from $\sqrt{2}$, but when α reaches the critical values in (2.23) such that $\omega(\mathbf{A}) = 0$, decay begins to accelerate. Instead of matching this behavior, (1.31a) continues to increase with departure from normality because $\kappa(\mathbf{V})$ grows without bound. The other bounds in Section 1.2 also have factors that grow large for nonnormal $\mathbf{A}$, as shown in Figure 1.1. Contrast this with the bound (3.13) illustrated in Figure 3.1, which decreases as $\omega(\mathbf{A})$ grows.

Chapter 3

Singular Value Decay and Hermitian Part Eigenvalues

The numerical abscissa, $\omega(\mathbf{A})$, was defined in (1.24). It is both the rightmost extent of the numerical range in the complex plane and the rightmost eigenvalue of the Hermitian part $H(\mathbf{A}) := (\mathbf{A} + \mathbf{A}^*)/2$. For small t, the numerical abscissa bounds the transient growth of the linear system $\dot{x}(t) = \mathbf{A}x(t)$ with $x(0) = x_0$

$$\max_{\substack{x_0 \in \mathbb{C}^n \\ \|x_0\|=1}} \frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|\Big|_{t=0} = \omega(\mathbf{A}),$$

see, for example, [23, Thm. 17.4]. Thus $\omega(\mathbf{A}) > 0$ is a necessary condition for solutions of $\dot{x}(t) = \mathbf{A}x(t)$ to exhibit transient growth—an important consequence of nonnormality in dynamical systems.

The subordinate eigenvalues of the Hermitian part reveal more information about the departure of $\mathbf{A}$ from normality. If $\mathbf{A}$ is close to normal, then the Hermitian part eigenvalues must be close to the spectrum of $\mathbf{A}$ as measured by $\Omega(\mathbf{A})$ in (2.12). Additionally, eigenvalues of the Hermitian part have recently been used to bound the number of Ritz values of $\mathbf{A}$ that can fall in subregions of $W(\mathbf{A})$ [6, Thm. 1.2]. Like $\omega(\mathbf{A})$, interior eigenvalues of $(\mathbf{A} + \mathbf{A}^*)/2$ can be positive even when $\mathbf{A}$ is stable. This chapter uses these eigenvalues to provide a new bound on the singular values of $\mathbf{X}$ that is fundamentally different from those in Section 1.2. The description of this bound uses and expands on the results from [2].

3.1 Hermitian Part Decay Bound

The following theorem bounds the eigenvalues of $(\mathbf{A}+\mathbf{A}^*)/2$ in terms of the singular values of $\mathbf{X}$. The result can be read from two different perspectives:

- given the singular values of $\mathbf{X}$, it bounds the level of nonnormality of those $\mathbf{A}$ that can support such solutions (Theorem 3.1 and Corollary 3.4);
- given $\mathbf{A}$, it bounds the decay of singular values of $\mathbf{X}$ and requires *faster* decay as the departure of $\mathbf{A}$ from normality increases (Corollary 3.3).

Theorem 3.1. *Let $\mathbf{X} \in \mathbb{C}^{n\times n}$ solve the Lyapunov equation* (1.3) *with $(\mathbf{A}, \mathbf{B})$ controllable. Then for all $1 \le j \le k \le n$*

$$\frac{s_{k+j-1}}{s_j} - 1 - \frac{\|\mathbf{B}\|^2}{2s_j\|\mathbf{A}\|} < \frac{\omega_k}{\|\mathbf{A}\|} \le 1 - \frac{s_{n-k+j}}{s_j}, \tag{3.1}$$

where ω_k denotes the kth rightmost eigenvalue of $(\mathbf{A}+\mathbf{A}^*)/2$ and s_k denotes the kth singular value of $\mathbf{X}$.

The right inequality of Theorem 3.1 can be interpreted as a statement about the rate of singular value decay across limited ranges of singular values. From the right inequality, each ω_k cannot be too far right if there is "stagnation," or little decay, across any $n-k+1$ consecutive singular values of $\mathbf{X}$, i.e., $s_{n-k+j} \approx s_j$ for some j. Thus, the *trailing* eigenvalues of $H(\mathbf{A})$ are controlled from the right by stagnation across just a few singular values of $\mathbf{X}$, while the *dominant* eigenvalues of $H(\mathbf{A})$ are limited on the right only when many singular values stagnate.

Proof. Write the solution as $\mathbf{X} = \xi(\mathbf{I}-\mathbf{E})$ for some $\xi > 0$ (to be chosen later) and Hermitian $\mathbf{E}$. Then since $\mathbf{X}$ solves the Lyapunov equation (1.3),

$$\frac{\mathbf{A}+\mathbf{A}^*}{2} = -\frac{1}{2\xi}\mathbf{B}\mathbf{B}^* + \frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}. \tag{3.2}$$

Weyl's inequalities for the eigenvalues of sums of Hermitian matrices (see, e.g., [14, Thm. 4.3.1]) imply

$$\lambda_n\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*\Big)+\lambda_k\Big(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)\le\lambda_k\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*+\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)$$

and

$$\lambda_k\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*+\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)\le\lambda_1\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*\Big)+\lambda_k\Big(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big).$$

Since $-\mathbf{B}\mathbf{B}^*/(2\xi)$ is Hermitian negative semidefinite,

$$\lambda_n\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*\Big)=-\frac{\|\mathbf{B}\|^2}{2\xi},\qquad\lambda_1\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*\Big)\le 0.$$

Now by equation (3.2),

$$\lambda_k\Big(-\frac{1}{2\xi}\mathbf{B}\mathbf{B}^*+\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)=\lambda_k\Big(\frac{\mathbf{A}+\mathbf{A}^*}{2}\Big)=:\omega_k.$$

