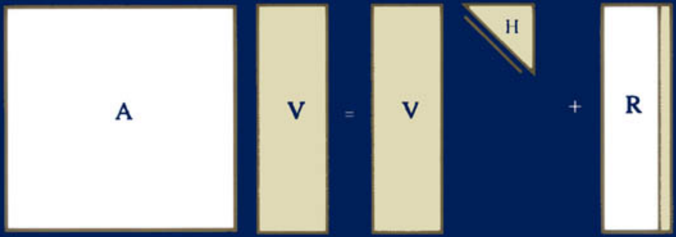
Approximation of Large-Scale Dynamical Systems



Athanasios C. Antoulas

Advances in Design and Control

siam

Approximation of Large-Scale Dynamical Systems

Advances in Design and Control

SIAM's Advances in Design and Control series consists of texts and monographs dealing with all areas of design and control and their applications. Topics of interest include shape optimization, multidisciplinary design, trajectory optimization, feedback, and optimal control. The series focuses on the mathematical and computational aspects of engineering design and control that are usable in a wide variety of scientific and engineering disciplines.

Series Volumes

Robinett, Rush D. III, Wilson, David G., Eisler, G. Richard, and Hurtado, John E., *Applied Dynamic Programming for Optimization of Dynamical Systems*
Huang, J., *Nonlinear Output Regulation: Theory and Applications*
Haslinger, J. and Mäkinen, R. A. E., *Introduction to Shape Optimization: Theory, Approximation, and Computation*
Antoulas, Athanasios C., *Approximation of Large-Scale Dynamical Systems*
Gunzburger, Max D., *Perspectives in Flow Control and Optimization*
Delfour, M. C. and Zolésio, J.-P., *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*
Betts, John T., *Practical Methods for Optimal Control Using Nonlinear Programming*
El Ghaoui, Laurent and Niculescu, Silviu-Iulian, eds., *Advances in Linear Matrix Inequality Methods in Control*
Helton, J. William and James, Matthew R., *Extending H^∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives*

Approximation of Large-Scale Dynamical Systems

Athanasios C. Antoulas
Rice University
Houston, Texas

Society for Industrial and Applied Mathematics
Philadelphia

Figures 2.2, 2.3, 2.4, and 2.5 reprinted courtesy of Martin Verlaan from the thesis *Efficient Kalman Filtering Algorithms for Hydrodynamic Models.*
Figure 2.6 reprinted with permission of Liley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.
Figure 2.7 reprinted courtesy of Draper Laboratories, Houston, TX.
Figure 2.8 reprinted courtesy of Robert Bosch Company.
Figures 2.9 and 2.10 reprinted with permission from P. Benner, "Solving large-scale control problems," Special Issue on Numerical Awareness in Control, *IEEE Control Systems Magazine*, Vol. 24, No. 1, pp. 44–59. © 2004 IEEE.

Library of Congress Cataloging-in-Publication Data

Antoulas, Athanasios C. (Athanasios Constantinos), 1950-
Approximation of large-scale dynamical systems / Athanasios C. Antoulas.
p. cm. – (Advances in design and control)
Includes bibliographical references and index.
ISBN 0-89871-529-6 (pbk.)
1. Differentiable dynamical systems. 2. Differential equations, Partial. 3. System analysis. 4. Approximation theory. 5. Linear systems. I. Title. II. Series.

QA614.9.A58 2005
515'.39–dc22 2005046507

Ἀφιερώνεται στούς γονείς μου
Θᾶνος Κ. Ἀντούλας

Dedicated to my parents
Thanos C. Antoulas

Μέ λογισμό καί μ' ὄνειρο
Οἱ Ἐλεύθεροι Πολιορκημένοι
Σχεδίασμα Γ'
Διονύσιος Σολωμός

With reflection and with vision
The Free Besieged
Third Draft
Dionysios Solomos

This page intentionally left blank

Contents

List of Figures

Foreword

Lectori Salutem.

This book deals with important problems and results on the interface between three central areas of (applied) mathematics: linear algebra, numerical analysis, and system theory. Whereas the first two of these fields have had a very fruitful symbiosis for a long time in the form of numerical linear algebra, the third area has remained somewhat distant from all this, until recently. The book in front of you is an example of this rapprochement.

At center stage in this book, we find the problem of *model reduction*. This problem can be formulated as follows. Given a linear time-invariant input/output system defined by the convolution integral, a transfer function, a state space representation, or their discrete-time counterparts, approximate this system by a simpler system.

As all approximation questions, this problem can be viewed as a trade-off between *complexity* and *misfit*. We are looking for a system of minimal complexity that approximates the original one (optimally) with a maximal allowed misfit, or, conversely, we are looking for a system that approximates the original one with minimal misfit within the class of systems with maximal admissible complexity.

A number of key questions arise:

1. What is meant by the complexity of a linear time-invariant system?
2. How is the complexity computed from the impulse response matrix, the transfer function, or from another system representation?
3. What is meant by the misfit between two systems?
4. How is this misfit computed?
5. What are the algorithms that compute optimal approximants?

The complexity issue leads to the theory of state space representations, commonly called *realization theory*. This theory, originally championed in the work of Kalman, is one of the most beautiful and useful parts of system theory.

The misfit issue leads to an in-depth discussion of matrix, operator, and system norms. Two important system norms emerge: the $\mathcal{L}_2$-induced, or $\mathcal{H}_\infty$-norm, and the *Hankel*-norm. There are numerous inequalities relating matrix norms, and some of these extend to system norms. Perhaps the most remarkable of these relations is the inequality that bounds the $\mathcal{H}_\infty$-norm of a system by twice the sum (without repetition) of its Hankel singular values.

The issue of system approximation centers around the singular values of the associated Hankel operator. Two effective approximation algorithms, both related to these Hankel singular values, are the following:

(i) Approximation by balancing,

(ii) AAK model reduction.

Model reduction by balancing is based on a very elegant method of finding a state space representation of a system with state components that are, so to speak, equally controllable as observable. This reduction method is heuristic, but it is shown that the resulting reduced system has very good properties. AAK model reduction is based on a remarkable result of Arov, Adamjan, and Krein, three Russian mathematicians who proved that a Hankel operator can be approximated equally well within the class of Hankel operators as in the class of general linear operators. While neither balancing nor AAK offers optimal reductions in the all-important $\mathcal{H}_\infty$-norm, very nice inequalities bounding the $\mathcal{H}_\infty$ approximation error can be derived.

Unfortunately, these singular value oriented methods are, computationally, rather complex, requiring of the order of n^3 operations, where n denotes the dimension of the state space of the system to be approximated. Together with accuracy considerations, this makes these methods applicable to systems of only modest dimension (a few hundred variables).

This book also introduces a second set of approximation methods based on moment matching. In system theory language, this moment matching is a generalization of the well-known partial realization problem. These methods can be iteratively implemented using standard algorithms from numerical linear algebra, namely, the Krylov iterative methods. These schemes were originally developed for computing eigenvalues and eigenvectors, but can be applied to model reduction via moment matching. Typically, these methods require only of the order of n^2 operations. Their disadvantage, however, is that stability of the reduced model is not guaranteed, and there is no known global error bound.

This brings us to the last part of the book, which aims at combining the singular value decomposition (SVD) based methods and the Krylov methods into what are called *SVD-Krylov methods.*

The SVD-based approach can be extended to nonlinear systems. The resulting method is known as POD (proper orthogonal decomposition) and is widely used by the PDE community.

The scope of this book is the complete theory of primarily linear system approximation. Special attention is paid to numerical aspects, simulation questions, and practical applications. It is hard to overestimate the importance of the theory presented in this book. I believe that its impact (for example, for numerical simulation of PDEs) has not yet been fully achieved. The mathematical ideas underlying the interplay of the SVD and linear system theory are of the most refined mathematical ideas in the field of system theory.

The book in front of you is unique in its coverage and promises to be a stimulating experience to everyone interested in mathematics and its relevance to practical problems.

Jan C. Willems
Leuven, May 3, 2003

Preface

In today's technological world, physical and artificial processes are mainly described by mathematical models, which can be used for simulation or control. These processes are dynamical systems, as their future behavior depends on their past evolution. The weather and very large scale integration (VLSI) circuits are examples, the former physical and the latter artificial. In simulation (control) one seeks to predict (modify) the system behavior; however, simulation of the full model is often not feasible, necessitating simplification of it. Due to limited computational, accuracy, and storage capabilities, system approximation—the development of simplified models that capture the main features of the original dynamical systems—evolved. Simplified models are used in place of original complex models and result in simulation (control) with reduced computational complexity. This book deals with what may be called the curse of complexity, by addressing the approximation of dynamical systems described by a finite set of differential or difference equations together with a finite set of algebraic equations. Our goal is to present approximation methods related to the singular value decomposition (SVD), to Krylov or moment matching methods, and to combinations thereof, referred to as SVD-Krylov methods.

Part I addresses the above in more detail. Part II is devoted to a review of the necessary mathematical and system theoretic prerequisites. In particular, norms of vectors and (finite) matrices are introduced in Chapter 3, together with a detailed discussion of the SVD of matrices. The approximation problem in the induced 2-norm and its solution given by the Schmidt–Eckart–Young–Mirsky theorem are tackled next. This result is generalized to linear dynamical systems in Chapter 8, which covers Hankel-norm approximation. Elements of numerical linear algebra are also presented in Chapter 3. Chapter 4 presents some basic concepts from linear system theory. Its first section discusses the external description of linear systems in terms of convolution integrals or convolution sums. The section following treats the internal description of linear systems. This is a representation in terms of first-order ordinary differential or difference equations, depending on whether we are dealing with continuous- or discrete-time systems. The associated structural concepts of reachability and observability are analyzed. Gramians, which are important tools for system approximation, are introduced in this chapter and their properties are explored. The last section of Chapter 4 is concerned with the relationship between internal and external descriptions, which is known as the realization problem. Finally, aspects of the more general problem of rational interpolation are displayed.

Chapter 5 introduces various norms of linear systems that are essential for system approximation and for the quantification of approximation errors. The Hankel operator is introduced, and its eigenvalues and singular values, together with those of the convolution operator, are computed. This leads to the concept of Hankel singular values and of the 8-norm of a system. After a more general discussion on induced norms of linear systems, we turn our attention to a brief review of system stability, followed by a discussion of 2-systems and all-pass 2-systems, which play an important role in Hankel-norm system approximation. The concept of dissipativity, which generalizes that of stability from autonomous systems to systems with external influences, is introduced; the special cases of bounded real and positive real systems are briefly explored. Chapter 6 is devoted to the study of two linear matrix equations, the Sylvester equation and the closely related Lyapunov equation. These equations are central to SVD-based methods, and various solution methods (from complex integration to the Cayley–Hamilton theorem to the sign function method) are discussed. Next, the inertia theorem for the Lyapunov equation is investigated. The chapter concludes with numerically reliable algorithms for the solution of these equations.

Following this presentation of the preparatory material, Part III commences with the exposition of the first class of approximation methods, namely, SVD-based approximation methods. Chapter 7 is devoted to approximation by balanced truncation. The ingredients are Lyapunov equations and the Hankel singular values. The main result is followed by a canonical form that can be applied to balanced systems. The last part of the chapter is involved with special types of balancing, including bounded real, positive real, frequency weighted balancing, which lead to methods for approximating unstable systems.

Chapter 8 presents the theory of optimal and suboptimal approximation in the induced 2-norm of the Hankel operator, which can be viewed as a refinement of approximation by balanced truncation. The final section of this chapter is devoted to the exposition of a polynomial approach that offers new insights into and connections between balancing and Hankel-norm approximation. Part III concludes with a chapter dedicated to special topics in SVD-based approximation methods. In this context, a brief description of the proper orthogonal decomposition method is given in section 9.1; its relation with balanced truncation is also mentioned. Approximation by modal truncation is discussed next. The latter part of the chapter is dedicated to a study of the decay rates of the Hankel singular values, which is of importance in predicting how well a given system can be approximated by a low-order system.

Part IV is concerned with Krylov-based approximation methods. These methods have their roots in numerical linear algebra and address the problem of providing good estimates for a few eigenvalues of a big matrix. Consequently, Chapter 10 gives an account of Lanczos and Arnoldi methods as they apply to eigenvalue problems. Chapter 11 discusses the application of these methods to system approximation. Krylov methods lead to approximants by matching moments. These methods turn out to be numerically efficient, although they lack other important properties, such as quantification of the approximation error. The connection with rational interpolation is discussed in some detail.

The final section, Part V, is concerned with the connections between SVD-based and Krylov-based approximation methods. In particular, a method that involves least squares combines attributes of both approaches. Furthermore, two iterative methods are presented that provide approximate solutions to Lyapunov equations and therefore can be used to obtain reduced-order systems that are approximately balanced. In Chapter 13, aspects of the approximation methods presented earlier are illustrated by means of numerical experiments, by applying them to various systems. Algorithms are compared in terms of both approximation error and computational effort. The book concludes with a chapter on projections, computational complexity, software, and open problems, followed by a collection of exercises (Chapter 15) appropriate for classroom use.

At this stage we would like to point to several related books: Obinata and Anderson [252] (model reduction for control system design, in particular, controller reduction); Gawronski [136] (model reduction of flexible structures); Fortuna, Nunnari, and Gallo [115] (model reduction with applications in electrical engineering); Datta [91] (comprehensive treatment of numerical issues in systems and control); Berkooz, Holmes, and Lumley [63] (dynamical systems described by partial differential equations); Banks [42] (control and estimation for distributed parameter systems); and Zhou et al. [370], [371] (comprehensive treatment of systems and control with emphasis on robustness issues). See also the special issue of *Control Systems Magazine* [347] on numerical awareness in control as well the collection of reprints [262] on numerical linear algebra and control. Insights into aspects mentioned earlier but not discussed further in this book can be found in the book by Skelton, Grigoriadis, and Iwasaki [299] (controller complexity) and in the surveys by Tempo and Dabbene [323] (randomized algorithms) and Blondel and Tsitsiklis [66] (complexity of algorithms in system theory).

There are numerous individuals without whose help this book would not have been completed. First, I would like to thank Jan Willems for his friendship and inspiration for more than a quarter-century. I would also like to thank Paul van Dooren for his advice on the book while I was visiting Louvain-la-Neuve—but I am even more thankful to him for introducing me to Dan Sorensen. Dan took it on himself to teach me numerical linear algebra, and not some watered-down version but the real deal. Meeting Dan was an event that added a new dimension to my research. Dan is also acknowledged for contributing part of the section on the decay rates. Many thanks go to Mark Embree for numerous discussions on pseudospectra and for his very careful reading and substantial comments on several chapters. The next recipient of my gratitude is Paolo Rapisarda, who was always willing to read critically and provide invaluable advice over extensive portions of the book. I would also like to thank Angelika Bunse-Gerstner, Peter Benner, and Caroline Boss for their comments. Next, my acknowledgements go to Yutaka Yamamoto, longtime friend and colleague from our years as graduate students under R. E. Kalman. Thanks also go to Brian Anderson, Alessandro Astolfi, Siep Weiland, Roberto Tempo, Matthias Heinkenschloss, Yunkai Zhou, Michael Hinze, and Stefan Volkwein for reading and commenting on various parts of the book. I would also like to thank the anonymous referees and several students who commented on the book at various stages. Special thanks go to Serkan Gugercin, who contributed in many ways over the last 3 years and to whom most numerical experiments and figures, as well as part of the last two chapters, are due. Finally, I would like to thank the editors at SIAM for their professional and efficient handling of this project.

Athanasios C. Antoulas

This page intentionally left blank

How to Use This Book

- A first course in model reduction would consist of the following sections:

  - Chapter 3, sections 3.1, 3.2.1–3.2.5, 3.3
  - Chapter 4, sections 4.1, 4.2, 4.3
  - Chapter 5, sections 5.1, 5.2, 5.3, 5.4, 5.5, 5.8.1, 5.8.2
  - Chapter 6
  - Chapter 7, sections 7.1, 7.2, 7.3
  - Chapter 10, sections 10.1, 10.3, 10.4
  - Chapter 11, sections 11.1, 11.2

- **Prerequisites**

  The most important prerequisite is familiarity with linear algebra. Knowledge of elementary system theory and numerical analysis is desirable.

  The target readership consists of graduate students and researchers in the fields of system and control theory, numerical analysis, theory of partial differential equations and computational fluid dynamics and anyone interested in model reduction.

- **Sections omitted at first reading**

  Sections marked with an asterisk contain material that can be omitted at first reading.

- **Notation**

$\mathbb{R}$ ($\mathbb{R}_+, \mathbb{R}_-$)	real numbers (positive, negative)	
$\mathbb{C}$ ($\mathbb{C}_+, \mathbb{C}_-$)	complex numbers (with positive, negative real part)	
$\mathbb{Z}$ ($\mathbb{Z}_+, \mathbb{Z}_-$)	integers (positive, negative)	
$\lambda_{\max}(\mathbf{M})$	largest (in magnitude) eigenvalue of $\mathbf{M} \in \mathbb{R}^{n\times n}$	
$\delta_{\max}(\mathbf{M})$	largest (in magnitude) diagonal entry of $\mathbf{M} \in \mathbb{R}^{n\times n}$	
adj $\mathbf{M}$	adjoint (matrix of cofactors) of $\mathbf{M} \in \mathbb{R}^{n\times n}$	
$\chi_M(s)$	characteristic polynomial of $\mathbf{M} \in \mathbb{R}^{n\times n}$	
rank $\mathbf{M}$	rank of $\mathbf{M} \in \mathbb{R}^{n\times m}$	
$\sigma_i(M)$	ith singular value of $\mathbf{M} \in \mathbb{R}^{n\times m}$	
$\kappa(\mathbf{M})$	condition number of $\mathbf{M} \in \mathbb{R}^{n\times m}$	
$\mathbf{M}^*$	transpose if $\mathbf{M} \in \mathbb{R}^{n\times m}$	
	complex conjugate transpose if $\mathbf{M} \in \mathbb{C}^{n\times m}$	
$\mathbf{M}_{(:,j)}$	jth column of $\mathbf{M} \in \mathbb{R}^{n\times m}$	
$\mathbf{M}_{(i,:)}$	ith row of $\mathbf{M} \in \mathbb{R}^{n\times m}$	
spec ($\mathbf{M}$)	spectrum (set of eigenvalues) of $\mathbf{M} \in \mathbb{R}^{n\times n}$	
spec$_\epsilon$ ($\mathbf{M}$)	ϵ-pseudo-spectrum of $\mathbf{M} \in \mathbb{R}^{n\times n}$	
in ($\mathbf{A}$)	inertia of $\mathbf{A} \in \mathbb{R}^{n\times n}$	(6.14)
$\boldsymbol{\Sigma}$	general dynamical system	(1.1)
$\mathbb{U}, \mathbb{X}, \mathbb{Y}$	input, state, output spaces	(1.2)
$\boldsymbol{\Sigma} = \left(\begin{array}{c\|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$	linear system matrices	(1.4), (4.13)
$\boldsymbol{\Sigma} = \left(\begin{array}{c\|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$	linear system with missing $\mathbf{D}$	(4.15)
$\boldsymbol{\Sigma} = \left(\begin{array}{c\|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$	linear system with missing $\mathbf{C}$ and $\mathbf{D}$	section 4.2.1
$\boldsymbol{\Sigma} = \left(\begin{array}{c\|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$	linear system with missing $\mathbf{B}$ and $\mathbf{D}$	section 4.2.2
$\boldsymbol{\Sigma}_+$	stable subsystem of $\boldsymbol{\Sigma}$	section 5.8.3, Chapter 8
$\boldsymbol{\Sigma}^*$	dual or adjoint of $\boldsymbol{\Sigma}$	section 4.2.3, (5.15)
Π	projection	(1.6)
$\hat{\boldsymbol{\Sigma}}$	reduced-order dynamical system	(1.7)
$\hat{\boldsymbol{\Sigma}} = \left(\begin{array}{c\|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \hat{\mathbf{D}} \end{array}\right)$	system matrices of linear reduced-order system	(1.8)
$\mathcal{S}$	convolution operator (discrete and continuous time)	
$\mathcal{S}^*$	adjoint of convolution operator	(5.14)
$\mathbf{h}_k \in \mathbb{R}^{p\times m}$	Markov parameters	(4.7), (4.23)
$\mathbf{h}(t)$	impulse response discrete-time systems	(4.3) and (4.21)
	impulse response continuous-time systems	(4.5) and (4.20)
$\mathbf{H}(s)$	transfer function discrete and continuous time	(4.8), (4.22)
$\mathbf{H}_+(s)$	transfer function of stable subsystem of Σ	section 5.8.3, Chapter 8
$\\|\mathbf{x}\\|_p$	p-norm of $\mathbf{x} \in \mathbb{R}^n$	(3.2)
$\\|\mathbf{A}\\|_{p,q}$	(p, q)-induced norm of $\mathbf{A}$	(3.4)
$fl(x)$	floating point representation of $x \in \mathbb{R}$	(3.21)
$e^{\mathbf{M}t}$	matrix exponential	(4.16)

$\mathcal{R}(\mathbf{A}, \mathbf{B})$	reachability matrix	(4.25)
$\mathcal{R}_n(\mathbf{A}, \mathbf{B})$	finite reachability matrix	(4.26)
$\mathcal{P}(t)$	(finite) reachability gramian continuous, discrete time	(4.28), (4.29)
$\mathcal{O}(\mathbf{C}, \mathbf{A})$	observability matrix	(4.25)
$\mathcal{O}_n(\mathbf{C}, \mathbf{A})$	finite observability matrix	(4.26)
$\mathcal{Q}(t)$	(finite) observability gramian continuous, discrete time	(4.40), (4.41)
$\mathcal{P}$	infinite reachability gramian continuous, discrete time	(4.43), (4.47)
	in the frequency domain	(4.51)
$\mathcal{Q}$	infinite reachability gramian continuous, discrete time	(4.44), (4.48)
	in the frequency domain	(4.52)
$\mathcal{X}$	cross gramian, continuous-time case	(4.60)
$\mathcal{H}$	Hankel matrix	(4.63)
$\mathbb{P}$	array of pairs of points for rational interpolation	(4.79), (4.80)
$\mathbb{L}$	Löwner matrix	(4.84)
$\ell(s)$	Lagrange interpolating polynomial	(4.81)
$\mathcal{R}_r$	generalized reachability matrix	(4.85)
$\mathcal{O}_p$	generalized observability matrix	(4.86)
$\mathbb{D}$	time series associated with an array	(4.93), (4.97)
Θ	generating system/Lyapunov function/storage function	(4.101), sections 5.8.2 and 5.9
$\mathbf{s}$	supply function	section 5.9
$\langle \cdot , \cdot \rangle$	inner product	(5.1)
$\ell_p^n(\mathcal{I}), \mathcal{L}_p^n(\mathcal{I})$	Lebesgue spaces of functions	section 5.1.2
$h_p^{q\times r}, \mathcal{H}_p^{q\times r}$	Hardy spaces of functions	section 5.1.3
$\Vert\mathbf{F}\Vert_{h_\infty}, \Vert\mathbf{F}\Vert_{\mathcal{H}_\infty}$	h_∞-norm, $\mathcal{H}_\infty$-norm	(5.5), (5.7)
$\Vert\mathbf{F}\Vert_{\ell_\infty}, \Vert\mathbf{F}\Vert_{\mathcal{L}_\infty}$	ℓ_∞-, $\mathcal{L}_\infty$-norm	(5.10)
$\Vert\boldsymbol{\Sigma}\Vert_2$	2-norm of the system $\boldsymbol{\Sigma}$	(5.16)
$\Vert\mathcal{S}\Vert_{2-\text{ind}}$	2-induced norm of the convolution operator $\mathcal{S}$	(5.16)
$\Vert\mathbf{H}\Vert_{\mathcal{H}_\infty}$	$\mathcal{H}_\infty$-norm of the transfer function	(5.16)
$\mathcal{H}$	Hankel operator of $\boldsymbol{\Sigma}$: discrete, continous time	(5.19), (5.20)
$\mathcal{H}^*$	adjoint of continuous-time Hankel operator	(5.23)
$\sigma_i(\boldsymbol{\Sigma})$	Hankel singular values	(5.22), (5.24)
$\Vert\boldsymbol{\Sigma}\Vert_H$	Hankel-norm of $\boldsymbol{\Sigma}$	Def. 5.7, formula (5.21)
$\Vert\mathcal{H}\Vert_{2-\text{ind}}$	2-induced norm of Hankel operator	Def. 5.7, formula (5.21)
$\Vert\boldsymbol{\Sigma}\Vert_{\mathcal{H}_2}$	$\mathcal{H}_2$-norm of $\boldsymbol{\Sigma}$	(5.27), (5.28), (5.29)
$\vert\vert\vert\ \mathbf{f}\ \vert\vert\vert_{(p,q)}$	mixed norm of vector-valued functions of time	(5.30)
$\vert\vert\vert\ \mathcal{T}\ \vert\vert\vert_{(p,q)}^{(r,s)}$	mixed induced operator norm	(5.32)
$\mathcal{P}_i$	reachability input-weighted gramian	(7.33)
$\mathcal{Q}_o$	observability output-weighted gramian	(7.34)
$\mathcal{P}(\omega)$	frequency-limited reachability gramian	(7.37)
$\mathcal{Q}(\omega)$	frequency-limited observability gramian	(7.38)
$\mathbf{S}(\omega)$	integral of log of resolvent $(s\mathbf{I} - \mathbf{A})^{-1}$	(7.40)
$\mathcal{P}(T)$	time-limited reachability gramian	(7.41)
$\mathcal{Q}(T)$	time-limited observability gramian	(7.41)
$\mathcal{C}(\mathbf{x}, \mathbf{y})$	Cauchy matrix	(9.2)
$\mathbf{d}(\mathbf{x}, \mathbf{y})$	diagonal of Cauchy matrix	(9.3)
$\lvert\gamma\rvert \prec_\mu \sigma$	multiplicative majorization	(9.10)
$\nabla \rho$	gradient	section 10.4.1
$\eta_k, \eta_k(s_0)$	moments, generalized moments of $\mathbf{h}$	(11.2), (11.4)

This page intentionally left blank

Part I

Introduction

This page intentionally left blank

Chapter 1

Introduction

In today's technological world, physical as well as artificial processes are described mainly by mathematical models. These models can be used to simulate the behavior of the processes in question. Sometimes, they are also used to modify or control the processes' behavior. The weather, on the one hand, and very large scale integration (VLSI) circuits, on the other, constitute examples of such processes, the former physical and the latter artificial. Furthermore, these are *dynamical systems*, as their future behavior depends on their past evolution. In this framework of mathematical models, there is an ever-increasing need for improved accuracy, which leads to models of high complexity.

The basic motivation for system approximation is the need for simplified models of dynamical systems, which capture the main features of the original complex model. This need arises from limited computational, accuracy, and storage capabilities. The simplified model is then used in place of the original complex model, for either *simulation* or *control*.

In the former case, simulation, one seeks to predict the system behavior. However, often simulation of the full model is not feasible. Consequently, an appropriate simplification of this model is necessary, resulting in simulation with reduced computational complexity. Prominent examples include weather prediction and air quality simulations. The complexity of models, measured in terms of the number of coupled first-order differential or difference equations, may reach the tens or hundreds of thousands. In particular, discretization in problems that arise from dynamical partial differential equations (PDEs) which evolve in three spatial dimensions can easily lead to 1 million equations. In such cases, reduced simulation models are essential for the quality and timeliness of the prediction. Other methods for accelerating the simulation time exist, like *parallelization* of the corresponding algorithm. (These aspects, however, are not addressed in this book.)

In the latter case, control, we seek to modify the system behavior to conform with certain desired performance specifications (e.g., we seek to control a CD player to decrease its sensitivity to disturbances (outside shocks)). Such modifications are achieved in the vast majority of cases by interconnecting the original system with a second dynamical system, called the *controller*. Generically, the complexity of the controller (the number of first-order differential or difference equations describing its behavior) is approximately the same as that of the system to be controlled. Hence, if the latter has high complexity, so will the controller. This, however, has three potential problems: *storage*—it may be hard to implement a high-order controller on a chip; *accuracy*—due to computational considerations (ill-conditioning), it may be impossible to compute such a high-order controller with any degree of accuracy; *computational speed*—due to limited computational speed, the time needed to compute the parameters of such a controller may be prohibitively large. The design of reduced-order controllers is a challenging problem, aspects of which have been investigated at least for systems of not-too-high complexity (see [252]). This will not be addressed in what follows. (See section 14.2 on open problems.)

In a broader context, this book deals with what is called the *curse of dimensionality* or, paraphrasing, the *curse of complexity*. In the computer science community, efficient algorithms are those that can be executed in polynomial time, that is, algorithms whose execution time grows polynomially with the size of the problem.

Here are examples of problems that can be solved in polynomial time (the complexity of a generic problem is of the order n^γ, where n is the size of the problem):

- set of linear equations, $\gamma = 3$;
- eigenvalue problem, $\gamma = 3$;
- linear matrix inequalities (LMI), $\gamma \approx 4 \cdots 6$.

On the other hand, the problems

- factorization of an integer into prime factors, and
- stabilization of a linear system with constant output feedback

can be solved in *exponential* but not polynomial time.

In our framework there are two additional constraints. First, the algorithm, besides being *efficient* in the sense mentioned above, must produce an *answer* in a given amount of time. This becomes problematic for sufficiently large complexities even if the underlying algorithm is polynomial in time. Second, the solution must be accurate enough, which is a problem given that numbers can be represented in a computer only with finite precision. We will see later on that a popular method for model reduction of dynamical systems, balanced truncation, requires of the order n^3 operations, where n is the complexity of the system to be approximated. It will also follow that methods that reduce the number of operations to $k \cdot n^2$ or $k^2 \cdot n$, where k is the complexity of the reduced model, represent considerable improvements. The reason is that with order n^3 operations one can deal with system complexities of a few hundred states, while with $k^2 \cdot n$, the complexity of the systems that can be dealt with climbs into the millions. It should be mentioned at this point that while the available *computing power* increases, this turns out to be a mixed blessing, since with increased computing power, the *numerical errors* increase as well.

There are numerous remedies against the curse of dimensionality. *Randomized algorithms* are one instance; the solution obtained, however, may fail to satisfy all constraints for some specified percentage of the cases. Other kinds of remedies mostly applicable to problems with polynomial time solution algorithms but very large complexity are *parallelization* methods, as mentioned earlier.

This book addresses the approximation of dynamical systems that are described by a finite set of differential or difference equations, together with a finite set of algebraic equations. Our goal is to present approximation methods related to the singular value decomposition (SVD), on one hand, and approximation methods related to Krylov or moment matching concepts, on the other. Roughly speaking, the former family preserves important properties of the original system, like stability, and in addition provides an explicit quantization of the approximation error. The latter family lacks these properties but leads to methods that can be implemented in a numerically much more efficient way. Thus, while the former family of methods can be applied to relatively low-dimensional systems (a few hundred states), the latter methods are applicable to problems whose complexity can be several orders of magnitude higher. The combination of these two basic approximation methods leads to a third one, which aims at merging their salient features and is referred to as *SVD-Krylov*-based approximation.

Finally, we present some thoughts that have guided our choice of topics concerning the dilemma *linear* versus *nonlinear*. The basic argument used is that real systems are nonlinear, and therefore methods addressing nonlinear system approximation should be primarily considered. However, we espouse the following arguments:

- All physical systems are locally linear; in applications, typically one linearizes around an operating point of interest. If the operating point cannot be fixed, linear time-varying models or piecewise linear models can be considered.
- Many physical laws, e.g., Newton's second law, Maxwell's equations, Kirchhoff's voltage laws (KVL), Kirchhoff's current laws (KCL), the diffusion equation, the wave equation, Schrödinger's equation, and probability laws (Markov equations), are linear, and this linearity holds for large ranges of the operating conditions.
- Linear theory is rich and extensive and offers a coherent picture.
- Artificial systems are sometimes designed to be linear.
- There are attempts in developing a nonlinear approximation theory, but they remain mostly ad hoc.

This book is dedicated to the presentation of primarily linear theory.

1.1 Problem set-up

The broader framework of the problems to be investigated is shown in Figure 1.1. The starting point is a physical or artificial system together with measured data. The *modeling phase* consists of deriving a set of ordinary differential equations (ODEs) or partial differential equations (PDEs). In the latter case, the equations are typically discretized in the space variables leading to a system of ODEs. This system will be denoted by $\mathbf{\Sigma}$. The *model reduction* step consists in developing a dynamical system $\hat{\mathbf{\Sigma}}$ by appropriately reducing the number of ODEs describing the system. $\hat{\mathbf{\Sigma}}$ is now used to simulate and possibly control $\mathbf{\Sigma}$. Sometimes the ODEs are discretized in time as well, yielding discrete-time dynamical systems.

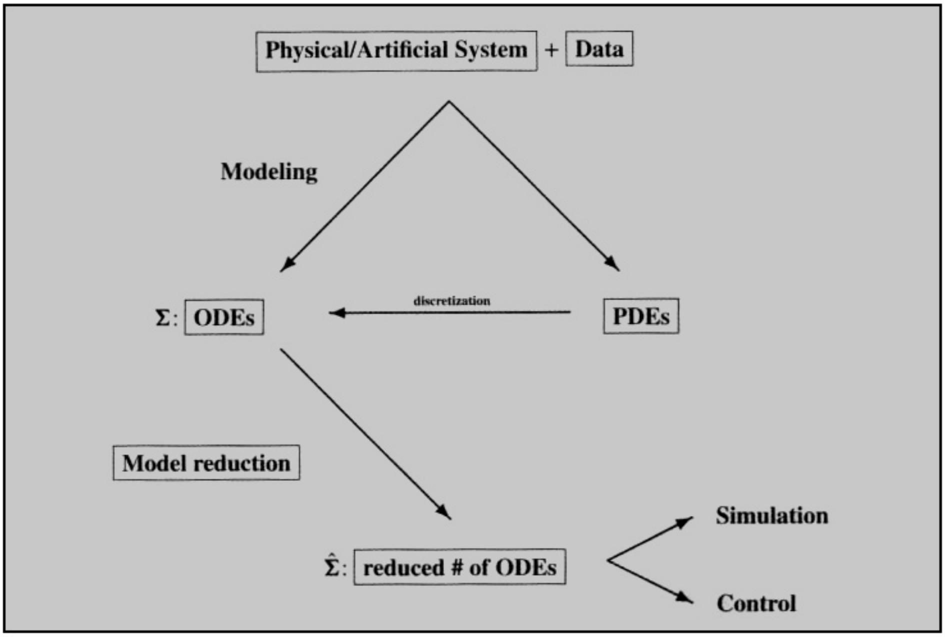


Figure 1.1. *The broad set-up.*

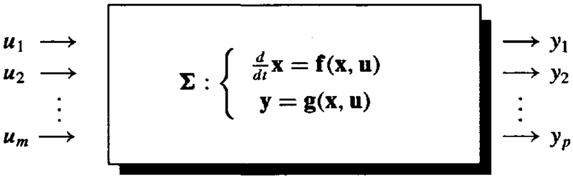

u_1
u_2
u_m
$\Sigma : \begin{cases} \frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \\ \mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{u}) \end{cases}$
y_1
y_2
y_p

Figure 1.2. *Explicit finite-dimensional dynamical system.*

Next, we will formalize the notion of a dynamical system $\boldsymbol{\Sigma}$ (see Figure 1.2). First, the time axis $\mathbb{T}$ is needed; we will assume for simplicity that $\mathbb{T} = \mathbb{R}$, the real numbers; other choices are $\mathbb{R}_+$ ($\mathbb{R}_-$), the positive (negative) real numbers, or $\mathbb{Z}_+$, $\mathbb{Z}_-$, $\mathbb{Z}$, the positive, negative, and all integers. In what follows we will assume that $\boldsymbol{\Sigma}$ consists of first-order ODEs together with the set of algebraic equations

$$\boxed{\boldsymbol{\Sigma} : \begin{cases} \frac{d}{dt}\mathbf{x} & = & \mathbf{f}(\mathbf{x},\ \mathbf{u}), \\ \mathbf{y} & = & \mathbf{g}(\mathbf{x},\ \mathbf{u}), \end{cases}} \tag{1.1}$$

where

$$\boxed{\mathbf{u} \in \mathbb{U} = \{\mathbf{u} : \ \mathbb{T} \to \mathbb{R}^m\}, \quad \mathbf{x} \in \mathbb{X} = \{\mathbf{x} : \ \mathbb{T} \to \mathbb{R}^n\}, \quad \mathbf{y} \in \mathbb{Y} = \{\mathbf{y} : \ \mathbb{T} \to \mathbb{R}^p\},}$$

(1.2)

are the input (excitation), an internal variable (usually the state), and the output (observation), respectively, while $\mathbf{f}$, $\mathbf{g}$ are vector-valued maps of appropriate dimensions. The *complexity* n of such systems is measured by the number of internal variables involved (assumed finite); that is, n is the size of $\mathbf{x} = (x_1 \cdots x_n)^*$.[1] Thus we will be dealing with dynamical systems that are finite-dimensional and described by a set of explicit first-order differential equations; the description is completed with a set of measurement or observation variables $\mathbf{y}$. Also, within the same framework, systems whose behavior is discrete in time can be treated equally well. In this case $\mathbb{T} = \mathbb{Z}$, the set of integers (or $\mathbb{Z}_+$, or $\mathbb{Z}_-$), and the first equation in (1.1) is replaced by the difference equation $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t))$. In what follows, however, we concentrate primarily on continuous-time systems. Often in practice, this explicit nonlinear system is linearized around some equilibrium trajectory (fixed point), with the resulting system being linear, parameter time-varying, and denoted by $\mathbf{\Sigma}_{\text{LPTV}}$. Finally, if this trajectory happens to be stationary (independent of time), we obtain a linear, time-invariant, system, denoted by $\mathbf{\Sigma}_{\text{LTI}}$:

$$\begin{aligned} \mathbf{\Sigma}_{\text{LPTV}} &: \left\{ \begin{array}{l} \dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \\ \mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t), \end{array} \right. \\ \mathbf{\Sigma}_{\text{LTI}} &: \left\{ \begin{array}{l} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t). \end{array} \right. \end{aligned} \tag{1.3}$$

A more general class of systems is obtained if we assume that the first equation in (1.1) is *implicit* in the derivative of $\mathbf{x}$, that is, $\mathbf{F}(\frac{d}{dt}\mathbf{x}, \mathbf{x}, \mathbf{u}) = 0$, for an appropriate vector-valued function $\mathbf{F}$. Such systems are known as *differential algebraic equation* (DAE) systems. Besides its occasional mention (see, e.g., Remark 11.3.1), this more general class of systems will not be addressed. See the book by Brenan, Campbell, and Petzold [74] for an account of this class of systems.