Together, these pieces imply

$$-\frac{\|\mathbf{B}\|^2}{2\xi}+\lambda_k\Big(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)\le\omega_k\le\lambda_k\Big(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big).\tag{3.3}$$

Note that $(\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*)/2$ is the Hermitian part of $\mathbf{A}\mathbf{E}$, and the kth singular value of a matrix bounds the kth eigenvalue of its Hermitian part [13, Cor. 3.1.5]. Applying this bound to both $\mathbf{A}\mathbf{E}$ and $-\mathbf{A}\mathbf{E}$ gives

$$-\varsigma_{n-k+1}(\mathbf{A}\mathbf{E})\le\lambda_k\Big(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)\le\varsigma_k(\mathbf{A}\mathbf{E}).$$

(Remember that $\varsigma_k(\cdot)$ is the kth largest singular value.) Using the singular value inequality [13, Thm. 3.3.16(d)],

$$\omega_k\le\lambda_k\Big(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\Big)\le\varsigma_k(\mathbf{A}\mathbf{E})\le\varsigma_1(\mathbf{A})\varsigma_k(\mathbf{E})=\|\mathbf{A}\|\varsigma_k(\mathbf{E}).$$

Applying the same results to the left-hand side of (3.3) gives

$$-\frac{\|\mathbf{B}\|^2}{2\xi}-\|\mathbf{A}\|\varsigma_{n-k+1}(\mathbf{E})\le-\frac{\|\mathbf{B}\|^2}{2\xi}-\varsigma_{n-k+1}(\mathbf{A}\mathbf{E})\le-\frac{\|\mathbf{B}\|^2}{2\xi}+\lambda_k\left(\frac{\mathbf{A}\mathbf{E}+\mathbf{E}\mathbf{A}^*}{2}\right)\le\omega_k,$$

which implies

$$-\frac{\|\mathbf{B}\|^2}{2\xi\|\mathbf{A}\|} - \varsigma_{n-k+1}(\mathbf{E}) \leq \frac{\omega_k}{\|\mathbf{A}\|} \leq \varsigma_k(\mathbf{E}). \tag{3.4}$$

Because $\mathbf{E}^*\mathbf{E} = \mathbf{E}^2 = (\mathbf{I} - \mathbf{X}/\xi)^2$ is a polynomial in $\mathbf{X}$, the eigenvalues of $\mathbf{E}^2$ are that polynomial in the eigenvalues of $\mathbf{X}$:

$$\sigma((\mathbf{I} - \mathbf{X}/\xi)^2) = \{(1 - s_j/\xi)^2 : j = 1, \ldots, n\}, \tag{3.5}$$

and the singular values of $\mathbf{E}$ are $|1 - s_j/\xi| = |\lambda_j(\mathbf{E})|$ for $j = 1, \ldots, n$.

Notice that some eigenvalues of $\mathbf{E}$ may be negative. Specifically if ξ lies between s_r and s_{r-1}

$$s_n \leq \cdots \leq s_r \leq \xi \leq s_{r-1} \leq \cdots \leq s_1,$$

then

$$\lambda_n \leq \cdots \leq \lambda_r \leq 0 \leq \lambda_{r-1} \leq \cdots \leq \lambda_1.$$

Consequently, the orders of the eigenvalues and singular values of $\mathbf{E}$ may not match after taking the absolute value, i.e., $\varsigma_k(\mathbf{E}) \neq |\lambda_k(\mathbf{E})|$ for some k. The new order is determined by the distance from the eigenvalues to ξ because $|1 - x/\xi|$ varies monotonically with $|x - \xi|$.

Now ξ should be chosen to make each side of (3.4) as sharp as possible. For any $1 \leq j \leq k$, consider the choice

$$\xi = \frac{s_j + s_{n-k+j}}{2}. \tag{3.6}$$

Singular values of $\mathbf{X}$ between s_{n-k+j} and s_j are closest to ξ, so they correspond to

the smallest singular values of $\mathbf{E}$. Specifically

$$\left|1-\frac{s_r}{\xi}\right| \le \left|1-\frac{s_j}{\xi}\right| \qquad \text{if } s_j \le s_r \le s_{n-k+j}, \tag{3.7a}$$

$$\left|1-\frac{s_r}{\xi}\right| = \left|1-\frac{s_j}{\xi}\right| \qquad \text{if } s_r = s_j \text{ or } s_r = s_{n-k+j}, \tag{3.7b}$$

and

$$\left|1-\frac{s_r}{\xi}\right| \ge \left|1-\frac{s_j}{\xi}\right| \qquad \text{if } s_r \le s_{n-k+j} \text{ or } s_r \ge s_j. \tag{3.7c}$$

There are at least $n-k+1$ values of r satisfying (3.7a) and $k+1$ values satisfying (3.7c), where any r such that $s_r = s_j$ or $s_r = s_{n-k+j}$ satisfies both with equality. Now (3.7) completely determines the position of $|\lambda_j(\mathbf{E})| = |\lambda_{n-k+j}(\mathbf{E})|$ among the singular values of $\mathbf{E}$:

$$\varsigma_k(\mathbf{E}) = \varsigma_{k+1}(\mathbf{E}) = |\lambda_j(\mathbf{E})| = |\lambda_{n-k+j}(\mathbf{E})|. \tag{3.8}$$

Next, it will be shown that, for some j, the right inequality of (3.4) is made as tight as possible by the choice (3.6). For ξ in a small neighborhood of (3.6), $\varsigma_k(\mathbf{E})$ is the larger of $|1-s_j/\xi|$ and $|1-(s_{n-k+j})/\xi|$, so (3.6) locally minimizes $\varsigma_k(\mathbf{E})$. The only other critical points of $\varsigma_k(\mathbf{E})$ as a function of ξ are local maxima where $\varsigma_k(\mathbf{E}) = \varsigma_{k-1}(\mathbf{E})$, specifically at $\xi = (s_j + s_{n-k+j+1})/2$ for $j = 1, \ldots, k-1$. It is not optimal to choose ξ beyond the extreme critical points (i.e., $\xi < (s_k + s_n)/2$ or $\xi > (s_1 + s_{n-k+1})/2$) because these are local minima.