Problem statement. Given $\mathbf{\Sigma} = (\mathbf{f}, \mathbf{g})$ with $\mathbf{u} \in \mathbb{U}$, $\mathbf{x} \in \mathbb{X}$, $\mathbf{y} \in \mathbb{Y}$, find

$$\hat{\mathbf{\Sigma}} = (\hat{\mathbf{f}}, \hat{\mathbf{g}}), \quad \mathbf{u} \in \mathbb{U}, \ \mathbf{y} \in \mathbb{Y}, \ \text{ and } \ \hat{\mathbb{X}} = \{\hat{\mathbf{x}} : \mathbb{T} \to \mathbb{R}^k\}, \ \text{ where } k < n,$$

such that (some of or all) the following conditions are satisfied:

(**1**) The approximation error is small—existence of global error bound

(**2**) Stability and passivity are preserved

(**3**) The procedure is computationally stable and efficient

(COND)

Special case: linear dynamical systems. If we consider linear, time-invariant dynamical systems $\mathbf{\Sigma}_{\text{LTI}}$ as in (1.3), denoted by

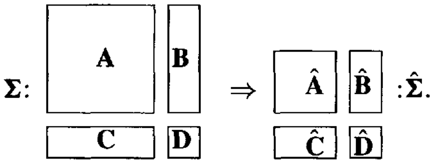
$$\boxed{\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array} \right) \in \mathbb{R}^{(n+p)\times(n+m)},} \tag{1.4}$$

[1] Given a vector or matrix with real entries, the superscript $*$ denotes its *transpose*. If the entries are complex, the same superscript denotes *complex conjugation with transposition*.

the **problem** consists in approximating $\mathbf{\Sigma}$ with

$$\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \hat{\mathbf{D}} \end{array}\right) \in \mathbb{R}^{(k+p)\times(k+m)}, \qquad k < n, \tag{1.5}$$

so that the above conditions are satisfied. Pictorially, we have the following:



That is, only the size of $\mathbf{A}$ is reduced, while the number of columns of $\mathbf{B}$ and the number of rows of $\mathbf{C}$ remain unchanged.

One possible *measure* for judging how well $\hat{\mathbf{\Sigma}}$ approximates $\mathbf{\Sigma}$ consists of comparing the outputs $\mathbf{y}$ and $\hat{\mathbf{y}}$ obtained by using the same excitation function $\mathbf{u}$ on $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$, respectively. We require, namely, that the size (norm) of the *worst* output error $\mathbf{y} - \hat{\mathbf{y}}$ be kept small or even minimized for all normalized inputs $\mathbf{u}$. Assuming that $\mathbf{\Sigma}$ is stable (that is, all eigenvalues of $\mathbf{A}$ are in the left half of the complex plane), this measure of fit is known as the $\mathcal{H}_\infty$-norm of the error. We will also make use of another norm for measuring approximation errors, the so-called $\mathcal{H}_2$-norm. This turns out to be the norm of the impulse response of the error system. For details on norms of linear systems, see Chapter 5.

1.1.1 Approximation by projection

Projections constitute a *unifying feature* of the approximation methods discussed in what follows. This feature is equivalent to simple *truncation* in an appropriate basis. Consider the change of basis $\mathbf{T} \in \mathbb{R}^{n\times n}$ in the state space $\bar{\mathbf{x}} = \mathbf{T}\mathbf{x}$. We define the following quantities by partitioning $\bar{\mathbf{x}}$, $\mathbf{T}$, and $\mathbf{T}^{-1}$:

$$\bar{\mathbf{x}} = \begin{pmatrix} \hat{\mathbf{x}} \\ \tilde{\mathbf{x}} \end{pmatrix}, \; \mathbf{T}^{-1} = [\mathbf{V} \;\; \mathbf{T}_1], \; \mathbf{T} = \begin{bmatrix} \mathbf{W}^* \\ \mathbf{T}_2^* \end{bmatrix}, \; \text{where } \hat{\mathbf{x}} \in \mathbb{R}^k, \; \tilde{\mathbf{x}} \in \mathbb{R}^{n-k}, \;\; \mathbf{V}, \; \mathbf{W} \in \mathbb{R}^{n\times k}.$$

Since $\mathbf{W}^*\mathbf{V} = \mathbf{I}_k$, it follows that

$$\boxed{\Pi = \mathbf{V}\mathbf{W}^* \in \mathbb{R}^{n\times n}} \tag{1.6}$$

is an *oblique projection* onto the k-dimensional subspace spanned by the columns $\mathbf{V}$ along the kernel of $\mathbf{W}^*$.

Substituting for $\mathbf{x}$ in (1.1) we obtain $\frac{d}{dt}\bar{\mathbf{x}} = \mathbf{T}\mathbf{f}(\mathbf{T}^{-1}\bar{\mathbf{x}}, \mathbf{u})$ and $\mathbf{y} = \mathbf{g}(\mathbf{T}^{-1}\bar{\mathbf{x}}, \mathbf{u})$. Retaining the first k differential equations leads to

$$\frac{d}{dt}\hat{\mathbf{x}} = \mathbf{W}^*\mathbf{f}(\mathbf{V}\hat{\mathbf{x}} + \mathbf{T}_1\tilde{\mathbf{x}}, \; \mathbf{u}), \;\; \mathbf{y} = \mathbf{g}(\mathbf{V}\hat{\mathbf{x}} + \mathbf{T}_1\tilde{\mathbf{x}}, \mathbf{u}).$$

These equations describe the evolution of the k-dimensional trajectory $\hat{\mathbf{x}}$ in terms of $\tilde{\mathbf{x}}$; notice that they are *exact*. The *approximation* occurs by neglecting the term $\mathbf{T}_1\tilde{\mathbf{x}}$. What results is a

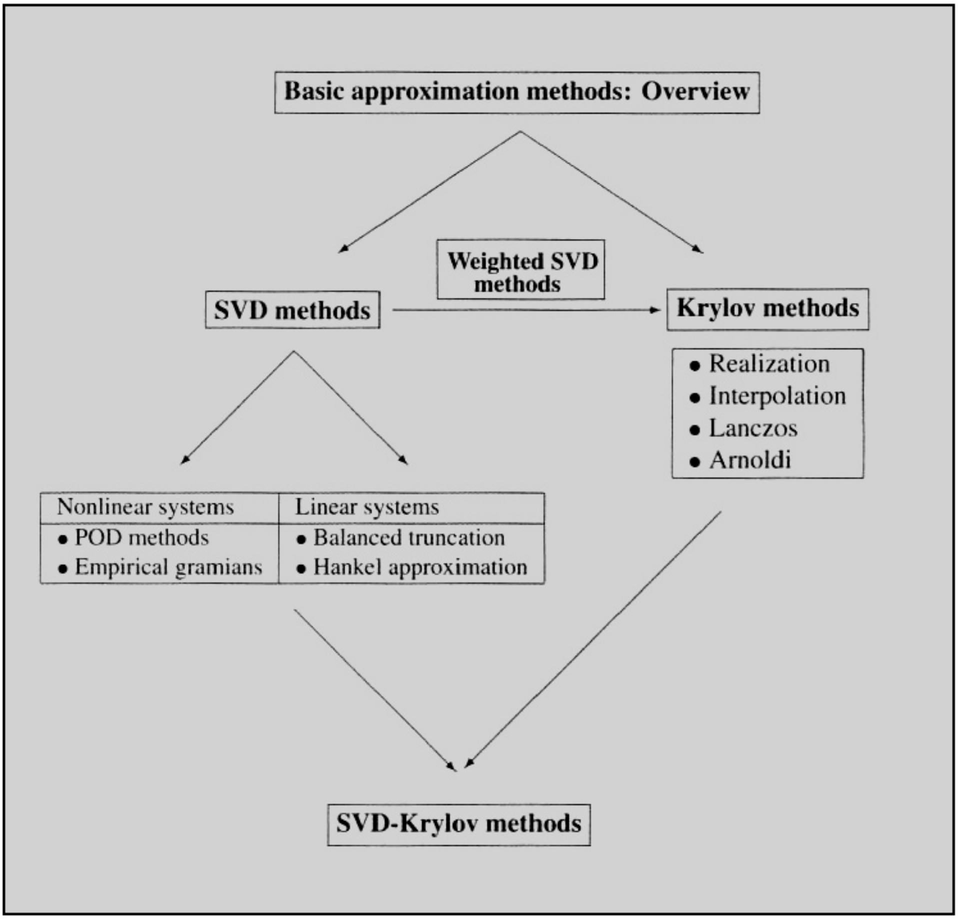


Figure 1.3. *Flowchart of approximation methods and their interconnections.*

dynamical system that evolves in a k-dimensional subspace obtained by restricting the full state as follows: $\hat{\mathbf{x}} = \mathbf{W}^*\mathbf{x}$. The resulting *approximant* $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ is

$$\hat{\Sigma} : \left\{ \begin{array}{rcl} \frac{d}{dt}\hat{\mathbf{x}}(t) & = & \mathbf{W}^*\mathbf{f}(\mathbf{V}\hat{\mathbf{x}}(t), \mathbf{u}(t)), \\ \mathbf{y}(t) & = & \mathbf{g}(\mathbf{V}\hat{\mathbf{x}}(t), \mathbf{u}(t)). \end{array} \right. \tag{1.7}$$

Consequently, if $\hat{\boldsymbol{\Sigma}}$ is to be a "good" approximation of $\boldsymbol{\Sigma}$, the influence of the *neglected* term $\mathbf{T}_1\tilde{\mathbf{x}}$ must be "small" in some appropriate sense. In the linear time-invariant case the resulting approximant is

$$\hat{\Sigma} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \mathbf{D} \end{array} \right) = \left(\begin{array}{c|c} \mathbf{W}^*\mathbf{A}\mathbf{V} & \mathbf{W}^*\mathbf{B} \\ \hline \mathbf{C}\mathbf{V} & \mathbf{D} \end{array} \right). \tag{1.8}$$

A flowchart of the contents of this book is given in Figure 1.3. Proper orthogonal decomposition (POD) and SVD are concepts that will be explained later.

1.2 Summary of contents

The purpose of this book is to describe certain families of approximation methods by projection. Three different methods are discussed for choosing the projection Π (that is, the matrices $\mathbf{V}$ and $\mathbf{W}$) so that some or all conditions (**COND**) are satisfied: (I) SVD-based methods, which are well known in the systems and control community and have good system theoretic properties; (II) Krylov-based methods, which are well known in the numerical analysis community and, to a lesser degree, in the system theory community and have good numerical properties; and (III) SVD-Krylov-based methods, which seek to merge the best attributes of (I) and (II). Furthermore, the class of weighted SVD methods, which establishes a link between (I) and (II), will also be discussed. We refer to the Preface for a description of the book contents.

Chapter 2

Motivating Examples

In this section we describe various applications in which large-scale dynamical systems arise. There are two categories of examples: those in which simulation and/or prediction of future behavior is of interest (examples 1, 2, 3, 4, 5, 6 in Figure 2.1), and those in which simulation and control are of primary interest (examples 7, 8, 9, 10).

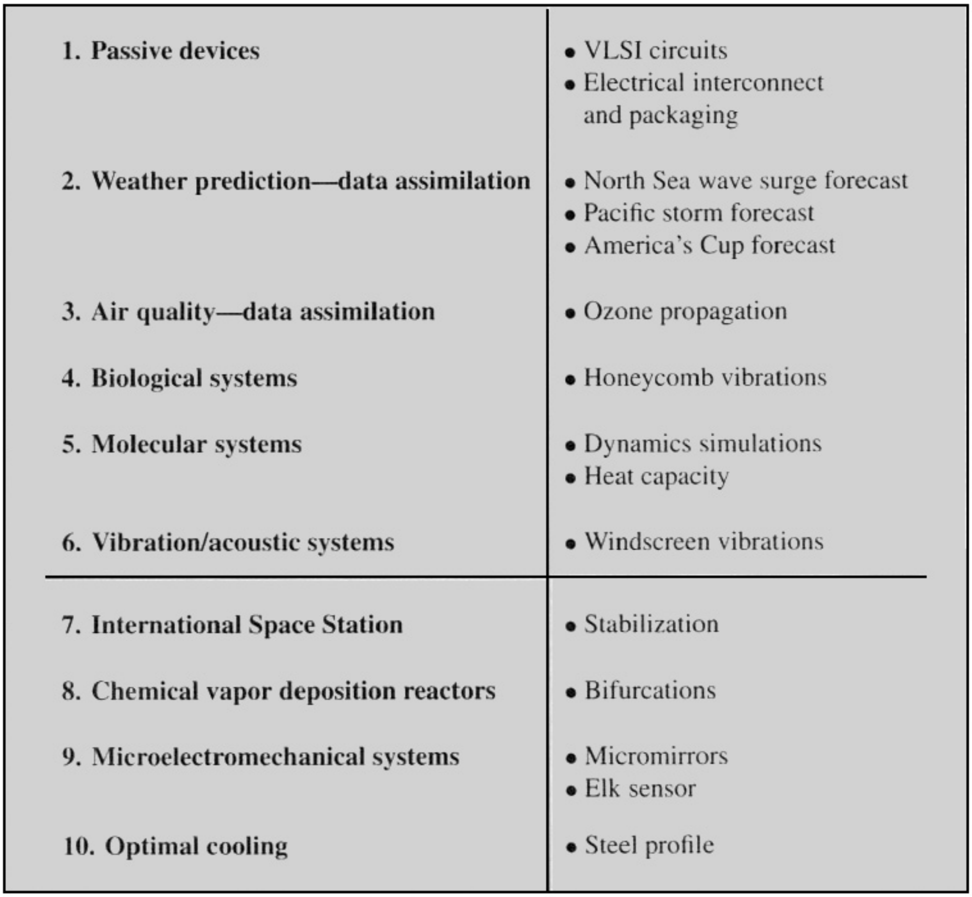
The examples given in Figure 2.1 are described briefly.

2.1 Passive devices

High-frequency, submicron VLSI circuits. The integrated circuit (IC) was introduced in the 1960s. Since then, the scaling trends in VLSI design are that (i) the decrease in feature size is greater than 10% per year; (ii) the increase in chip size is greater than 10% per year; and (iii) there is an increase in operating frequency which now reaches the gigahertz range. As a consequence, the chip *complexity* has been increasing by at least 50% each year. A comparison between the Intel® 4004 processor, released in 1971, and the Intel Pentium® IV processor, released in 2001, shows that the feature size has decreased from 10μ to 0.18μ, the number of components has increased from 2300 to 42 million, and the speed has increased from 64 KHz to 2 GHz; in addition, the length of all interconnections in the Pentium IV totals approximately 2 km, and the components are arranged in seven layers.

These trends impact physical parameters due to the increase of interconnect length and interconnect resistance. Furthermore, capacitance and inductance effects influence the chip, and there is a decrease of metal width and dielectric thickness. The resulting chips are multilayered, and the passive parts thus correspond to three-dimensional resistor-inductor-capacitor (RLC) circuits.

The design phase of a VLSI circuit is followed by the *physical verification* phase, where potential design flaws are discovered. Simulations are thus required to verify that internal electromagnetic fields do not significantly delay or distort circuit signals. This requires the solution of Maxwell's equations for three-dimensional circuits with interconnections



1. Passive devices	• VLSI circuits • Electrical interconnect and packaging
2. Weather prediction—data assimilation	• North Sea wave surge forecast • Pacific storm forecast • America's Cup forecast
3. Air quality—data assimilation	• Ozone propagation
4. Biological systems	• Honeycomb vibrations
5. Molecular systems	• Dynamics simulations • Heat capacity
6. Vibration/acoustic systems	• Windscreen vibrations
7. International Space Station	• Stabilization
8. Chemical vapor deposition reactors	• Bifurcations
9. Microelectromechanical systems	• Micromirrors • Elk sensor
10. Optimal cooling	• Steel profile

Figure 2.1. *Applications leading to large-scale dynamical systems.*

of the order of several kilometers, and submicron scale geometric resolution. One method for deriving a model in this case is known as the partial element equivalent circuit (PEEC), which works by spatial discretization of Maxwell's equations for three-dimensional geometries. The complexity of the resulting model reaches unmanageable proportions and can be anywhere from $n \approx 10^5$ to 10^6. Therefore, reduced-order modeling is of great importance. We thus seek to generate models that are as simple as possible but are nevertheless capable of generating the actual chip behavior. For details, see van der Meijs [330]. See also [279] and [202].

There is a general-purpose simulation program for electric circuits known as SPICE (Simulation Program with Integrated Circuit Emphasis), which was developed in the 1970s at the University of California at Berkeley. This allows the following components: resistors, capacitors, inductors, independent sources, dependent sources, transmission lines, diodes, and transistors. SPICE can handle complexities of a few hundred such elements. It thus becomes inadequate for complexities of the order mentioned above.

The first attempts to use ***model reduction*** to simplify such circuits were made by Pillage and Rohrer [269], who proposed the method of ***asymptotic waveform evaluation*** (AWE). This consists of computing some of the ***moments*** of the corresponding circuits (see section 11.1 for the definition of moments and details on moment matching methods); the AWE method is described in [87]. It turns out, however, that computing moments is numerically ill-conditioned, and the fact that such problems can be solved efficiently using Krylov methods was soon proposed for solving AWE problems [125]. The next step along this line of approach to circuit simulation came with the work of Feldmann and Freund, who proposed the ***Padé via Lanczos*** approach. There are numerous references on this topic; we give just a few: [90], [194], [114], [33], [34], [31], [117], and more recently [27].

A more general problem is the ***electromagnetic modeling of packages and interconnects***; see [316] for an overview. Again, the starting point is ***Maxwell's equations***, and the PEEC method is used to obtain a finite-dimensional approximation. A related problem was described in [80].

2.2 Weather prediction—data assimilation

2.2.1 North Sea wave surge forecast

Because part of The Netherlands is below sea level, it is important to monitor wave surges at river openings. In the case of such a surge, water barriers can be closed to prevent flooding. Since these rivers are in many cases important waterways, the barriers must stay closed only while the surge lasts. Furthermore, the warning has to come about 6 hours in advance.

The equations governing the evolution of the wave surge are in this case the ***shallow water*** equations, which are PDEs. In Figure 2.2, the horizontal and vertical axes indicate the number of discretization points, while the color bar on the right-hand side indicates the depth of the sea at various locations of interest (justifying the use of ***shallow*** water equations).

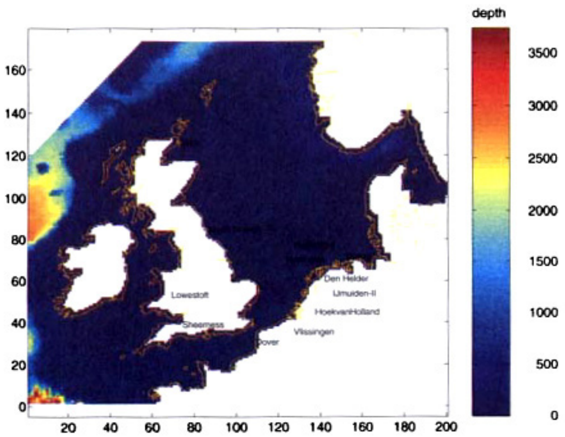


Figure 2.2. *Wave surge prediction problem: depth of the North Sea.*

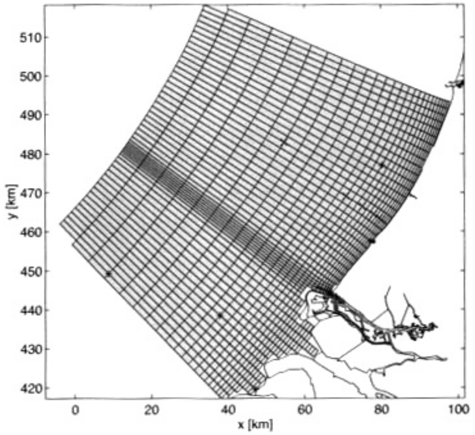


Figure 2.3. *Wave surge prediction problem: discretization grid close to the coast.*

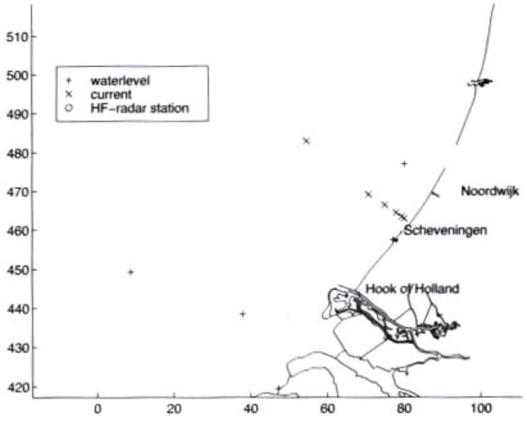


Figure 2.4. *Wave surge prediction problem: measurement locations.*

The discretization grid used in this case is shown in Figure 2.3. The problem in this case is not just prediction of the wave surge based on initial conditions. There are several locations where the wave surge is measured (see Figure 2.4). There are also locations where the movement of the sea currents is measured. The problem thus becomes *data assimilation*, as one wishes to predict the wave surge based on both the model and the provided measurements. This is achieved by means of a Kalman filter.

The *finite element* (FE) discretization of the shallow water equations yields approximately 60,000 equations, and the resulting computational time is several times the allowed limit of 6 hours. Therefore, reduced-order models are necessary. Figure 2.5 shows the error covariance of water level prediction in two cases: first, with wind disturbance and

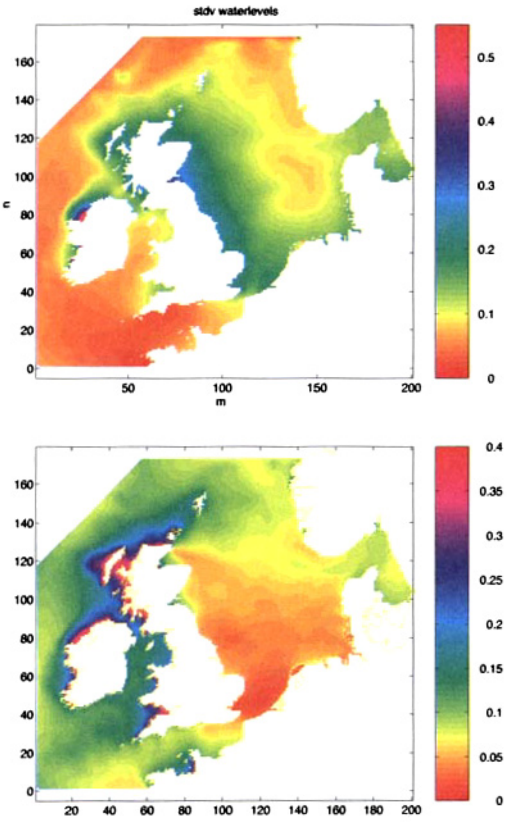


Figure 2.5. *Wave surge prediction problem: error covariance of the estimated water level without measurements (top) and with assimilation of eight water level measurements (bottom). The bar on the right shows the color-coding of the error.*

no additional measurements, and second, by assimilating the measurements from the eight locations. This problem has been studied by Verlaan and Heemink at Delft University of Technology in The Netherlands; for details see [348], [168].

2.2.2 Pacific storm tracking

The issue here is to determine the sensitivity of atmospheric equilibria to perturbations. In particular, we wish to determine the initial perturbation that produces the greatest perturbation growth over some specified interval of time. In [109], perturbations to the vorticity equation of a Couette flow are studied. These are governed by the Orr–Sommerfeld equation; assuming harmonic perturbations in the wind velocity of the form $\Phi(x, y, t) = \phi(y, t)e^{ikx}$, we have

$$\frac{\partial\phi(y,t)}{\partial t} = \mathbf{A}\phi(y,t) = -iky\frac{\partial^2\phi(y,t)}{\partial y^2} + \frac{1}{\mathrm{Re}}\left(\frac{\partial^2\phi(y,t)}{\partial y^2} - k^2\phi(y,t)\right)^2,$$

where Re denotes the Reynolds number. Discretization in y yields the set of ODEs:

$$\frac{d\hat{\phi}(t)}{dt} = \hat{\mathbf{A}}\hat{\phi}(t), \qquad \hat{\mathbf{A}} \in \mathbb{R}^{n\times n}.$$

We assume that this system is influenced by perturbations; in particular, we assume that (i) random inputs are affecting *all* variables $\hat{\phi}_i$, and (ii) *all* these variables are measured (observed). The discretized system is thus a linear system having the same number of inputs m, state variables n, and outputs p:

$$\Sigma = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \mathbf{I}_n \\ \hline \mathbf{I}_n & 0 \end{array}\right) \Rightarrow m = p = n.$$

Such models are used for storm tracking in the midlatitude Pacific. For data assimilation in this context, see [110].

2.2.3 America's Cup

The America's Cup is a race of sailing boats that takes place every 4 years. The 31st competition since 1848 took place in 2003 in New Zealand between Team NZ and the Swiss team Alinghi. Much technological know-how goes into the construction of the boats. Less well known, however, is that the contestants set up *weather teams*. These are meant to advise on the weather (in particular, the wind direction and speed) that will prevail during the approximately 8 hours of the race. It is important to be able to predict wind shifts. The goal is to choose the right sails and develop appropriate strategies for the race. The 2003 winner, the Alinghi team, set up a strong weather forecasting group, lead by Dr. J. Katzfey, an expert in weather forecasting at the Commonwealth Scientific and Industrial Research Organization in Australia. Furthermore, the Alinghi team set up eight weather stations, which provided data for the data assimilation part of the model. At 7:00 a.m. before each race, the weather team presented a weather prediction for the next 6 to 8 hours of sailing. This strategy turned out to be an important factor for the Alinghi team; in the third regatta, for instance, last-minute updates brought Alinghi 200 m ahead of the New Zealand boat, which proved decisive for the winners.

2.3 Air quality simulations—data assimilation

Air pollution was thought to be a local phenomenon until the 1970s and a regional one during the 1980s. At present it is recognized that air pollution extends from urban to regional and global scales. Emissions from fossil fuel combustion and biomass burning and the resulting photochemical production of tropospheric ozone and climate warming are considered global problems.

A current challenge to the atmospheric science community is to quantify the impact of human activities on global atmospheric photochemistry. This is achieved by means of *air quality models* (AQMs), which provide a mechanism for elucidating the underlying physical and chemical processes responsible for the formation, accumulation, transport, and removal of air pollutants. AQMs are designed to calculate concentrations of ozone and other pollutants and their variations in space and time in response to particular emission inputs and for specified meteorological scenarios. The AQM is the only prognostic tool available to the policy-making community, i.e., a tool capable of quantitatively estimating future air quality outcomes for conditions or emissions different from those that have existed in the past. Because of this unique capability, AQMs have come to play a central role in determining how pollutant emissions should be managed to achieve air quality goals.

A variety of AQMs are being applied on urban, regional, and global scales. Many models share common features. In particular, AQMs are based on solving the same species conservation equations which describe the formation, transport, and fate of air pollutants:

$$\frac{\partial c_i}{\partial t} + \nabla \cdot (\mathbf{u} c_i) - \nabla \cdot (\mathbf{K} \nabla c_i) = R_i(c_1, c_2, \ldots, c_s) + S_i, \qquad i = 1, \ldots, s.$$

Here, c_i is the (averaged) concentration of species i; $\mathbf{u}(\mathbf{x}, t)$ is the wind velocity vector at location $\mathbf{x}$ and time t; $\mathbf{K}(\mathbf{x}, t)$ is the turbulence diffusivity tensor; R_i is the rate of concentration change of species i by chemical reactions; $S_i(\mathbf{x}, t)$ is the source/sink of species i; and s is the number of predicted species. R_i can also be a function of meteorological variables (e.g., temperature). The source/sink term S_i can include emissions of a species as well as its loss due to various processes, such as dry deposition and rainout. Ensemble averaging is used to dispense with the need to capture the extremely small-scale fluctuations due to turbulence. With appropriate initial and boundary conditions, the system described by the above equation represents the continuum chemistry transport model (CTM). A CTM system is composed of four basic components: a chemical kinetic mechanism, a source emissions inventory, a description of pollutant transport and removal, and a set of numerical algorithms for integrating the governing equations. Differences between various AQMs stem from alternative choices made by their developers in characterizing these physical and chemical processes, procedures for their numerical solution, and the approach taken to adapt the model to the computational domain of interest.

After spatial discretization of the CTM, we obtain

$$\frac{d\mathbf{c}}{dt} = \mathbf{f}(\mathbf{c}) = \mathbf{f}_a(\mathbf{c}) + \mathbf{f}_d(\mathbf{c}) + \mathbf{f}_r(\mathbf{c}) + \mathbf{u}, \quad \mathbf{c}(t_0) = \mathbf{c}_0, \quad \mathbf{c} = [c_1, \ldots, c_s].$$

The grid in the x and y directions is 1 km $=$ 100 points; in the z direction it is 10 km $=$ 30 points. This results in 300,000 equations.

Many measurements of pollutants exist. In the past few years, the MOPITT satellite was launched by NASA to take pictures of pollutant concentrations. Thus, measurements consist mainly of satellite images, and the problem becomes, once more, one of *data assimilation*.

For details, see [103], [246], [292], [293]. An important recent development in three-dimensional chemistry transport modeling is MOZART (Model of Ozone and Related Tracers); it has been developed in the framework of National Center for Atmospheric Research community climate model (CCM) [179].

2.4 Biological systems: Honeycomb vibrations

The honeybee dance language, in which foragers perform dances containing information about the distance and direction to a food source, is an example of symbolic communication in nonprimates. Honeybees and human beings possess an *abstract* system of communication. It was noted by Aristotle that honeybees recruit nestmates and lead them to a food source. In the 1960s, K. von Frisch (1973 Nobel Prize winner in pysiology and medicine) postulated that this recruitment takes place by means of the so-called waggle dance performed on the honeycomb. This dance consists of a looping figure eight movement with a central straight waggle run. During this dance, distance and direction information about the food source are transmitted. The possible mechanisms involved in this unique communication are mechanical and chemical; there are no optical signals involved given the darkness of the hive. The mechanical mechanism results from the vibration of the honeycomb, while the chemical one results by transmission of pollen or nectar.

During the waggle dance, the dancer bee waggles her body at 15 Hz and vibrates her wings intermittently at 200–300 Hz. The transmitted vibrations have an amplitude of about 1.4 μm. The question thus arises as to the ability of the bees to detect weak vibrations in a noisy environment (the hive).

Experimental measurements have shown that the vibrations occur only in the horizontal direction to the plane of the combs; furthermore, the comb seems to amplify vibrations in frequencies around 250 Hz. It has also been experimentally observed that the combs exhibit an impedance minimum to horizontal vibrations at 230–270 Hz. Most of the experimental investigations of this problem have been performed by Dr. J. Tautz at the Universität Würzburg.

The goal is therefore to find a model of the honeycomb that explains the phenomena observed and measured and that provides new insights into these phenomena. In particular, we would like to know to what extent a honeycomb is appropriate as a medium for the transmission of information through vibrations. For more details, see [102] and the more recent article [321].

2.5 Molecular dynamics

This example involves *simulation* in molecular dynamics, and in particular protein substate modeling and identification. The main tool used for this example is the SVD. For details see [275].

Proteins are dynamic entities primarily due to the thermal motion of their atoms. They have different states which are attained by thermal motion and determine their biological properties. Proteins can exist in a number of different conformational substates, a conformational substate being a collection of structures that are energetically degenerate. The thermal motion of the atoms drives the transitions between the different substates accessible to a protein. The distribution of these substates and their transitions are major factors in determining the biological properties of proteins. The equations describing the motion of proteins are of the type $\frac{d^2}{dt^2}\mathbf{x}(t) = -\nabla\phi(\mathbf{x}(t))$, where ϕ is a potential function, and $\mathbf{x} \in \mathbb{R}^{3n}$, where n is the number of atoms in the protein.

To find the most important protein configurations, an SVD of snapshots of $\mathbf{x}$ is used. This method is known as POD and is described in section 9.1. This SVD provides a way of decomposing a molecular dynamics trajectory into fundamental modes of atomic motion. The left singular vectors describe the direction in which each atom prefers to

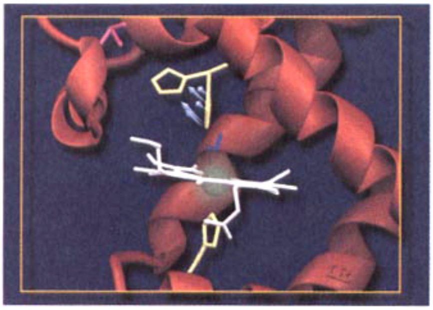


Figure 2.6. *Molecular dynamics: myoglobin (protein); heme: active site; histidine: part that opens and closes, catching oxygen molecules.*

move. The right singular vectors provide temporal information; they are projections of the protein conformations onto these modes showing the protein motion in a generalized low-dimensional basis.

If an atom were constrained to move along only one of the left singular vectors, then its motion in Cartesian space can be projected onto this vector, giving a curve. The elements of the right singular vector can be thought of as scaling the size of the left singular vector to describe where the atom is at each time point or in each conformation.

Figure 2.6 depicts a protein called *myoglobin* (more precisely, F46V mutant bioglobin). The active site that catches oxygen molecules is called the *heme* (this is shown in a yellow closing and a white closed state); finally, oxygen molecules (shown in green) are captured by means of the *active site*, which is called *histidine*.

The first left singular vector of the (distal) histidine in this F46V mutant bioglobin describes more than 90% of the total motion of the histidine. This histidine also ranked second on the list of residues. Thus, (i) incorporating conformational substate information improves the refinement model; (ii) multiconformer refinement appears to be the better method for overall improvement of the model; and (iii) the SVD provides a powerful tool for visualizing complex high-dimensional systems in a low-dimensional space.

Heat capacity of molecular systems

The next application is concerned with the determination of the heat capacity of a molecular system. This involves the calculation of the following integral:

$$C_v = \int_0^\infty f(\omega)g(\omega)d\omega,$$

where $g(\omega)d\omega$ gives the number of vibrational frequencies in the interval $(\omega, \omega+d\omega)$. Since the molecular system is discrete we have $\sigma(\omega) = \int_0^\omega g(\tau)d\tau = \sum_{i=1}^n \mathbb{I}(\omega - \omega_i)$, where n is the number of particles in the molecular system and $\mathbb{I}$ is the Heaviside step function. This requires *all* the fundamental frequencies of the system to be computed in advance. Instead, *Gauss quadrature* is used; this involves the *Lanczos algorithm*, where **A** is the Hessian of ϕ and **b** is arbitrary. For details, see [364].

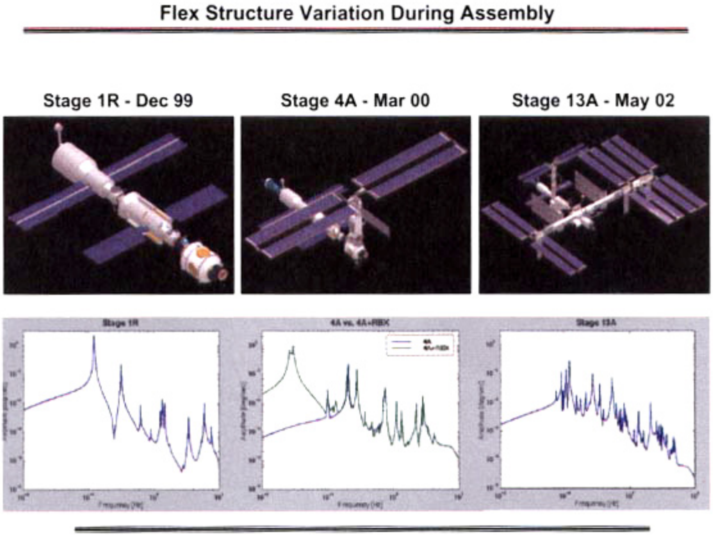


Figure 2.7. *Frequency response of the ISS as complexity increases. Figure courtesy of Draper Laboratories, Houston.*

2.6 International Space Station

The International Space Station (ISS) is a complex structure composed of many modules; these modules are contributed by NASA and other space agencies. The flex modes of each module are described in terms of $n \approx 10^3$ state variables. The goal is to develop controllers that are implemented onboard the space station. Consequently, controllers of *low* complexity are needed because of hardware, radiation, throughput, and testing issues. In Chapter 13, model reduction of the flex models for two specific modules, namely, the 1R (Russian service module) and the 12A (second left-side truss segment), are discussed. Figure 2.7 shows the frequency response (amplitude Bode plot) as more components are added to the space station. The complexity of the resulting model is reflected in the number of spikes present in the frequency response. These models were provided by Draper Laboratories in Houston. For details on the ISS and its assembly see http://spaceflight.nasa.gov/station/.

2.7 Vibration/acoustic systems

Consider a car windscreen subject to an acceleration load. The problem consists of computing the noise generated at points away from the window. The first step in solving this problem is the PDE which describes the deformation of the windscreen of a specific material. The finite element discretization gives, in a specific case, 7564 nodes (three layers of 60-by-30 elements); the material is glass with Young modulus $7 \cdot 10^{10}\,\mathrm{N/m^2}$, density $2490\,\mathrm{kg/m^3}$, and Poisson ratio 0.23. These parameters help determine the coefficients of the resulting FE model experimentally. Finally, the windscreen is subjected to a point force at some given point (node 1891), and the goal is to compute the displacement at the same point. The discretized problem in this particular case has a dimension of 22,692. Notice that the last two problems (the windscreen and the ISS) yield second-order equations:

$$\mathbf{M}\frac{d^2}{dt^2}\mathbf{x}(t) + \mathbf{C}\frac{d}{dt}\mathbf{x}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{f}(t),$$

where $\mathbf{x}$ is position, $\frac{d}{dt}\mathbf{x}$ is velocity of the windscreen at the grid points chosen, and $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$ are the *mass, damping,* and *stiffness* matrices. Since this is a second-order system, its complexity is twice as high (45,384 states). For details on eigenvalue problems for second-order systems, see [325]. This problem was provided by Karl Meerbergen of Free Field Technologies, Leuven, Belgium. See also [235].

2.8 CVD reactors

An important problem in semiconductor manufacturing is the control of chemical reactors, for instance, CVD reactors. This problem is addressed in the literature using POD methods. For details, see the work of Banks and coworkers [199], [200], [49]. The dimension of the resulting linear systems is on the order of a few thousand state variables.

Another issue concerning CVD reactors is the determination of the stability of steady states. To address this problem, the transient behavior is linearized around a steady state. This leads to a generalized eigenvalue problem. The eigenvalues with largest real part are calculated using the Arnoldi iteration. A model problem of three-dimensional incompressible flow and heat transfer in a rotating disk CVD reactor is used to analyze the effect of parameter change on the performance of the eigenvalue algorithm. The description of this system requires a full three-dimensional Navier–Stokes model. The calculation of leading eigenvalues for matrix systems of order 4 million to 16 million is required. These calculations lead to the critical Grashof, Rayleigh, and Reynolds numbers for a Hopf bifurcation. For details, see [226].

2.9 Microelectromechanical devices

Microelectromechanical systems (MEMS) are integrated systems combining electrical and mechanical components. They are usually fabricated using IC techniques and can range in size from micrometers to millimeters. In their most general form, MEMS consist of mechanical microstructures, microsensors, microactuators, and electronics, all integrated onto the same chip.

Finite element model (FEM) simulation is often used in MEMs and results in complex systems. System-level models with reduced order and accuracy have to be generated as a basis of system simulation [291].

SUGAR is a simulation tool for MEMS devices based on nodal analysis techniques of integrated circuit simulation. Beams, electrostatic gaps, circuit elements, etc., are modeled by small, coupled systems of differential equations. For a description, see [89] and http://www-bsac.eecs.berkeley.edu/cadtools/sugar/sugar/.

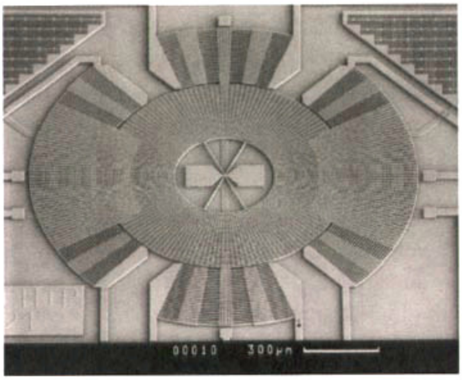


Figure 2.8. *MEMS angular velocity sensor used in car navigation systems and for rollover detection in passenger restraint systems. Picture courtesy of Robert Bosch Co.*

2.9.1 Micromirrors

Two such MEMS devices are the micromirror and micromirror arrays. They are used for precision light manipulation, e.g., as an *optical switch* in fiberoptics. A mirror tilts to reflect light from one fiber to another. Large arrays, up to 1000 by 1000 mirrors, will be needed for telecommunication applications. As the fiber-to-mirror distance increases, strict precision of mirror tilt control is necessary. For a distance of 500 microns, at least 0.10 degree of precision is required. A second application of micromirrors is *maskless lithography* (virtual masks). Feedback control is needed to achieve precision positioning of a mirror of 2 degrees of freedom. Advantages of feedback control as an optical switch are that it is faster, smaller, and cheaper than electrical switches.

2.9.2 Elk sensor

A few years ago, production of the Mercedes-Benz A Class cars had to be stopped just after their launch because they failed the *elk test*. This test consists of forcing the car to take a sharp turn to avoid an obstacle that suddenly appears on the road. To remedy this situation, the company incorporated a rollover (elk) sensor that could detect turning movement and apply the brakes to slow the rotational movement. The first rollover sensors were mechanical. Subsequently, Bosch AG developed a microelectromechanical sensor at reduced cost and reduced size. Now, the elk sensor is standard equipment in many cars. A similar angular velocity sensor is shown in Figure 2.8.

Once such a device has been designed, the next issue consists in testing its performance by simulation. One method is a physically oriented modeling (see, e.g., Schwarz [290] and Teegarden, Lorenz, and Neul [322]), using appropriate simulation packages, as described in [291]. The more detailed the modeling, the higher the complexity (i.e., the number of differential equations) of the resulting model. As the available simulation packages are built to handle low complexities, there is a need for simplification of the model through model reduction.

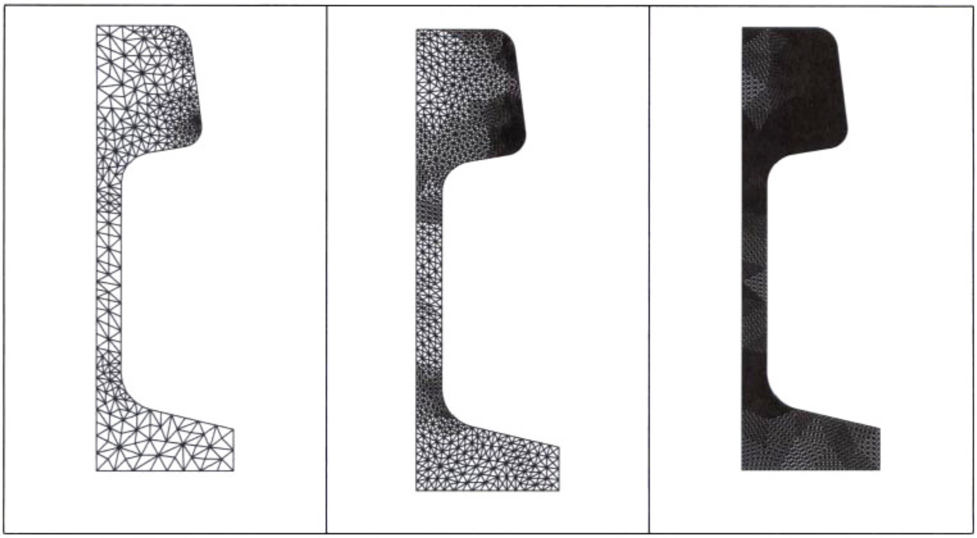


Figure 2.9. *Optimal cooling: discretization grid.*

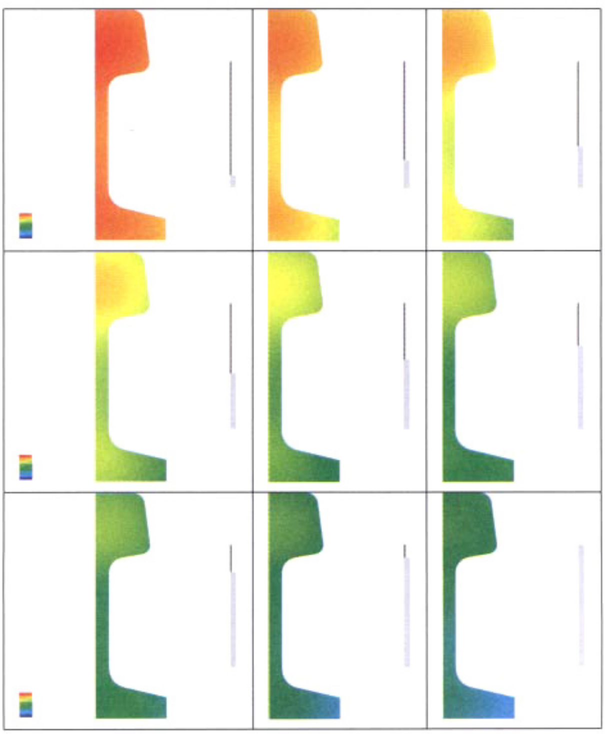


Figure 2.10. *Progression of the cooling of a steel profile (from left to right and from top to bottom); the bars are cooled from* 1000°*C to* 500°*C.*

2.10 Optimal cooling of steel profile

Many problems in control are both structured and computationally intensive. An application arising in tractor steering can be found in [61]. For an overview, see Fassbender [112]. Here we will describe an application worked out by Benner [60]. In rolling mills, steel bars have to be heated and cooled quickly and uniformly to achieve a fast production rate. Cooling takes place by spraying the bars with cooling fluids. The problem consists of devising and applying an optimal cooling strategy.

The heat equation is used to model the cooling process. The steel bar is assumed to have infinite length, thus reducing the problem to two dimensions. The domain is obtained by cutting the steel bar vertically; this domain can be further halved due to symmetry. It is assumed that there are eight nozzles spraying cooling liquid uniformly on the boundary of the bar. Thus a two-dimensional boundary control problem results.

The heat-diffusion equation with Neumann boundary conditions is discretized in the spatial variables using the FEM described in the package ALBERT. The initial mesh leads to a system of order $n = 106$. Subsequently, the mesh is refined, leading to systems of orders $n = 371$, 1357, and 5177 (see Figure 2.9). The resulting mass and stiffness matrices are sparse (with approximately $3 \cdot 10^5$ and $3 \cdot 10^4$ nonzero elements, respectively). The control law is obtained by means of linear quadratic regulator (LQR) design.

Finally, the model uses water at 20°C and has the spraying intensity as control parameter. Figure 2.10 shows the progression of the cooling.

Part II

Preliminaries

This page intentionally left blank

Chapter 3
Tools from Matrix Theory

The broad topic of this book is approximation. To be able to discuss approximation problems, we need to be able to measure the sizes of various objects, such as the size of the elements belonging to certain linear function spaces, as well as the size of operators acting between these spaces. The most commonly used norm for this purpose is the *2-norm* for elements of function spaces and the associated *2-induced norm* for measuring the size of operators between these spaces.

The 2-norm of the n-vector $\mathbf{x} = (x_1 \cdots x_n)^* \in \mathbb{R}^n$ is the usual Euclidean norm: $\| \mathbf{x} \|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$. The 2-induced norm of $\mathbf{A} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is defined as

$$\| \mathbf{A} \|_{2-\text{ind}} = \sup_{\mathbf{x} \neq 0} \frac{\| \mathbf{Ax} \|_2}{\| \mathbf{x} \|_2}.$$

Furthermore, the *complexity* of $\mathbf{A}$ is defined as its *rank*.

In this simple case, one *approximation problem* is as follows. Given a matrix (operator) $\mathbf{A}$ of complexity n, find a matrix (operator) $\mathbf{X}_k$ of lower complexity (say, $k < n$) so that the 2-induced norm of the *error* is minimized:

$$\mathbf{X}_k = \arg \min_{\text{rank}\, \mathbf{X} \leq k} \| \mathbf{A} - \mathbf{X} \|_{2-\text{ind}} \, .$$

The problem just defined is a *nonconvex* optimization problem. Despite the lack of convexity, however, it can be solved explicitly by means of the singular value decomposition (SVD); this is a decomposition of $\mathbf{A}$ in terms of two unitary matrices $\mathbf{U}$, $\mathbf{V}$ and a diagonal matrix with nonnegative entries $\sigma_1, \ldots, \sigma_n$ on the diagonal, such that $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$, where $\Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_n)$. The σ_i are the *singular values* of $\mathbf{A}$, which are the square roots of the largest n eigenvalues of $\mathbf{A}^*\mathbf{A}$ or $\mathbf{A}\mathbf{A}^*$. Furthermore, $\sigma_1 = \| \mathbf{A} \|_{2-\text{ind}}$.

This decomposition is one of the most useful tools in applied linear algebra. It can be efficiently computed and is an important tool for both theoretical and computational considerations. For an overview of the role of matrix decompositions in computing, see [315], which lists the ***big six matrix decompositions*** that form the foundations of matrix computations: (i) the Cholesky decomposition, (ii) the pivoted LU (lower-upper) decomposition, (iii) the QR (orthogonal-upper triangular) algorithm, (iv) the spectral decomposition, (v) the Schur decomposition, and (vi) the SVD.

The solution of the approximation problem introduced above can now be obtained as follows. First, one notices that the minimum value of the error is

$$\min_{\mathbf{X}, \mathrm{rank}\mathbf{X} \leq k} \| \mathbf{A} - \mathbf{X} \|_{2-\mathrm{ind}} = \sigma_{k+1}(\mathbf{A}).$$

Next, it is easy to be convinced that this lower bound is achieved by simple truncation of the SVD of **A** and that

$$\mathbf{X}_k = \mathbf{U}\Sigma_k\mathbf{V}^*, \quad \text{where } \Sigma_k = \mathrm{diag}\,(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$$

is an optimal solution. This is a powerful result relating the complexity k of the approximant with the $(k+1)$st largest eigenvalue of $\mathbf{A}\mathbf{A}^*$ (or $\mathbf{A}^*\mathbf{A}$). Details will be discussed.

This chapter contains some fundamental results from linear algebra and numerical linear algebra. The first section lists various *norms* of vectors and matrices. The second discusses an important decomposition of matrices and operators, the SVD. Then we briefly review two concepts from numerical analysis that are essential in assessing the accuracy of computations: the *condition number* of a given problem and the *stability* of an algorithm to solve the problem.

More material from linear algebra is presented in Chapter 10. This material concerns the *Krylov Iteration*, which turns out to be important in both linear algebra and approximation of linear dynamical systems. Also in Chapter 10, related issues dealing with eigenvalue computations and pseudospectra are briefly discussed.

Most of the material presented in this chapter is standard and can be found in many textbooks. For the material of the first three sections, see, for example, Golub and Van Loan [144], Horn and Johnson [181], Stewart and Sun [314], Higham [171], Trefethen and Bau [328], Meyer [238], and Thompson [324]. For a treatment with more functional analytic flavor, see Bhatia [65] and Lax [225]. For the material in section 3.3, see the notes by Sorensen [308].

3.1 Norms of finite-dimensional vectors and matrices

Let X be a linear space over the field of reals $\mathbb{R}$ or complex numbers $\mathbb{C}$. A *norm* on X is a function

$$\nu : X \to \mathbb{R}$$

such that the following three properties are satisfied:

$$\left.\begin{array}{rll} \text{strict positiveness:} & \nu(\mathbf{x}) \geq 0 & \forall\, \mathbf{x} \in X \text{ with equality iff } x = 0, \\ \text{triangle inequality:} & \nu(\mathbf{x}+\mathbf{y}) \leq \nu(\mathbf{x}) + \nu(\mathbf{y}) & \forall\, \mathbf{x}, \mathbf{y} \in X, \\ \text{positive homogeneity:} & \nu(\alpha\mathbf{x}) = |\alpha|\nu(\mathbf{x}) & \forall\, \alpha \in \mathbb{C},\ \forall\, \mathbf{x} \in X. \end{array}\right\} \tag{3.1}$$

For vectors $\mathbf{x} \in \mathbb{C}^n$ the Hölder or p-norms are defined as follows:

$$\|\mathbf{x}\|_p = \begin{cases} \left(\sum_{i=1}^{n} |\, x_i \,|^p\right)^{\frac{1}{p}}, & 1 \le p < \infty, \\ \max_{i\in\{1\cdots n\}} |\, x_i \,|, & p = \infty, \end{cases} \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}. \tag{3.2}$$

These norms satisfy *Hölder's inequality*:

$$|\mathbf{x}^*\mathbf{y}| \le \|\mathbf{x}\|_p \|\mathbf{y}\|_q \text{ for } \frac{1}{p} + \frac{1}{q} = 1. \tag{3.3}$$

For $p = q = 2$ this becomes the *Cauchy–Schwarz inequality*,

$$|\mathbf{x}^*\mathbf{y}| \le \|\mathbf{x}\|_2 \|\mathbf{y}\|_2,$$

with equality holding if and only if $\mathbf{y} = c\mathbf{x}$, $c \in \mathbb{C}$. An important property of the 2-norm is that it is invariant under *unitary* (orthogonal) transformations. Let $\mathbf{U}$ be $n \times n$ and $\mathbf{U}^*\mathbf{U} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. It follows that $\|\mathbf{U}\mathbf{x}\|_2^2 = \mathbf{x}^*\mathbf{U}^*\mathbf{U}\mathbf{x} = \mathbf{x}^*\mathbf{x} = \|\mathbf{x}\|_2^2$. This holds also if $\mathbf{U}$ has size $n \times m$, where $m \le n$; in this case $\mathbf{U}$ is sometimes called *suborthogonal* (*subunitary*).

The following relationship between the Hölder norms for $p = 1, 2, \infty$ holds:

$$\|\mathbf{x}\|_\infty \le \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\, \|\mathbf{x}\|_2, \qquad \|\mathbf{x}\|_2 \le \sqrt{n}\, \|\mathbf{x}\|_\infty.$$

The unit balls in the 1-,2-, and ∞-norms are shown in Figure 3.1 (*diamond*, *circle*, and outer *square*). Notice that the unit ball for the p-norm, $1 < p < 2$, is a circular figure lying between the diamond and the circle, and for $p > 2$ it is also a circular figure lying between the circle and the outer square.

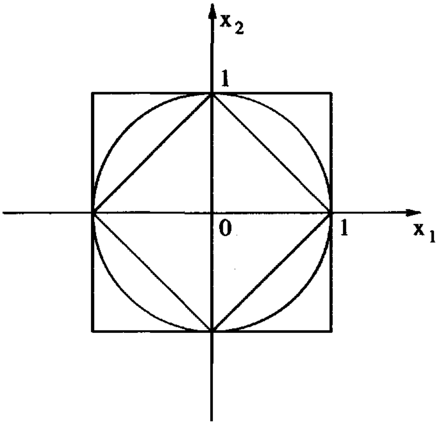


Figure 3.1. *Unit balls in $\mathbb{R}^2$: 1-norm (inner square), 2-norm (circle), ∞-norm (outer square).*

An important class of matrix norms are those that are *induced* by the vector p-norms defined above. More precisely, for $\mathbf{A} = (A_{ij}) \in \mathbb{C}^{n\times m}$,

$$\|\mathbf{A}\|_{p,q} = \sup_{x\neq 0} \frac{\|\mathbf{Ax}\|_q}{\|\mathbf{x}\|_p} \tag{3.4}$$

is the *induced p, q-norm* of $\mathbf{A}$. In particular, for *equi-induced* norms, i.e., $p = q = 1, 2, \infty$, the following expressions hold:

$$\begin{aligned}
\|\mathbf{A}\|_1 &= \max_{j\in\{1\cdots m\}} \sum_{i=1}^{n} |A_{ij}|,\\
\|\mathbf{A}\|_\infty &= \max_{i\in\{1\cdots n\}} \sum_{j=1}^{m} |A_{ij}|,\\
\|\mathbf{A}\|_2 &= \lambda_{\max}(\mathbf{AA}^*)^{\frac{1}{2}} = \lambda_{\max}(\mathbf{A}^*\mathbf{A})^{\frac{1}{2}}.
\end{aligned}$$

More generally, the following expressions hold for mixed induced norms [86]:

$$\begin{aligned}
\|\mathbf{A}\|_{1,p} &= \max_j \|\mathbf{A}_{:,j}\|_p,\\
\|\mathbf{A}\|_{p,\infty} &= \max_i \|\mathbf{A}_{i,:}\|_{\bar{p}}, \qquad \bar{p} = \frac{p}{p-1},
\end{aligned}$$

for $p \in [1, \infty]$, where $\mathbf{A}_{:,j}$, $\mathbf{A}_{i,:}$ denote the jth column, and ith row of $\mathbf{A}$, respectively. In particular, the following special cases hold:

$$\|\mathbf{A}\|_{1,2} = \delta_{\max}(\mathbf{AA}^*)^{\frac{1}{2}}, \quad \|\mathbf{A}\|_{2,\infty} = \delta_{\max}(\mathbf{A}^*\mathbf{A})^{\frac{1}{2}}, \quad \|\mathbf{A}\|_{1,\infty} = \max_{i,j} |\mathbf{A}_{i,j}|,$$

where $\delta_{\max}(\mathbf{M})$ denotes the largest diagonal entry of the matrix $\mathbf{M}$.

There exist other norms besides the induced matrix norms. An example is the *Schatten p-norms*. These noninduced norms are unitarily invariant. To define them, we introduce the *singular values* of $\mathbf{A}$, denoted by $\sigma_i(\mathbf{A})$, $i = 1, \ldots, \min(n, m)$. In the next section we describe the singular values in greater detail, but for now, it suffices to say that $\sigma_i(\mathbf{A})$ is the square root of the ith largest eigenvalue of $\mathbf{AA}^*$. Then for $m \leq n$,

$$\|\mathbf{A}\|_{S,p} = \left(\sum_{i=1}^{m} \sigma_i^p(\mathbf{A})\right)^{\frac{1}{p}}, \qquad 1 \leq p < \infty. \tag{3.5}$$

It follows that the Schatten norm for $p = \infty$ is

$$\|\mathbf{A}\|_{S,\infty} = \sigma_{max}(\mathbf{A}), \tag{3.6}$$

which is the same as the 2-induced norm of $\mathbf{A}$. For $p = 1$ we obtain the *trace norm*

$$\|\mathbf{A}\|_{S,1} = \sum_{i=1}^{m} \sigma_i(\mathbf{A}).$$

For $p = 2$ the resulting norm is also known as the Frobenius norm, the Schatten 2-norm, or the Hilbert–Schmidt norm of $\mathbf{A}$:

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{\min(m,n)} \sigma_i^2(\mathbf{A})\right)^{\frac{1}{2}} = \left[\operatorname{tr}(\mathbf{A}^*\mathbf{A})\right]^{\frac{1}{2}}, \tag{3.7}$$

where tr $(\cdot)$ denotes the *trace* of a matrix.

All the matrix norms discussed above satisfy the *submultiplicativity* property:

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|. \tag{3.8}$$

Notice that there exist matrix norms that do *not* satisfy this relationship. As an example, consider the matrix norm $\|\mathbf{A}\| = \max |A_{ij}|$.

3.2 The SVD

The SVD is one of the most useful tools in applied linear algebra. It can be efficiently computed and is an important tool for both theoretical and computational considerations. For an overview of the role of matrix decompositions in computing, see Stewart [315], which lists the *big six matrix decompositions* that form the foundations of matrix computations:

1. the Cholesky decomposition,
2. the pivoted LU decomposition,
3. the QR algorithm,
4. the spectral decomposition,
5. the Schur decomposition, and
6. the SVD.

It is safe to say that if we were to keep only one of these decompositions, it would be the SVD.

Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$, $n \leq m$, let the ordered nonnegative numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ be the positive square roots of the eigenvalues of $\mathbf{AA}^*$; let also $\mathbf{I}_k$ denote the $k \times k$ identity matrix. There exist unitary matrices $\mathbf{U} \in \mathbb{C}^{n\times n}$, $\mathbf{UU}^* = \mathbf{I}_n$, and $\mathbf{V} \in \mathbb{C}^{m\times m}$, $\mathbf{VV}^* = \mathbf{I}_m$, such that

$$\boxed{\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*,} \tag{3.9}$$

where Σ is an $n \times m$ matrix with $\Sigma_{ii} = \sigma_i$, $i = 1, \ldots, n$, and zero elsewhere. Thus if $n = m$, Σ is a square diagonal matrix with the ordered σ_i on the diagonal. The decomposition (3.9) is called the SVD of the matrix $\mathbf{A}$; σ_i are the *singular values* of $\mathbf{A}$, while the columns of $\mathbf{U}$ and $\mathbf{V}$,

$$\mathbf{U} = (\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_n),\quad \mathbf{V} = (\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_m),$$

are called the *left* and *right singular vectors* of $\mathbf{A}$, respectively. These singular vectors are the *eigenvectors* of $\mathbf{AA}^*$ and $\mathbf{A}^*\mathbf{A}$, respectively. It readily follows that

$$\mathbf{Av}_i = \sigma_i\mathbf{u}_i, \qquad i = 1, \ldots, n.$$

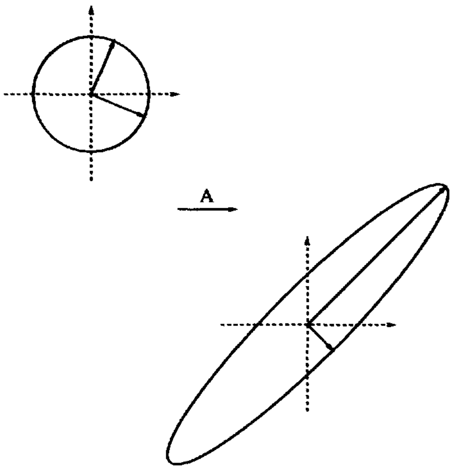


Figure 3.2. *Quantities describing the SVD in* $\mathbb{R}^2$.

Example 3.1. Consider the matrix $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & \sqrt{2} \end{pmatrix}$. It readily follows that the eigenvalue decomposition of the matrices $\mathbf{AA}^* = \begin{pmatrix} 2 & \sqrt{2} \\ \sqrt{2} & 2 \end{pmatrix}$ and $\mathbf{A}^*\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}$ are

$$\mathbf{AA}^* = \mathbf{U}\Sigma^2\mathbf{U}^* \text{ and } \mathbf{A}^*\mathbf{A} = \mathbf{V}\Sigma^2\mathbf{V}^*, \text{ where}$$

$$\mathbf{U} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \; \Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}, \; \mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{2}\sigma_1} & \frac{1}{\sqrt{2}\sigma_2} \\ \frac{1+\sqrt{2}}{\sqrt{2}\sigma_1} & \frac{1-\sqrt{2}}{\sqrt{2}\sigma_2} \end{pmatrix},$$

$$\sigma_1 = \sqrt{2+\sqrt{2}}, \; \sigma_2 = \sqrt{2-\sqrt{2}}.$$

Notice that $\mathbf{A}$ maps $\mathbf{v}_1 \mapsto \sigma_1\mathbf{u}_1$ and $\mathbf{v}_2 \mapsto \sigma_2\mathbf{u}_2$ (see Figure 3.2). Since $\mathbf{A} = \sigma_1\mathbf{u}_1\mathbf{v}_1^* + \sigma_2\mathbf{u}_2\mathbf{v}_2^*$, it follows from Theorem 3.6 that $\mathbf{X} = \sigma_2\mathbf{u}_2\mathbf{v}_2^*$ is a perturbation of smallest 2-norm (equal to σ_2) such that $\mathbf{A} - \mathbf{X}$ is singular:

$$\mathbf{X} = \frac{1}{2}\begin{pmatrix} 1 & 1-\sqrt{2} \\ -1 & \sqrt{2}-1 \end{pmatrix} \Rightarrow \mathbf{A} - \mathbf{X} = \frac{1}{2}\begin{pmatrix} 1 & 1+\sqrt{2} \\ 1 & 1+\sqrt{2} \end{pmatrix}.$$

In other words, the distance of $\mathbf{A}$ to singularity is σ_2.

The singular values of $\mathbf{A}$ are *unique*. The left and right singular vectors corresponding to singular values of multiplicity one are also uniquely determined up to simultaneous sign change. Thus the SVD is *unique* when the matrix $\mathbf{A}$ is *square* and the singular values have *multiplicity one*.

Lemma 3.2. *The largest singular value of a matrix* $\mathbf{A}$ *is equal to its induced 2-norm:* $\sigma_1 = \|\mathbf{A}\|_2$.

Proof. By definition,

$$\|\mathbf{A}\|_2^2 = \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{Ax}\|_2^2}{\|\mathbf{x}\|_2^2} = \sup_{\mathbf{x}\neq 0} \frac{\mathbf{x}^*\mathbf{A}^*\mathbf{Ax}}{\mathbf{x}^*\mathbf{x}}.$$

Let $\mathbf{y}$ be defined as $\mathbf{y} = \mathbf{V}^*\mathbf{x}$, where $\mathbf{V}$ is the matrix containing the eigenvectors of $\mathbf{A}^*\mathbf{A}$, i.e., $\mathbf{A}^*\mathbf{A} = \mathbf{V}\Sigma^*\Sigma\mathbf{V}^*$. Substituting in the above expression, we obtain

$$\sigma_n^2 \le \frac{\mathbf{x}^*\mathbf{A}^*\mathbf{Ax}}{\mathbf{x}^*\mathbf{x}} = \frac{\sigma_1^2 y_1^2 + \cdots + \sigma_n^2 y_n^2}{y_1^2 + \cdots + y_n^2} \le \sigma_1^2.$$

This latter expression is maximized and equals σ_1^2 for $\mathbf{y} = \mathbf{e}_1$ (the first canonical unit vector $[1\ \ 0\ \cdots\ 0]^*$), that is, $\mathbf{x} = \mathbf{v}_1$, where $\mathbf{v}_1$ is the first column of $\mathbf{V}$. □

We are now ready to state the existence of the SVD for all matrices.

Theorem 3.3. *Every matrix* $\mathbf{A}$ *with entries in* $\mathbb{C}$ *has an SVD.*

3.2.1 Three proofs

Given the importance of the SVD, we provide three proofs.

First proof. This proof is based on the lemma above. Let σ_1 be the 2-norm of $\mathbf{A}$; there exist unit length vectors $\mathbf{x}_1 \in \mathbb{C}^m$, $\mathbf{x}_1^*\mathbf{x}_1 = 1$, and $\mathbf{y}_1 \in \mathbb{C}^n$, $\mathbf{y}_1^*\mathbf{y}_1 = 1$, such that $\mathbf{Ax}_1 = \sigma_1\mathbf{y}_1$. Define the unitary matrices $\mathbf{V}_1, \mathbf{U}_1$ so that their first column is $\mathbf{x}_1, \mathbf{y}_1$, respectively: $\mathbf{V}_1 = [\mathbf{x}_1\ \ \hat{\mathbf{V}}_1]$, $\mathbf{U}_1 = [\mathbf{y}_1\ \ \hat{\mathbf{U}}_1]$. It follows that

$$\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1 = \begin{pmatrix} \sigma_1 & \mathbf{w}^* \\ 0 & \mathbf{B} \end{pmatrix} = \mathbf{A}_1, \quad \text{where } \mathbf{w} \in \mathbb{C}^{m-1},$$

and consequently,

$$\mathbf{U}_1^*\mathbf{A}\mathbf{A}^*\mathbf{U}_1 = \mathbf{A}_1\mathbf{A}_1^* = \begin{pmatrix} \sigma_1^2 + \mathbf{w}^*\mathbf{w} & \mathbf{w}^*\mathbf{B}^* \\ \mathbf{Bw} & \mathbf{BB}^* \end{pmatrix}.$$

Since the 2-norm of every matrix is greater than or equal to the norm of any of its submatrices, we conclude that

$$\sigma_1^2 + \mathbf{w}^*\mathbf{w} \le \|\mathbf{AA}^*\| = \sigma_1^2.$$

This implies that $\mathbf{w}$ must be the zero vector, $\mathbf{w} = \mathbf{0}$. Thus

$$\mathbf{U}_1^*\mathbf{A}\mathbf{V}_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \mathbf{B} \end{pmatrix}.$$

The procedure is now repeated for $\mathbf{B}$, which has size $(n-1) \times (m-1)$. Again, since the norm of any submatrix is no greater than the norm of the whole matrix, $\sigma_2 = \|\mathbf{B}\| \leq \sigma_1$. □

Second proof. The second proof is based on the eigenvalue decomposition of the positive semidefinite Hermitian matrix $\mathbf{A}\mathbf{A}^*$:

$$\mathbf{A}\mathbf{A}^* = \mathbf{U}\Lambda\mathbf{U}^*, \quad \mathbf{U}\mathbf{U}^* = \mathbf{I}_n, \quad \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \in \mathbb{R}^{n \times n},$$

where $\lambda_i \geq \lambda_{i+1}$. Set $\Sigma = \Lambda^{\frac{1}{2}}$. Since $n \leq m$, for the rest of the proof we assume for simplicity that $\lambda_n \neq 0$. If $n > m$, follow the same procedure with $\mathbf{A}^*\mathbf{A}$. The above relationship implies

$$\mathbf{K}_1\mathbf{K}_1^* = \mathbf{I}_n, \quad \text{where } \mathbf{K}_1 = \Lambda^{-\frac{1}{2}}\mathbf{U}^*\mathbf{A} \in \mathbb{C}^{n \times m}.$$

Therefore $\mathbf{K}_1$ can be extended to a unitary matrix, say, $\mathbf{V} = (\mathbf{K}_1^* \ \ \mathbf{K}_2^*)$ of size $m \times m$; consequently

$$\mathbf{U}[\Lambda^{\frac{1}{2}} \ \ \mathbf{0}]\mathbf{V}^* = \mathbf{A},$$

which completes the proof. □

Third proof (*see* [357]). First recall that the SVD of $\mathbf{A} \in \mathbb{R}^{n \times m}$ corresponds to the existence of orthonormal bases $(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ of $\mathbb{R}^n$ and $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ of $\mathbb{R}^m$ such that $\mathbf{A}\mathbf{v}_k = \sigma_k\mathbf{u}_k$, $k = 1, 2, \ldots, n$, assuming $n \leq m$, where σ_k are nonnegative real numbers.

(i) If $\mathbf{A} = \mathbf{0}$, take for $\mathbf{U}$ and $\mathbf{V}$ any orthogonal (unitary) matrices, and put $\Sigma = \mathbf{0}$.

(ii) Assume that $\mathbf{A} \neq \mathbf{0}$ and $n \leq m$ (otherwise consider $\mathbf{A}^*$). The proof goes by induction on n.

(ii.1) For $n = 1$, $\mathbf{A} = \mathbf{a}$ is a row vector; take $\mathbf{U} = 1$, $\mathbf{V}$ any orthogonal matrix with first column $\frac{\mathbf{a}}{\|\mathbf{a}\|}$, and $\Sigma = (\|\mathbf{a}\| \ \ 0 \ \cdots \ 0)$.

(ii.2) Assume that the result holds for matrices having $n-1$ rows. We show that it holds for matrices having n rows. Let $\mathbf{w} \neq 0$ be an eigenvector of $\mathbf{A}^*\mathbf{A}$ corresponding to the eigenvalue $\lambda > 0$. Define the spaces

$$\mathcal{V}_{\mathbf{w}}^{\perp} = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^*\mathbf{w} = 0\}, \quad \mathcal{V}_{\mathbf{Aw}}^{\perp} = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y}^*\mathbf{A}\mathbf{w} = 0\}.$$

Since $\mathbf{x} \in \mathcal{V}_{\mathbf{w}}^{\perp}$ implies $\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{w}\rangle = \langle \mathbf{x}, \mathbf{A}^*\mathbf{A}\mathbf{w}\rangle = \lambda\langle \mathbf{x}, \mathbf{w}\rangle = 0$, we conclude that $\mathbf{A}\mathcal{V}_{\mathbf{w}}^{\perp} \subset \mathcal{V}_{\mathbf{Aw}}^{\perp}$. By the induction hypothesis, since these spaces have dimension $m-1, n-1$, respectively, there are orthonormal bases $\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_n$ and $\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_n$ for (a subspace of) $\mathcal{V}_{\mathbf{w}}^{\perp}$ and $\mathcal{V}_{\mathbf{Aw}}^{\perp}$, respectively, such that $\mathbf{A}\mathbf{u}_k = \sigma_k\mathbf{v}_k$, $k = 2, \ldots, n$. Now define

$$\mathbf{u}_1 = \frac{\mathbf{w}}{\|\mathbf{w}\|}, \quad \mathbf{v}_1 = \frac{\mathbf{A}\mathbf{w}}{\|\mathbf{A}\mathbf{w}\|}, \quad \sigma_1 = \frac{\|\mathbf{A}\mathbf{w}\|}{\|\mathbf{w}\|}.$$

Since $\mathbf{A}\frac{\mathbf{w}}{\|\mathbf{w}\|} = \frac{\|\mathbf{A}\mathbf{w}\|}{\|\mathbf{w}\|}\frac{\mathbf{A}\mathbf{w}}{\|\mathbf{A}\mathbf{w}\|}$, we have $\mathbf{A}\mathbf{u}_1 = \frac{\|\mathbf{A}\mathbf{w}\|}{\|\mathbf{w}\|}\mathbf{v}_1 = \sigma_1\mathbf{v}_1$. Therefore, $\mathbf{u}_k$, $\mathbf{v}_k$ for $k = 1, 2, \ldots, n$ satisfy the desired requirements.

(ii.3) If the resulting $\sigma_1, \ldots, \sigma_n$ are not ordered in decreasing order, apply a suitable permutation to the bases to achieve this, and the third proof is complete. □

3.2.2 Properties of the SVD

The SVD has important properties, which are stated next. Assume that in (3.9) $\sigma_r > 0$, while $\sigma_{r+1} = 0$; the matrices $\mathbf{U}$, Σ, $\mathbf{V}$ are partitioned compatibly in two blocks, the first having r columns:

$$\mathbf{U} = [\mathbf{U}_1 \ \ \mathbf{U}_2], \ \ \Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix} \in \mathbb{R}^{n\times m} \text{ and } \mathbf{V} = [\mathbf{V}_1 \ \ \mathbf{V}_2], \tag{3.10}$$

$$\Sigma_1 = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{pmatrix} > \mathbf{0}, \qquad \Sigma_2 = \mathbf{0} \in \mathbb{R}^{(n-r)\times(m-r)},$$

where $\mathbf{U}_1, \mathbf{U}_2$ have $r, n-r$ columns and $\mathbf{V}_1, \mathbf{V}_2$ have $r, m-r$ columns, respectively.

Corollary 3.4. *Given* (3.9) *and* (3.10), *the following statements hold:*

1. Rank $\mathbf{A} = r$.
2. The four fundamental spaces associated with $\mathbf{A}$ are

$$\begin{aligned} \text{span col } \mathbf{A} &= \text{span col } \mathbf{U}_1, & \ker \mathbf{A}^* &= \text{span col } \mathbf{U}_2, \\ \text{span col } \mathbf{A}^* &= \text{span col } \mathbf{V}_1, & \ker \mathbf{A} &= \text{span col } \mathbf{V}_2. \end{aligned}$$

3. **Dyadic decomposition.** $\mathbf{A}$ can be decomposed as a sum of r outer products of rank one:

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^* + \ \cdots \ + \sigma_r \mathbf{u}_r \mathbf{v}_r^*. \tag{3.11}$$

4. The orthogonal projection onto the span of the columns of $\mathbf{A}$ is $\mathbf{U}_1\mathbf{U}_1^*$.
5. The orthogonal projection onto the kernel of $\mathbf{A}^*$ is $\mathbf{I}_n - \mathbf{U}_1\mathbf{U}_1^* = \mathbf{U}_2\mathbf{U}_2^*$.
6. The orthogonal projection onto the span of the columns of $\mathbf{A}^*$ is $\mathbf{V}_1\mathbf{V}_1^*$.
7. The orthogonal projection onto the kernel of $\mathbf{A}$ is $\mathbf{I}_m - \mathbf{V}_1\mathbf{V}_1^* = \mathbf{V}_2\mathbf{V}_2^*$.
8. The Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_n^2}$.

Remark 3.2.1. The *short form* of the SVD in (3.10) is

$$\mathbf{A} = \mathbf{U}_1\Sigma_1\mathbf{V}_1^*, \qquad \mathbf{U}_1 \in \mathbb{R}^{n\times r}, \ \Sigma_1 \in \mathbb{R}^{r\times r}, \ \mathbf{V}_1 \in \mathbb{R}^{m\times r},$$

where r is the rank of $\mathbf{A}$. It follows that the outer products in (3.11) are unique, and thus, given a pair of left and right singular vectors $(\mathbf{u}_i, \mathbf{v}_i)$, $i = 1, \ldots, r$, the only other option for this pair is $(-\mathbf{u}_i, -\mathbf{v}_i)$. On the other hand, the columns of $\mathbf{U}_2$ are arbitrary subject to the constraint that they be linearly independent, normalized, and orthogonal to the columns of $\mathbf{U}_1$. Similarly, the columns of $\mathbf{V}_2$ are arbitrary, subject to linear independence, normalization, and orthogonality with the columns of $\mathbf{V}_1$. Thus $\mathbf{U}_2, \mathbf{V}_2$ are not necessary for the computation of the SVD of $\mathbf{A}$.

In MATLAB® the command `svd(A)` computes the full SVD of $\mathbf{A}$, while the command `svds(A,k)` computes a short SVD containing k terms, that is, the first k singular values and singular vectors. The use of the short SVD is recommended for $\min(n, m) \gg 1$.