Now use (3.6) to find

$$\varsigma_k(\mathbf{E}) = |1-s_j/\xi| = \frac{1-(s_{n-k+j})/s_j}{1+(s_{n-k+j})/s_j} \le 1-\frac{s_{n-k+j}}{s_j} \tag{3.9}$$

which, combined with (3.4), proves the right inequality of (3.1).

Optimizing the left inequality of (3.4) proceeds similarly; for any $1 \le j \le n-k$, choose $\xi = (s_j + s_{j+k-1})/2$ so that $\varsigma_{n-k+1}(\mathbf{E}) = \varsigma_{n-k+2}(\mathbf{E})$ and $\varsigma_{n-k+1}(\mathbf{E})$ is locally minimized. This choice of ξ also locally maximizes the left expression of (3.4) and

gives

$$\begin{aligned}\varsigma_{n-k+1}(\mathbf{E}) &= |1 - s_j/\xi| \\ &= |1 - s_{j+k-1}/\xi| \\ &= \frac{s_j - s_{j+k-1}}{s_j + s_{j+k-1}}.\end{aligned} \tag{3.10}$$

Substituting this into the left-hand side of (3.4),

$$\begin{aligned}\frac{\|\mathbf{B}\|^2}{2\xi\|\mathbf{A}\|} + \varsigma_{n-k+1}(\mathbf{E}) &= \frac{\|\mathbf{B}\|^2}{(s_j + s_{j+k-1})\|\mathbf{A}\|} + \frac{s_j - s_{j+k-1}}{s_j + s_{j+k-1}} \\ &= \frac{1 - (s_{j+k-1})/s_j + \|\mathbf{B}\|^2/(s_j\|\mathbf{A}\|)}{1 + (s_{j+k-1})/s_j} \\ &< 1 - \frac{s_{j+k-1}}{s_j} + \frac{\|\mathbf{B}\|^2}{s_j\|\mathbf{A}\|}.\end{aligned} \tag{3.11}$$

With (3.4), this proves the left inequality of (3.1). □

Notice that because of the simplifying relaxation in (3.9), the right inequality of (3.1) must also be strict unless $s_{n-k+j} = s_j$. The left inequality is strict only because of (3.11).

Choosing $j = 1$ in (3.1) gives a bound involving the decay of the dominant singular values, which are the most important in evaluating low-rank approximations of $\mathbf{X}$. This result was included in [2].

Corollary 3.2. *For controllable* $(\mathbf{A}, \mathbf{B})$*, the singular values of the solution* $\mathbf{X} \in \mathbb{C}^{n\times n}$ *to the Lyapunov equation* (1.3) *satisfy*

$$\frac{s_k}{s_1} - 1 - \frac{\|\mathbf{B}\|^2}{2s_1\|\mathbf{A}\|} < \frac{\omega_k}{\|\mathbf{A}\|} \leq 1 - \frac{s_{n-k+1}}{s_1}, \qquad k = 1, \ldots, n. \tag{3.12}$$

Rearranging the right bound in Corollary 3.2 gives an upper bound on the decay of the trailing singular values of $\mathbf{X}$ that is distinguished from the bounds of Section 1.2 by being independent of the rank of $\mathbf{B}$ and the choice of ADI shifts.

Corollary 3.3. *For controllable* $(\mathbf{A}, \mathbf{B})$*, the singular values of the solution* $\mathbf{X} \in \mathbb{C}^{n\times n}$ *to the Lyapunov equation* (1.3) *satisfy*

$$\frac{s_{n-k+1}}{s_1} \leq 1 - \frac{\omega_k}{\|\mathbf{A}\|}, \qquad k = 1, \ldots, n. \tag{3.13}$$

The importance of $\omega(\mathbf{A}) = \omega_1$ calls for a separately stated result. Choosing $k = 1$ in Corollary 3.2 gives bounds on the rightmost extent of any numerical range that can support a solution $\mathbf{X}$ with extreme singular values s_1 and s_n.

Corollary 3.4. *For controllable* $(\mathbf{A}, \mathbf{B})$*, the numerical abscissa* $\omega(\mathbf{A})$ *is bounded by the extreme singular values of the solution* $\mathbf{X} \in \mathbb{C}^{n\times n}$ *to the Lyapunov equation* (1.3)*:*

$$-\frac{\|\mathbf{B}\|^2}{2\|\mathbf{A}\|s_1} < \frac{\omega(\mathbf{A})}{\|\mathbf{A}\|} \leq 1 - \frac{s_n}{s_1}. \tag{3.14}$$

3.2 Analysis of Corollary 3.3

Corollary 3.3 provides a decay bound in the case that $\mathbf{A}$ is sufficiently far from normal that some eigenvalues of $H(\mathbf{A})$ are nonnegative even though $\mathbf{A}$ is stable. As Figures 3.1 and 3.2 illustrate, (3.13) can be very pessimistic but still be better than previous bounds when $\omega(\mathbf{A}) > 0$. Corollary 3.3 has several advantages over the bounds surveyed in Section 1.2.

- The bound (3.13) requires faster decay with greater nonnormality rather than allowing slower decay.

- The bound (3.13) is parameterless, while the other bounds depend on the choice of ADI shifts. This advantage is offset by the fact that (3.13) only promises the *existence* of a low-rank factorization of $\mathbf{X}$ with a certain accuracy. It does not suggest an algorithm to produce such a factorization, whereas if ADI shifts can

be found that make the bounds of Section 1.2 small, these same shifts may be used for ADI iteration with fast convergence.

- The rank of $\mathbf{B}$ does not feature in (3.13), whereas the other bounds allows slower decay as the rank of $\mathbf{B}$ increases. In particular, (1.30) is meaningless when $\mathbf{B}$ is full rank ($p = n$), which prevents the inequalities (1.31) from giving any information about the singular values of $\mathbf{X}$.

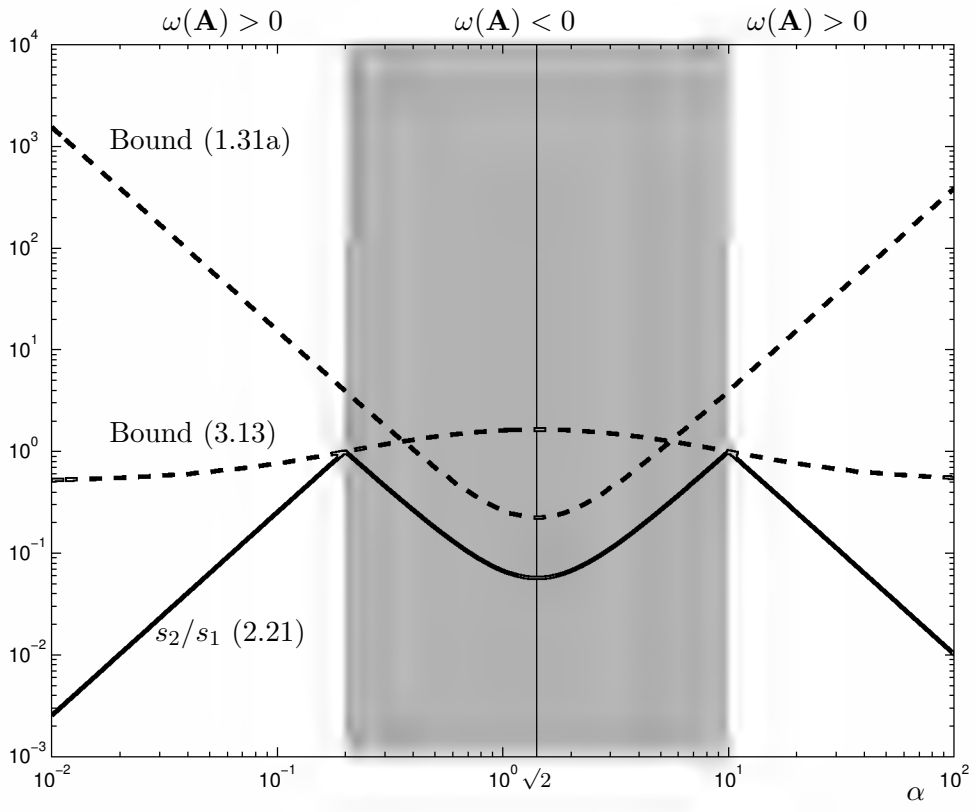

Figure 3.1 : Example (2.17) is revisited, which was previously illustrated in Figure 2.3. This example demonstrates the main advantage of the bound (3.13) over the bounds in Section 1.2, such as (1.31a), which increase with departure of $\mathbf{A}$ from normality. Beyond the $\omega(\mathbf{A}) = 0$ threshold, decay accelerates. This is better described by (3.13), which decreases for extreme departure from normality. Both (1.31a) and (3.13) fail to remain below the trivial bound $s_2/s_1 \leq 1$ for some values of α: (1.31a) is greater than 1 for highly nonnormal $\mathbf{A}$ because $\kappa(\mathbf{V}) \to \infty$ as $\alpha \to 0$ or $\alpha \to \infty$, and (3.13) is greater than 1 when $\omega(\mathbf{A}) < 0$ (the gray strip).

If $\mathbf{A}$ exhibits moderate departure from normality, Corollary 3.3 can hold with equality, as in the case of no decay ($s_1 = s_n$) considered in Section 2.2. Remember that, when $\text{rank}(\mathbf{B}) < n$, this case implies that $\omega_1 = 0 = 1 - s_n/s_1$, so Corollary 3.3 with $k = 1$ is sharp. If $\mathbf{A}$ is far from normal, i.e., $0 < \omega_k \approx \|\mathbf{A}\|$, Corollary 3.3 requires that the kth lowest singular value of $\mathbf{X}$ be small.

Corollary 3.3 is not useful when $\|\mathbf{A}\|$ is instead controlled by eigenvalues far in the left half-plane. If $\|\mathbf{A}\| \approx |\omega_n|$ is much larger than ω_k, the right-hand side of (3.13) may be almost 1 while s_{n-k+1}/s_1 may be much smaller. In particular, when $\omega_k < 0$ (as must occur for all k when $\mathbf{A}$ is stable and normal), the bound in (3.13) is vacuous.

Even if $\mathbf{A}$ is far from normal, the rate of decay could be even faster than indicated by Corollary 3.3. For the 2×2 Jordan block considered in Section 1.3,

$$\omega_1 = \alpha/2 - 1, \qquad \|\mathbf{A}\| = \sqrt{1 + \alpha^2/2 + \alpha\sqrt{\alpha^2/4 + 1}},$$

so Corollary 3.3 gives the bound

$$\frac{s_2}{s_1} \le 1 - \frac{\omega_1}{\|\mathbf{A}\|} \to 1/2, \qquad \alpha \to \infty,$$

whereas Section 1.3 showed that $s_2/s_1 \to 0$ as $\alpha \to \infty$ for this example.

Additionally, Proposition 3.5 shows there is a trade-off between the strength of Corollary 3.3 and the number of k values for which it is meaningful.