3.2.3 Comparison with the eigenvalue decomposition

We will now point out some similarities and differences between the *eigenvalue decomposition* (*EVD*) and the SVD. More about the EVD, in particular as it relates to the *Krylov methods*, will be discussed in Chapter 10; for instance, its sensitivity to perturbations, as formalized by means of the concept of *pseudospectra*, will be briefly discussed there.

Given a square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, there exist matrices $\mathbf{X}$ and Λ, with $\det \mathbf{X} \neq 0$, such that

$$\mathbf{A} = \mathbf{X}\Lambda\mathbf{X}^{-1}. \tag{3.12}$$

Λ is in *Jordan form*, i.e., it is a direct sum of Jordan blocks Λ_k of size m_k:

$$\Lambda_k = \lambda_k \mathbf{I}_{m_k} + \mathbf{J}_{m_k} \in \mathbb{C}^{m_k \times m_k},$$

where $\mathbf{J}_{m_k}$ is a nilpotent matrix with ones on the superdiagonal and zeros everywhere else. Every square matrix has an EVD as above, where the λ_k are the *eigenvalues* of $\mathbf{A}$ and the columns of $\mathbf{X}$ are the *eigenvectors* and *generalized eigenvectors* of $\mathbf{A}$. *Symmetric* matrices are *diagonalizable*, i.e., Λ is a diagonal matrix (each Jordan block is scalar); in addition, the set of eigenvectors can be chosen to form an *orthonormal set* in $\mathbb{C}^n$, i.e., $\mathbf{X}$ is unitary:

$$\mathbf{A} = \mathbf{A}^* \quad \Rightarrow \quad \mathbf{X}\mathbf{X}^* = \mathbf{I}_n \text{ and } \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \qquad \lambda_i \in \mathbb{R}.$$

If we require that $\mathbf{X}$ be unitary even if $\mathbf{A}$ is not Hermitian, we obtain the *Schur decomposition*:

$$\mathbf{A} = \mathbf{U}\mathbf{T}\mathbf{U}^*, \quad \mathbf{U}\mathbf{U}^* = \mathbf{I},$$

where $\mathbf{U}$ is unitary and $\mathbf{T}$ is upper triangular with the eigenvalues on the diagonal. Thus the EVD, the SVD, and the Schur decomposition are decompositions of the general form $\mathbf{A}\mathbf{Y} = \mathbf{Z}\Phi$, where Φ is either diagonal or close to diagonal (i.e., Jordan) or upper triangular:

$$\begin{aligned} &\textbf{EVD}: \ \mathbf{A}\mathbf{X} = \mathbf{X}\Lambda, && \det \mathbf{X} \neq 0, \\ &\textbf{SVD}: \ \mathbf{A}\mathbf{V} = \mathbf{U}\Sigma, && \mathbf{V}\mathbf{V}^* = \mathbf{I}_m, \ \ \mathbf{U}\mathbf{U}^* = \mathbf{I}_n. \end{aligned}$$

By forcing $\mathbf{Y} = \mathbf{Z} = \mathbf{X}$, it sometimes happens that Φ is *not* diagonal and $\mathbf{X}$ is not orthogonal. In the SVD, the constraint $\mathbf{Y} = \mathbf{Z}$ is relaxed and replaced by the requirement that $\mathbf{Y}$ and $\mathbf{Z}$ be unitary. This leads to a $\Phi = \Sigma$ matrix which is always diagonal and furthermore its elements are nonnegative. By not requiring that $\mathbf{Y} = \mathbf{Z}$, the condition that $\mathbf{A}$ be square is also relaxed. Hence, in contrast to the EVD, even rectangular matrices have an SVD. In the Schur decomposition, the unitarity of $\mathbf{X}$ forces the matrix $\mathbf{U}^*\mathbf{A}\mathbf{U}$ to be upper triangular with the eigenvalues on the diagonal. A further difference between the EVD and the SVD is illustrated by the following example.

Example 3.5. Consider the 2×2 matrix

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

The eigenvalues of this matrix are 1 and 0 *irrespective* of the value of ϵ. The SVD of $\mathbf{A}$ is

$$\mathbf{A} = I_2 \begin{pmatrix} \sqrt{1+\epsilon^2} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix},$$

$$\cos\theta = \frac{1}{\sqrt{1+\epsilon^2}} \quad \text{and} \quad \sin\theta = \frac{\epsilon}{\sqrt{1+\epsilon^2}}.$$

Hence the largest singular value, i.e., the 2-norm of $\mathbf{A}$, is equal to $\sqrt{1+\epsilon^2}$ and depends on the value of ϵ. This example shows that big changes in the entries of a matrix may have *no effect* on the eigenvalues. This is not true, however, with the singular values. This matrix is revisited in section 10.2 from the point of view of *pseudospectra*.

SVD for symmetric matrices

For symmetric matrices, the SVD can be directly obtained from the EVD. Let the EVD of a given $\mathbf{A} \in \mathbb{R}^{n\times n}$ be

$$\mathbf{A} = \mathbf{A}^* = \mathbf{V}\Lambda\mathbf{V}^*.$$

Define by $\mathbf{S} = \text{diag}\,(\text{sgn}\lambda_1, \ldots, \text{sgn}\lambda_n)$, where $\text{sgn}\lambda$ is the *signum* function; it equals $+1$ if $\lambda \geq 0$, -1 if $\lambda < 0$. Then the SVD of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*, \quad \text{where} \quad \mathbf{U} = \mathbf{V}\mathbf{S} \quad \text{and} \quad \Sigma = \text{diag}\,(|\lambda_1|, \ldots, |\lambda_n|).$$

3.2.4 Optimal approximation in the 2-induced norm

The SVD is the tool that leads to the solution of the problem of *approximating* a matrix by one of lower rank, *optimally* in the 2-induced norm.

Problem. Given $\mathbf{A} \in \mathbb{C}^{n\times m}$, rank $\mathbf{A} = r \leq n \leq m$, find $\mathbf{X} \in \mathbb{C}^{n\times m}$, rank $\mathbf{X} = k < r$, such that the 2-norm of the *error* matrix $\mathbf{E} = \mathbf{A} - \mathbf{X}$ is minimized.

The solution of this problem is given in the following result, due to four researchers. Schmidt derived a version of it in the early 1900s in connection with integral equations [289]. Then two researchers from the quantitative social sciences, Eckart and Young, published another version of this result in the mid 1930s [101]. Finally, Mirsky in the 1960s derived a general version that is valid for all unitarily invariant norms [240]. A detailed account of this result can be found in the book by Stewart and Sun [314].

Theorem 3.6. Schmidt–Eckart–Young–Mirsky. *With the notation introduced above,*

$$\min_{\mathbf{X},\ \text{rank}\ \mathbf{X}=k} \|\mathbf{A} - \mathbf{X}\|_2 = \sigma_{k+1}(\mathbf{A}), \tag{3.13}$$

provided that $\sigma_k > \sigma_{k+1}$. A (nonunique) minimizer $\mathbf{X}_*$ is obtained by truncating the dyadic decomposition (3.11) *to contain the first k terms:*

$$\mathbf{X}_* = \sigma_1\mathbf{u}_1\mathbf{v}_1^* + \sigma_2\mathbf{u}_2\mathbf{v}_2^* + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^*. \tag{3.14}$$

The proof of this result is based on the next lemma.

Lemma 3.7. *Given* $\mathbf{A}$ *of rank r for all* $\mathbf{X}$ *of rank less than or equal to k, there holds*

$$\|\mathbf{A}-\mathbf{X}\|_2 \geq \sigma_{k+1}(\mathbf{A}). \tag{3.15}$$

Proof of lemma. Let $\mathbf{y}_i \in \mathbb{C}^m$, $i = 1, \ldots, m-k$, be a basis for the kernel of $\mathbf{X}$

$$\ker \mathbf{X} = \operatorname{span}\{\mathbf{y}_1, \ldots, \mathbf{y}_{m-k}\}.$$

From a dimension argument follows that the intersection of the two spans is nontrivial:

$$\operatorname{span}\{\mathbf{y}_1, \ldots, \mathbf{y}_{m-k}\} \cap \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{k+1}\} \neq \{0\}.$$

Let $\mathbf{z} \in \mathbb{C}^m$, $\mathbf{z}^*\mathbf{z} = 1$, belong to this intersection. It follows that $(\mathbf{A}-\mathbf{X})\mathbf{z} = \mathbf{A}\mathbf{z}$; thus

$$\|\mathbf{A}-\mathbf{X}\|_2^2 \geq \|(\mathbf{A}-\mathbf{X})\mathbf{z}\|_2^2 = \|\mathbf{A}\mathbf{z}\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2(\mathbf{v}_i^*\mathbf{z})^2 \geq \sigma_{k+1}^2.$$

This completes the proof. □

Proof of the theorem. There remains to show that the lower bound obtained in the above lemma is attained. The minimizer (3.14) does precisely this. □

Remark 3.2.2. **(a)** *Optimal approximation in the Frobenius norm.* As it turns out, the Schmidt–Mirsky–Eckart–Young theorem provides the solution to the optimal approximation problem in the Frobenius norm as well. The minimum value of the error (3.13) becomes in this case

$$\min_{\mathbf{X},\,\operatorname{rank}\mathbf{X}=k} \|\mathbf{A}-\mathbf{X}\|_F = \left(\sum_{i=k+1}^{m} \sigma_i^2(\mathbf{A})\right)^{\frac{1}{2}}.$$

Furthermore, provided that $\sigma_k > \sigma_{k+1}$, there is a *unique* minimizer given by (3.14).

(b) The importance of the Schmidt–Mirsky result is the relationship between the rank k of the approximant and the $(k+1)$st singular value of $\mathbf{A}$.

(c) The minimizer (3.14) given in the Schmidt–Eckart–Young–Mirsky theorem is not unique in the 2-induced norm. A class of minimizers is given as follows:

$$\mathbf{X}(\eta_1, \ldots, \eta_k) = \sum_{i=1}^{k}(\sigma_i - \eta_i)\mathbf{u}_i\mathbf{v}_i^*, \quad \text{where } 0 \leq |\eta_i| \leq \sigma_{k+1}. \tag{3.16}$$

This property is important in the Hankel-norm approximation problem. See, e.g., formula (8.23) in Example 8.9.

(d) It is easy to see that the rank of the sum (or the linear combination) of two matrices is in general not equal to the sum of their ranks. This is also true if the linear combination is *convex*, that is, the coefficients are nonnegative and their sum is equal to one. As a consequence, the problem of minimizing the 2-induced norm of $\mathbf{A}-\mathbf{X}$ over all matrices $\mathbf{X}$

of rank (at most) k is a *nonconvex* problem, and there is little hope of solving it by applying general optimization algorithms.

(e) To obtain a larger set of solutions to the problem of optimal approximation in the 2-induced norm, one may relax the optimality condition (3.13). Instead, one may wish to obtain all matrices $\mathbf{X}$ of rank k satisfying

$$\sigma_{k+1}(\mathbf{A}) < \|\mathbf{A}-\mathbf{X}\|_2 < \epsilon,$$

where $\sigma_{k+1}(\mathbf{A}) < \epsilon < \sigma_k(\mathbf{A})$. This is the *suboptimal* approximation problem in the 2-induced norm, which was solved in [331].

Example 3.8. Here we consider the 4 × 3 matrix

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

which has rank 3. We will compute optimal approximations of rank 1 and 2.

The singular values of $\mathbf{T}$ are $\sigma_1 = 3 + \sqrt{7} = 2.38$, $\sigma_2 = 1$, $\sigma_3 = 3 - \sqrt{7} = 0.59$. The approximants T_1, T_2, of rank 1, 2, respectively, obtained by means of the dyadic decomposition, turn out to be

$$\mathbf{T}_1 = \begin{bmatrix} 0.69 & 1.07 & 0.38 \\ 0.42 & 0.65 & 0.23 \\ 0.42 & 0.65 & 0.23 \\ 0.84 & 1.30 & 0.46 \end{bmatrix}, \quad \mathbf{T}_2 = \begin{bmatrix} 0.69 & 1.07 & 0.38 \\ 0.08 & 0.98 & -0.10 \\ 0.08 & 0.98 & -0.10 \\ 1.17 & 0.96 & 0.79 \end{bmatrix}.$$

The approximants shown in Figure 3.3 should be interpreted as follows. A pixel with value 0/1 represents black/white; in-between values represent levels of gray. All pixels with value bigger than 1 and smaller than 0 are approximated by 1 (white) and 0 (black), respectively.

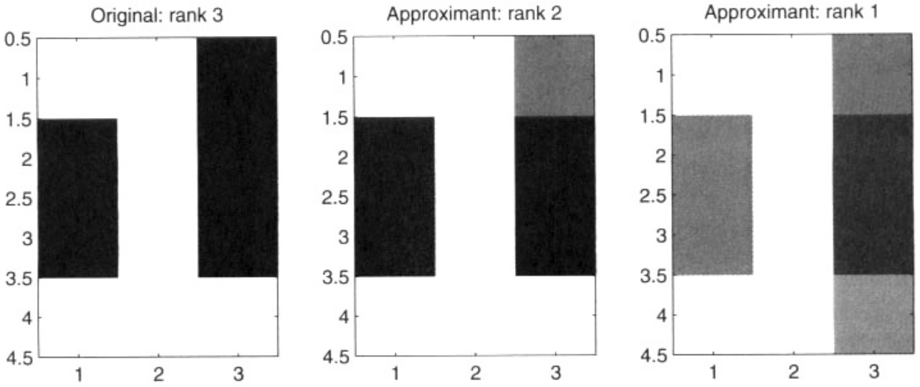


Figure 3.3. *Approximation of an image using the SVD. Left: original. Middle: rank 2. Right: rank 1 approximants.*

Notice that the norm of the errors, i.e., the largest singular value of $\mathbf{E}_2 = \mathbf{T} - \mathbf{T}_2$, is $\sigma_2 = 1$ and that of $\mathbf{E}_1 = \mathbf{T} - \mathbf{T}_1$ is $\sigma_3 = .59$. It also follows that no entry of $\mathbf{E}_2$ can exceed σ_3, while no entry of $\mathbf{E}_1$ can exceed $\sqrt{\sigma_2^2 + \sigma_3^2} = 1.16$.

We notice that, to be able to represent the results graphically, **T** must be considered as a 12-pixel image of the number one. In MATLAB each pixel is depicted as a gray square; there are 128 levels of gray, with 1 being white and 0 being black; numbers that are not integer multiples of $\frac{1}{128}$ are approximated by the nearest integer multiple of $\frac{1}{128}$; finally, numbers less than zero are depicted by black pixels, while numbers bigger than one are depicted by white pixels. Figure 3.3 shows a pictorial representation in MATLAB of the original figure and its approximants.

3.2.5 Further applications of the SVD

The SVD can be used to tackle a number of other problems, some of which we mention here. First is the determination of the *rank* of a matrix. Because the SVD is well-conditioned, it can be used to determine both the *numerical* and the *actual* rank of a matrix. This is done by counting the number of singular values that are above a certain threshold. The threshold is zero for the actual rank and some small number determined by the user according to the application at hand for the *numerical rank.*

The second application of the SVD is the calculation of the *Moore–Penrose pseudoinverse*. Given a not necessarily invertible or even square matrix $\mathbf{A} \in \mathbb{C}^{n\times m}$, a *pseudoinverse* or *generalized inverse* is defined as a matrix $\mathbf{X} \in \mathbb{C}^{m\times n}$ that satisfies the relationships

$$\mathbf{AXA} = \mathbf{A}, \quad \mathbf{XAX} = \mathbf{X}, \quad (\mathbf{AX})^* = \mathbf{AX}, \quad (\mathbf{XA})^* = \mathbf{XA}.$$

The problem can also be formulated equivalently as follows. Given **A** as above, find **X** that solves

$$\min_{\mathbf{X}} \|\mathbf{AX} - \mathbf{I}_n\|_F.$$

The solution can now be given using the SVD of **A**. Let $\bar{\Sigma} = \text{diag}(\Sigma_1, 0)$, where $\Sigma_1 > 0$. It follows that

$$\mathbf{X} = \mathbf{V}\begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}\mathbf{U}^*. \tag{3.17}$$

Often the pseudoinverse of **A** is denoted by $\mathbf{A}^+$.

The least squares problem is closely related to the pseudoinverse of a matrix. Given are $\mathbf{A} \in \mathbb{C}^{n\times m}$, $n \geq m$, and a vector $\mathbf{b} \in \mathbb{C}^n$. We wish to find $\mathbf{x} \in \mathbb{C}^m$, which solves

$$\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2.$$

We denote the norm of this residual by ρ_{LS}. The solution is

$$\mathbf{x}_{LS} = \mathbf{Xb} = \sum_{i=1}^{m} \frac{\mathbf{u}_i^*\mathbf{b}}{\sigma_i}\mathbf{v}_i, \quad \rho_{LS}^2 = \sum_{i=m+1}^{n} (\mathbf{u}_i^*\mathbf{b})^2.$$

The final problem discussed here is that of the stability radius in robust control. Given **M**, the *complex stability radius* is the inverse of the smallest 2-norm of a complex Δ such that

$$\mathbf{I}_n - \Delta\mathbf{M}, \qquad \Delta \in \mathbb{C}^{m\times n}, \ \mathbf{M} \in \mathbb{C}^{n\times m},$$

is singular. Applying the Schmidt–Eckart–Young–Mirsky theorem, it can be shown that the (complex) stability radius is equal to $\sigma_1(\mathbf{M})$. It was recently shown that the *real* stability radius (i.e., $\Delta \in \mathbb{R}^{n\times n}$) can be computed as the second singular value of a real matrix derived from the complex matrix $\mathbf{M}$.

3.2.6 The semidiscrete decomposition

A variant of the SVD is the semidiscrete decomposition (SDD) of a matrix. This decomposition was introduced in [255]. Recall the dyadic decomposition (3.11). The rank k approximation of a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ by means of the SDD is given by a different dyadic decomposition:

$$\mathbf{A}_k = \delta_1\mathbf{x}_1\mathbf{y}_1^* + \delta_2\mathbf{x}_2\mathbf{y}_2^* + \cdots + \delta_k\mathbf{x}_k\mathbf{y}_k^* = \mathbf{X}_k\Delta_k\mathbf{Y}_k^*.$$

In this expression the entries of the vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ are restricted to belong to the set $\{-1, 0, 1\}$, while $\delta_i \in \mathbb{R}$ are real numbers. The first remark is that this decomposition does not reproduce $\mathbf{A}$ even for $k = n$ or $k = m$. The advantage obtained comes from the fact that the rank k approximation requires storage of only $2k(n+m)$ bits plus k scalars as opposed to $(n+m+1)k$ scalars for the rank k approximation of $\mathbf{A}$ using the SVD.

The computation of this decomposition proceeds iteratively. Given $\mathbf{A}_{k-1}$ we seek $\mathbf{x}, \mathbf{y}, \delta$ such that the Frobenius norm of the error matrix $\mathbf{E}_{k-1}$ is minimized:

$$\|\mathbf{E}_{k-1} - \delta\mathbf{x}\mathbf{y}^*\|_F, \ \mathbf{E}_{k-1} = \mathbf{A} - \mathbf{A}_{k-1}.$$

This problem is solved iteratively by fixing $\mathbf{y}$ and solving for $\mathbf{x} \in \{-1, 0, 1\}^n$ and $\delta \in \mathbb{R}$ and, subsequently, by fixing $\mathbf{x}$ and solving for $\mathbf{y} \in \{-1, 0, 1\}^m$ and $\delta \in \mathbb{R}$. The iteration is stopped when the difference in two successive iterates of δ drops below a specified tolerance.

Example 3.9. Applying the iterative SDD approximation to Example 3.8, we obtain approximants $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{T}}_2$ of rank one and two, respectively:

$$\hat{\mathbf{T}}_1 = \frac{3}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \ \hat{\mathbf{T}}_2 = \hat{\mathbf{T}}_1 - \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 3 & 3 & 0 \\ -1 & 3 & 0 \\ -1 & 3 & 0 \\ 3 & 3 & 0 \end{bmatrix}.$$

The corresponding largest singular value of the error matrices $\mathbf{T} - \hat{\mathbf{T}}_1$, $\mathbf{T} - \hat{\mathbf{T}}_2$ is 1.188, 1.0877.

3.3 Basic numerical analysis

In this section, we present a brief introduction to issues of accuracy in numerical computations. Two concepts are involved. First, the *condition* of the problem is a measure of the sensitivity of the solution to perturbations in the data. The condition of a problem does not depend on the algorithm used to compute an answer. Second, *error analysis*, for a *given* solution algorithm, quantifies the error due to *floating point* arithmetic. A *forward error analysis* is concerned with how close the computed solution is to the exact solution, while a *backward error analysis* is concerned with whether the computed inexact solution can be interpreted as the exact solution of the same problem with perturbed initial data. For problems of linear algebra, it is often advantageous to perform a backward error analysis. A knowledge of the backward error, if it exists, together with a knowledge of the sensitivity of the problem allow us to measure the forward error.

Besides references listed at the beginning of this chapter, we also recommend the lecture notes by Sorensen [308] and Van Dooren [334], [335], [337], [338] and the papers by Higham et al. [172] and Smale [301]. See also [206].

3.3.1 Condition numbers

First we examine the problem of estimating the effect of perturbations on *input data*. This leads to the concept of *condition number* of the problem at hand. A problem is ill-conditioned if *small* changes in the data cause relatively *large* changes in the solution. Otherwise a problem is well-conditioned.

First, let us look at the problem of the sensitivity of evaluating a given function f at a point x_0. Assume that $f : \mathbb{R} \rightarrow \mathbb{R}$ is twice differentiable in a neighborhood of x_0. If x_0 is perturbed to $x_0 + \delta$, a Taylor series expansion of $f(x)$ around x_0 gives $f(x_0+\delta) = f(x_0) + f'(x_0)\delta + O(|\delta|^2)$, where prime denotes the derivative with respect to x. The *absolute condition* number is precisely this derivative $f'(x_0)$. It tells us the amount by which infinitesimal perturbations in x_0 will be amplified to produce the perturbation in the evaluation. For the *relative condition number*, notice that

$$\frac{|f(x_0+\delta) - f(x_0)|}{|f(x_0)|} = \frac{|f'(x_0)| \cdot |x_0|}{|f(x_0)|} \cdot \frac{|\delta|}{|x_0|} + O(|\delta^2|).$$

This condition number is defined as the absolute value of the ratio of the *relative change of* $f(x)$ over the *relative change in* x:

$$\kappa_f(x_0) = \frac{|f'(x_0)| \cdot |x_0|}{|f(x_0)|}.$$

The interpretation of the condition number is that given a small relative error in x_0, the resulting relative error in $f(x_0)$ will be amplified by $\kappa_f(x_0)$.

Next we will examine the condition number of evaluating the vector-valued function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ at a point $\mathbf{x}_0 \in \mathbb{R}^n$. The Taylor series expansion becomes in this case

$$\mathbf{f}(\mathbf{x}_0 + \delta) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J}^*(\mathbf{x}_0)\delta + O(\|\delta\|^2),$$

where $\delta \in \mathbb{R}^n$ is a vector perturbation and $\mathbf{J}(\mathbf{x}_0)$ is the *Jacobian* of $\mathbf{f}$ evaluated at $\mathbf{x}_0$: $\mathbf{J}_{i,j}(\mathbf{x}_0) = \frac{\partial \mathbf{f}_j}{\partial \mathbf{x}_i}|_{\mathbf{x}=\mathbf{x}_0}$. In this case, a perturbation δ in $\mathbf{x}_0$ will result in a perturbation $\mathbf{J}(\mathbf{x}_0)\delta$ in $\mathbf{f}(\mathbf{x}_0)$, which depends on the *direction* of δ. Thus assuming that 2-norms are used, the amplification factor for the perturbations lies between the largest and the smallest singular values of the Jacobian $\mathbf{J}(\mathbf{x}_0)$. The *absolute condition number* is defined as the largest singular value, i.e., the (2-induced) norm of the Jacobian $\|\mathbf{J}(\mathbf{x}_0)\|_2$; more generally, it is the norm $\|\mathbf{J}(\mathbf{x}_0)\|$ induced by the vector norms in the domain and the range of $\mathbf{f}$. The *relative condition number* is defined as

$$\kappa_{\mathbf{f}}(\mathbf{x}_0) = \frac{\|\mathbf{J}(\mathbf{x}_0)\| \cdot \|\mathbf{x}_0\|}{\|\mathbf{f}(\mathbf{x}_0)\|},$$

where the matrix norm in the above expression is induced by the vector norm used. Again, given a relative error in $\mathbf{x}_0$, the largest resulting relative error in $\mathbf{f}(\mathbf{x}_0)$ will be amplified by $\kappa_\mathbf{f}(\mathbf{x}_0)$.

In the next section we will make use of the *condition number of the inner product*. For $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, we wish to compute the condition number of the inner product $\mathbf{a}^* \cdot \mathbf{x}$ at $\mathbf{x} = \mathbf{x}_0$. Since the Jacobian in this case is $\mathbf{J} = \mathbf{a} \in \mathbb{R}^n$, the earlier formula yields

$$\kappa = \frac{\|\mathbf{a}\|\|\mathbf{x}_0\|}{|\mathbf{a}^*\mathbf{x}_0|}. \tag{3.18}$$

The condition number of a matrix-vector product $\mathbf{Ab}$ can be determined as follows. Starting from the equality $\mathbf{b} = \mathbf{AA}^{-1}\mathbf{b}$, we derive the inequality $\|\mathbf{b}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{Ab}\|$ and hence $\frac{\|\mathbf{b}\|}{\|\mathbf{A}^{-1}\|} \leq \|\mathbf{Ab}\|$. Using this last inequality together with the relative error we obtain the following upper bound:

$$\frac{\|\mathbf{Ab} - \mathbf{A}\hat{\mathbf{b}}\|}{\|\mathbf{A}\hat{\mathbf{b}}\|} \leq \frac{\|\mathbf{A}\|\|\mathbf{b} - \hat{\mathbf{b}}\|}{\frac{1}{\|\mathbf{A}^{-1}\|} \cdot \|\hat{\mathbf{b}}\|} = \underbrace{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}_{\kappa} \frac{\|\mathbf{b} - \hat{\mathbf{b}}\|}{\|\hat{\mathbf{b}}\|}.$$

Thus the condition number, i.e., the largest amplification of perturbations in $\mathbf{b}$, is $\kappa = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$. This is called the *condition number* of the matrix $\mathbf{A}$. This formula implies that the condition number of $\mathbf{A}$ is the same as that of its inverse $\mathbf{A}^{-1}$; therefore, the condition number for solving the set of linear equation $\mathbf{Ax} = \mathbf{b}$, $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, due to perturbations in $\mathbf{b}$, is the same as that for $\mathbf{Ab}$.

Finally, we turn our attention to the condition number of the linear system of equations $\mathbf{Ax} = \mathbf{b}$, where both $\mathbf{b}$ and $\mathbf{A}$ are perturbed. The system actually solved is $\hat{\mathbf{A}}\hat{\mathbf{x}} = \hat{\mathbf{b}}$, where the norms of $\mathbf{A} - \hat{\mathbf{A}}$ and $\mathbf{b} - \hat{\mathbf{b}}$ are small. It can be shown (see, e.g., [171, p. 145]) that the relative error in the solution has the upper bound

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa \left(\frac{\|\mathbf{A} - \hat{\mathbf{A}}\|}{\|\mathbf{A}\|} + \frac{\|\mathbf{b} - \hat{\mathbf{b}}\|}{\|\mathbf{b}\|} \right), \quad \text{where } \kappa = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|.$$

Hence the relative error in the solution can be amplified up to κ times with respect to the relative error in the data. Again, if the 2-induced norm is used, it follows from the SVD of $\mathbf{A}$ that $\|\mathbf{A}\|_2 = \sigma_1$, $\|\mathbf{A}^{-1}\|_2 = \frac{1}{\sigma_n}$, and thus the condition number

$$\kappa = \frac{\sigma_1(\mathbf{A})}{\sigma_n(\mathbf{A})} \tag{3.19}$$

is the ratio of the largest to the smallest singular values of $\mathbf{A}$.

Condition number of EVD and SVD

A problem of considerable importance is that of computing the eigenvalue or the singular value decompositions of a given matrix. The sensitivity of these calculations to uncertainty in the data must therefore be studied. Given the square matrix $\mathbf{A}$, let $\mathbf{x}$, $\mathbf{y}$ be the right, left eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ:

$$\mathbf{Ax} = \lambda\mathbf{x}, \quad \mathbf{y}^*\mathbf{A} = \lambda\mathbf{y}^*.$$

If $\mathbf{A}$ is perturbed to $\mathbf{A}+\Delta\mathbf{A}$, where $\|\Delta\mathbf{A}\|$ is small, the other quantities in the above equation will be perturbed too:

$$(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\delta\mathbf{x}) = (\lambda+\delta\lambda)(\mathbf{x}+\delta\mathbf{x}).$$

Our goal is to describe the change $\delta\lambda$ for infinitesimal perturbations in $\mathbf{A}$. Expanding and neglecting second-order terms in the perturbations, followed by left multiplication by $\mathbf{y}^*$, we obtain $\delta\lambda = \frac{\mathbf{y}^*\Delta\mathbf{A}\mathbf{x}}{\mathbf{y}^*\mathbf{x}}$. This implies that $\frac{|\lambda|}{\|\Delta\mathbf{A}\|} \le \frac{\|\mathbf{y}\|\|\mathbf{x}\|}{|\mathbf{y}^*\mathbf{x}|}$. The maximum of this expression, namely,

$$\kappa_\lambda = \frac{\|\mathbf{y}\|\,\|\mathbf{x}\|}{|\mathbf{y}^*\mathbf{x}|},$$

is called the *absolute condition number of the eigenvalue* λ. For details, see Chapter 2 of the book by Wilkinson [355].

When there is a nontrivial Jordan block associated with λ, then $\mathbf{x} \perp \mathbf{y}$, so their inner product is zero, and hence the conditioning is the worst possible. For this reason, the numerical computation of the Jordan canonical form is an ill-posed problem. In this case, the linearization is invalid, and the eigenvalues associated with the Jordan block split as $\|\Delta\mathbf{A}\|^{1/i}$, where i is the index of λ.

In the special case where the matrix is *Hermitian*, $\mathbf{A} = \mathbf{A}^*$, we have that $\mathbf{x} = \mathbf{y}$, and therefore the condition number of computing eigenvalues of symmetric matrices is perfect $\kappa_\lambda = 1$. Consequently, since the singular values of a matrix are the square roots of the eigenvalues of symmetric (and positive semidefinite) matrices, namely, $\mathbf{A}\mathbf{A}^*$ and $\mathbf{A}^*\mathbf{A}$, the conditioning of the singular values is also perfect: $\kappa_\sigma = 1$.

To study the condition number of the *eigenvectors*, we need to study the sensitivity of the corresponding eigenspaces. These results will not be discussed here; see, e.g., Golub and Van Loan [144] for details. We quote one simple case that gives the type of results obtained. Suppose that λ is an eigenvalue of algebraic multiplicity one. Let $\hat{\lambda}$ denote the eigenvalue of the matrix under consideration that is closest to λ. Then a small perturbation in $\mathbf{A}$ will result in a perturbation of the eigenvector that is inversely proportional to the difference $|\lambda - \hat{\lambda}|$. Thus the smaller the gap between the eigenvalue of interest and the remaining ones is, the more sensitive the calculation of the corresponding eigenvector becomes. A similar result holds for eigenvectors of symmetric matrices and hence of singular vectors of matrices.

3.3.2 Stability of solution algorithms

The stability of a method for solving a problem (algorithm) is concerned with the sensitivity of the method to *rounding errors* in the solution process. A method that guarantees as accurate a solution as the data warrants is said to be *stable*; otherwise the method is unstable.

Error analysis is concerned with establishing whether an algorithm is stable for the problem at hand. A forward error analysis is concerned with how close the computed solution is to the exact solution. A backward error analysis is concerned with how well the computed solution satisfies the problem to be solved. The purpose of error analysis is the discovery of the factors determining the stability of an algorithm.

Numbers in a computer are represented by means of finitely many bits. Therefore the precision is also finite. There exist many varieties of *floating point arithmetic*. By far, the most commonly used is the ANSI/IEEE 745-1985 standard for binary floating point arithmetic [186], [171], [256]. In this arithmetic, *double precision* corresponds to a quantization error,

$$\boxed{\epsilon = 2^{-53} \approx 1.1 \cdot 10^{-16},} \tag{3.20}$$

which is also referred to as *machine precision*. Moreover, only real numbers x in the interval between $[-N_{\max}, N_{\max}]$ can be represented, where for double precision arithmetic $N_{\max} = 2^{1024} \approx 1.79 \cdot 10^{308}$. Given a real number x in the above interval, we will denote by $fl(x)$ its representation in floating point arithmetic, where $|x - fl(x)| \leq \epsilon |x|$, that is,

$$fl(x) = x \cdot (1 + \delta), \quad \text{where } |\delta| \leq \epsilon. \tag{3.21}$$

The operations of floating point addition and multiplication in IEEE arithmetic yield a result that is up to ϵ close to the true result, namely,

$$fl(x \pm y) = (x \pm y) \cdot (1 + \delta), \quad fl(x \cdot y) = (x \cdot y) \cdot (1 + \delta), \quad \text{and} \quad fl(x/y) = (x/y) \cdot (1 + \delta).$$

There are two ways to interpret the error due to floating point arithmetic. Let our goal be to compute $\mathbf{f}(\mathbf{x})$, i.e., evaluate the function $\mathbf{f}$ at the point $\mathbf{x}$. Using our favorite algorithm denoted by A, due to floating point errors during its execution, the result obtained will be $\mathbf{f}_A(\mathbf{x}) \neq \mathbf{f}(\mathbf{x})$. It is now assumed that there is a point $\mathbf{x}_A$ such that $\mathbf{f}(\mathbf{x}_A) = \mathbf{f}_A(\mathbf{x})$; this means that there exists a point that is mapped to the actual result under the function $\mathbf{f}$. In the evaluation of the function above at the given point, four quantities are involved, namely, $\mathbf{f}$, $\mathbf{f}_A$, $\mathbf{x}$, $\mathbf{x}_A$. Forward stability is defined using the first two, while backward stability is defined using the latter two.

Forward error is defined as $\|\mathbf{f}_A(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|$. In other words, the forward error aims at estimating how close the final result is to the exact result.

Backward error is defined as $\|\mathbf{x}_A - \mathbf{x}\|$, that is, it interprets this same result as an exact calculation on perturbed data $\mathbf{x}_A$ and seeks to determine how close $\mathbf{x}_A$ is to $\mathbf{x}$.

$$\begin{array}{lcll}
\mathbf{x} & \mapsto & \mathbf{f}(\mathbf{x}), & \text{exact arithmetic on given input data,} \\
\mathbf{x} & \mapsto & \mathbf{f}_A(\mathbf{x}) = fl(\mathbf{f}(\mathbf{x})), & \text{floating point arithmetic on given input data,} \\
\mathbf{x}_A & \mapsto & \mathbf{f}(\mathbf{x}_A) = \mathbf{f}(fl(\mathbf{x})), & \text{exact arithmetic on different input data,}
\end{array}$$

$$\begin{array}{rl}
\|\mathbf{f}_A(\mathbf{x}) - \mathbf{f}(\mathbf{x}_A)\| : & \text{forward error,} \\
\|\mathbf{x}_A - \mathbf{x}\| : & \text{backward error.}
\end{array}$$

We are now ready to define stability of an algorithm.

The algorithm $\mathbf{f}_A(\mathbf{x})$ for computing $\mathbf{f}(\mathbf{x})$ is *forward stable* if

$$\mathbf{f}_A(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \mathbf{e}_f, \quad \text{where } \|\mathbf{e}_f\| = O(\epsilon)\|\mathbf{f}(\mathbf{x})\|. \tag{3.22}$$

The same algorithm is called *backward stable* if

$$\mathbf{f}_A(\mathbf{x}) = \mathbf{f}(\mathbf{x} + \mathbf{e}_x), \quad \text{where } \|\mathbf{e}_x\| = O(\epsilon)\|\mathbf{x}\|, \tag{3.23}$$

where $O(\epsilon)$ is interpreted as a polynomial of modest degree in the size n of the problem times ϵ.

The two types of error introduced above are related through a Taylor series expansion,

$$\mathbf{f}_A(\mathbf{x}) = \mathbf{f}(\mathbf{x}_A) = \mathbf{f}(\mathbf{x}) + \mathbf{J}^*(\mathbf{x})\,(\mathbf{x}_A - \mathbf{x}) + O(\|\mathbf{x}_A - \mathbf{x}\|^2).$$

Thus if the norm of the difference $\mathbf{x}_A - \mathbf{x}$ is small (say, on the order of machine precision), the forward error can be upper bounded as follows:

$$\|\mathbf{f}_A(\mathbf{x}) - \mathbf{f}(\mathbf{x}_A)\| \le \kappa_{\mathbf{f}}(\mathbf{x})\|\mathbf{x}_A - \mathbf{x}\|.$$

Therefore, if the backward error is small, the forward error is bounded from above by the product of the backward error and the condition number (see [171, p. 10]). We thus obtain the following important rule of thumb.

Rule of Thumb: Forward Error ≤ Condition Number × Backward Error

The inner product in floating point arithmetic

As an illustration of numerical stability, we will now compute the floating point inner product of two vectors $\mathbf{x} = (x_i)$, $\mathbf{y} = (y_i) \in \mathbb{R}^n$; with $s_i = x_i \cdot y_i$, the desired inner product is obtained by summing these quantities, $\mathbf{x}^* \cdot \mathbf{y} = \sum_{i=1}^n x_i \cdot y_i$.

In floating point arithmetic the *order* of summation plays a role. Here is a simple example.

Example 3.10. The following experiment is conducted in MATLAB. Let $a = \frac{1}{2}$, $b = -\frac{1}{3}$, $c = -\frac{1}{6}$. In exact arithmetic the sum of these three fractions is zero, and that holds for summation in any order. However, in floating point arithmetic we get $(a + c) + b = (c+a)+b = 5.55 \cdot 10^{-17}$, $(a+b)+c = (b+a)+c = 2.77 \cdot 10^{-17}$, $(b+c)+a = (c+b)+a = 0$. The above differences are due to the fact that $fl(a) = 5.000000000000000 \cdot 10^{-1}$, $fl(b) = -3.333333333333333 \cdot 10^{-1}$, $fl(c) = -1.666666666666667 \cdot 10^{-1}$; thus a is represented without error, while the sum $b+c$ happens to be exactly representable in floating point arithmetic, as well.