Proposition 3.5. *For any stable* $\mathbf{A}$*, let* M^+ *be the number of positive eigenvalues of* $H(\mathbf{A})$*, that is, the number of* j *such that* $\omega_j > 0$*. Then for any* $k \le M^+$*, the right-hand side of* (3.13) *satisfies the bound*

$$1 - \frac{\omega_k}{\|\mathbf{A}\|} > 1 - \frac{n - M^+}{k} \tag{3.15a}$$

$$\textit{and} \qquad 1 - \frac{\omega_k}{\|\mathbf{A}\|} > 2 - \frac{n}{k}. \tag{3.15b}$$

For $k > M^+$*, the bound is trivial since* $1 - \omega_k/\|\mathbf{A}\| \ge 1$.

Stated another way, if only a few eigenvalues of $H(\mathbf{A})$ are positive, M^+ is small and (3.15a) permits $1 - \omega_k/\|\mathbf{A}\|$ to be small (i.e., a good bound on $(s_{n-k+1})/s_1$), but only for those few k less than M^+. Alternatively, if M^+ is large ($M^+ \approx n$), then there is a non-vacuous bound $(s_{n-k+1})/s_1 \leq 1 - \omega_k/\|\mathbf{A}\| < 1$ for most values of k, but the bound is weak because $1 - \omega_k/\|\mathbf{A}\| > 1 - (n - M^+)/k \approx 1$.

Proof. Consider that

$$\sum_{j=1}^{n} \omega_j = \operatorname{tr}(\mathbf{A} + \mathbf{A}^*)/2 = \operatorname{Re}(\operatorname{tr}\mathbf{A}) = \sum_{\lambda \in \sigma(\mathbf{A})} \operatorname{Re}\lambda < 0 \tag{3.16}$$

because $\mathbf{A}$ is stable. So the negative eigenvalues of $H(\mathbf{A})$ have a greater absolute sum than the positive eigenvalues

$$\sum_{\omega_j > 0} \omega_j < -\sum_{\omega_j < 0} \omega_j = \sum_{\omega_j < 0} |\omega_j|. \tag{3.17}$$

Because $\|\mathbf{A}\| \geq |\omega_j|$ for all j,

$$k\omega_k \leq \sum_{\omega_j > 0} \omega_j < \sum_{\omega_j < 0} |\omega_j| \leq \|\mathbf{A}\|(n - M^+) \tag{3.18}$$

and when $k \leq M^+$, (3.18) implies

$$1 - \frac{\omega_k}{\|\mathbf{A}\|} > 1 - \frac{n - M^+}{k} \geq 1 - \frac{n-k}{k} = 2 - \frac{n}{k} \tag{3.19}$$

as promised. □

In summary, Corollary 3.3 is considerably better than previous results in some highly nonnormal cases, but singular values may decay much more quickly than even this improved bound. With so much room for progress, this topic is open to additional study.