Therefore our algorithm for computing the inner product will consist of choosing a particular order of summation, namely, $S_k = s_k + S_{k-1}$, $2 \le k \le n$, $S_1 = s_1$, which implies that S_n is the desired result (or more precisely, an approximation thereof). In floating point arithmetic, the following sums and products are computed:

$$\begin{aligned}
fl(s_i) &= fl(x_i \cdot y_i) = x_i \cdot y_i \cdot (1 + \epsilon_i),\ i = 1, \ldots, n,\ |\epsilon_i| \le \epsilon,\\
fl(S_2) &= fl(fl(s_2) + fl(s_1)) = [s_1 \cdot (1 + \epsilon_1) + s_2 \cdot (1 + \epsilon_2)] \cdot (1 + \delta_2),\ |\delta_2| \le \epsilon,\\
fl(S_3) &= fl(fl(s_3) + fl(S_2)) = [s_3 \cdot (1 + \epsilon_3) + fl(S_2)] \cdot (1 + \delta_3),\\
fl(S_n) &= fl(fl(s_n) + fl(S_{n-1}))\\
&= [s_n(1 + \epsilon_n) + fl(S_{n-1})] \cdot (1 + \delta_n)\\
&= \sum_{i=1}^n x_i y_i \pi_i, \quad \text{where } \pi_k = (1 + \epsilon_k) \prod_{i=k}^n (1 + \delta_i),\ k = 1, \ldots, n,\ \delta_1 = 0.
\end{aligned}$$

Let $\mu_k = \pi_k - 1$. Then we can write

$$fl(S_n) = \sum_{i=1}^{n} x_i y_i (1 + \underbrace{\mu_i}_{\mathbf{e_x}}) = \sum_{i=1}^{n} x_i y_i + \underbrace{\sum_{i=1}^{n} x_i y_i \mu_i}_{\mathbf{e_f}},$$

where $\mathbf{e_x}$ is the backward error and $\mathbf{e_f}$ is the forward error. We will now estimate μ_k. Since ϵ_i and δ_i are bounded in absolute value by machine precision, $|\pi_k| \leq (1 + \epsilon)^n = 1 + n\epsilon + n(n-1)\frac{\epsilon^2}{2} + \cdots$. Assuming that $n\epsilon < 1$, it can be shown that the largest μ_k, denoted by μ, is:

$$\mu_k \leq \mu = n\epsilon \left[1 + \frac{n\epsilon}{2}\right].$$

The assumption $n\epsilon < 1$ implies that the number of terms in the inner product cannot be too big.

As indicated in the previous section, the error can be interpreted in two different ways—one resulting from perturbation of the data and another resulting from perturbation of the final result. We will estimate both of these errors. In the former case, we assume that there is no perturbation in $\mathbf{x}$, while there is perturbation in $\mathbf{y}$: $\hat{\mathbf{y}} = (\mu_i y_i)$ and $\mathbf{y} = (y_i)$. Thus $\|\hat{\mathbf{y}} - \mathbf{y}\| \leq \mu \|\mathbf{y}\|$, and the relative error in $\mathbf{y}$ has the upper bound μ:

$$\frac{\|\hat{\mathbf{y}} - \mathbf{y}\|}{\|\mathbf{y}\|} \leq \mu.$$

This is the *backward error*. Since $\mu \approx n\epsilon$, the algorithm used is backward stable. The *forward error* $\sum_{i=1}^{n} x_i y_i \mu_i$ can be bounded as $|\sum_{i=1}^{n} x_i y_i \mu_i| \leq \mu \sum_{i=1}^{n} |x_i||y_i|$, and thus the relative error has the upper bound

$$\mu\kappa, \quad \text{where } \kappa = \frac{\sum_{i=1}^{n} |x_i||y_i|}{\left|\sum_{i=1}^{n} x_i y_i\right|}.$$

Recall from the previous section that κ is the condition number of the inner product. Thus, as expected, the forward error is bounded from above by the product of the backward error times the condition number of the problem at hand.

Example 3.11. *Backward and forward stable inner product calculation.* A simple example in MATLAB illustrates the issues discussed above. Consider the column vector of length 1 million: `b=ones(1000000,1);` let `a` be the row vector of the same length: `a=exp(1)*b'`. We compute the inner product `a*b`. The exact result gives $\mathbf{a} \cdot \mathbf{b} = e \cdot 10^6$. The result in MATLAB gives

$$\texttt{2.718281828476833} \cdot \texttt{10}^6,$$

while `exp(1)` = `2.718281828459054`. Therefore, we have lost five significant digits of accuracy. This can be explained as follows. According to the considerations above, the backward error is of the order $n\epsilon \approx 10^6 \cdot 10^{-16} = 10^{-10}$. The condition number for this problem, given by (3.18), is 1. Therefore according to the rule of thumb given above, the forward error is at most equal to the backward error. Indeed, the backward error implies a loss of precision of six digits, while the actual forward error shows a loss of precision of five digits.

Example 3.12. *Ill-conditioned inner product gives large forward error.* We will investigate the computation of the exponentials $\mathtt{E}_\alpha = \mathtt{exp}(\alpha\mathtt{t})$, for $t > 0$, and $\alpha = \pm 1$, using the truncated power series expansion:

$$\mathtt{exp}(\alpha\mathtt{t}) = 1 + \frac{\alpha\mathtt{t}}{1!} + \frac{\alpha^2\mathtt{t}^2}{2!} + \cdots + \frac{\alpha^{\mathtt{n}-1}\mathtt{t}^{\mathtt{n}-1}}{(\mathtt{n}-1)!} = \mathbf{x}^* \cdot \mathbf{y},$$

where the ith entries of these vectors are $(\mathbf{x})_i = \alpha^{i-1}$ and $(\mathbf{y})_i = \frac{t^{i-1}}{(i-1)!}$, $i = 1, \dots, n$. As shown earlier in this section, the inner product is backward stable since $\|\mathbf{e}_\mathbf{x}\| \leq n\epsilon$; to determine forward stability, we need the condition number of the inner product, which was computed in (3.18), namely, $\kappa = \frac{\|\mathbf{x}\|\|\mathbf{y}\|}{|\mathbf{x}^* \cdot \mathbf{y}|}$. Thus for $\alpha = 1$, it follows that $\kappa \leq \sqrt{n}$. For $\alpha = -1$, however, this is approximately $\kappa \approx \sqrt{n} \cdot e^{2t}$. Thus while for positive exponent the computation is forward stable (i.e., the result will be accurate to within $n^{\frac{3}{2}} \cdot \epsilon$), for negative exponent, the forward error is $n^{\frac{3}{2}} \cdot \epsilon \cdot \mathtt{exp}(2\mathtt{t})$, which for $t > 1$ will lose $\mathtt{exp}(2\mathtt{t})$ digits of precision and thus may become unstable. For example, with $n = 100$, $\mathtt{exp}(-20) \approx \mathtt{2.0} \cdot \mathtt{10}^{-9}$, the forward error is of the same order of magnitude, namely, $-2.1 \cdot 10^{-9}$.

Conclusion. Taken as an algorithm for computing e^{-z}, $z > 0$, this series expansion is an unstable algorithm. On the other hand, computing e^z and inverting $\frac{1}{e^z}$ gives an algorithm for computing e^{-z}, which is both forward and backward stable.

Remark 3.3.1. *When is a numerical result accurate?* The answer to this question depends on two issues: the problem has to have a good condition number (i.e., not be too sensitive to perturbations in the data), and the algorithm used must be backward stable; this means that the errors due to floating point arithmetic (round-off errors) must be attributable to execution of the algorithm chosen on data that are close to the initial data. Thus,

Backward stability by itself does not imply accurate answers.

We conclude this section with the forward error for a collection of important expressions. For this purpose we need to introduce notation. Given the matrix $\mathbf{A} = (a_{ij})$, its *absolute value* is equal to the matrix of absolute values $|\mathbf{A}| = (|a_{ij}|)$. Furthermore, given a second matrix $\mathbf{B}$, the notation $|\mathbf{A}| \leq |\mathbf{B}|$ means $|a_{ij}| \leq |b_{ij}|$ for all i, j. Finally, note that $|\mathbf{AB}| \leq |\mathbf{A}||\mathbf{B}|$.

The following are forward errors for various vector-matrix operations:

- $fl(\mathbf{Ax}) = \mathbf{Ax} + \mathbf{z}$, where $|\mathbf{z}| \leq \mu|\mathbf{A}||\mathbf{x}|$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mu \approx n\epsilon$;
- $fl(\mathbf{AB}) = \mathbf{AB} + \mathbf{Z}$, where $|\mathbf{Z}| \leq \mu|\mathbf{A}||\mathbf{B}|$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times k}$, $\mu \approx n\epsilon$;
- $fl(\|\mathbf{x}\|_2) = \|\mathbf{x}\|_2 \cdot (1 + \delta)$, where $|1 + \delta| \leq \sqrt{1 + \mu}(1 + \epsilon)$, $\mathbf{x} \in \mathbb{R}^n$, $\mu \approx n\epsilon$.

3.3.3 Two consequences

We will now briefly discuss two consequences of the above considerations: the stability of the SVD and the ill-posedness of the rank of a matrix. Both are at the foundations of subsequent developments.

Stability of the SVD

An important property of the SVD is that it can be computed in a backward stable way. That is, in MATLAB notation, let $[\mathbf{U}, \Sigma, \mathbf{V}] = svd(\mathbf{A})$ be the exact decomposition, and let $[\hat{\mathbf{U}}, \hat{\Sigma}, \hat{\mathbf{V}}] = fl(svd(\mathbf{A}))$ be the one obtained by taking into account floating point errors. It can be shown (see, e.g., [144]) that the following holds:

$$\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\Sigma}\hat{\mathbf{V}}^* = \mathbf{A} + \mathbf{E}, \quad \text{where } \|\mathbf{E}\| = O(\epsilon)\|\mathbf{A}\|. \tag{3.24}$$

As before, $O(\epsilon)$ is a polynomial of modest degree in the size of the problem, times machine precision ϵ. In other words, (3.24) says that the SVD can be computed using a *backward stable* algorithm.

Furthermore, it follows from the Fan inequalities (3.27) that given any $\mathbf{E}$,

$$|\sigma_i(\mathbf{A}+\mathbf{E}) - \sigma_i(\mathbf{A})| = |\hat{\sigma}_i - \sigma_i| \le \sigma_1(\mathbf{E}) = \|\mathbf{E}\|_2 \qquad \forall\, i.$$

Thus $|\hat{\sigma}_i - \sigma_i| \le \|\mathbf{E}\| = O(\epsilon)\|\mathbf{A}\|$, which implies that the computation of the singular values can be accomplished in a *forward stable* manner as well. The above discussion is summarized in the following result.

Lemma 3.13. **The SVD of a matrix is *well-conditioned* with respect to perturbations of its entries.**

This property does not hold for the EVD. There exist matrices in which a small change in the parameters causes a large change in the eigenvalues. Consider, for instance, a 10×10 matrix $\mathbf{A}$ in companion form with all 10 eigenvalues equal to 1. A small change of the order of 10^{-3} in one of the entries of $\mathbf{A}$ causes the eigenvalues to change drastically. The reason is that the condition number of the eigenvalues of this matrix is of the order of 10^{13}. Furthermore, the condition number of $\mathbf{A}$ (which is the distance to singularity) is of the order of 10^5. Such phenomena motivate the use of *pseudospectra* (see Chapter 10.2).

Ill-posedness of rank computation

The *rank* of a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $n \le m$, can be considered as a map ρ between the space of all $n \times m$ matrices and the set of natural numbers $\mathbf{I} = \{0, 1, \ldots, n\}$,

$$\rho : \mathbb{R}^{n\times m} \longrightarrow \mathbf{I}.$$

We will show that for rank deficient matrices, this map is not differentiable. The Fréchet derivative of ρ denoted by $\mathbf{D}_\rho$ is defined by means of the equation

$$|\rho(\mathbf{A}+\mathbf{H}) - \rho(\mathbf{A}) - \mathbf{D}_\rho(\mathbf{H})| = O(\|\mathbf{H}\|),$$

where $\mathbf{D}_\rho(\mathbf{H})$ is linear in $\mathbf{H}$. We distinguish two cases.

If $\mathbf{A}$ has full rank, $\rho(\mathbf{A}) = n$ for $\mathbf{H}$ of sufficiently small norm (less than the smallest singular value of $\mathbf{A}$) $\rho(\mathbf{A}+\mathbf{H}) = \rho(\mathbf{A})$. Thus $\mathbf{D}_\rho$ can be chosen as the zero function.

If $\rho(\mathbf{A}) = k < n$ for almost all $\mathbf{H}$ of infinitesimally small norm, the rank of the perturbed matrix becomes full $\rho(\mathbf{A}+\mathbf{H}) = n$. Therefore, there exists no Fréchet derivative $\mathbf{D}_\rho$ (the defining equation cannot be satisfied). Therefore, we have the following.

Lemma 3.14. The determination of the rank of a matrix is an ill-posed problem.

The numerical rank

According to Corollary 3.4 on page 35, the *rank* of a matrix is equal to the number of nonzero singular values. To define the numerical rank, we need to define a *tolerance* $\delta > 0$ (e.g., $\delta = \epsilon\|\mathbf{A}\|_2 = \epsilon\sigma_1$, where ϵ is machine precision). The *numerical rank* of a matrix is now defined as the number of singular values that are *bigger* than δ:

$$\text{rank}_\delta \mathbf{A} = \{k : \sigma_k(\mathbf{A}) > \delta,\ \sigma_{k+1}(\mathbf{A}) \leq \delta\}.$$

Although this is still an ill-posed problem, the *tolerance* is *user specified*. Therefore, it can be adapted to the problem at hand so that the determination of the numerical rank becomes well-posed. For instance, if there is a sufficient gap between, say, σ_k and σ_{k+1}, the tolerance may be set as $\delta = \frac{\sigma_k+\sigma_{k+1}}{2}$, in which case the numerical rank will be k, and this remains constant for sufficiently small perturbations.

Remark 3.3.2. A basic problem in numerical analysis is trying to decide when a small floating point number should be considered as zero—or trying to decide when a cluster of eigenvalues should be regarded as one multiple eigenvalue or as nearby but distinct eigenvalues. In terms of backward error analysis, the computed singular values of $\mathbf{A}$ are the exact singular values of the perturbed matrix $\mathbf{A}+\mathbf{E}$.

3.3.4 Distance to singularity

The remedy to the ill-posedness of the rank determination problem is to consider instead the *distance to singularity*. For a given $\mathbf{A} \in \mathbb{R}^{n\times m}$, $n \leq m$, the distance to singularity in a given norm is defined as the smallest norm of a matrix $\mathbf{H} \in \mathbb{R}^{n\times m}$ such that the rank of the perturbed matrix $\mathbf{A}-\mathbf{H}$ is less than n. In a similar fashion, the *distance to matrices of rank* k can be defined as the smallest norm of $\mathbf{H}$ such that $\mathbf{A}-\mathbf{H}$ has rank at most k.

These distance problems in the 2-norm can be solved explicitly by means of the SVD and are therefore well-conditioned. From the dyadic decomposition (3.11) of $\mathbf{A}$, the perturbation $\mathbf{H} = \sum_{i=k+1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^*$ is such that $\mathbf{A}-\mathbf{H}$ has k nonzero singular values and hence rank k. The minimality follows from Lemma 3.7. We thus have the following result.

Lemma 3.15. *Let* $\mathbf{A} \in \mathbb{R}^{n\times m}$, $n \leq m$, *have singular values* $\sigma_i(\mathbf{A})$, $i = 1, \ldots, n$. *The distance of* $\mathbf{A}$ *to the set of rank* k *matrices in the* 2*-norm is* $\sigma_{k+1}(\mathbf{A})$, *the* $(k+1)$*st singular value of* $\mathbf{A}$. *Therefore the distance to singularity is* $\sigma_n(\mathbf{A})$, *the smallest singular value of* $\mathbf{A}$.

Often, the *relative distance to singularity* is of importance. It is the distance to singularity scaled by the norm of **A**. Thus this distance in the 2-norm is the quotient of the smallest to the largest singular values which, according to (3.19), is the inverse of the 2-condition number

$$\text{relative distance to singularity} = \frac{\sigma_n(\mathbf{A})}{\sigma_1(\mathbf{A})} = \frac{1}{\kappa(\mathbf{A})}.$$

We can say in general that the *reciprocal of the relative distance to ill-posed problems is equal to the* 2*-condition number*; see Demmel [94] for an analysis of this issue.

Example 3.16. The SVD of $\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ is $\mathbf{U\Sigma V^*}$, where $\sigma_1 = \frac{\sqrt{5}+1}{2} = 1.618$, $\sigma_2 = \frac{\sqrt{5}-1}{2} = .618$, while

$$\mathbf{U} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \ \mathbf{V} = \begin{pmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{pmatrix}, \text{ and } \tan\theta = \sigma_2 \Rightarrow \theta = 31.72^\circ.$$

It follows that

$$\mathbf{H}_2 = \sigma_2\mathbf{u}_2\mathbf{v}_2^* = \frac{\sigma_2}{1+\sigma_2^2}\begin{pmatrix} \sigma_2 & -\sigma_2^2 \\ -1 & \sigma_2 \end{pmatrix} = \begin{pmatrix} .276 & -.171 \\ -.447 & .276 \end{pmatrix}$$

is a matrix of least 2-norm (i.e., σ_2) such that the perturbed $\mathbf{A} - \mathbf{H}$ is singular. For an account of other matrix distances to singularity, see Rump [282] and references therein. For matrix nearness problems, see Higham [169]. This issue can be understood in terms of *pseudospectra*.

Example 3.17. Consider the square matrix of size n, with positive ones on the diagonal and negative ones in all entries above the diagonal and zeros elsewhere. The absolute and the relative distances to singularity decrease with increasing n:

n	σ_n	σ_{n-1}	$\frac{\sigma_n}{\sigma_1}$
$n = 5$	$9.29 \cdot 10^{-2}$	1.5094	$3.40 \cdot 10^{-2}$
$n = 10$	$2.92 \cdot 10^{-3}$	1.5021	$5.12 \cdot 10^{-4}$
$n = 15$	$9.15 \cdot 10^{-5}$	1.5009	$1.05 \cdot 10^{-5}$
$n = 20$	$2.86 \cdot 10^{-6}$	1.5005	$2.41 \cdot 10^{-7}$
$n = 50$	$2.65 \cdot 10^{-15}$	1.5001	$8.60 \cdot 10^{-17}$

Thus for $n = 50$ the matrix is singular, for machine precision tolerance. We also notice that the distance of the above matrix to matrices of rank $n-2$ is practically independent of n. Such phenomena (small σ_n but bounded σ_{n-1}) are typical of Toeplitz matrices. For details, see [70].

3.3.5 LAPACK software

LAPACK [7] is a state-of-the-art software package for the numerical solution of problems in dense and banded linear algebra. It allows users to assess the accuracy of their solutions.

It provides, namely, error bounds for most quantities computed by LAPACK (see [7, Chapter 4]. Here is a list of its salient features:

- solution of systems of linear equations,
- solution of linear least squares problems.
- solution of eigenvalue and singular value problems, including generalized problems,
- matrix factorizations,
- condition and error estimates,
- BLAS (basic linear algebra subprograms) as a portability layer,
- linear algebra package for high-performance computers, and
- dense and banded linear algebra for shared memory.

Starting with Version 6, MATLAB has incorporated LAPACK. Therefore, calls to dense linear algebra algorithms (e.g., LU, QR, SVD, EVD) use LAPACK.

3.4 General rank additive matrix decompositions*

In this section we investigate matrix decompositions

$$\mathbf{Q} = \mathbf{R} + \mathbf{S}, \quad \text{where } \operatorname{rank}\mathbf{Q} = \operatorname{rank}\mathbf{R} + \operatorname{rank}\mathbf{S},$$

that satisfy the property that the rank of the sum is equal to the sum of the ranks. From our earlier discussion it follows that the SVD provides such a decomposition. Below, a general characterization of such decompositions is given, following the developments in [88]; see also [185].

Lemma 3.18. *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times k}$, $\mathbf{C} \in \mathbb{R}^{k\times n}$, $\mathbf{D} \in \mathbb{R}^{k\times k}$ *with* $\operatorname{rank}\mathbf{B} = \operatorname{rank}\mathbf{C} = \operatorname{rank}\mathbf{D} = k$. *Then*

$$\operatorname{rank}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) = \operatorname{rank}\mathbf{A} - \operatorname{rank}(\mathbf{B}\mathbf{D}^{-1}\mathbf{C}) \tag{3.25}$$

if and only if there exist matrices $\mathbf{X} \in \mathbb{R}^{n\times k}$ *and* $\mathbf{Y} \in \mathbb{R}^{k\times n}$ *such that*

$$\mathbf{B} = \mathbf{A}\mathbf{X}, \quad \mathbf{C} = \mathbf{Y}\mathbf{A}, \quad \textit{and} \quad \mathbf{D} = \mathbf{Y}\mathbf{A}\mathbf{X}. \tag{3.26}$$

Proof. First recall the identity

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_n & \mathbf{B}\mathbf{D}^{-1} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix} \begin{pmatrix} \mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{I}_k \end{pmatrix}.$$

Therefore, $\operatorname{rank}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \operatorname{rank}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) + \operatorname{rank}\mathbf{D}$; $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ is called the *Schur complement* of the block matrix in question. If conditions (3.26) hold, then

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{A}\mathbf{X} \\ \mathbf{Y}\mathbf{A} & \mathbf{Y}\mathbf{A}\mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{Y} \end{pmatrix} \mathbf{A} \begin{pmatrix} \mathbf{I} & \mathbf{X} \end{pmatrix}.$$

Therefore, the rank of this matrix is equal to the rank of $\mathbf{A}$; this implies rank $\mathbf{A}$ = rank $(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$ + rank $\mathbf{D}$, and since rank $\mathbf{D}$ = rank $(\mathbf{B}\mathbf{D}^{-1}\mathbf{C})$, this proves the sufficiency of (3.26). Conversely, let (3.25) be satisfied. From the above analysis it follows that

$$\text{rank}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \text{rank}\,\mathbf{A}.$$

Then the fact that the rank of the first block row and the first block column of this matrix both equal the rank of $\mathbf{A}$ implies the existence of $\mathbf{X}$, $\mathbf{Y}$, such that the first two conditions in (3.26) are satisfied; the third condition follows then from the fact that the whole matrix should have rank equal to that of $\mathbf{A}$. □

The following result for the rank-one case was discovered by Wedderburn in the early 1930s [353].

Corollary 3.19. Rank-one reduction. *The rank of the difference* $\mathbf{A} - \frac{1}{\pi}\mathbf{v}\mathbf{w}^*$, $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, *is one less than the rank of* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *if and only if* $\mathbf{v} = \mathbf{A}\mathbf{x}$, $\mathbf{w}^* = \mathbf{y}^*\mathbf{A}$, *and* $\pi = \mathbf{y}^*\mathbf{A}\mathbf{x}$ *for some vectors* $\mathbf{x}, \mathbf{y}$ *of appropriate dimension.*

Least squares

Given is the set of data (measurements) $\mathbf{x}_k = \begin{pmatrix} \mathbf{y}_k \\ \mathbf{z}_k \end{pmatrix} \in \mathbb{R}^n$, $k = 1, 2, \ldots, N \geq n$, where $\mathbf{y}_k \in \mathbb{R}^{n_1}$, $\mathbf{z}_k \in \mathbb{R}^{n-n_1}$. These data are arranged in matrix form:

$$\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_N] = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Z} \end{pmatrix} \in \mathbb{R}^{n\times N}, \quad \text{where } \mathbf{Y} = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_N] \in \mathbb{R}^{n_1\times N},$$

$$\mathbf{Z} = [\mathbf{z}_1 \ \cdots \ \mathbf{z}_N] \in \mathbb{R}^{(n-n_1)\times N}.$$

The least squares problem consists of finding $\mathbf{A} \in \mathbb{R}^{(n-n_1)\times n_1}$ such that $\|\mathbf{A}\mathbf{Y} - \mathbf{Z}\|_F^2$ is minimized, where the subscript F denotes the Frobenius norm defined in (3.7). This latter expression can also be written in the more familiar form $\sum_{i=1}^N \|\mathbf{A}\mathbf{y}_i - \mathbf{z}_i\|_2^2$. The least squares solution leads to the following *rank additive* decomposition of the data matrix $\mathbf{X} = \hat{\mathbf{X}} + \tilde{\mathbf{X}}$, where

$$\hat{\mathbf{X}} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{Y}\mathbf{Y}^+\mathbf{Z} \end{pmatrix} \text{ and } \tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{0} \\ [\mathbf{I}_{n-n_1} - \mathbf{Y}\mathbf{Y}^+]\,\mathbf{Z} \end{pmatrix},$$

where $\mathbf{Y}^+$ denotes a generalized inverse of $\mathbf{Y}$ (see (3.17)). Notice also that the decomposition is orthogonal since $\langle\tilde{\mathbf{X}}, \hat{\mathbf{X}}\rangle = \tilde{\mathbf{X}}^*\hat{\mathbf{X}} = \mathbf{0}$. It follows from the above considerations that in least squares, the first n_1 components of each measurement vector are assumed to be error free, an assumption that may not always be justified.

3.5 Majorization and interlacing*

In this section we review some relationships between the eigenvalues, the singular values, and the diagonal entries of a matrix. These relationships are known as *majorization inequalities*. For details, see Marshall and Olkin [233], Horn and Johnson [182], and Thompson [324]. Majorization inequalities will be used in section 9.4 to express an average rate of decay of the so-called Hankel singular values.

Definition 3.20. *Given two vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, *with entries arranged in decreasing order,* $x_i \geq x_{i+1}$, $y_i \geq y_{i+1}$, *we say that* $\mathbf{y}$ majorizes $\mathbf{x}$ *additively, or multiplicatively, and write* $\mathbf{x} \prec_{\Sigma} \mathbf{y}$, $\mathbf{x} \prec_{\pi} \mathbf{y}$, *respectively, if the following inequalities hold:*

$\boxed{x \prec_{\Sigma} y}$	$\boxed{x \prec_{\pi} y}$
$\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i$	$\Pi_{i=1}^{k} x_i \leq \Pi_{i=1}^{k} y_i$
$\sum_{i=1}^{n} x_i = \sum_{i=1}^{k} y_i$	$\Pi_{i=1}^{n} x_i = \Pi_{i=1}^{n} y_i$

We say that $\mathbf{y}$ weakly majorizes $\mathbf{x}$ *and write* $\mathbf{x} \prec_{\Sigma,w} \mathbf{y}$, $\mathbf{x} \prec_{\pi,w} \mathbf{y}$, *if the last relationship above is an inequality:* $\sum_{i=1}^{n} x_i < \sum_{i=1}^{n} y_i$, $\Pi_{i=1}^{n} x_i < \Pi_{i=1}^{n} y_i$, *respectively.*

Given the above definitions, the following holds.

Proposition 3.21. *Given two real vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *with ordered entries as above, let* $\mathbf{x} = \mathbf{S}\mathbf{y}$, *for some* $\mathbf{S} \in \mathbb{R}^{n \times n}$. *Then* $\mathbf{y}$ *majorizes* $\mathbf{x}$ *additively if and only if* $\mathbf{S}$ *is a doubly stochastic matrix, i.e., it has nonnegative entries and the sum of its rows and columns are equal to one.*

Next, consider matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{B} = \mathbf{U}\mathbf{A}\mathbf{V}^*$ (all matrices are square); let $\beta = [B_{11} \cdots B_{nn}]^*$, $\alpha = [A_{11} \cdots A_{nn}]^*$ denote the vectors containing the diagonal entries of each matrix. A straightforward computation shows that

$$\beta = (\mathbf{U} \odot \mathbf{V})\,\alpha = \begin{pmatrix} u_{11}v_{11} & \cdots & u_{1n}v_{1n} \\ u_{21}v_{21} & & u_{2n}v_{2n} \\ \vdots & \ddots & \vdots \\ u_{n1}v_{n1} & \cdots & u_{nn}v_{nn} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}.$$

The matrix $\mathbf{U} \odot \mathbf{V}$ is the *Hadamard product* of $\mathbf{U}$ and $\mathbf{V}$ and is defined as the elementwise multiplication of these two matrices: $(\mathbf{U} \odot \mathbf{V})_{ij} = U_{ij} V_{ij}$.

Consider the symmetric matrix $\mathbf{A} = \mathbf{A}^* \in \mathbb{R}^{n \times n}$. Let its eigenvalue decomposition be $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^*$; finally, let α denote the vector of diagonal entries of $\mathbf{A}$, and $\lambda = (\lambda_1 \cdots \lambda_n)^*$, the vector composed of the eigenvalues of $\mathbf{A}$. We assume without loss of generality that the entries of α and λ are arranged in decreasing order. Simple algebra reveals that

$$\alpha = (\mathbf{U} \odot \mathbf{U})\,\lambda,$$

and since $\mathbf{U}$ is unitary, $\mathbf{U} \odot \mathbf{U}$ is doubly stochastic (actually, it is called *orthostochastic*), and therefore $\alpha \prec_{\Sigma} \lambda$.

Proposition 3.22. *The ordered vector of eigenvalues of a symmetric matrix majorizes the ordered vector of diagonal entries* $\alpha \prec_{\Sigma} \lambda$. *Conversely, given any set of vectors satisfying this majorization relationship, there exists a symmetric matrix with the prescribed diagonal entries and eigenvalues.*

A similar result holds for the singular values and the absolute values of the diagonal entries of an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.

Proposition 3.23. *Let $\alpha \in \mathbb{R}^n$ be such that $\alpha_i = A_{ii}$ and $|\alpha_i| \geq |\alpha_{i+1}|$. Furthermore, let $\sigma \in \mathbb{R}^n$ be such that $\sigma_i^2 = \lambda(\mathbf{AA}^*)$ is the ith singular value of* **A**. *Then σ weakly majorizes $|\alpha|$ and in addition satisfies the inequality*

$$|\alpha_1| + \cdots + |\alpha_{n-1}| - |\alpha_n| \leq \sigma_1 + \cdots + \sigma_{n-1} - \sigma_n.$$

These conditions are also sufficient. In other words, given vectors α and σ satisfying the above inequalities, there exists a matrix A with the prescribed diagonal entries and singular values.

We now turn our attention to the relationship between eigenvalues and singular values. For a symmetric (Hermitian) matrix, $\sigma_i = |\lambda_i|$, i.e., the singular values are equal to the absolute values of the eigenvalues. For arbitrary matrices, both additive and multiplicative versions of the majorization inequalities hold.

Proposition 3.24. *Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, let λ_i, σ_i denote the ith eigenvalue and singular value, respectively. Let also $\mu_i = \lambda_i\left[\frac{\mathbf{A}+\mathbf{A}^*}{2}\right]$, $i = 1, \ldots, n$. These quantities are ordered as follows: $|\lambda_i| \geq |\lambda_{i+1}|$, $|\sigma_i| \geq |\sigma_{i+1}|$, and $\mu_i \geq \mu_{i+1}$. The following majorizations hold:*

(a) $(|\lambda_1|, \ldots, |\lambda_n|) \prec_\pi (\sigma_1, \ldots, \sigma_n)$;

(b) $(|\lambda_1|, \ldots, |\lambda_n|) \prec_{\Sigma,w} (\sigma_1, \ldots, \sigma_n)$;

(c) $(|\lambda_1|^2, \ldots, |\lambda_n|^2) \prec_\Sigma (\sigma_1^2, \ldots, \sigma_n^2)$;

(d) $(\mathcal{R}e(\lambda_1), \ldots, \mathcal{R}e(\lambda_n)) \prec_\Sigma (\mu_1, \ldots, \mu_n)$;

(e) $|\mu_i| \leq \sigma_i$.

The next result is concerned with sums of matrices.

Proposition 3.25. *Consider the matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$. *Let $\sigma(\ \cdot\ )$ denote the vector of ordered singular values. Then*

$$\sigma(\mathbf{A} + \mathbf{B}) \prec_{\Sigma,w} \sigma(\mathbf{A}) + \sigma(\mathbf{B}).$$

Furthermore, in the opposite direction, we have $\sum_{i=1}^k \sigma_i(\mathbf{A}+\mathbf{B}) \geq \sum_{i=1}^k \left[\sigma_i(\mathbf{A}) + \sigma_{n-i+1}(\mathbf{B})\right]$, $k = 1, \ldots, n$. For Hermitian matrices, let $\lambda(\ \cdot\ )$ denote the vector of ordered eigenvalues. Then the symmetry $\mathbf{A} = \mathbf{A}^*$, $\mathbf{B} = \mathbf{B}^*$ implies

$$\lambda(\mathbf{A} + \mathbf{B}) \prec_\Sigma \lambda(\mathbf{A}) + \lambda(\mathbf{B}).$$

Finally, the following inequality is due to Fan:

$$\sigma_{r+s+1}(\mathbf{A} + \mathbf{B}) \leq \sigma_{r+1}(\mathbf{A}) + \sigma_{s+1}(\mathbf{B}), \tag{3.27}$$

where $r, s \geq 0$, $r + s + 1 \leq n$.

Remark 3.5.1. Inequality (3.27) can be used to derive the lower bound given by (3.15). For $r = 0$ and $s = k$, we obtain $\sigma_{k+1}(\mathbf{A} + \mathbf{B}) \leq \sigma_1(\mathbf{A}) + \sigma_{k+1}(\mathbf{B})$. Thus by relabeling the quantities involved, we get $\sigma_{k+1}(\mathbf{A}) \leq \sigma_1(\mathbf{A} - \mathbf{X}) + \sigma_{k+1}(\mathbf{X})$, where $\mathbf{X}$ is a rank-k approximation of $\mathbf{A}$. Since the rank of $\mathbf{X}$ is at most k, $\sigma_{k+1}(\mathbf{X}) = 0$, and the desired inequality (3.15), $\sigma_{k+1}(\mathbf{A}) \leq \sigma_1(\mathbf{A} - \mathbf{X})$ follows; namely, provided that the approximant has rank at most k, the norm of the error is lower-bounded by the $(k+1)$st singular value of $\mathbf{A}$.

We conclude with a discussion of the relationship between the spectral properties of matrices and some submatrices. The classical result in this area is known as the *Cauchy interlacing property*. It states that if λ_i, $i = 1, \ldots, n$, are the eigenvalues of a given symmetric matrix $\mathbf{A}$, arranged in decreasing order, and $\hat{\lambda}_i$, $i = 1, \ldots, n-1$, are the eigenvalues of the principal submatrix $\mathbf{A}_{[k]}$ of $\mathbf{A}$ obtained by deleting column k and row k from $\mathbf{A}$, then

$$\lambda_1 \geq \hat{\lambda}_1 \geq \lambda_2 \geq \hat{\lambda}_2 \geq \cdots \geq \lambda_{n-1} \geq \hat{\lambda}_{n-1} \geq \lambda_n.$$

A more general result states the following.

Proposition 3.26. *Given $\mathbf{A} = \mathbf{A}^* \in \mathbb{R}^{n\times n}$ and $\mathbf{U} \in \mathbb{R}^{n\times k}$ such that $\mathbf{U}^*\mathbf{U} = \mathbf{I}_k$, let $\mathbf{B} = \mathbf{U}^*\mathbf{A}\mathbf{U} \in \mathbb{R}^{k\times k}$. Denote the ordered eigenvalues of $\mathbf{A}$, $\mathbf{B}$ by α_i, β_i. The following interlacing inequalities hold:*

$$\alpha_i \geq \beta_i \geq \alpha_{i+n-k}, \qquad i = 1, \ldots, k.$$

Conversely, given $\mathbf{A}$ and k real numbers β_i satisfying the above inequalities, there exists a matrix $\mathbf{U}$ composed of k orthonormal columns such that the eigenvalues of $\mathbf{B} = \mathbf{U}^*\mathbf{A}\mathbf{U}$ are precisely the desired β_i.

The above result has an analogue for arbitrary matrices, where the singular values take the place of the eigenvalues. However, no interlacing holds.

Proposition 3.27. *Given $\mathbf{A} \in \mathbb{R}^{n\times n}$, let $\mathbf{B}$ be obtained by deleting column k and row k from $\mathbf{A}$. The singular values α_i and β_i of $\mathbf{A}$, $\mathbf{B}$, respectively, satisfy the inequalities*

$$\alpha_i \geq \beta_i \geq \alpha_{i+2}, \ i = 1, \ldots, n-2, \ \textit{and} \ \beta_{n-1} \geq \alpha_{n-1}.$$

Combining the above results, we obtain a result concerning the singular values of the product of two matrices.

Corollary 3.28. *Given two matrices $\mathbf{A} \in \mathbb{R}^{n\times k}$ and $\mathbf{B} \in \mathbb{R}^{k\times m}$, $k \leq n$, m, there hold*

$$\sigma(\mathbf{AB}) \prec_{\Sigma,w} \sigma(\mathbf{A}) \odot \sigma(\mathbf{B}) \prec_{\Sigma,w} \frac{\lambda(\mathbf{AA}^*) + \lambda(\mathbf{BB}^*)}{2},$$

together with a multiplicative majorization:

$$\sum_{i=1}^{r} \sigma_i(\mathbf{AB}) \leq \sum_{i=1}^{r} \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) \leq \sum_{i=1}^{r} \frac{1}{2}\left[\sigma_i^2(\mathbf{A}) + \sigma_i^2(\mathbf{B})\right], \qquad r = 1, \ldots, k,$$

$$\Pi_{i=1}^{k}\sigma_i(\mathbf{AB}) \leq \Pi_{i=1}^{k}\sigma_i(\mathbf{A})\sigma_i(\mathbf{B}), \ k = 1, \ldots, n-1, \ \textit{and} \ \Pi_{i=1}^{n}\sigma_i(\mathbf{AB}) = \Pi_{i=1}^{n}\sigma_i(\mathbf{A})\sigma_i(\mathbf{B}).$$

3.6 Chapter summary

The focus of this chapter is on tools from matrix analysis, which will be used subsequently. The first section introduces norms of vectors and matrices, an important ingredient for carrying out an analysis of the quality of approximations and of the numerical accuracy of computations. The second section introduces the SVD which, as mentioned, is one of the most important factorizations in matrix analysis, if not *the* most important. Associated with the SVD is the *Schmidt–Eckart–Young–Mirsky* theorem, which provides the solution of the problem of approximating matrices (operators) by ones that are of lower rank and are optimal in the 2-induced norm. This is a prototype of the kind of approximation result one would like to obtain, with the singular values of the original matrix providing a trade-off between achievable accuracy and desired complexity. The third section's goal is to briefly discuss the issue of *accuracy* of computations by providing an analysis of the various sources of error. *Backward stability* and *forward stability* are important concepts in this regard. The dyadic decomposition (3.11) leads to rank additive decompositions of matrices. The fourth section, which can be omitted on a first reading, gives a general account of rank additive matrix decompositions; it is also shown that the popular least squares data fitting method induces a rank additive decomposition of the data matrix. The final section (which can also be omitted at first reading) lists a collection of results on *majorization* inequalities.

This page intentionally left blank

Chapter 4

Linear Dynamical Systems: Part 1

In this chapter we review some basic results concerning linear dynamical systems, which are geared toward the main topic of this book, namely, approximation of large-scale systems. General references for the material in this chapter are [280], [304], [370], [371], [76]. For an introduction to linear systems from basic principles, see the book by Polderman and Willems [270]. Here it is assumed that the external variables have been partitioned into *input variables* $\mathbf{u}$ and *output variables* $\mathbf{y}$, and we consider *convolution systems*, i.e., systems where the relation between $\mathbf{u}$ and $\mathbf{y}$ is given by a convolution sum or integral

$$\mathbf{y} = \mathbf{h} * \mathbf{u}, \tag{4.1}$$

where $\mathbf{h}$ is an appropriate *weighting pattern*. This is called the *external description*. We are also concerned with systems in which in addition to the input and output variables, the *state* $\mathbf{x}$ has been defined as well. Furthermore, the relationship between $\mathbf{x}$ and $\mathbf{u}$ is given by means of a set of first-order difference or differential equations with constant coefficients, while that of $\mathbf{y}$ with $\mathbf{x}$ and $\mathbf{u}$ is given by a set of linear algebraic equations. It is assumed that $\mathbf{x}$ lives in a finite-dimensional space:

$$\sigma\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \ \ \mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \tag{4.2}$$

where σ is the derivative operator or shift operator and $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are linear constant maps. This is called the *internal description*.

The first section is devoted to the discussion of systems governed by (4.1), while the next section investigates some structural properties of systems described by (4.2), namely, reachability and observability. Closely related is the concept of gramians for linear systems, which is central in subsequent developments. The third section discusses the equivalence of the external and internal descriptions. As it turns out, going from the latter to the former involves the elimination of $\mathbf{x}$ and is thus straightforward. The converse, however, is far from trivial as it involves the *construction of state*. It is called the *realization problem*.

4.1 External description

Let $\mathbb{U} = \{\mathbf{u} : \mathbb{Z} \to \mathbb{R}^m\}$, $\mathbb{Y} = \{\mathbf{y} : \mathbb{Z} \to \mathbb{R}^p\}$. A *discrete-time linear system* $\boldsymbol{\Sigma}$ with m input and p output channels can be viewed as an operator between the *input space* $\mathbb{U}$ and the *output space* $\mathbb{Y}$, $\mathcal{S} : \mathbb{U} \longrightarrow \mathbb{Y}$, which is linear. There exists a sequence of matrices $\mathbf{h}(i, j) \in \mathbb{R}^{p\times m}$ such that

$$\mathbf{u} \longmapsto \mathbf{y} = \mathcal{S}(\mathbf{u}), \;\; \mathbf{y}(i) = \sum_{j\in\mathbb{Z}} \mathbf{h}(i, j)\mathbf{u}(j), \qquad i \in \mathbb{Z}.$$

This relationship can be written in matrix form as follows:

$$\begin{pmatrix} \vdots \\ \mathbf{y}(-2) \\ \mathbf{y}(-1) \\ \hline \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \end{pmatrix} = \left(\begin{array}{ccc|ccc} \ddots & \vdots & \vdots & \vdots & \vdots & \cdot^{\cdot^{\cdot}} \\ \cdots & \mathbf{h}(-2,-2) & \mathbf{h}(-2,-1) & \mathbf{h}(-2,0) & \mathbf{h}(-2,1) & \cdots \\ \cdots & \mathbf{h}(-1,-2) & \mathbf{h}(-1,-1) & \mathbf{h}(-1,0) & \mathbf{h}(-1,1) & \cdots \\ \hline \cdots & \mathbf{h}(0,-2) & \mathbf{h}(0,-1) & \mathbf{h}(0,0) & \mathbf{h}(0,1) & \cdots \\ \cdots & \mathbf{h}(1,-2) & \mathbf{h}(1,-1) & \mathbf{h}(1,0) & \mathbf{h}(1,1) & \cdots \\ \cdot^{\cdot^{\cdot}} & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right) \begin{pmatrix} \vdots \\ \mathbf{u}(-2) \\ \mathbf{u}(-1) \\ \hline \mathbf{u}(0) \\ \mathbf{u}(1) \\ \vdots \end{pmatrix}.$$

The system $\boldsymbol{\Sigma}$ described by $\mathcal{S}$ is called *causal* if

$$\mathbf{h}(i, j) = \mathbf{0}, \qquad i \le j,$$

and *time-invariant* if

$$\mathbf{h}(i, j) = \mathbf{h}_{i-j} \in \mathbb{R}^{p\times m}.$$

For a time-invariant system $\boldsymbol{\Sigma}$ we can define the sequence of $p \times m$ constant matrices,

$$\mathbf{h} = (\ldots, \mathbf{h}_{-2}, \mathbf{h}_{-1}, \mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \ldots).$$

This sequence is called the *impulse response* of $\boldsymbol{\Sigma}$. In the *single-input, single-output* (SISO) case $m = p = 1$, it is the output obtained in response to a unit pulse,

$$\mathbf{u}(t) = \delta(t) = \begin{cases} 1, & t = 0, \\ 0, & t \neq 0. \end{cases}$$

In the *multi-input, multi-output* (MIMO) case, the subsequence of $\mathbf{h}$ composed of the kth column of each entry $\mathbf{h}_i$ is produced by applying the input $\mathbf{e}_k\delta(t)$, where $\mathbf{e}_k$ is the kth canonical unit vector (all entries are zero except the kth, which is 1). The operation of $\mathcal{S}$ can now be represented as a *convolution sum*:

$$\mathcal{S} : \mathbf{u} \longmapsto \mathbf{y} = \mathcal{S}(\mathbf{u}) = \mathbf{h} * \mathbf{u}, \;\; \text{where } (\mathbf{h} * \mathbf{u})(t) = \sum_{k=-\infty}^{\infty} \mathbf{h}_{t-k}\mathbf{u}(k), \qquad t \in \mathbb{Z}. \qquad (4.3)$$

The convolution sum is also known as a *Laurent operator* in the theory of Toeplitz matrices (see, e.g., [70]). Moreover, the matrix representation of $\mathcal{S}$ in this case is a (doubly infinite) block Toeplitz matrix,

$$\begin{pmatrix} \vdots \\ \mathbf{y}(-2) \\ \mathbf{y}(-1) \\ \hline \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \end{pmatrix} = \left(\begin{array}{ccc|ccc} \ddots & \vdots & \vdots & \vdots & \vdots & \iddots \\ \cdots & \mathbf{h}_0 & \mathbf{h}_{-1} & \mathbf{h}_{-2} & \mathbf{h}_{-3} & \cdots \\ \cdots & \mathbf{h}_1 & \mathbf{h}_0 & \mathbf{h}_{-1} & \mathbf{h}_{-2} & \cdots \\ \hline \cdots & \mathbf{h}_2 & \mathbf{h}_1 & \mathbf{h}_0 & \mathbf{h}_{-1} & \cdots \\ \cdots & \mathbf{h}_3 & \mathbf{h}_2 & \mathbf{h}_1 & \mathbf{h}_0 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right) \begin{pmatrix} \vdots \\ \mathbf{u}(-2) \\ \mathbf{u}(-1) \\ \hline \mathbf{u}(0) \\ \mathbf{u}(1) \\ \vdots \end{pmatrix}. \tag{4.4}$$

In what follows we will restrict our attention to causal and time-invariant linear systems. The matrix representation of $\mathcal{S}$ in this case is lower triangular and Toeplitz. In this case

$$\mathbf{\Sigma}: \quad \mathbf{y}(t) = \mathbf{h}_0\mathbf{u}(t) + \mathbf{h}_1\mathbf{u}(t-1) + \cdots + \mathbf{h}_k\mathbf{u}(t-k) + \cdots, \qquad t \in \mathbb{Z}.$$

The first term, $\mathbf{h}_0\mathbf{u}(t)$, denotes the *instantaneous* action of the system. The remaining terms denote the *delayed* or *dynamic* action of $\mathbf{\Sigma}$.

In analogy to the discrete-time case, let $\mathbb{U} = \{\mathbf{u} : \mathbb{R} \to \mathbb{R}^m\}$, $\mathbb{Y} = \{\mathbf{y} : \mathbb{R} \to \mathbb{R}^p\}$. A *continuous-time linear system* $\mathbf{\Sigma}$ with m input and p output channels can be viewed as an operator $\mathcal{S}$ mapping the input space $\mathbb{U}$ to the output space $\mathbb{Y}$, which is linear. In particular, we will be concerned with systems for which $\mathcal{S}$ can be expressed by means of an integral

$$\mathcal{S} : \mathbf{u} \longmapsto \mathbf{y}, \; \mathbf{y}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t,\tau)\mathbf{u}(\tau)\, d\tau, \qquad t \in \mathbb{R},$$

where $\mathbf{h}(t,\tau)$ is a matrix-valued function called the *kernel* or *weighting pattern* of $\mathbf{\Sigma}$. The system just defined is *causal* if

$$\mathbf{h}(t,\tau) = 0, \qquad t \leq \tau,$$

and *time-invariant* if $\mathbf{h}$ depends on the difference of the two arguments,

$$\mathbf{h}(t,\tau) = \mathbf{h}(t-\tau).$$

In this case $\mathcal{S}$ is a *convolution operator*

$$\mathcal{S} : \mathbf{u} \mapsto \mathbf{y} = \mathcal{S}(\mathbf{u}) = \mathbf{h} * \mathbf{u}, \; \text{ where } \; (\mathbf{h} * \mathbf{u})(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\tau)\mathbf{u}(\tau)\, d\tau, \qquad t \in \mathbb{R}. \tag{4.5}$$

It is assumed from now on that $\mathcal{S}$ is both causal and time-invariant, which means that the upper limit of integration can be replaced by t. In addition, as in the discrete-time case, we will distinguish between *instantaneous* and purely *dynamic* action, that is, we will express the output as a sum of two terms, the first being the instantaneous and the second the dynamic action:

$$\mathbf{y}(t) = \mathbf{h}_0\mathbf{u}(t) + \int_{-\infty}^{t} \mathbf{h}_a(t-\tau)\mathbf{u}(\tau)\, d\tau,$$

where $\mathbf{h}_0 \in \mathbb{R}^{p\times m}$ and $\mathbf{h}_a$ is a smooth kernel. In particular, this requirement implies that $\mathbf{h}$ can be expressed as

$$\mathbf{h}(t) = \mathbf{h}_0\delta(t) + \mathbf{h}_a(t), \qquad t \geq 0, \tag{4.6}$$

where δ denotes the δ-distribution. It readily follows that $\mathbf{h}$ is the response of the system to the impulse δ and is therefore termed the *impulse response* of Σ.

In what follows we will assume that $\mathbf{h}_a$ is an *analytic* function. This assumption implies that $\mathbf{h}_a$ is uniquely determined by the coefficients of its Taylor series expansion at $t = 0^+$:

$$\mathbf{h}_a(t) = \mathbf{h}_1 + \mathbf{h}_2 \frac{t}{1!} + \mathbf{h}_3 \frac{t^2}{2!} + \cdots + \mathbf{h}_k \frac{t^{k-1}}{(k-1)!} + \cdots, \qquad \mathbf{h}_k \in \mathbb{R}^{p \times m}.$$

It follows that if (4.6) is satisfied, the output $\mathbf{y}$ is at least as smooth as the input $\mathbf{u}$, and Σ is consequently called a *smooth* system. Hence, just like in the case of discrete-time systems, smooth continuous-time linear systems can be described by means of the infinite sequence of $p \times m$ matrices $\mathbf{h}_i$, $i \geq 0$. We formalize this conclusion next.

Definition 4.1. *The* external description *of a time-invariant, causal, and smooth continuous-time system and that of a time-invariant, causal, discrete-time linear system with m inputs and p outputs is given by an infinite sequence of $p \times m$ matrices,*

$$\Sigma = (\mathbf{h}_0, \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_k, \ldots), \qquad \mathbf{h}_k \in \mathbb{R}^{p \times m}. \tag{4.7}$$

The matrices $\mathbf{h}_k$ *are often referred to as the* Markov parameters *of the system* Σ.

Notice that by abuse of notation, we use Σ to denote both the system operator and the underlying sequence of Markov parameters. It should be clear from the context which of the two cases applies.

The (continuous- or discrete-time) Laplace transform of the impulse response yields the transfer function of the system

$$\mathbf{H}(\xi) = (\mathcal{L}\mathbf{h})(\xi). \tag{4.8}$$

The Laplace variable is denoted by ξ for both continuous- and discrete-time systems. It readily follows that $\mathbf{H}$ can be expanded in a formal power series in ξ:

$$\mathbf{H}(\xi) = \mathbf{h}_0 + \mathbf{h}_1\xi^{-1} + \mathbf{h}_2\xi^{-2} + \cdots + \mathbf{h}_k\xi^{-k} + \cdots.$$

This can also be regarded as a Laurent expansion of $\mathbf{H}$ around infinity. Consequently, (4.3) and (4.5) can be written as

$$\mathbf{Y}(\xi) = \mathbf{H}(\xi)\mathbf{U}(\xi).$$

Remark 4.1.1. *The behavioral framework.* In the classical framework (see, e.g., Kalman, Falb, and Arbib [192, Chapter 1]), a dynamical system is viewed as a mapping which transforms inputs $\mathbf{u}$ into outputs $\mathbf{y}$. Two basic considerations express the need for a framework at a more fundamental level. First, in many cases (think, for example, of electrical circuits), the distinction between inputs and outputs is not a priori clear; instead, it should follow as a consequence of the modeling. Second, it is desirable to be able to treat the different representations of a given system (for example, input-output and state-space representations) in a unified way.

In the behavioral setting, the basic variables considered are the external or manifest variables $\mathbf{w}$, which consist of $\mathbf{u}$ and $\mathbf{y}$, without distinguishing between them. The collection of trajectories describing the evolution of $\mathbf{w}$ over time defines a dynamical system. It turns out that this definition provides the right level of abstraction, necessary for accommodating the two considerations laid out above. This establishes the foundations of a parameter-free theory of dynamical systems, the advantages of representation-independent results—or, vice versa, the disadvantages of representation-dependent results—being well recognized. The resulting central object is the most powerful unfalsified model (MPUM) derived from the data, which, again, is a space of trajectories. Subsequently, inputs and outputs can be introduced and the corresponding input-output operator recovered. For details on the behavioral framework, see [270]; see also [272]. ■

4.2 Internal description

Alternatively, we can characterize a linear system via its *internal description*, which in addition to the input $\mathbf{u}$ and the output $\mathbf{y}$ uses the state $\mathbf{x}$. Again, for a first-principles treatment of the concept of state, see the book by Willems and Poldeman [270]. For our purposes, three linear spaces are given: the *state space* $\mathbb{X}$, the *input space* $\mathbb{U}$, and the *output space* $\mathbb{Y}$, containing functions taking values in $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^p$, respectively. The *state equations* describing a linear system are a set of first-order linear *differential* or *difference* equations, according to whether we are dealing with a continuous- or a discrete-time system:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad t \in \mathbb{R}, \quad \text{or} \tag{4.9}$$

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad t \in \mathbb{Z}. \tag{4.10}$$

In both cases, $\mathbf{x} \in \mathbb{X}$ is the *state* of the system, while $\mathbf{u} \in \mathbb{U}$ is the input function. Moreover,

$$\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^n, \quad \mathbf{A} : \mathbb{R}^n \to \mathbb{R}^n$$

are (constant) linear maps; the first is called the *input* map, while the second describes the *dynamics* or *internal evolution* of the system. Equations (4.9) and (4.10) can be written in a unified way,

$$\sigma\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \tag{4.11}$$

where σ denotes the derivative operator for continuous-time systems and the (backward) shift operator for discrete-time systems.

The *output equations*, for both discrete- and continuous-time linear systems, are composed of a set of linear algebraic equations,

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}, \tag{4.12}$$

where $\mathbf{y}$ is the output function (response) and

$$\mathbf{C} : \mathbb{R}^n \to \mathbb{R}^p, \quad \mathbf{D} : \mathbb{R}^m \to \mathbb{R}^p$$

are (constant) linear maps; $\mathbf{C}$ is called the *output* map. It describes how the system interacts with the outside world.

In what follows, the term *linear system* in the internal description is used to denote a linear, time-invariant, continuous- or discrete-time system which is finite-dimensional. Linear means that $\mathbb{U}$, $\mathbb{X}$, $\mathbb{Y}$ are linear spaces, and **A**, **B**, **C**, **D** are linear maps; finite-dimensional means that m, n, p are all finite positive integers; time-invariant means that **A**, **B**, **C**, **D** do not depend on time; their matrix representations are constant $n \times n$, $n \times m$, $p \times n$, $p \times m$ matrices. By a slight abuse of notation, the linear maps **A**, **B**, **C**, **D** as well as their matrix representations (in some appropriate basis) are denoted with the same symbols. We are now ready to give a definition.

Definition 4.2. (a) *A* linear system *in* internal *or* state space *description is a quadruple of linear maps (matrices)*

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right), \qquad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m},\ \mathbf{C} \in \mathbb{R}^{p\times n},\ \mathbf{D} \in \mathbb{R}^{p\times m}. \tag{4.13}$$

The dimension *of the system is defined as the dimension of the associated state space:*

$$\dim \Sigma = n. \tag{4.14}$$

(b) $\mathbf{\Sigma}$ *is called* stable *if the eigenvalues of* **A** *have negative real parts (for continuous-time systems) or lie inside the unit disk (for discrete-time systems).*

The concept of stability is introduced formally in the above definition. For a more detailed discussion, see section 5.8. We will also use the notation

$$\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right), \qquad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m},\ \mathbf{C} \in \mathbb{R}^{p\times n}. \tag{4.15}$$

It denotes a linear system where either $\mathbf{D} = \mathbf{0}$ or **D** is irrelevant for the argument pursued.

Example 4.3. We consider the dynamical system $\mathbf{\Sigma}$ shown in Figure 4.1. The external variables are the voltage applied at the terminals denoted by **u** and the voltage across the resistor denoted by **y**. The former is the *input* or *excitation function* and the latter the *output* or *measured variable* of $\mathbf{\Sigma}$. One choice for the internal or state variables is to pick the *current* through the inductor, denoted by $\mathbf{x}_1$, and the voltage across the capacitor, denoted by $\mathbf{x}_2$. The state equations are thus $\mathbf{u} = R\mathbf{x}_1 + L\dot{\mathbf{x}}_1 + \mathbf{x}_2$ and $C\dot{\mathbf{x}}_2 = \mathbf{x}_1$, while the output equation is $\mathbf{y} = R\mathbf{x}_1$. Consequently, in (4.9), $\mathbf{x} = (\mathbf{x}_1,\ \mathbf{x}_2)^*$,

$$\mathbf{A} = \begin{pmatrix} -\frac{R}{L} & -\frac{1}{L} \\ \frac{1}{C} & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \frac{1}{L} \\ 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} R & 0 \end{pmatrix}, \quad \mathbf{D} = 0.$$

The system has dimension $n = 2$, and assuming that R, L, C are positive, it is stable since the characteristic polynomial $\chi_{\mathbf{A}}(s) = s^2 + \frac{R}{L}s + \frac{1}{CL}$ of **A** has roots with negative real parts.

Solution of the state equations

We will now give the solution of (4.11). For this we will need the *matrix exponential*; given $\mathbf{M} \in \mathbb{R}^{n\times n}$ and $t \in \mathbb{R}$, we define the matrix exponential by means of the same series

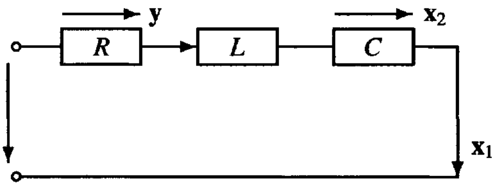


Figure 4.1. *An RLC circuit.*

representation as the scalar exponential, namely,

$$e^{\mathbf{M}t} = \mathbf{I}_n + \frac{t}{1!}\mathbf{M} + \frac{t^2}{2!}\mathbf{M}^2 + \cdots + \frac{t^k}{k!}\mathbf{M}^k + \cdots . \tag{4.16}$$

Let $\phi(\mathbf{u}; \mathbf{x}_0; t)$ denote the solution of the state equations (4.11), i.e., the state of the system at time t attained from the initial state $\mathbf{x}_0$ at time t_0, under the influence of the input $\mathbf{u}$. In particular, for the continuous-time state equations (4.9),

$$\phi(\mathbf{u}; \mathbf{x}_0; t) = e^{\mathbf{A}(t-t_0)}\mathbf{x}_0 + \int_{t_0}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau, \qquad t \geq t_0, \tag{4.17}$$

while for the discrete-time state equations (4.10),

$$\phi(\mathbf{u}; \mathbf{x}_0; t) = \mathbf{A}^{t-t_0}\mathbf{x}_0 + \sum_{j=t_0}^{t-1} \mathbf{A}^{t-1-j}\mathbf{B}\mathbf{u}(j), \qquad t \geq t_0. \tag{4.18}$$

For both discrete- and continuous-time systems, it follows that the output is given by

$$\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{u}; \mathbf{x}(t_0); t) + \mathbf{D}\mathbf{u}(t) = \mathbf{C}\phi(\mathbf{0}; \mathbf{x}(t_0); t) + \mathbf{C}\phi(\mathbf{u}; \mathbf{0}; t) + \mathbf{D}\mathbf{u}(t). \tag{4.19}$$

If we compare the above expressions for $t_0 = -\infty$ and $\mathbf{x}_0 = 0$ with (4.3) and (4.5), it follows that the *impulse response* $\mathbf{h}$ has the form below. For continuous-time systems,

$$\mathbf{h}(t) = \begin{cases} \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \delta(t)\mathbf{D}, & t \geq 0, \\ \mathbf{0}, & t < 0, \end{cases} \tag{4.20}$$

where δ denotes the δ-distribution. For discrete-time systems,

$$\mathbf{h}(t) = \begin{cases} \mathbf{C}\mathbf{A}^{t-1}\mathbf{B}, & t > 0, \\ \mathbf{D}, & t = 0, \\ \mathbf{0}, & t < 0. \end{cases} \tag{4.21}$$

Finally, by (4.8) the Laplace transform of the impulse response, which is called the transfer function of $\mathbf{\Sigma}$, is

$$\mathbf{H}_{\Sigma}(\xi) = \mathbf{D} + \mathbf{C}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}, \tag{4.22}$$

where $\xi = s$ (continuous-time Laplace transform) and $\xi = z$ (discrete-time Laplace or $\mathcal{Z}$-transform). Expanding the transfer function in a Laurent series for large ξ, i.e., in the neighborhood of infinity, we get

$$\mathbf{H}_\Sigma(\xi) = \mathbf{D} + \mathbf{CB}\xi^{-1} + \mathbf{CAB}\xi^{-2} + \cdots + \mathbf{CA}^{k-1}\mathbf{B}\xi^{-k} + \cdots,$$

and the corresponding external description given by the Markov parameters (4.7) is

$$\boldsymbol{\Sigma} = (\mathbf{D}, \mathbf{CB}, \mathbf{CAB}, \mathbf{CA}^2\mathbf{B}, \ldots, \mathbf{CA}^{k-1}\mathbf{B}, \ldots). \tag{4.23}$$

Sometimes it is advantageous to describe the system from a point of view different from the original one. In our case, since the external variables (i.e., the input and the output) are fixed, only the state variables can be transformed. In particular, if the new state is

$$\tilde{\mathbf{x}} = \mathbf{Tx}, \quad \det \mathbf{T} \neq 0,$$

the corresponding matrices describing the system will change. More precisely, given the *state transformation* **T**, (4.11) and (4.12) become

$$\sigma\tilde{\mathbf{x}} = \underbrace{\mathbf{TAT}^{-1}}_{\tilde{\mathbf{A}}}\tilde{\mathbf{x}} + \underbrace{\mathbf{TB}}_{\tilde{\mathbf{B}}}\mathbf{u}, \quad \mathbf{y} = \underbrace{\mathbf{CT}^{-1}}_{\tilde{\mathbf{C}}}\tilde{\mathbf{x}} + \underbrace{\mathbf{D}}_{\tilde{\mathbf{D}}}\mathbf{u},$$

where **D** remains unchanged. The corresponding system triples are called *equivalent*. Put differently, $\boldsymbol{\Sigma}$ and $\tilde{\boldsymbol{\Sigma}}$ are equivalent if

$$\left(\begin{array}{c|c} \mathbf{T} & \\ \hline & \mathbf{I}_p \end{array}\right)\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} \tilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \hline \tilde{\mathbf{C}} & \tilde{\mathbf{D}} \end{array}\right)\left(\begin{array}{c|c} \mathbf{T} & \\ \hline & \mathbf{I}_m \end{array}\right) \tag{4.24}$$

for some invertible matrix **T**. If $\boldsymbol{\Sigma}$ and $\tilde{\boldsymbol{\Sigma}}$ are equivalent with equivalence transformation **T**, it readily follows that

$$\begin{aligned}\mathbf{H}_\Sigma(\xi) &= \mathbf{D} + \mathbf{C}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \mathbf{D} + \mathbf{CT}^{-1}\mathbf{T}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{T}^{-1}\mathbf{TB} \\ &= \mathbf{D} + \mathbf{CT}^{-1}(\xi\mathbf{I} - \mathbf{TAT}^{-1})^{-1}\mathbf{TB} = \tilde{\mathbf{D}} + \tilde{\mathbf{C}}(\xi\mathbf{I} - \tilde{\mathbf{A}})^{-1}\tilde{\mathbf{B}} = \mathbf{H}_{\tilde{\Sigma}}(\xi).\end{aligned}$$

This immediately implies that $\mathbf{h}_k = \tilde{\mathbf{h}}_k$, $k = 1, 2, \ldots$. We have thus proved the following.

Proposition 4.4. *Equivalent triples have the same transfer function and consequently the same Markov parameters.*

Example 4.5. *Continuation of Example* 4.3. The first five Markov parameters of $\boldsymbol{\Sigma}$ are

$$0, \ \frac{R}{L}, \ -\frac{R^2}{L^2}, \ (CR^2 - L)\frac{R}{CL^3}, \ -(CR^2 - 2L)\frac{R^2}{C^2L^5}.$$

Assuming that $R = 1$, $L = 1$, $C = 1$, the matrix exponential is

$$e^{\mathbf{A}t} = e^{-\frac{t}{2}}\cos\left[\frac{t\sqrt{3}}{2}\right]\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{3}e^{-\frac{t}{2}}\sin\left[\frac{t\sqrt{3}}{2}\right]\begin{pmatrix} -1 & -2 \\ 2 & 1 \end{pmatrix}.$$

Thus, the impulse response is $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} = e^{-\frac{t}{2}}\cos\left[\frac{t\sqrt{3}}{2}\right] - \frac{1}{3}e^{-\frac{t}{2}}\sin\left[\frac{t\sqrt{3}}{2}\right]$, $t \geq 0$, while the transfer function (in terms of R, L, C) is

$$\mathbf{H}_\Sigma(s) = \frac{\frac{R}{L}s}{s^2 + \frac{R}{L}s + \frac{1}{RC}}.$$

Finally, if the state is changed to $\tilde{\mathbf{x}} = \mathbf{T}\mathbf{x}$, where $\mathbf{T} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, the new state space representation of $\boldsymbol{\Sigma}$ is

$$\tilde{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1} = \frac{-1}{2LC}\begin{pmatrix} RC-L+C & RC-L-C \\ RC+L+C & RC+L-C \end{pmatrix}, \quad \tilde{\mathbf{B}} = \mathbf{T}\mathbf{B} = \frac{1}{L}\begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\tilde{\mathbf{C}} = \mathbf{C}\mathbf{T}^{-1} = \frac{R}{2}(1 \quad 1).$$

4.2.1 The concept of reachability

In this subsection we introduce and discuss the fundamental concept of reachability of a linear system $\boldsymbol{\Sigma}$. This concept allows us to identify the extent to which the state of the system $\mathbf{x}$ can be manipulated through the input $\mathbf{u}$. The related concept of *controllability* is discussed subsequently. Both concepts involve only the state equations. Consequently, for this subsection and the next, $\mathbf{C}$ and $\mathbf{D}$ will be ignored: $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \end{array}\right)$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$. For a survey of reachability and observability (which is introduced later), see [16].

Definition 4.6. *Given is* $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \end{array}\right)$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$. *A state* $\bar{\mathbf{x}} \in \mathbb{X}$ *is* reachable *from the zero state if there exist an input function* $\bar{\mathbf{u}}(t)$, *of finite energy, and a time* $\bar{T} < \infty$, *such that*

$$\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; 0; \bar{T}).$$

The reachable subspace $\mathbb{X}^{\text{reach}} \subset \mathbb{X}$ *of* $\boldsymbol{\Sigma}$ *is the set containing all reachable states of* $\boldsymbol{\Sigma}$. *The system* $\boldsymbol{\Sigma}$ *is* (completely) reachable *if* $\mathbb{X}^{\text{reach}} = \mathbb{X}$. *Furthermore,*

$$\mathcal{R}(\mathbf{A}, \mathbf{B}) = [\mathbf{B} \quad \mathbf{A}\mathbf{B} \quad \mathbf{A}^2\mathbf{B} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{B} \quad \cdots] \tag{4.25}$$

is the reachability matrix *of* $\boldsymbol{\Sigma}$.

By the Cayley–Hamilton theorem, the rank of the reachability matrix and the span of its columns are determined (at most) by the first n terms, i.e., $\mathbf{A}^t\mathbf{B}$, $t = 0, 1, \ldots, n-1$. Thus for computational purposes the following (finite) reachability matrix is of importance:

$$\mathcal{R}_n(\mathbf{A}, \mathbf{B}) = [\mathbf{B} \quad \mathbf{A}\mathbf{B} \quad \mathbf{A}^2\mathbf{B} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{B}]. \tag{4.26}$$

The *image* of a linear map $\mathbf{L}$ is denoted by $\operatorname{im}\mathbf{L}$. The fundamental result concerning reachability is the following.

Theorem 4.7. *Given* $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \end{array}\right)$ *for both the continuous- and the discrete-time case,* $\mathbb{X}^{\text{reach}}$ *is a linear subspace of* $\mathbb{X}$*, given by the formula*

$$\mathbb{X}^{\text{reach}} = \text{im } \mathcal{R}(\mathbf{A}, \mathbf{B}). \tag{4.27}$$

As mentioned in section 10.4, expression (10.7) is referred to as the *Krylov subspace* in the numerical linear algebra community, while it is known as the reachability space in the systems community.

Corollary 4.8. **(a)** $\mathbf{A}\mathbb{X}^{\text{reach}} \subset \mathbb{X}^{\text{reach}}$. **(b)** Σ *is (completely) reachable if and only if* rank$\mathcal{R}(\mathbf{A}, \mathbf{B}) = n$. **(c)** *Reachability is basis independent.*

Proof. We will first prove Corollary 4.8. **(a)** $\mathbf{A}\,\mathbb{X}^{\text{reach}} = \mathbf{A}\,\text{im}\,\mathcal{R}(\mathbf{A}, \mathbf{B}) = \text{im}\,\mathbf{A}\mathcal{R}(\mathbf{A}, \mathbf{B}) = \text{im}\,(\mathbf{AB}\ \mathbf{A}^2\mathbf{B}\ \cdots) \subset \text{im}\,(\mathbf{B}\ \mathbf{AB}\ \cdots) = \mathbb{X}^{\text{reach}}$. **(b)** The result follows by noticing that Σ is (completely) reachable if and only if $\text{im}\,\mathcal{R}(\mathbf{A}, \mathbf{B}) = \mathbb{R}^n$. **(c)** Let $\mathbf{T}$ be a nonsingular transformation in $\mathbb{X}$, i.e., $\det \mathbf{T} \neq 0$. It follows from (4.24) that the pair $\mathbf{A}, \mathbf{B}$ is transformed into the pair $\mathbf{TAT}^{-1}, \mathbf{TB}$. It is readily checked that

$$\mathcal{R}(\mathbf{TAT}^{-1}, \mathbf{TB}) = \mathbf{T}\mathcal{R}(\mathbf{A}, \mathbf{B}),$$

which shows that the ranks of the original and the transformed reachability matrices are the same. □

Before proceeding with the proof of the theorem, some remarks are in order. In general, reachability is an *analytic* concept. The above theorem, however, shows that for linear, finite-dimensional, time-invariant systems, reachability reduces to an *algebraic* concept depending only on properties of $\mathbf{A}$ and $\mathbf{B}$ and, in particular, on the rank of the reachability matrix $\mathcal{R}(\mathbf{A}, \mathbf{B})$ but *independent* of time and the input function. It is also worthwhile to notice that formula (4.27) is valid for both continuous- and discrete-time systems. This, together with a similar result on observability (4.39), has as a consequence the fact that many tools for studying linear systems are algebraic. It should be noticed, however, that the physical significance of $\mathbf{A}$ and $\mathbf{B}$ is different for the discrete- and continuous-time cases; if we discretize, for instance, the continuous-time system $\dot{\mathbf{x}}(t) = \mathbf{A}_{\text{cont}}\mathbf{x}(t) + \mathbf{B}_{\text{cont}}\mathbf{u}(t)$ to $\mathbf{x}(t+1) = \mathbf{A}_{\text{discr}}\mathbf{x}(t) + \mathbf{B}_{\text{discr}}\mathbf{u}(t)$, then $\mathbf{A}_{\text{discr}} = e^{\mathbf{A}_{\text{cont}}}$.

A very useful concept is that of the *reachability gramian*. It is used in the proof of the theorem above and extensively in later chapters.

Definition 4.9. *The finite* reachability gramians *at time* $t < \infty$ *are defined for continuous-time systems as*

$$\mathcal{P}(t) = \int_0^t e^{\mathbf{A}\tau}\mathbf{BB}^* e^{\mathbf{A}^*\tau}d\tau, \qquad t \in \mathbb{R}_+, \tag{4.28}$$

and for discrete-time systems as

$$\mathcal{P}(t) = \mathcal{R}_t(\mathbf{A}, \mathbf{B})\mathcal{R}_t^*(\mathbf{A}, \mathbf{B}) = \sum_{k=0}^{t-1} \mathbf{A}^k\mathbf{BB}^*(\mathbf{A}^*)^k, \qquad t \in \mathbb{Z}_+. \tag{4.29}$$

A Hermitian matrix $\mathbf{X} = \mathbf{X}^*$ is called *positive semidefinite* (*positive definite*) if its eigenvalues are nonnegative (positive). The difference of two Hermitian matrices $\mathbf{X}$, $\mathbf{Y}$ satisfies $\mathbf{X} \geq \mathbf{Y}$, ($\mathbf{X} > \mathbf{Y}$) if the eigenvalues of the difference $\mathbf{X} - \mathbf{Y}$ are nonnegative (positive).

Proposition 4.10. *The reachability gramians have the following properties:* **(a)** $\mathcal{P}(t) = \mathcal{P}^*(t) \geq 0$ *and* **(b)** *their columns span the reachability subspace, i.e.,*

$$\operatorname{im} \mathcal{P}(t) = \operatorname{im} \mathcal{R}(\mathbf{A}, \mathbf{B}).$$

This relationship holds for continuous-time systems for all $t > 0$ and for discrete-time systems (at least) for $t \geq n$.

Corollary 4.11. $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ *is reachable if and only if $\mathcal{P}(t)$ is positive definite for some* $t > 0$.

Proof. Next we will prove Proposition 4.10. **(a)** The symmetry and semidefiniteness of the reachability gramian follows by definition. **(b)** To prove the second property, it is enough to show that $\mathbf{q} \perp \mathcal{P}(t)$ if and only if $\mathbf{q} \perp \mathcal{R}(\mathbf{A}, \mathbf{B})$. Since $\mathcal{P}(t)$ is symmetric and semidefinite, $\mathbf{q} \perp \mathcal{P}(t)$ if and only if $\mathbf{q}^*\mathcal{P}(t)\mathbf{q} = 0$. Moreover,

$$\mathbf{q}^*\mathcal{P}(t)\mathbf{q} = \int_0^t \|\mathbf{B}^* e^{\mathbf{A}^*(t-\tau)}\mathbf{q}\|^2 \, d\tau = 0$$

is equivalent to $\mathbf{B}^* e^{\mathbf{A}^* t}\mathbf{q} = 0$ for all $t \geq 0$. Since the exponential is an analytic function, this condition is equivalent to the function and all its derivatives being zero at $t = 0$, i.e., $\mathbf{B}^*(\mathbf{A}^*)^{i-1}\mathbf{q} = 0$, $i > 0$. This in turn is equivalent to $\mathbf{q} \perp \mathbf{A}^{i-1}\mathbf{B}$, $i > 0$, i.e., $\mathbf{q} \perp \mathcal{R}(\mathbf{A}, \mathbf{B})$. The proof for discrete-time systems is similar. □

Proof. We now turn our attention to Theorem 4.7. First we show that $\mathbb{X}^{\text{reach}}$ is a linear space, i.e.,

$$\text{if } \mathbf{x}_i = \phi(\mathbf{u}_i; \mathbf{0}; T_i) \in \mathbb{X}^{\text{reach}}, \ i = 1, 2, \ \text{ then } \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 \in \mathbb{X}^{\text{reach}}$$

for all $\alpha_1, \alpha_2 \in \mathbb{R}$. Let $T_1 \geq T_2$. Define the input function

$$\hat{\mathbf{u}}_2(t) = \begin{cases} \mathbf{0}, & t \in [0, T_1 - T_2], \\ \mathbf{u}_2(t - T_1 + T_2), & t \in [T_1 - T_2, T_1]. \end{cases}$$

It is readily checked that for both continuous- and discrete-time systems,

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 = \phi(\alpha_1\mathbf{u}_1 + \alpha_2\hat{\mathbf{u}}_2; 0; T_1).$$

Next we prove (4.27) for discrete-time systems. Consider

$$\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; \bar{T}) = \sum_{j=0}^{\bar{T}-1} \mathbf{A}^{\bar{T}-1-j}\mathbf{B}\bar{\mathbf{u}}(j).$$

Clearly, $\bar{\mathbf{x}} \in \operatorname{im}\mathcal{R}_{\bar{T}}(\mathbf{A}, \mathbf{B}) \subset \operatorname{im}\mathcal{R}(\mathbf{A}, \mathbf{B})$. Conversely, consider an element $\bar{\mathbf{x}} \in \operatorname{im}\mathcal{R}(\mathbf{A}, \mathbf{B})$. By the Cayley–Hamilton theorem, this implies $\bar{\mathbf{x}} \in \operatorname{im}\mathcal{R}_n(\mathbf{A}, \mathbf{B})$; thus, there exist elements

$\bar{\mathbf{u}}(j) \in \mathbb{R}^m$, $j = 0, 1, \ldots, n-1$, such that $\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; n-1)$. To prove (4.27) for continuous-time systems, we make use of the expansion (4.16), i.e., $e^{\mathbf{A}t} = \sum_{i>0} \frac{t^{i-1}}{(i-1)!}\mathbf{A}^{i-1}$. Let $\bar{\mathbf{x}} \in \mathbb{X}^{\text{reach}}$. Then, for some $\bar{\mathbf{u}}$, $\bar{T}$ we have

$$\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; \bar{T}) = \sum_{i>0} \mathbf{A}^{i-1}\mathbf{B} \int_0^t \frac{(t-\tau)^{i-1}}{(i-1)!}\mathbf{u}(\tau)\,d\tau,$$

which shows that $\bar{\mathbf{x}} \in \operatorname{im} \mathcal{R}(\mathbf{A}, \mathbf{B})$.