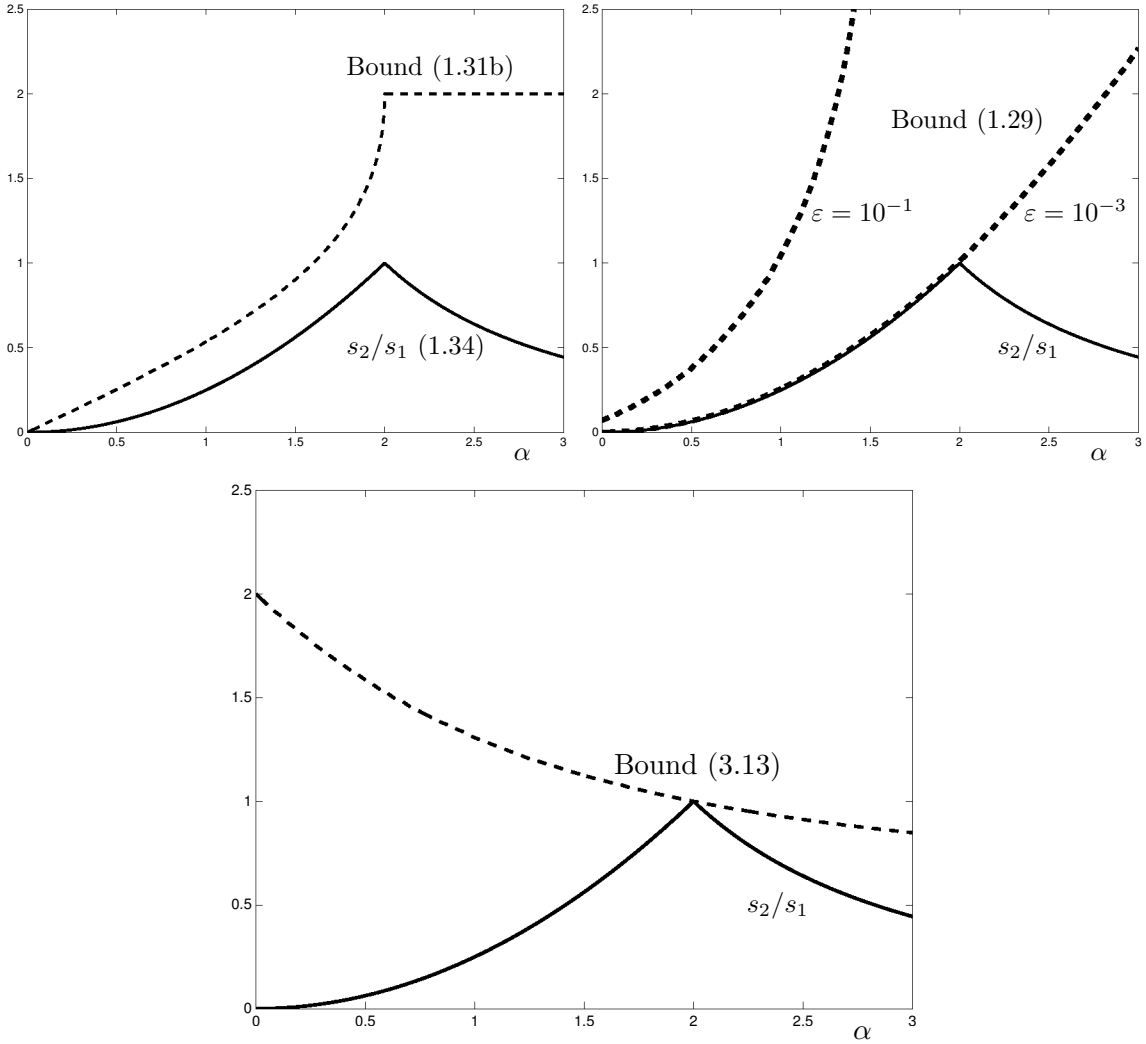

Figure 3.2 : Figure 1.1 (top) is reproduced, showing bounds (1.31b) and (1.31c) applied to the Jordan block example (1.33). As in the diagonalizable example shown in Figure 3.1, the new bound (3.13) (bottom) decreases for increasingly nonnormal Jordan blocks whereas the other bounds increase.

Chapter 4

Concluding Observations

Chapter 2 provided examples of parameterized Lyapunov equations suggesting that as the coefficient $\mathbf{A}$ departs from normality, the singular values of the solution $\mathbf{X}$ must decay more quickly. Chapter 3 makes this idea precise by proving the bound (3.13) which requires faster decay when $\mathbf{A}$ is far from normal. Like Chapter 2, Section 4.1 gives examples of Sylvester equations with solutions exhibiting accelerating singular value decay. This acceleration is not known to occur in general, but the examples suggest that a result similar to Theorem 3.1 may exist for Sylvester equations. This chapter ends with a concluding summary.

4.1 Sylvester Equations

The continuous time Lyapunov equation (1.6) is a specific case of the continuous time Sylvester equation

$$\mathbf{A}_1\mathbf{X} + \mathbf{X}\mathbf{A}_2 = \mathbf{G}. \tag{4.1}$$

Assume that $\mathbf{A}_1$ and $-\mathbf{A}_2$ have no eigenvalues in common, which is more general than assuming that $\mathbf{A}$ is stable in (1.6). This condition is necessary and sufficient for the existence and uniqueness of the solution $\mathbf{X}$; however, unlike the Lyapunov case, $\mathbf{X}$ may not be positive definite, Hermitian, or even square.

The solution of the particular Sylvester equation

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A} = -\mathbf{B}\mathbf{C}^* \tag{4.2}$$

is called the "cross Gramian" of the control system (1.1) introduced in [9]. The cross Gramian is of interest in model order reduction since it may be used—instead of the controllability and observability Gramians together—to compute a balanced reduction of (1.1) [1, Sec. 12.3]. Additionally—for a system that is stable, controllable, observable, and symmetric (i.e., $\mathbf{A}\Psi = \Psi\mathbf{A}^*$, $\mathbf{B}\Psi = \Psi\mathbf{C}^*$ for some Ψ)—the absolute values of the eigenvalues of the cross Gramian happen to be the singular values of the Hankel operator (1.7) [1, Prop. 5.9], which are known to measure the compressibility of the system [1, Thm. 7.9].

As with the Lyapunov equation, one can reasonably solve the Sylvester equation with direct methods only for small n. The Bartels–Stewart algorithm is suitable for this [3]. For large problems, limited storage and processing power require iterative methods that do not involve constructing the large dense solution matrix. Such algorithms can take a variety of forms [4, 5, 15], but they all construct factors of a low-rank approximation of $\mathbf{X}$. The singular values s_k of $\mathbf{X}$ in (4.1) have the same importance as in the Lyapunov case: they give the optimal convergence rate for *any* low-rank solution method.

For Lyapunov equations, Theorem 3.1 gives a bound on s_k that tightens as $\mathbf{A}$ departs from the normality beyond a threshold. But an example similar to that of Section 1.3 demonstrates that the nonnormality of coefficients does not have such a direct relationship with singular value decay for solutions of Sylvester equations. Using the same Jordan block $\mathbf{A}$ and rank-1 $\mathbf{G}$ matrix of (1.33), consider

$$\begin{bmatrix} -1 & \alpha \\ 0 & -1 \end{bmatrix}\mathbf{X} + \mathbf{X}\begin{bmatrix} -1 & \alpha \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} -t^2 & -t \\ -t & -1 \end{bmatrix} \tag{4.3}$$

which has the solution

$$\mathbf{X} = \frac{1}{4}\begin{bmatrix} \alpha t + 2t^2 & \alpha + 2t + \alpha^2 t + \alpha t^2 \\ 2t & 2 + \alpha t \end{bmatrix}. \tag{4.4}$$

As in Sections 1.3 and 2.4, it is illuminating to consider the right-hand side (t value) which makes decay slowest. When $\alpha < 2$, then $\omega(\mathbf{A}) < 0$ and the choice $t = -1$ results in the slowest decay. But when $\alpha > 2$, then $\omega(\mathbf{A}) > 0$ and no decay occurs ($s_2 = s_1$) with the choice

$$t = -\frac{1}{2}\Big(\alpha + \sqrt{\alpha^2 - 4}\Big).$$

The decay for these pessimal t values is

$$\max_t \frac{s_2}{s_1} = \begin{cases} \frac{\alpha^2}{2} - 2\alpha + 5 + \frac{1}{2\alpha^2}\big(16 - 16\alpha + (\alpha - 2)(\alpha^2 - 2\alpha + 4)\sqrt{4 + \alpha^2}\big) \\ \qquad\qquad \text{if } 0 \le \alpha \le 2; \\ 1 \qquad \text{if } \alpha \ge 2. \end{cases} \tag{4.5}$$

In other words, for every $\alpha \ge 2$, there is a rank-1 $\mathbf{G}$ such that the singular values of $\mathbf{X}$ do not decay at all.

Thus, the worst-case singular value decay need not accelerate with departure from normality for Sylvester equations in general. This is different from the Lyapunov case where $\max_t s_2/s_1$ reached 1 when $\omega(\mathbf{A}) = 0$ (for a certain $\mathbf{G}$) but necessarily decreased for greater α. However, for each fixed t, greater nonnormality does in fact cause singular value decay to accelerate. Figure 4.1 illustrates this with plots of decay for several fixed values of t, as well as the slowest possible decay for any t.

An experiment with random matrices suggests that this phenomenon is widespread. For a Sylvester equation with random coefficient entries, Figure 4.2 shows singular value decay that accelerates as $\mathbf{A}_1$, $\mathbf{A}_2$, or both depart from normality. So far, there is no concrete explanation for this behavior.

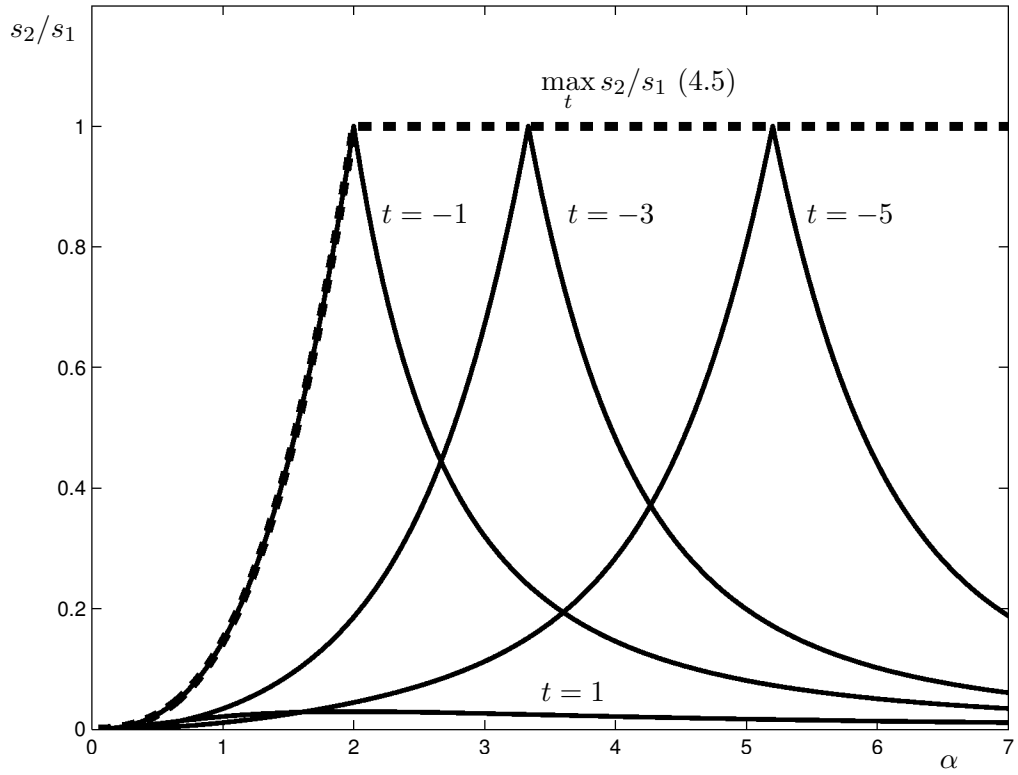

Figure 4.1 : Singular value decay of the solution $\mathbf{X}$ of the Sylvester equation (4.1) with $\mathbf{A}$ as the 2×2 Jordan block (1.33a). For every $\alpha \geq 2$, there is a right-hand side that gives *no* decay. However, for each fixed right-hand side, increasing nonnormality beyond some threshold causes decay to accelerate sharply. This behavior is similar to the decay observed for solutions of the Lyapunov equation illustrated in Figure 1.1. So although (4.5) is the optimal bound across all rank-one $\mathbf{G}$, it appears that a much better bound that is similar to (3.13) but depends on $\mathbf{G}$ could be developed.

4.2 Conclusion

It was observed that previously existing bounds on the convergence rate of ADI for Lyapunov equations are not qualitatively descriptive of the fastest possible convergence rate, i.e., the decay rate of the singular values of the solution. The singular values of $\mathbf{X}$ typically decay very quickly when the coefficient matrix is far from normal, but none of the bounds in Section 1.2 predict this. When the coefficient matrix is far from normal, the new bound of Theorem 3.1 matches the actual singular value