For the converse inclusion, we use the proposition given above, which asserts that for every $\bar{\mathbf{x}} \in \operatorname{im} \mathcal{R}(\mathbf{A}, \mathbf{B})$, there exists $\bar{\xi}$ such that

$$\bar{\mathbf{x}} = \mathcal{P}(\bar{T})\bar{\xi}, \tag{4.30}$$

where $\mathcal{P}$ is the reachability gramian and $\bar{T}$ is any positive real number for the continuous-time case and at least n for the discrete-time case. Choose

$$\bar{\mathbf{u}}(t) = \mathbf{B}^* e^{\mathbf{A}^*(\bar{T}-t)}\bar{\xi} \tag{4.31}$$

for the continuous-time case and

$$\bar{\mathbf{u}}(t) = \mathbf{B}^*(\mathbf{A}^*)^{(\bar{T}-t)}\bar{\xi} \tag{4.32}$$

for the discrete-time case. (Recall that * denotes transposition if the matrix or vector is real and denotes complex conjugation and transposition if the matrix or vector is complex.) It follows that $\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; 0; \bar{T}) \in \mathbb{X}^{\text{reach}}$. This concludes the proof of the theorem. □

Remark 4.2.1. *A formula for the matrix exponential.* Consider the square matrix $\mathbf{A} \in \mathbb{R}^{\nu\times\nu}$, with eigenvalues λ_i, $i = 1, \ldots, \nu$. One way to compute the matrix exponential of $\mathbf{A}$ given by (4.16) is

$$e^{\mathbf{A}t} = f_\nu(t)\mathbf{A}^{\nu-1} + f_{\nu-1}(t)\mathbf{A}^{\nu-2} + \cdots + f_2(t)\mathbf{A} + f_1(t)\mathbf{I}_\nu, \quad \text{where}$$
$$[f_1(t) \cdots f_\nu(t)] = [\phi_1(t) \cdots \phi_\nu(t)]\,\mathbf{V}(\lambda_1, \ldots, \lambda_\nu)^{-1}.$$

If the eigenvalues of $\mathbf{A}$ are distinct, the functions ϕ_i are $\phi_i(t) = e^{\lambda_i t}$, $i = 1, \ldots, \nu$, and $\mathbf{V}$ is the Vandermonde matrix,

$$\mathbf{V}(\lambda_1, \ldots, \lambda_\nu) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & & \vdots \\ \lambda_1^{\nu-1} & \lambda_2^{\nu-1} & \cdots & \lambda_n^{\nu-1} \end{pmatrix}.$$

If the eigenvalues are not distinct, the functions ϕ_i are a fundamental set of solutions of the autonomous differential equation $q(D)f = 0$, where $D = \frac{d}{dt}$ and q is the characteristic polynomial of $\mathbf{A}$. In this case the Vandermonde matrix has to be modified accordingly.

From a numerical viewpoint, the computation of the matrix exponential is a challenging proposition. A method known as *scaling and squaring* yields the best results for a wide variety of matrices $\mathbf{A}$. A survey on this topic can be found in Moler and Van Loan [241].

The definition of the inner product for vector-valued sequences and functions is given in (5.12). The *energy* or *norm* of the sequence or function $\mathbf{f}$ denoted by $\| \mathbf{f} \|$ is thus defined as its 2-norm, i.e.,

$$\| \mathbf{f} \|^2 = \langle \mathbf{f}, \mathbf{f} \rangle = \int_0^T \mathbf{f}^*(t)\mathbf{f}(t)\, dt.$$

The input function $\bar{\mathbf{u}}$ defined by (4.30) and (4.31) is a *minimal energy input* which steers the system to the desired state at a given time.

Proposition 4.12. *Consider* $\bar{\mathbf{u}}$ *defined by* (4.30) *and* (4.31), *and let* $\hat{\mathbf{u}}$ *be any input function which reaches the state* $\bar{\mathbf{x}}$ *at time* $\bar{T}$, *i.e.,* $\phi(\hat{\mathbf{u}}; \mathbf{0}; \bar{T}) = \bar{\mathbf{x}}$. *Then*

$$\| \hat{\mathbf{u}} \| \geq \| \bar{\mathbf{u}} \| . \tag{4.33}$$

Furthermore, the minimal energy required to reach the state $\bar{\mathbf{x}}$ *at time* $\bar{T}$ *is equal to the energy of the input function* $\bar{\mathbf{u}}$, *which is equal to* $\bar{\xi}^*\mathcal{P}(\bar{T})\bar{\xi}$; *if the system is reachable, this formula becomes*

$$\| \bar{\mathbf{u}} \|^2 = \bar{\mathbf{x}}^*\mathcal{P}(\bar{T})^{-1}\bar{\mathbf{x}}. \tag{4.34}$$

Proof. The proof is based on the fact that the inner product of $\bar{\mathbf{u}}$ with $\hat{\mathbf{u}} - \bar{\mathbf{u}}$ is zero. □

From the above considerations, we can quantify the time needed to arrive at a given reachable state.

Proposition 4.13. *Given is* $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$. **(a)** *For discrete-time systems, every reachable state can be reached in at most n time steps.* **(b)** *For continuous-time systems, every reachable state can be reached arbitrarily fast.*

The second part of the proposition implies that the delay in reaching a given state can be attributed to the nonlinearities present in the system.

Proof. Part **(a)** follows immediately from the Cayley–Hamilton theorem together with (4.27). In the latter part of the proof of Theorem 4.7, we showed that for any $\bar{\mathbf{x}} \in \mathbb{X}^{\text{reach}}$ we have $\bar{\mathbf{x}} = \phi(\bar{\mathbf{u}}; \mathbf{0}; \bar{T})$, where $\bar{\mathbf{u}}$ is defined by (4.30) and (4.31), while $\bar{T}$ is an arbitrary positive real number. This establishes claim **(b)**. □

Next we show that a nonreachable system can be decomposed in a canonical way into two subsystems: one whose states are all reachable and a second whose states are all unreachable.

Lemma 4.14. Reachable canonical decomposition. *Given is* $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$. *There exists a basis in* $\mathbb{X}$ *such that* $\mathbf{A}, \mathbf{B}$ *have the following matrix representations:*

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right) = \left(\begin{array}{cc|c} \mathbf{A}_r & \mathbf{A}_{r\bar{r}} & \mathbf{B}_r \\ \mathbf{0} & \mathbf{A}_{\bar{r}} & \mathbf{0} \\ \hline & & \end{array}\right), \tag{4.35}$$

where the subsystem $\Sigma_r = \left(\begin{array}{c|c} \mathbf{A}_r & \mathbf{B}_r \\ \hline & \end{array}\right)$ *is reachable.*

Proof. Let $\mathbb{X}' \subset \mathbb{X}$ be such that $\mathbb{X} = \mathbb{X}^{\text{reach}} + \mathbb{X}'$ and $\dim \mathbb{X}' = n - q$. Choose a basis $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of $\mathbb{X}$ so that $\mathbf{x}_1, \ldots, \mathbf{x}_q$ is a basis of $\mathbb{X}^{\text{reach}}$ and $\mathbf{x}_{q+1}, \ldots, \mathbf{x}_n$ is a basis for $\mathbb{X}'$. Since $\mathbb{X}^{\text{reach}}$ is $\mathbf{A}$-invariant, the matrix representation of $\mathbf{A}$ in the above basis has the form given by formula (4.35). Moreover, since $\operatorname{im} \mathbf{B} \subset \mathbb{X}^{\text{reach}}$, the matrix representation of $\mathbf{B}$ in the above basis has the form given by the same formula. Finally, to prove that $\mathbf{A}_r \in \mathbb{R}^{q \times q}$, $\mathbf{B}_r \in \mathbb{R}^{q \times m}$ is a reachable pair, it suffices to notice that

$$\operatorname{rank} \mathcal{R}(\mathbf{A}_r, \mathbf{B}_r) = \operatorname{rank} \mathcal{R}(\mathbf{A}, \mathbf{B}) = \dim \mathbb{X}^{\text{reach}} = q.$$

This concludes the proof of the lemma. □

Thus every system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ can be decomposed in a subsystem $\mathbf{\Sigma}_r = \left(\begin{array}{c|c} \mathbf{A}_r & \mathbf{B}_r \\ \hline & \end{array}\right)$, which is reachable, and in a subsystem $\mathbf{\Sigma}_{\bar{r}} = \left(\begin{array}{c|c} \mathbf{A}_{\bar{r}} & \mathbf{0} \\ \hline & \end{array}\right)$, which is completely unreachable, i.e., it cannot be influenced by outside forces. The interaction between $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}_{\bar{r}}$ is given by $\mathbf{A}_{r\bar{r}}$. Since $\mathbf{A}_{\bar{r}r} = 0$, it follows that the unreachable subsystem $\mathbf{\Sigma}_{\bar{r}}$ influences the reachable subsystem $\mathbf{\Sigma}_r$ but not vice versa. It should be noticed that although $\mathbb{X}'$ in the proof above is not unique, the form (block structure) of the reachable decomposition (4.35) is unique.

We conclude this subsection by stating various equivalent conditions for reachability.

Theorem 4.15. Reachability conditions. *The following are equivalent:*

1. *The pair* $(\mathbf{A}, \mathbf{B})$, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, *is reachable.*
2. *The rank of the reachability matrix is full:* $\operatorname{rank} \mathcal{R}(\mathbf{A}, \mathbf{B}) = n$.
3. *The reachability gramian is positive definite* $\mathcal{P}(t) > 0$ *for some* $t > 0$.
4. *No left eigenvector* $\mathbf{v}$ *of* $\mathbf{A}$ *is in the left kernel of* $\mathbf{B}$: $\mathbf{v}^*\mathbf{A} = \lambda \mathbf{v}^* \;\Rightarrow\; \mathbf{v}^*\mathbf{B} \neq 0$.
5. $\operatorname{rank}\,(\mu \mathbf{I}_n - \mathbf{A}, \quad -\mathbf{B}) = n$ *for all* $\mu \in \mathbb{C}$
6. *The polynomial matrices* $s\mathbf{I} - \mathbf{A}$ *and* $\mathbf{B}$ *are left coprime.*

The fourth and fifth conditions in the theorem are known as the Popov–Belevich–Hautus (PBH) tests for reachability. The last condition of the theorem is given for completeness; the concept of *left coprimeness* of polynomial matrices is not used further in this book; for a definition, see [123].

Proof. The equivalence of the first three statements has already been proved. The equivalence between conditions 4 and 5 is straightforward, and 6 can be considered as a different way of stating 5. We will prove the equivalence between conditions 1 and 4.

If there exists some nonzero $\mathbf{v}$ for which $\mathbf{v}^*\mathbf{A} = \lambda \mathbf{v}^*$ and $\mathbf{v}^*\mathbf{B} = \mathbf{0}$, clearly $\mathbf{v}^*\mathcal{R}(\mathbf{A}, \mathbf{B}) = \mathbf{0}$; this implies the lack of reachability of $(\mathbf{A}, \mathbf{B})$. Conversely, let $(\mathbf{A}, \mathbf{B})$ be unreachable; there exists a basis in the state space such that $\mathbf{A}$ and $\mathbf{B}$ have the form given by (4.35).

Let $\mathbf{v}_2 \neq \mathbf{0}$ be a left eigenvector of $\mathbf{A}_{\bar{r}}$. Then $\mathbf{v} = (\mathbf{0} \;\; \mathbf{v}_2^*)^*$ (where $\mathbf{0}$ is the zero vector of appropriate dimension) is a left eigenvector of $\mathbf{A}$ that is also in the left kernel of $\mathbf{B}$. This concludes the proof. □

After introducing the reachability property, we introduce a weaker concept, which is sufficient for many problems of interest. Recall the canonical decomposition (4.35) of a pair $(\mathbf{A}, \mathbf{B})$.

Definition 4.16. *The pair* $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$ *is* stabilizable *if* $\mathbf{A}_{\bar{r}}$ *is stable, i.e., all its eigenvalues either have negative real parts or are inside the unit disk, depending on whether we are dealing with continuous- or discrete-time systems.*

Remark 4.2.2. Reachability is a *generic* property. This means, intuitively, that almost every $n \times n, n \times m$ pair of matrices $\mathbf{A}, \mathbf{B}$ satisfies

$$\operatorname{rank} \mathcal{R}_n(\mathbf{A}, \mathbf{B}) = n.$$

Put differently, in the space of all $n \times n, n \times m$ pairs of matrices, the unreachable pairs form a hypersurface (of "measure" zero).

A concept that is closely related to reachability is that of *controllability*. Here, instead of driving the zero state to a desired state, a given nonzero state is steered to the zero state. More precisely, we have the next definition.

Definition 4.17. *Given* $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$, *a (nonzero) state* $\bar{\mathbf{x}} \in \mathbb{X}$ *is* controllable *to the zero state if there exist an input function* $\bar{\mathbf{u}}(t)$ *and a time* $\bar{T} < \infty$, *such that*

$$\phi(\bar{\mathbf{u}}; \bar{\mathbf{x}}; \bar{T}) = \mathbf{0}.$$

The controllable subspace $\mathbb{X}^{\text{contr}}$ *of* Σ *is the set of all controllable states. The system* Σ *is* (completely) controllable *if* $\mathbb{X}^{\text{contr}} = \mathbb{X}$.

The next theorem shows that for continuous-time systems the concepts of reachability and controllability are equivalent, while for discrete-time systems the latter is weaker. For this reason, only the notion of reachability is used in what follows.

Theorem 4.18. *Given is* $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline & \end{array}\right)$. **(a)** *For continuous-time systems,* $\mathbb{X}^{\text{contr}} = \mathbb{X}^{\text{reach}}$. **(b)** *For discrete-time systems,* $\mathbb{X}^{\text{reach}} \subset \mathbb{X}^{\text{contr}}$; *in particular,* $\mathbb{X}^{\text{contr}} = \mathbb{X}^{\text{reach}} + \ker \mathbf{A}^n$.

Proof. **(a)** By definition, $\mathbf{x} \in \mathbb{X}^{\text{contr}}$ implies $\phi(\mathbf{u}; \mathbf{x}; T) = \mathbf{0}$ for some $\mathbf{u}$; this implies

$$e^{\mathbf{A}T}\mathbf{x} = -\int_0^T e^{\mathbf{A}(T-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau \in \mathbb{X}^{\text{reach}}.$$

Thus, $\mathbf{x} \in e^{-\mathbf{A}T}\mathbb{X}^{\text{reach}} \subset \mathbb{X}^{\text{reach}}$; the latter inclusion follows because by Corollary 4.8, $\mathbb{X}^{\text{reach}}$ is $\mathbf{A}$-invariant. Thus, $\mathbb{X}^{\text{contr}} \subset \mathbb{X}^{\text{reach}}$. Conversely, let $\mathbf{x} \in \mathbb{X}^{\text{reach}}$; there exist $\mathbf{u}$ and T such that

$-\mathbf{x} = \phi(\mathbf{u}; \mathbf{0}; T)$. It follows that $-e^{\mathbf{A}T}\mathbf{x} \in \mathbb{X}^{\text{reach}}$, which in turn implies $\phi(\mathbf{u}; \mathbf{x}; T) = \mathbf{0}$. Thus, $\mathbf{x} \in \mathbb{X}^{\text{contr}}$, i.e., $\mathbb{X}^{\text{reach}} \subset \mathbb{X}^{\text{contr}}$. This completes the proof of (a).

(b) Let $\mathbf{x} \in \mathbb{X}^{\text{reach}}$. Then, $-\mathbf{x} = \phi(\mathbf{u}; \mathbf{0}; T)$. Since $\mathbb{X}^{\text{reach}}$ is $\mathbf{A}$-invariant, $\mathbf{A}^T\mathbf{x} \in \mathbb{X}^{\text{reach}}$. But $\mathbf{x}$ satisfies $\phi(\mathbf{u}, \mathbf{x}, T) = \mathbf{0}$. Thus, $\mathbf{x} \in \mathbb{X}^{\text{contr}}$. The converse does not hold true in general, as $\mathbf{A}$ may be singular. □

Remark 4.2.3. From the above results it follows that for any two states $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}^{\text{reach}}$ there exist $\mathbf{u}_{12}$, T_{12} such that $\mathbf{x}_1 = \phi(\mathbf{u}_{12}; \mathbf{x}_2; T_{12})$. To see this, note that since $\mathbf{x}_2$ is reachable it is also controllable; thus there exist $\mathbf{u}_2$, T_2 such that $\phi(\mathbf{u}_2; \mathbf{x}_2; T_2) = \mathbf{0}$. Finally, the reachability of $\mathbf{x}_1$ implies the existence of $\mathbf{u}_1$, T_1 such that $\mathbf{x}_1 = \phi(\mathbf{u}_1; \mathbf{0}; T_1)$. The function $\mathbf{u}_{12}$ is then the concatenation of $\mathbf{u}_2$ with $\mathbf{u}_1$, while $T_{12} = T_1 + T_2$. In general, if $\mathbf{x}_1, \mathbf{x}_2$ are not reachable, there is a trajectory passing through the two points if and only if

$$\mathbf{x}_2 - \mathbf{f}(\mathbf{A}, T)\mathbf{x}_1 \in \mathbb{X}^{\text{reach}} \quad \text{for some} \quad T,$$

where $\mathbf{f}(\mathbf{A}, T) = e^{\mathbf{A}T}$ for continuous-time systems and $\mathbf{f}(\mathbf{A}, T) = \mathbf{A}^T$ for discrete-time systems. This shows that if we start from a reachable state $\mathbf{x}_1 \neq 0$, the states that can be attained are also within the reachable subspace.

Distance to reachability/controllability

Following the considerations in section 3.3.3, the numerical computation of rank is an *ill-posed* problem. Therefore, the same holds for the numerical determination of reachability (controllability) of a given pair $(\mathbf{A}, \mathbf{B})$. One could consider instead the *numerical rank* of the reachability matrix $\mathcal{R}(\mathbf{A}, \mathbf{B})$ or of the reachability gramian $\mathcal{P}(T)$ or, if the system is stable, of the infinite gramian $\mathcal{P}$.

A measure of reachability that is well-posed is the *distance* of the pair to the set of unreachable/uncontrollable ones, denoted by $\delta_r(\mathbf{A}, \mathbf{B})$. Following part 5 of Theorem 4.15, this distance is defined as follows:

$$\delta_r(\mathbf{A}, \mathbf{B}) = \inf_{\mu \in \mathbb{C}} \sigma_{\min}[\mu\mathbf{I}_n - \mathbf{A}, \ \ \mathbf{B}]. \tag{4.36}$$

In other words, this is the infimum over all complex μ of the smallest singular value of $[\mu\mathbf{I}_n - \mathbf{A}, \ \ \mathbf{B}]$.

4.2.2 The state observation problem

To be able to modify the dynamical behavior of a system, very often the state $\mathbf{x}$ needs to be available. Typically, however, the state variables are inaccessible and only certain linear combinations $\mathbf{y}$ thereof, given by the output equations (4.12), are known. Thus we need to discuss the problem of reconstructing the state $\mathbf{x}(T)$ from observations $\mathbf{y}(\tau)$, where τ is in some appropriate interval. If $\tau \in [T, T + t]$, we have the *state observation problem*, while if $\tau \in [T - t, T]$, we have the *state reconstruction problem*.

The observation problem is discussed first. Without loss of generality, we assume that $T = 0$. Recall (4.17), (4.18), and (4.19). Since the input $\mathbf{u}$ is known, the latter two terms in (4.19) are also known for $t \geq 0$. Therefore, in determining $\mathbf{x}(0)$ we may assume without loss of generality that $\mathbf{u}(\cdot) = 0$. Thus, the observation problem reduces to the following: given $\mathbf{C}\phi(\mathbf{0}; \mathbf{x}(0); t)$ for $t \geq 0$, find $\mathbf{x}(0)$. Since $\mathbf{B}$ and $\mathbf{D}$ are irrelevant, for this subsection,

$$\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right), \qquad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{C} \in \mathbb{R}^{p\times n}.$$

Definition 4.19. *A state $\bar{x} \in \mathbb{X}$ is* unobservable *if* $\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{0}; \bar{\mathbf{x}}; t) = 0$ *for all $t \geq 0$, i.e., if $\bar{\mathbf{x}}$ is indistinguishable from the zero state for all $t \geq 0$. The* unobservable subspace $\mathbb{X}^{\text{unobs}}$ *of $\mathbb{X}$ is the set of all unobservable states of Σ. Σ is (completely)* observable *if* $\mathbb{X}^{\text{unobs}} = \{\mathbf{0}\}$. *The* observability matrix *of Σ is*

$$\mathcal{O}(\mathbf{C}, \mathbf{A}) = (\mathbf{C}^* \ \ \mathbf{A}^*\mathbf{C}^* \ \ (\mathbf{A}^*)^2\mathbf{C}^* \ \ \cdots \ )^*. \tag{4.37}$$

Again by the Cayley–Hamilton theorem, the kernel of $\mathcal{O}(\mathbf{C}, \mathbf{A})$ is determined by the first n terms, i.e., $\mathbf{C}\mathbf{A}^{i-1}$, $i = 1, \ldots, n$. Therefore, for computational purposes, the finite version

$$\mathcal{O}_n(\mathbf{C}, \mathbf{A}) = (\mathbf{C}^* \ \ \mathbf{A}^*\mathbf{C}^* \ \ \cdots \ \ (\mathbf{A}^*)^{n-1}\mathbf{C}^*)^* \tag{4.38}$$

of the observability matrix is used. We are now ready to state the main theorem.

Theorem 4.20. *Given* $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$ *for both $t \in \mathbb{Z}$ and $t \in \mathbb{R}$, $\mathbb{X}^{\text{unobs}}$ is a linear subspace of $\mathbb{X}$ given by*

$$\mathbb{X}^{\text{unobs}} = \ker \mathcal{O}(\mathbf{C}, \mathbf{A}) = \{\mathbf{x} \in \mathbb{X} : \ \mathbf{C}\mathbf{A}^{i-1}\mathbf{x} = \mathbf{0}, \ i > 0\}. \tag{4.39}$$

An immediate consequence of the above formula is the following corollary.

Corollary 4.21. **(a)** *The unobservable subspace $\mathbb{X}^{\text{unobs}}$ is $\mathbf{A}$-invariant.* **(b)** *Σ is observable if and only if* rank $\mathcal{O}(\mathbf{C}, \mathbf{A}) = n$. **(c)** *Observability is basis independent.*

Remark 4.2.4. Given $\mathbf{y}(t)$, $t \geq 0$, let $\mathbf{Y}_0$ denote the following $np \times 1$ vector:

$$\mathbf{Y}_0 = (\mathbf{y}^*(0) \ \ D\mathbf{y}^*(0) \ \ \cdots \ \ D^{n-1}\mathbf{y}^*(0))^* \ \text{ for continuous-time systems,}$$
$$\mathbf{Y}_0 = (\mathbf{y}^*(0) \ \ \mathbf{y}^*(1) \ \ \cdots \ \ \mathbf{y}^*(n-1))^* \qquad \text{for discrete-time systems,}$$

where $D = \frac{d}{dt}$. The observation problem reduces to the solution of the set of linear equations,

$$\mathcal{O}_n(\mathbf{C}, \mathbf{A})\mathbf{x}(0) = \mathbf{Y}_0.$$

This set of equations is solvable for all initial conditions $\mathbf{x}(0)$, i.e., it has a *unique* solution if and only if Σ is observable. Otherwise, $\mathbf{x}(0)$ can be determined only modulo $\mathbb{X}^{\text{unobs}}$, i.e., up to an arbitrary linear combination of unobservable states.

Proof. Next, we give the proof of Theorem 4.20. Let $\mathbf{x}_1, \mathbf{x}_2$ be unobservable states. Then

$$\mathbf{C}\phi(\mathbf{0}; \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2; t) = \alpha_1\mathbf{C}\phi(\mathbf{0}; \mathbf{x}_1; t) + \alpha_2\mathbf{C}\phi(\mathbf{0}; \mathbf{x}_2; t) = \alpha_1\mathbf{y}_1(t) + \alpha_2\mathbf{y}_2(t) = 0$$

for all constants α_1, α_2 and $t \geq 0$. This proves the linearity of the unobservable subspace.

For continuous-time systems, by definition, $\mathbf{x}$ is unobservable if $\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x} = \mathbf{0}$, $t \geq 0$. Since $\mathbf{C}e^{\mathbf{A}t}$ is analytic, it is completely determined by all its derivatives at $t = 0$. This implies (4.39). For discrete-time systems, formula (4.39) follows from the fact that the unobservability of $\mathbf{x}$ is equivalent to $\mathbf{y}(i) = \mathbf{C}\mathbf{A}^i\mathbf{x} = \mathbf{0}$, $i \geq 0$. □

Definition 4.22. *Let* $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$. *The finite* observability gramians *at time* $t < \infty$ *are*

$$\mathcal{Q}(t) = \int_0^t e^{\mathbf{A}^*\tau}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}\tau}\,d\tau, \qquad t \in \mathbb{R}_+, \tag{4.40}$$

$$\mathcal{Q}(t) = \mathcal{O}_t^*(\mathbf{C}, \mathbf{A})\mathcal{O}_t(\mathbf{C}, \mathbf{A}), \qquad t \in \mathbb{Z}_+. \tag{4.41}$$

It readily follows that $\ker \mathcal{Q}(t) = \ker \mathcal{O}(\mathbf{C}, \mathbf{A})$. As in the case of reachability, this relationship holds for continuous-time systems for $t > 0$ and for discrete-time systems, at least for $t \geq n$. The *energy* of the output function $\mathbf{y}$ at time T caused by the initial state $\mathbf{x}$ is denoted by $\| y \|$. In terms of the observability gramian, this energy can be expressed as

$$\| \mathbf{y} \|^2 = \mathbf{x}^*\mathcal{Q}(T)\mathbf{x}. \tag{4.42}$$

Remark 4.2.5. For completeness, we now briefly turn our attention to the reconstructibility problem. A state $\bar{\mathbf{x}} \in \mathbb{X}$ is *unreconstructible* if $\mathbf{y}(t) = \mathbf{C}\phi(\mathbf{0}; \bar{\mathbf{x}}; t) = 0$ for all $t \leq 0$, i.e., if $\bar{\mathbf{x}}$ is indistinguishable from the zero state for all $t \leq 0$. The *unreconstructible subspace* $\mathbb{X}^{\text{unrecon}}$ of $\mathbb{X}$ is the set of all unreconstructible states of Σ. Σ is (completely) *reconstructible* if $\mathbb{X}^{\text{unrec}} = \{\mathbf{0}\}$.

Given is the pair $(\mathbf{C}, \mathbf{A})$. For continuous-time systems $\mathbb{X}^{\text{unrec}} = \mathbb{X}^{\text{unobs}}$. For discrete-time systems, $\mathbb{X}^{\text{unrec}} \supset \mathbb{X}^{\text{unobs}}$, in particular, $\mathbb{X}^{\text{unobs}} = \mathbb{X}^{\text{unrec}} \cap \operatorname{im} \mathbf{A}^n$. This shows that while for continuous-time systems the concepts of observability and reconstructibility are equivalent, for discrete-time systems the latter is weaker. For this reason, only the concept of observability is used here.

4.2.3 The duality principle in linear systems

Let $\mathbf{A}^*$, $\mathbf{B}^*$, $\mathbf{C}^*$, $\mathbf{D}^*$, be the dual maps of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, respectively. The *dual* system Σ^* of $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ is formally defined as

$$\Sigma^* = \left(\begin{array}{c|c} -\mathbf{A}^* & -\mathbf{C}^* \\ \hline \mathbf{B}^* & \mathbf{D}^* \end{array}\right) \in \mathbb{R}^{(n+m)\times(n+p)},$$

i.e., the input map is given by $-\mathbf{C}^*$, the output map by $\mathbf{B}^*$, and the dynamics by $-\mathbf{A}^*$. The matrix representations of $\mathbf{A}^*$, $\mathbf{C}^*$, $\mathbf{B}^*$, $\mathbf{D}^*$ are the complex conjugate transposes of $\mathbf{A}, \mathbf{C}, \mathbf{B}, \mathbf{D}$, respectively, computed in appropriate dual bases. One may think of the dual system Σ^* as the system Σ but with the role of the inputs and outputs interchanged, or with the flow of causality reversed and time running backward. In section 5.2, it is shown that the dual system is also the adjoint system defined by (5.15), with respect to the standard inner product. The main result is the *duality principle*.

Theorem 4.23. *The orthogonal complement of the reachable subspace of* $\mathbf{\Sigma}$ *is equal to the unobservable subspace of its dual* $\mathbf{\Sigma}^*$*:* $(\mathbb{X}_{\Sigma}^{\text{reach}})^{\perp} = \mathbb{X}_{\Sigma^*}^{\text{unobs}}$. *The system* $\mathbf{\Sigma}$ *is reachable if and only if its dual* $\mathbf{\Sigma}^*$ *is observable.*

Proof. The result follows immediately from formulas (4.25) and (4.37), on recalling that for a linear map $\mathbf{M}$ there holds $(\text{im}\,\mathbf{M})^{\perp} = \ker \mathbf{M}^*$. □

In a similar way, one can prove that controllability and reconstructibility are dual concepts. Since $(\mathbf{A}, \mathbf{B})$ is reachable if and only if $(\mathbf{B}^*, \mathbf{A}^*)$ is observable, we obtain the following results. Their proof follows by duality from the corresponding results for reachability and is omitted.

Lemma 4.24. Observable canonical decomposition. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$. *There exists a basis in* $\mathbb{X}$ *such that* $\mathbf{A}$, $\mathbf{C}$ *have the following matrix representations:*

$$\left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cc|c} \mathbf{A}_{\bar{o}} & \mathbf{A}_{\bar{o}o} & \\ \mathbf{0} & \mathbf{A}_{o} & \\ \hline \mathbf{0} & \mathbf{C}_{o} & \end{array}\right),$$

where $\mathbf{\Sigma}_o = \left(\begin{array}{c|c} \mathbf{A}_o & \\ \hline \mathbf{C}_o & \end{array}\right)$ *is observable.*

The reachable and observable canonical decompositions given in Lemmas 4.14 and 4.24 can be combined to obtain the following decomposition of the triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$.

Lemma 4.25. Reachable-observable canonical decomposition. *Given is* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$. *There exists a basis in* $\mathbb{X}$ *such that* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$ *have the following matrix representations:*

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccc|c} \mathbf{A}_{r\bar{o}} & \mathbf{A}_{12} & \mathbf{A}_{13} & \mathbf{A}_{14} & \mathbf{B}_{r\bar{o}} \\ \mathbf{0} & \mathbf{A}_{ro} & \mathbf{0} & \mathbf{A}_{24} & \mathbf{B}_{ro} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{\bar{r}\bar{o}} & \mathbf{A}_{34} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A}_{\bar{r}o} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{C}_{ro} & \mathbf{0} & \mathbf{C}_{\bar{r}o} & \end{array}\right),$$

where the triple $\mathbf{\Sigma}_{ro} = \left(\begin{array}{c|c} \mathbf{A}_{ro} & \mathbf{B}_{ro} \\ \hline \mathbf{C}_{ro} & \end{array}\right)$ *is both reachable and observable.*

The dual of stabilizability is detectability. $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \\ \hline \mathbf{C} & \end{array}\right)$ is *detectable* if $\mathbf{A}_{\bar{o}}$ in the observable canonical decomposition is stable, i.e., has eigenvalues either in the left half of the complex plane or inside the unit disk, depending on whether we are dealing with a continuous- or a discrete-time system.

We conclude this subsection by stating the dual to Theorem 4.15.

Theorem 4.26. Observability conditions. *The following are equivalent:*

1. *The pair* $(\mathbf{C}, \mathbf{A})$, $\mathbf{C} \in \mathbb{R}^{p\times n}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, *is observable.*
2. *The rank of the observability matrix is full:* $\operatorname{rank} \mathcal{O}(\mathbf{C}, \mathbf{A}) = n$.
3. *The observability gramian is positive definite:* $\mathcal{Q}(t) > 0$ *for some* $t > 0$.
4. *No right eigenvector* $\mathbf{v}$ *of* $\mathbf{A}$ *is in the right kernel of* $\mathbf{C}$: $\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \;\; \Rightarrow \;\; \mathbf{C}\mathbf{v} \neq \mathbf{0}$.
5. $\operatorname{rank}\,(\mu\mathbf{I}_n - \mathbf{A}^*, \;\; \mathbf{C}^*) = n$ *for all* $\mu \in \mathbb{C}$.
6. *The polynomial matrices* $s\mathbf{I} - \mathbf{A}$ *and* $\mathbf{C}$ *are right coprime.*

Again, by section 3.3.3, the numerical determination of the observability of a given pair $(\mathbf{C}, \mathbf{A})$ is an ill-posed problem. Therefore, as in the reachability case the *distance* of the pair to the set of unobservables, denoted by $\delta_o(\mathbf{C}, \mathbf{A})$, will be used instead. This distance is defined by means of the distance to reachability of the dual pair $(\mathbf{A}^*, \mathbf{C}^*)$, which from (4.36) is

$$\delta_o(\mathbf{C}, \mathbf{A}) = \delta_r(\mathbf{A}^*, \mathbf{C}^*) = \inf_{\mu\in\mathbb{C}} \sigma_{\min}[\mu\mathbf{I}_n - \mathbf{A}^*, \;\; \mathbf{C}^*] = \inf_{\mu\in\mathbb{C}} \sigma_{\min} \begin{bmatrix} \mu\mathbf{I}_n - \mathbf{A} \\ \mathbf{C} \end{bmatrix}.$$

4.3 The infinite gramians

Consider a continuous-time linear system $\mathbf{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ that is *stable*, i.e., all eigenvalues of $\mathbf{A}$ have *negative real part*. In this case, both (4.28) and (4.40) are defined for $t = \infty$:

$$\boxed{\mathcal{P} = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^*\tau}\, d\tau,} \tag{4.43}$$

$$\boxed{\mathcal{Q} = \int_0^\infty e^{\mathbf{A}^*\tau}\mathbf{C}^*\mathbf{C} e^{\mathbf{A}\tau}\, d\tau.} \tag{4.44}$$

$\mathcal{P}$ and $\mathcal{Q}$ are the *infinite reachability* and *infinite observability gramians* associated with $\mathbf{\Sigma}_c$. These gramians satisfy the following linear matrix equations, called Lyapunov equations.

Proposition 4.27. *Given the stable, continuous-time system* $\mathbf{\Sigma}_c$ *as above, the associated infinite reachability gramian* $\mathcal{P}$ *satisfies the continuous-time Lyapunov equation*

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}, \tag{4.45}$$

while the associated infinite observability gramian satisfies

$$\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}. \tag{4.46}$$

Proof. It is readily checked that due to stability,

$$\begin{aligned}\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* &= \int_0^\infty \left[\mathbf{A}e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*\tau} + e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*\tau}\mathbf{A}^*\right] d\tau \\ &= \int_0^\infty d(e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*\tau}) = -\mathbf{B}\mathbf{B}^*.\end{aligned}$$

This proves (4.45); (4.46) is proved similarly. □

The matrices $\mathcal{P}$ and $\mathcal{Q}$ are indeed gramians in the following sense. Recall that the impulse response of a continuous-time system $\mathbf{\Sigma}_c$ is $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$, $t > 0$. Now, consider the following two maps:

input-to-state map $\xi(t) = e^{\mathbf{A}t}\mathbf{B}$ and

state-to-output map $\eta(t) = \mathbf{C}e^{\mathbf{A}t}$.

If the input to the system is the impulse $\delta(t)$, the resulting state is $\xi(t)$; moreover, if the initial condition of the system is $\mathbf{x}(0)$, in the absence of a forcing function $\mathbf{u}$, the resulting output is $\mathbf{y}(t) = \eta(t)\mathbf{x}(0)$. The *gramians* corresponding to $\xi(t)$ and $\eta(t)$ for time running from 0 to T are

$$\mathcal{P} = \int_0^T \xi(t)\xi(t)^*\, dt = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^*e^{\mathbf{A}^*t}\, dt$$

and

$$\mathcal{Q} = \int_0^\infty \eta(t)^*\eta(t)\, dt = \int_0^T e^{\mathbf{A}^*t}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}t}\, dt.$$

These are the expressions that we have encountered earlier as (finite) gramians.

Similarly, if the discrete-time system $\mathbf{\Sigma}_d = \left(\begin{array}{c|c}\mathbf{A} & \mathbf{B}\\ \hline \mathbf{C} & \mathbf{D}\end{array}\right)$ is *stable*, i.e., all eigenvalues of $\mathbf{A}$ are inside the unit disk, the gramians (4.29) as well as (4.41) are defined for $t = \infty$:

$$\mathcal{P} = \mathcal{R}\mathcal{R}^* = \sum_{k=0}^{\infty} \mathbf{A}^k\mathbf{B}\mathbf{B}^*(\mathbf{A}^*)^k, \tag{4.47}$$

$$\mathcal{Q} = \mathcal{O}^*\mathcal{O} = \sum_{k=0}^{\infty} (\mathbf{A}^*)^k\mathbf{C}^*\mathbf{C}\mathbf{A}^k. \tag{4.48}$$

Notice that $\mathcal{P}$ can be written as $\mathcal{P} = \mathbf{B}\mathbf{B}^* + \mathbf{A}\mathcal{P}\mathbf{A}^*$; moreover, $\mathcal{Q} = \mathbf{C}^*\mathbf{C} + \mathbf{A}^*\mathcal{Q}\mathbf{A}$. These are the so-called discrete-time Lyapunov or Stein equations.