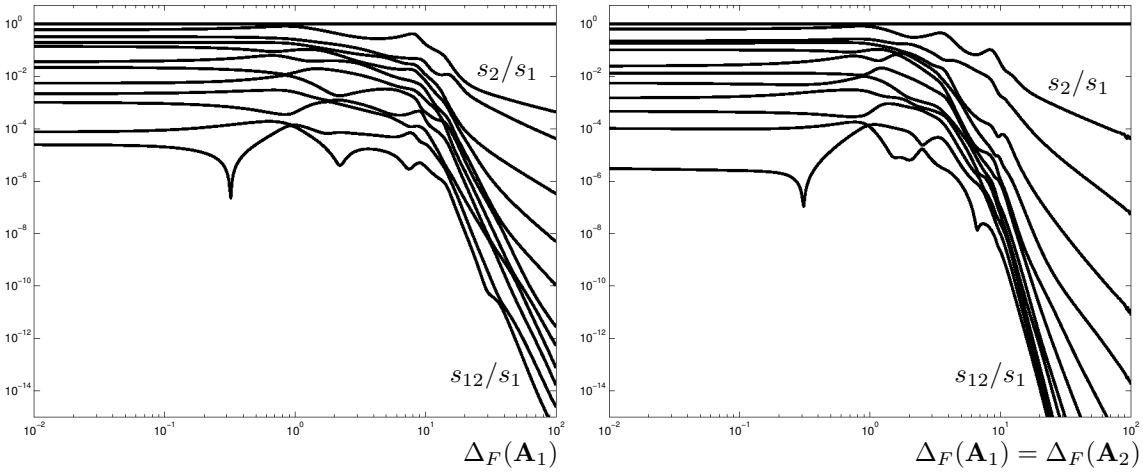

Figure 4.2 : The singular value decay of $\mathbf{X}$ eventually accelerates as one or both coefficient matrices depart from normality. For this figure, $n = 12$; $p = 3$; $\mathbf{A}_1$, $\mathbf{A}_2$, and $\mathbf{B}$ have normally distributed independently random entries; and $\mathbf{G} = -\mathbf{B}\mathbf{B}^*$. The Schur factorization method of (2.15) was applied to each coefficient matrix to create families of Sylvester equations. The off-diagonal scaling parameter is proportional to Henrici's measure of nonnormality (2.5), so the horizontal axes are labelled with $\Delta_F(\mathbf{A}_1)$ and $\Delta_F(\mathbf{A}_2)$. If $\mathbf{A}_1$ departs from normality while $\mathbf{A}_2$ remain fixed (left) or if both coefficients are adjusted together (right), the trailing singular values of $\mathbf{X}$ shrink rapidly in the long run.

decay of the solution better, as illustrated by the examples of Section 3.2.

Several unsolved challenges remain. First, nonnormality typically causes ADI to converge slowly but causes singular values to decay quickly. It remains to find an algorithm to compute nearly optimal low-rank solutions in these cases. Second, it may be possible to improve Theorem 3.1, which can only be highly informative for a few singular values, and was not tight even in an asymptotic sense for the given examples. Furthermore, this bound is derived quite abstractly and does not fully clarify the mechanisms of accelerating singular value decay. Finally, Sylvester equations exhibits similar decay, but Theorem 3.1 has no obvious extension to this case. Further investigation into the subtle effects of coefficient nonnormality may illuminate these issues.

Bibliography

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

[2] J. Baker, M. Embree, and J. Sabino. Fast singular value decay for Lyapunov solutions with nonnormal coefficients. *SIAM J. Matrix Anal. Appl.*, 36(2):656–668, 2015.

[3] R. H. Bartels and G. W. Stewart. Solution of the matrix equation $ax + xb = c$. *Commun. ACM*, 15(9):820–826, September 1972.

[4] P. Benner and P. Kürschner. Computing real low-rank solutions of Sylvester equations by the factored ADI method. *Computers & Mathematics with Applications*, 67(9):1656–1672, 2014.

[5] P. Benner and E. S. Quintana-Ortí. Solving stable generalized Lyapunov equations with the matrix sign function. *Numerical Algorithms*, 20(1):75–100, 1999.

[6] R. Carden and M. Embree. Ritz value localization for non-Hermitian matrices. *SIAM J. Matrix Anal. Appl.*, 33:1320–1338, 2012.

[7] M. Crouzeix. Numerical range and functional calculus in Hilbert space. *J. Functional Analysis*, 244(2):668–690, 2007.

[8] L. Elsner and M. H. C. Paardekooper. On measures of nonnormality of matrices. *Linear Algebra Appl.*, 92:107–123, 1987.

[9] K. Fernando and H. Nicholson. On the structure of balanced and other principal representations of siso systems. *Automatic Control, IEEE Transactions on*, 28(2):228–231, 1983.

[10] R. Grone, C. R. Johnson, E. M. Sa, and H. Wolkowicz. Normal matrices. *Linear Algebra App.*, 87:213–225, 1987.

[11] S. J. Hammarling. Numerical solution of the stable, non-negative definite Lyapunov equation. *IMA J. Numer. Anal.*, 2(3):303–323, 1982.

[12] P. Henrici. Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. *Numer. Math.*, 4(1):24–40, 1962.

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

[14] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[15] D. Y. Hu and L. Reichel. Krylov-subspace methods for the Sylvester equation. *Linear Algebra and its Applications*, 172:283–313, 1992.

[16] T. Penzl. A cyclic low rank Smith method for large sparse Lyapunov equations with applications in model reduction and optimal control. *SIAM J. Sci. Comput.*, 21:1401–1418, 1998.

[17] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Systems & Control Letters*, 40(2):139–144, 2000.

[18] A. Ruhe. Closest normal matrix finally found! *BIT Numerical Mathematics*, 27(4):585–598, 1987.

[19] J. Sabino. *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*. PhD thesis, Rice University, 2006.

[20] V. Simoncini. Computational methods for linear matrix equations. Technical report, University of Bologna, March 2013.

[21] R. A. Smith. Matrix equation $XA+BX = C$. *SIAM J. Applied Math*, 16(1):198–201, 1968.

[22] D. C. Sorensen and Y. Zhou. Bounds on eigenvalue decay rates and sensitivity of solutions to Lyapunov equations. Technical Report TR 02-07, Rice University, Department of Computational and Applied Mathematics, June 2002.

[23] L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, 2005.

[24] A. van der Sluis. Condition numbers and equilibration of matrices. *Numer. Math.*, 14:14–23, 1969.

[25] E. L. Wachspress. Iterative solution of the Lyapunov matrix equation. *Applied Math Letters*, 1:87–90, 1988.

[26] E. L. Wachspress. Alternating direction implicit iteration for systems with complex spectra. *SIAM J. Numer. Anal.*, 28(3):859–870, 1991.