Proposition 4.28. *Given the stable, discrete-time system $\mathbf{\Sigma}_d$ as above, the associated infinite reachability gramian $\mathcal{P}$ satisfies the discrete-time Lyapunov equation*

$$\mathbf{A}\mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathcal{P}, \tag{4.49}$$

while the associated infinite observability gramian $\mathcal{Q}$ satisfies

$$\mathbf{A}^*\mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathcal{Q}. \tag{4.50}$$

The infinite gramians in the frequency domain

The infinite gramians can also be expressed in the frequency domain. In particular, applying Plancherel's theorem[2] to (4.43) we obtain

$$\boxed{\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty}(i\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I}-\mathbf{A}^*)^{-1}\,d\omega;} \tag{4.51}$$

similarly, (4.44) yields

$$\boxed{\mathcal{Q} = \frac{1}{2\pi}\int_{-\infty}^{\infty}(-i\omega\mathbf{I}-\mathbf{A}^*)^{-1}\mathbf{C}^*\mathbf{C}(i\omega\mathbf{I}-\mathbf{A})^{-1}\,d\omega.} \tag{4.52}$$

In the discrete-time case the infinite gramians defined by (4.47) and (4.48) can be expressed as

$$\mathcal{P} = \frac{1}{2\pi}\int_{0}^{2\pi}(e^{i\theta}\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(e^{-i\theta}\mathbf{I}-\mathbf{A}^*)^{-1}\,d\theta, \tag{4.53}$$

$$\mathcal{Q} = \frac{1}{2\pi}\int_{0}^{2\pi}(e^{-i\theta}\mathbf{I}-\mathbf{A}^*)^{-1}\mathbf{C}^*\mathbf{C}(e^{i\theta}\mathbf{I}-\mathbf{A})^{-1}\,d\theta. \tag{4.54}$$

These expressions will be useful in model reduction methods involving *frequency weighting* (section 7.6).

4.3.1 The energy associated with reaching/observing a state

An important consideration in model reduction is the ability to classify states according to their *degree of reachability* or their *degree of observability*. Recall (4.33), (4.34), and (4.42), valid for both discrete- and continuous-time systems. From the definition of the gramians, it follows that

$$\mathcal{P}(t_2) \geq \mathcal{P}(t_1),\ \ \mathcal{Q}(t_2) \geq \mathcal{Q}(t_1), \qquad t_2 \geq t_1,$$

irrespective of whether we are dealing with discrete- or continuous-time systems. Hence from (4.34) it follows that the minimal energy for the transfer from state $\mathbf{0}$ to $\mathbf{x}_r$ is obtained as $\bar{T} \to \infty$; hence, assuming stability and (complete) reachability, the gramian is positive definite, and this minimal energy is

$$\mathbf{x}_r^*\mathcal{P}^{-1}\mathbf{x}_r. \tag{4.55}$$

Similarly, the largest observation energy produced by the state $\mathbf{x}_o$ is also obtained for an infinite observation interval and is equal to

$$\mathbf{x}_o^*\mathcal{Q}\mathbf{x}_o. \tag{4.56}$$

[2] In the theory of Fourier transform, Plancherel's theorem states that the inner product of two (matrix-valued) functions in the time domain and in the frequency domain is (up to a constant) the same. In continuous time we have $2\pi\int_{-\infty}^{\infty}\mathbf{g}^*(t)\mathbf{f}(t)\,dt = \int_{-\infty}^{\infty}\mathbf{F}^*(-i\omega)\mathbf{G}(i\omega)\,d\omega$, while in discrete-time there holds $2\pi\sum_{-\infty}^{\infty}\mathbf{g}^*(t)\mathbf{f}(t) = \int_0^{2\pi}\mathbf{F}^*(e^{-i\theta})\mathbf{G}(e^{i\theta})\,d\theta$.

We summarize these results as follows.

Lemma 4.29. *Let $\mathcal{P}$ and $\mathcal{Q}$ denote the infinite gramians of a stable linear system Σ.*

(a) *The minimal energy required to steer the state of the system from* $\mathbf{0}$ *to* $\mathbf{x}_r$ *is given by* (4.55).

(b) *The maximal energy produced by observing the output of the system whose initial state is* $\mathbf{x}_o$ *is given by* (4.56).

This lemma provides a way to determine the *degree of reachability* or the *degree of observability* of the states of Σ. The states that are the most difficult, i.e., require the most energy to reach, are (have a significant component) in the span of those eigenvectors of $\mathcal{P}$ which correspond to small eigenvalues. Furthermore, the states that are difficult to observe, i.e., produce small observation energy, are (have a significant component) in the span of those eigenvectors of $\mathcal{Q}$ which correspond to small eigenvalues.

The above conclusion is at the heart of the concept of balancing, discussed in Chapter 7. Recall the definition of equivalent systems (4.24). Under equivalence, the gramians are transformed as follows:

$$\tilde{\mathcal{P}} = \mathbf{T}\mathcal{P}\mathbf{T}^*,\quad \tilde{\mathcal{Q}} = \mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1} \quad\Rightarrow\quad \tilde{\mathcal{P}}\tilde{\mathcal{Q}} = \mathbf{T}(\mathcal{P}\mathcal{Q})\mathbf{T}^{-1}. \qquad (4.57)$$

Therefore, the product of the two gramians of equivalent systems is related by similarity transformation, and hence has the same eigenvalues. Quantities that are invariant under state-space transformation are called *input-output invariants* of the associated system Σ.

Proposition 4.30. *The eigenvalues of the product of the reachability and of the observability gramians are input-output invariants.*

Remark 4.3.1. As discussed in Lemma 5.8 and (5.24), the eigenvalues of $\mathcal{P}\mathcal{Q}$ are important invariants called *Hankel singular values* of the system. They turn out to be equal to the singular values of the *Hankel operator* introduced in section 5.1.

Remark 4.3.2. *A formula for the reachability gramian.* Given a continuous-time system described by the pair $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, the reachability gramian is defined by (4.28). If the eigenvalues of $\mathbf{A}$ are assumed to be distinct, $\mathbf{A}$ is diagonalizable. Let the EVD be

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}, \quad \text{where } \mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n],\ \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n);$$

$\mathbf{v}_i$ denotes the eigenvector corresponding to the eigenvalue λ_i. Notice that if the ith eigenvalue is complex, the corresponding eigenvector is also complex. Let $\mathbf{W} = \mathbf{V}^{-1}\mathbf{B} \in \mathbb{C}^{n\times m}$, and denote by $\mathbf{W}_i \in \mathbb{C}^{1\times m}$ the ith row of $\mathbf{W}$. With the notation introduced above, the following formula holds:

$$\mathcal{P}(T) = \mathbf{V}\mathcal{R}(T)\mathbf{V}^*, \quad \text{where } [\mathcal{R}(T)]_{ij} = \frac{-\mathbf{W}_i\mathbf{W}_j^*}{\lambda_i + \lambda_j^*}\left(1 - \exp\left[(\lambda_i + \lambda_j^*)T\right]\right) \in \mathbb{C}.$$

Furthermore, if $\lambda_i + \lambda_j^* = 0$, $[\mathcal{R}(T)]_{ij} = (\mathbf{W}_i\mathbf{W}_j^*)\,T$. If in addition $\mathbf{A}$ is stable, the infinite gramian (4.43) is given by $\mathcal{P} = \mathbf{V}\mathcal{R}\mathbf{V}^*$, where $\mathcal{R}_{ij} = \frac{-\mathbf{W}_i\mathbf{W}_j^*}{\lambda_i+\lambda_j^*}$. This formula accomplishes both the computation of the exponential and the integration implicitly, in terms of the EVD of $\mathbf{A}$.

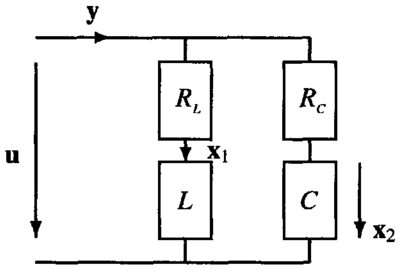


Figure 4.2. *Parallel connection of a capacitor and an inductor.*

Example 4.31. Consider the electric circuit consisting of the parallel connection of two branches, as shown in Figure 4.2. The input is the voltage $\mathbf{u}$ applied to this parallel connection, while the output is the current $\mathbf{y}$; we choose as states the current through the inductor $\mathbf{x}_1$, and the voltage across the capacitor $\mathbf{x}_2$.

The state equations are $L\dot{\mathbf{x}}_1 = -R_L\mathbf{x}_1 + \mathbf{u}$, $CR_C\dot{\mathbf{x}}_2 = -\mathbf{x}_2 + \mathbf{u}$, while the output equation is $R_C\mathbf{y} = R_C\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{u}$. Thus

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cc|c} -\tau_L & 0 & \frac{1}{R_L}\tau_L \\ 0 & -\tau_C & \tau_C \\ \hline 1 & -\frac{1}{R_C} & \frac{1}{R_C} \end{array}\right], \tag{4.58}$$

$$\text{where } \tau_L = \frac{R_L}{L},\ \tau_C = \frac{1}{R_C C},$$

are the time constants of the two branches of the circuit. Therefore, the impulse response is

$$\mathbf{h}(t) = \frac{\tau_L}{R_L}e^{-\tau_L t} - \frac{\tau_C}{R_C}e^{-\tau_C t} + \frac{1}{R_C}\delta(t), \qquad t \geq 0.$$

It readily follows that this system is reachable and observable if the two time constants are different ($\tau_L \neq \tau_C$).

Assuming that the values of these elements are $L = 1$, $R_L = 1$, $C = 1$, $R_C = \frac{1}{2}$:

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \Rightarrow e^{\mathbf{A}\bar{t}}\mathbf{B} = \begin{bmatrix} e^{-\bar{t}} \\ 2e^{-2\bar{t}} \end{bmatrix}.$$

Reachability in this case inquires about the existence of an input voltage $\mathbf{u}$ which will steer the state of the system to some desired $\bar{\mathbf{x}}$, at a given time $T > 0$. In this case, since the system is reachable (for positive values of the parameters), any state can be reached. We choose $\mathbf{x}^1 = [1\ \ 0]^*$, $\mathbf{x}^2 = [0\ \ 1]^*$. The gramian $\mathcal{P}(T)$ and the infinite gramian $\mathcal{P}$ are

$$\mathcal{P}(T) = \begin{bmatrix} -\frac{1}{2}e^{-2T} + \frac{1}{2} & -\frac{2}{3}e^{-3T} + \frac{2}{3} \\ -\frac{2}{3}e^{-3T} + \frac{2}{3} & -e^{-4T} + 1 \end{bmatrix}, \ \mathcal{P} = \lim_{T\to\infty}\mathcal{P}(T) = \begin{bmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{bmatrix}.$$

The corresponding inputs valid for $\bar{t}$ between 0 and T are

$$\mathbf{u}_T^1(\bar{t}) = -\frac{6e^{-\bar{t}}\left(3e^{-4T} - 3 - 4e^{-\bar{t}-3T} + 4e^{-\bar{t}}\right)}{e^{-6T} - 9e^{-2T} - 9e^{-4T} + 1 + 16e^{-3T}},$$

$$\Rightarrow \mathbf{u}_\infty^1(\bar{t}) = \lim_{T\to\infty} u_{1,T}(t) = 18e^{-\bar{t}} - 24e^{-2\bar{t}}.$$

$$\mathbf{u}_T^2(\bar{t}) = \frac{6e^{-\bar{t}}\left(2e^{-3T} - 2 - 3e^{-\bar{t}-2T} + 3e^{-\bar{t}}\right)}{e^{-6T} - 9e^{-2T} - 9e^{-4T} + 1 + 16e^{-3T}},$$

$$\Rightarrow \mathbf{u}_\infty^2(\bar{t}) = \lim_{T\to\infty} u_{2,T}(t) = -12e^{-\bar{t}} + 18e^{-2\bar{t}}.$$

In the above expressions, $\bar{t} = T - t$, where t is the time; in the upper plot of Figure 4.3, the time axis is $\bar{t}$, i.e., time runs backward from T to 0; in the lower plot, the time axis is t, running from 0 to T. Both plots show the minimum energy inputs required to steer the system to $\mathbf{x}_1$ for $T = 1,\ 2,\ 10$ units of time. Notice that for $T = 10$ the input function is zero for most of the interval, starting with $t = 0$; consequently, for $T \to \infty$, the activity occurs close to $T = \infty$ and the input function can thus be plotted only in the $\bar{t}$ axis. If the system is stable, i.e., $\mathcal{R}e(\lambda_i(\mathbf{A})) < 0$, the reachability gramian is defined for $T = \infty$, and it satisfies (4.45). Hence, the infinite gramian can be computed as the solution to this linear matrix equation; explicit calculation of the matrix exponentials, multiplication, and subsequent integration are not required. In MATLAB, if in addition the pair $(\mathbf{A}, \mathbf{B})$ is reachable, we have

$$\texttt{P} = \texttt{lyap}\,(\texttt{A}, \texttt{B} * \texttt{B}').$$

For the matrices defined earlier, using the `lyap` command in the format `short e`, we get

$$\mathcal{P} = \begin{bmatrix} 0.50000 & 0.66666 \\ 0.66666 & 1.00000 \end{bmatrix}.$$

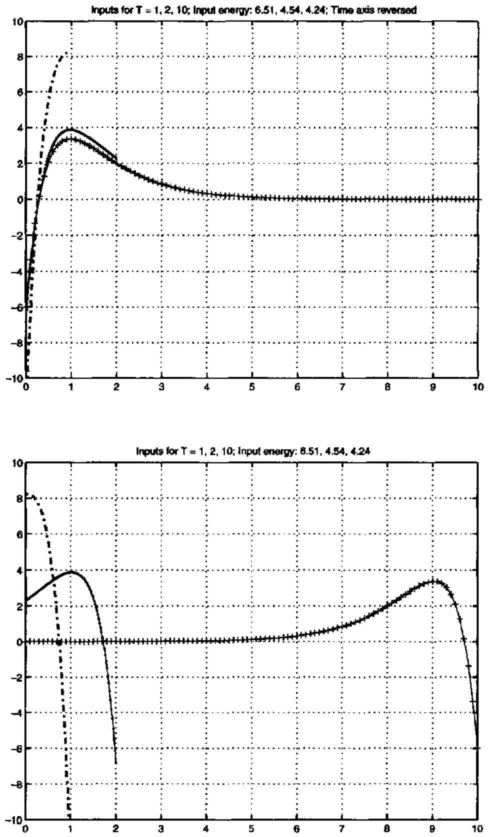
We conclude this example with the computation of the reachability gramian in the frequency domain:

$$\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I}_2 - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I}_2 - \mathbf{A}^*)^{-1}\,d\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \begin{bmatrix} \frac{1}{\omega^2+1} & \frac{1}{\omega^2+i\omega+2} \\ \frac{1}{\omega^2-i\omega+2} & \frac{4}{\omega^2+4} \end{bmatrix} d\omega$$

$$= \frac{1}{2\pi}\begin{bmatrix} 2\arctan\omega & \frac{4}{3}\arctan\frac{\omega}{2} + \frac{4}{3}\arctan\omega \\ \frac{4}{3}\arctan\frac{\omega}{2} + \frac{4}{3}\arctan\omega & 4\arctan\frac{\omega}{2} \end{bmatrix}_{\omega=\infty}$$

$$= \frac{1}{2\pi}\begin{bmatrix} \pi & \frac{4\pi}{3} \\ \frac{4\pi}{3} & 2\pi \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{2}{3} & 1 \end{bmatrix}$$


Inputs for T = 1, 2, 10; Input energy: 6.51, 4.54, 4.24; Time axis reversed

Inputs for T = 1, 2, 10; Input energy: 6.51, 4.54, 4.24

Figure 4.3. *Electric circuit example. Minimum energy inputs steering the system to* $\mathbf{x}_1 = [1 \;\; 0]^*$ *for* $T = 1, 2, 10$. *Top plot: time axis running backward; bottom plot: time axis running forward.*

Example 4.32. A second simple example is the following:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}, \; \mathbf{B} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \;\Rightarrow\; e^{\mathbf{A}t} = \begin{pmatrix} -e^{-2t} + 2e^{-t} & e^{-t} - e^{-2t} \\ -2e^{-t} + 2e^{-2t} & 2e^{-2t} - e^{-t} \end{pmatrix}.$$

This implies

$$\mathcal{P}(T) = \begin{pmatrix} -\frac{1}{2}e^{-2T} + \frac{2}{3}e^{-3T} - \frac{1}{4}e^{-4T} + \frac{1}{12} & -e^{-3T} + \frac{1}{2}e^{-2T} + \frac{1}{2}e^{-4T} \\ -e^{-3T} + \frac{1}{2}e^{-2T} + \frac{1}{2}e^{-4T} & -e^{-4T} + \frac{4}{3}e^{-3T} - \frac{1}{2}e^{-2T} + \frac{1}{6} \end{pmatrix}.$$

And finally, the infinite gramian is

$$\texttt{P} = \texttt{lyap}\,(\texttt{A}, \texttt{B} * \texttt{B}') = \begin{pmatrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{6} \end{pmatrix}.$$

In the frequency domain

$$\begin{aligned}
\mathcal{P} &= \frac{1}{2\pi}\int_{-\infty}^{\infty}(i\omega\mathbf{I}_2-\mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I}_2-\mathbf{A}^*)^{-1}\,d\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\begin{bmatrix} \frac{1}{\omega^4+5\omega^2+4} & \frac{-i\omega}{\omega^4+5\omega^2+4} \\ \frac{i\omega}{\omega^4+5\omega^2+4} & \frac{\omega^2}{\omega^4+5\omega^2+4}\end{bmatrix} d\omega \\
&= \frac{1}{2\pi}\begin{bmatrix} \frac{2}{3}\arctan\omega-\frac{1}{3}\arctan\frac{\omega}{2} & 0 \\ 0 & \frac{4}{3}\arctan\frac{\omega}{2}-\frac{2}{3}\arctan\omega\end{bmatrix}_{\omega=\infty} \\
&= \frac{1}{2\pi}\begin{bmatrix} \frac{\pi}{6} & 0 \\ 0 & \frac{\pi}{3}\end{bmatrix} = \begin{bmatrix} \frac{1}{12} & 0 \\ 0 & \frac{1}{6}\end{bmatrix}.
\end{aligned}$$

Example 4.33. We will now compute the gramian of a simple discrete-time, second-order system,

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & \frac{1}{2}\end{pmatrix},\ \mathbf{B} = \begin{pmatrix} 0 \\ 1\end{pmatrix}.$$

To compute the reachability gramian in the time domain, we make use of (4.47):

$$\mathcal{P} = \mathcal{R}\mathcal{R}^* = \sum_{k=0}^{\infty}\mathbf{A}^k\mathbf{B}\mathbf{B}^*(\mathbf{A}^*)^k = \begin{bmatrix} 1+\frac{1}{4}+\frac{1}{16}+\cdots & \frac{1}{2}+\frac{1}{8}+\frac{1}{32}+\cdots \\ \frac{1}{2}+\frac{1}{8}+\frac{1}{32}+\cdots & 1+\frac{1}{4}+\frac{1}{16}+\cdots \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2\end{bmatrix}.$$

For the frequency domain computation, we make use of formula (4.53):

$$\begin{aligned}
\mathcal{P} &= \frac{1}{2\pi}\int_0^{2\pi}\begin{bmatrix} \frac{4e^{-i\theta}}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} & \frac{4}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} \\ \frac{4}{(e^{-i\theta}-2)(2e^{-i\theta}-1)} & \frac{4e^{-i\theta}}{(e^{-i\theta}-2)(2e^{-i\theta}-1)}\end{bmatrix} d\theta \\
&= \frac{1}{2\pi}\begin{bmatrix} \frac{4i}{3}\left[\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] & \frac{2i}{3}\left[4\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] \\ \frac{2i}{3}\left[4\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right] & \frac{4i}{3}\left[\ln(e^{i\theta}-2)-\ln(2e^{i\theta}-1)\right]\end{bmatrix}_{\theta=0}^{\theta=2\pi} \\
&= \frac{2}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2\end{bmatrix}.
\end{aligned}$$

4.3.2 The cross gramian

In addition to the reachability and observability gramians, a third one, the *cross gramian*, is used. This concept is first defined for *discrete-time* systems $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$. Given the (infinite) reachability matrix $\mathcal{R}(\mathbf{A},\mathbf{B})$ (4.25) and the observability matrix $\mathcal{O}(\mathbf{C},\mathbf{A})$ (4.37), the *cross gramian* is the $n \times n$ matrix defined by $\mathcal{X} = \mathcal{R}\mathcal{O}$. Thus, summarizing, the three (infinite) gramians of Σ are

$$\mathcal{P} = \mathcal{R}\mathcal{R}^*,\ \mathcal{Q} = \mathcal{O}^*\mathcal{O},\ \mathcal{X} = \mathcal{R}\mathcal{O} \in \mathbb{R}^{n\times n}.$$

Notice that these gramians are the three finite matrices that can be formed from the reachability matrix (which has infinitely many columns) and the observability matrix (which has infinitely many rows).

These first two gramians satisfy the Stein equations (4.49) and (4.50). The cross gramian satisfies a Stein equation as well, but its form depends on the number of inputs and outputs m, p of $\mathbf{\Sigma}$. If $m = p$, a moment's reflection shows that $\mathcal{X}$ satisfies the following Sylvester equation:

$$\mathbf{A}\mathcal{X}\mathbf{A} + \mathbf{BC} = \mathcal{X}.$$

If $m = 2p$, let $\mathbf{B} = [\mathbf{B}_1 \;\; \mathbf{B}_2]$, $\mathbf{B}_i \in \mathbb{R}^{n\times p}$, $i = 1, 2$; it can be verified that

$$\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2, \quad \text{where } \mathcal{X}_1 = \mathcal{R}(\mathbf{B}_1, \mathbf{A})\mathcal{O}(\mathbf{C}, \mathbf{A}^2) \text{ and } \mathcal{X}_2 = \mathcal{R}(\mathbf{B}_2, \mathbf{A})\mathcal{O}(\mathbf{CA}, \mathbf{A}^2).$$

Combining these expressions we obtain the Stein equation satisfied by $\mathcal{X}$:

$$\mathbf{A}\mathcal{X}\mathbf{A}^2 + \mathbf{B}\begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \end{pmatrix} = \mathcal{X}.$$

For general m and p, the Stein equation involves $\mathbf{A}^r$, where r is the least common multiple of m, p.

As in the discrete-time case, if the number of inputs of the stable continuous-time system $\mathbf{\Sigma}$ is equal to the number of outputs $m = p$, the **cross gramian** $\mathcal{X}$ is defined as the solution to the Sylvester equation

$$\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{BC} = \mathbf{0}. \tag{4.59}$$

Similarly to (4.43) and (4.44) in Proposition 4.27, it can readily be shown that $\mathcal{X}$ can be expressed as

$$\mathcal{X} = \int_0^\infty e^{\mathbf{A}t}\mathbf{BC}e^{\mathbf{A}t}\,dt. \tag{4.60}$$

All three gramians are related to the eigenvalues and singular values of the *Hankel operator*, which will be introduced later. Under a state-space transformation $\mathbf{T}$, the three gramians are transformed to $\mathbf{T}\mathcal{P}\mathbf{T}^*$, $\mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1}$, $\mathbf{T}\mathcal{X}\mathbf{T}^{-1}$, respectively. Therefore, while the eigenvalues of the reachability and observability gramians are *not* input-output invariants, both the product $\mathcal{PQ}$ and $\mathcal{X}$ are transformed by similarity. Their eigenvalues are input-output invariants for the associated $\mathbf{\Sigma}$, both for discrete- and continuous-time systems. As will be shown in section 5.4, the eigenvalues and the singular values of the *Hankel operator* $\mathcal{H}$ associated with Σ are given by these eigenvalues, namely,

$$\boxed{\lambda_i(\mathcal{H}) = \lambda_i(\mathcal{X}), \quad \sigma_i(\mathcal{H}) = \sqrt{\lambda_i(\mathcal{PQ})}\,.}$$

The cross gramian for SISO systems was introduced in [113].

Example 4.34. Consider the circuit shown in Figure 4.2. The system matrices are given by (4.58). Thus the reachability, observability, and cross gramians are:

$\mathcal{P} =$	$\mathcal{Q} =$	$\mathcal{X} =$
$\tau_L\tau_C\begin{bmatrix} \frac{1}{R_L^2}\frac{1}{2\tau_C} & \frac{1}{R_L}\frac{1}{\tau_L+\tau_C} \\ \frac{1}{R_L}\frac{1}{\tau_L+\tau_C} & \frac{1}{2\tau_L} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2\tau_L} & -\frac{1}{R_C}\frac{1}{\tau_L+\tau_C} \\ -\frac{1}{R_C}\frac{1}{\tau_L+\tau_C} & \frac{1}{R_C^2}\frac{1}{2\tau_C} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2R_L} & -\frac{1}{R_LR_C}\frac{\tau_L}{\tau_L+\tau_C} \\ \frac{\tau_C}{\tau_L+\tau_C} & -\frac{1}{2R_C} \end{bmatrix}$.

Notice that $\mathcal{P}$ and $\mathcal{Q}$ become semidefinite if the two time constants are equal, $\tau_L = \tau_R$; if in addition $R_L = R_C$, the product $\mathcal{P}\mathcal{Q} = \mathbf{0}$, which is reflected in the fact that $\mathcal{X}$ in this case has two zero eigenvalues.

4.3.3 A transformation between continuous- and discrete-time systems

Often it is advantageous to transform a given problem in a way that its solution becomes easier, either theoretically or computationally. A transformation that is of interest in the present context is the ***bilinear transformation***. We will mention here some cases in which this transformation is important.

The theory of optimal approximation in the Hankel-norm discussed in Chapter 8 is easier to formulate for discrete-time systems, while it is easier to solve for continuous-time systems. Thus given a discrete-time system, the bilinear transformation is used to obtain the solution in continuous time and then transform back. (See Example 8.9.) Second, as stated in the next proposition, the gramians remain invariant under the bilinear transformation. In section 12.2, this fact is used to iteratively solve a continuous-time Lyapunov equation in discrete time, that is, by solving the corresponding Stein equation.

The bilinear transformation is defined by $z = \frac{1+s}{1-s}$ and maps the open left half of the complex plane onto the inside of the unit disc and the imaginary axis onto the unit circle. In particular, the transfer function $\mathbf{H}_c(s)$ of a continuous-time system is obtained from that of the discrete-time transfer function $\mathbf{H}_d(z)$ as follows:

$$\mathbf{H}_c(s) = \mathbf{H}_d\left(\frac{1+s}{1-s}\right).$$

Consequently, the matrices

$$\boldsymbol{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right), \quad \boldsymbol{\Sigma}_d = \left(\begin{array}{c|c} \mathbf{F} & \mathbf{G} \\ \hline \mathbf{H} & \mathbf{J} \end{array}\right)$$

of these two systems are related as given in the following table:

Continuous time		**Discrete time**
$\mathbf{A},\ \mathbf{B},\ \mathbf{C},\ \mathbf{D}$	$z = \frac{1+s}{1-s}$	$\left\{\begin{array}{l} \mathbf{F} = (\mathbf{I}+\mathbf{A})(\mathbf{I}-\mathbf{A})^{-1} \\ \mathbf{G} = \sqrt{2}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} \\ \mathbf{H} = \sqrt{2}\mathbf{C}(\mathbf{I}-\mathbf{A})^{-1} \\ \mathbf{J} = \mathbf{D}+\mathbf{C}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} \end{array}\right.$
$\left.\begin{array}{l} \mathbf{A} = (\mathbf{F}+\mathbf{I})^{-1}(\mathbf{F}-\mathbf{I}) \\ \mathbf{B} = \sqrt{2}(\mathbf{F}+\mathbf{I})^{-1}\mathbf{G} \\ \mathbf{C} = \sqrt{2}\mathbf{H}(\mathbf{F}+\mathbf{I})^{-1} \\ \mathbf{D} = \mathbf{J}-\mathbf{H}(\mathbf{F}+\mathbf{I})^{-1}\mathbf{G} \end{array}\right\}$	$s = \frac{z-1}{z+1}$	$\mathbf{F},\ \mathbf{G},\ \mathbf{H},\ \mathbf{J}$

Proposition 4.35. *Given the stable continuous-time system* $\boldsymbol{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ *with infinite gramians* $\mathcal{P}_c$, $\mathcal{Q}_c$, *let* $\boldsymbol{\Sigma}_d = \left(\begin{array}{c|c} \mathbf{F} & \mathbf{G} \\ \hline \mathbf{H} & \mathbf{J} \end{array}\right)$, *with infinite gramians* $\mathcal{P}_d$, $\mathcal{Q}_d$, *be the stable discrete-time system obtained by means of the bilinear transformation given above. It follows that this bilinear transformation preserves the gramians:*

$$\mathcal{P}_c = \mathcal{P}_d \text{ and } \mathcal{Q}_c = \mathcal{Q}_d.$$

Consequently, the Hankel-norm of Σ_c *and* Σ_d, *defined by* (5.7), *is the same. Furthermore, the transformation preserves the infinity norms as defined by* (5.8) *and* (5.9).

The above result implies that the bilinear transformation between discrete- and continuous-time systems *preserves* balancing; this concept is discussed in section 7.

4.4 The realization problem

In the preceding sections, we presented two ways of describing linear systems: the internal and the external. The former makes use of the inputs $\mathbf{u}$, states $\mathbf{x}$, and outputs $\mathbf{y}$. The latter makes use *only* of the inputs $\mathbf{u}$ and the outputs $\mathbf{y}$. The question thus arises as to the relationship between these two descriptions.

In one direction, this problem is trivial. Given the internal description $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ of a system, the external description is readily derived. The transfer function of the system is given by (4.22)

$$\mathbf{H}_{\boldsymbol{\Sigma}}(\xi) = \mathbf{D} + \mathbf{C}(\xi\mathbf{I} - \mathbf{A})^{-1}\mathbf{B},$$

while from (4.23), the Markov parameters are given by

$$\mathbf{h}_0 = \mathbf{D},\ \mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B} \in \mathbb{R}^{p\times m}, \qquad k = 1, 2, \ldots. \tag{4.61}$$

The converse problem, i.e., given the external description, derive the internal one, is far from trivial. This is the realization problem: given the external description of a linear system, construct an internal or state variable description. In other words, given the impulse response $\mathbf{h}$ or, equivalently, the transfer function $\mathbf{H}$, or the Markov parameters $\mathbf{h}_k$ of a system, construct $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ such that (4.61) holds. It readily follows without computation that $\mathbf{D} = \mathbf{h}_0$. Hence the following problem results.

Definition 4.36. *Given the sequence of* $p \times m$ *matrices* $\mathbf{h}_k$, $k > 0$, *the* realization problem *consists of finding a positive integer* n *and constant matrices* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *such that*

$$\mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B}, \qquad \mathbf{C} \in \mathbb{R}^{p\times n},\ \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m},\ k = 1, 2, \ldots. \tag{4.62}$$

The triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *is then called a* realization *of the sequence* $\mathbf{h}_k$, *and the latter is called a* realizable *sequence.* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *is a* minimal *realization if among all realizations of the sequence, its dimension is the smallest possible.*

The realization problem is sometimes referred to as the problem of construction of state for linear systems described by convolution relationships.

Remark 4.4.1. *Realization* was formally introduced in the 1960s (see Kalman, Falb, and Arbib [192]), and eventually two approaches crystallized: the state-space and the polynomial. (See Fuhrmann [121] for an overview of the interplay between these two approaches in linear system theory.) The state-space method uses the Hankel matrix as a main tool and will be presented next. The polynomial approach has the Euclidean division algorithm as a focal point; see, e.g., Kalman [193], Fuhrmann [123], Antoulas [9], and van Barel and Bultheel [329]. Actually, Antoulas [9] presents the complete theory of recursive realization for multi-input, multi-output systems.

Example 4.37. Consider the following (scalar) sequences:

$$\Sigma_1 = \{1, 1, 1, 1, 1, 1, 1, 1, 1, \ldots\},$$

$$\Sigma_2 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, \ldots\} \text{ natural numbers},$$

$$\Sigma_3 = \{1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots\} \text{ Fibonacci numbers},$$

$$\Sigma_4 = \{1, 2, 3, 5, 7, 11, 13, 17, 19, \ldots\} \text{ primes},$$

$$\Sigma_5 = \left\{\frac{1}{1!}, \frac{1}{2!}, \frac{1}{3!}, \frac{1}{4!}, \frac{1}{5!}, \frac{1}{6!}, \ldots\right\} \text{ inverse factorials.}$$

It is assumed that for all sequences, $\mathbf{h}_0 = \mathbf{D} = 0$. Which sequences are realizable? This question will be answered in the example of section 4.43.

Problems. The following problems arise:

(a) Existence: given a sequence $\mathbf{h}_k$, $k > 0$, determine whether there exist a positive integer n and a triple of matrices $\mathbf{C}$, $\mathbf{A}$, $\mathbf{B}$ such that (4.62) holds.

(b) Uniqueness: in case such an integer and triple exist, are they unique in some sense?

(c) Construction: in case of existence, find n and give an algorithm to construct such a triple.

The main tool for answering the above questions is the matrix $\mathcal{H}$ of Markov parameters:

$$\mathcal{H} = \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_k & \mathbf{h}_{k+1} & \cdots \\ \mathbf{h}_2 & \mathbf{h}_3 & \cdots & \mathbf{h}_{k+1} & \mathbf{h}_{k+2} & \cdots \\ \vdots & \vdots & \iddots & \vdots & \vdots & \iddots \\ \mathbf{h}_k & \mathbf{h}_{k+1} & \cdots & \mathbf{h}_{2k-1} & \mathbf{h}_{2k} & \cdots \\ \mathbf{h}_{k+1} & \mathbf{h}_{k+2} & \cdots & \mathbf{h}_{2k} & \mathbf{h}_{2k+1} & \cdots \\ \vdots & \vdots & \iddots & \vdots & \vdots & \end{pmatrix}. \tag{4.63}$$

This is the Hankel matrix; it has infinitely many rows, infinitely many columns, and block Hankel structure, i.e., $(\mathcal{H})_{i,j} = \mathbf{h}_{i+j-1}$, for $i, j > 0$. We start by listing conditions related to the realization problem.

Lemma 4.38. *The following statements are equivalent:*

(a) *The sequence* $\mathbf{h}_k$, $k > 0$, *is* realizable.

(b) *The formal power series* $\sum_{k>0} \mathbf{h}_k s^{-k}$ *is* rational.

(c) *The sequence* $\mathbf{h}_k$, $k > 0$, *satisfies a* recursion *with constant coefficients, i.e., there exist a positive integer* r *and constants* α_i, $0 \leq i < r$, *such that*

$$\mathbf{h}_{r+k} = -\alpha_0 \mathbf{h}_k - \alpha_1 \mathbf{h}_{k+1} - \alpha_2 \mathbf{h}_{k+2} - \cdots - \alpha_{r-2}\mathbf{h}_{r+k-2} - \alpha_{r-1}\mathbf{h}_{r+k-1}, \quad k > 0. \tag{4.64}$$

(d) *The rank of* $\mathcal{H}$ *is* finite.

Proof. **(a) ⇒ (b).** Realizability implies (4.62). Hence

$$\sum_{k>0} \mathbf{h}_k s^{-k} = \sum_{k>0} \mathbf{C}\mathbf{A}^{k-1}\mathbf{B}s^{-k} = \mathbf{C}\left(\sum_{k>0} \mathbf{A}^{k-1}s^{-k}\right)\mathbf{B} = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}.$$

This proves (b). Notice that the quantity in parentheses is a *formal* power series and convergence is not an issue.

(b) ⇒ (c). Let $\det(s\mathbf{I} - \mathbf{A}) = \alpha_0 + \alpha_1 s + \cdots + \alpha_{r-1}s^{r-1} + s^r = \chi_{\mathbf{A}}(s)$. The previous relationship implies

$$\chi_{\mathbf{A}}(s)\left(\sum_{k>0} \mathbf{h}_k s^{-k}\right) = \mathbf{C}\left[\operatorname{adj}(s\mathbf{I}-\mathbf{A})\right]\mathbf{B},$$

where adj (**M**) denotes the *matrix adjoint* of the **M**, i.e., the matrix of *cofactors*. (For the definition and properties of the cofactors and of the adjoint, see Chapter 6 of Meyer's book [238].) On the left-hand side are terms having both positive and negative powers of s, while on the right-hand side are only terms having positive powers of s. Hence the coefficients of the negative powers of s on the left-hand side must be identically zero; this implies precisely (4.64).

(c) ⇒ (d). Relationships (4.64) imply that the $(r + 1)$st block column of $\mathcal{H}$ is a linear combination of the previous r block columns. Furthermore, because of the block Hankel structure, every block column of $\mathcal{H}$ is a subcolumn of the previous one; this implies that all block columns after the rth are linearly dependent on the first r, which in turn implies the finiteness of the rank of $\mathcal{H}$. □

The following lemma describes a fundamental property of $\mathcal{H}$; it also provides a direct proof of the implication (a) ⇒ (d).

Lemma 4.39. Factorization of $\mathcal{H}$. *If the sequence of Markov parameters is* realizable *by means of the triple* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$, *$\mathcal{H}$ can be factored,*

$$\mathcal{H} = \mathcal{O}(\mathbf{C}, \mathbf{A})\mathcal{R}(\mathbf{A}, \mathbf{B}). \tag{4.65}$$

Consequently, if the sequence of Markov parameters is realizable, the rank of $\mathcal{H}$ is finite.

Proof. If the sequence $\{\mathbf{h}_n,\ n = 1, 2, \ldots\}$ is realizable, the relationships $\mathbf{h}_n = \mathbf{C}\mathbf{A}^{n-1}\mathbf{B}$ hold. Hence,

$$\mathcal{H} = \begin{pmatrix} \mathbf{CB} & \mathbf{CAB} & \cdots \\ \mathbf{CAB} & \mathbf{CA}^2\mathbf{B} & \cdots \\ \vdots & \vdots & \\ & & \end{pmatrix} = \mathcal{O}(\mathbf{C}, \mathbf{A})\mathcal{R}(\mathbf{A}, \mathbf{B}).$$

It follows that $\operatorname{rank} \mathcal{H} \leq \max\{\operatorname{rank} \mathcal{O}, \operatorname{rank} \mathcal{R}\} \leq \dim(\mathbf{A})$. □

To discuss the uniqueness issue of realizations, we need to recall the concept of equivalent systems defined by (4.24). In particular, Proposition 4.4 asserts that equivalent triples $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ have the same Markov parameters. Hence the best one can hope for the uniqueness question is that realizations be equivalent. Indeed, as shown in the next section, this holds for realizations with the smallest possible dimension.

4.4.1 The solution of the realization problem

We are now ready to answer the three questions posed at the beginning of this section. In the process we prove the implication (d) $\Rightarrow$ (a) and hence the equivalence of the statements in Lemma 4.38.

Theorem 4.40. Main Result.

(1) *The sequence* $\mathbf{h}_k$, $k > 0$, *is realizable if and only if* $\operatorname{rank} \mathcal{H} = n < \infty$.

(2) *The state-space dimension of any solution is at least n. All realizations that are minimal are both reachable and observable. Conversely, every realization that is reachable and observable is minimal.*

(3) *All minimal realizations are equivalent.*

Lemma 4.39 proves part (1) of the main theorem in one direction. To prove (1) in the other direction we will actually construct a realization assuming that the rank of $\mathcal{H}$ is finite. For this we need to define the shift σ. It acts on the columns on the Hankel matrix; if $(\mathcal{H})_k$ denotes the kth column of $\mathcal{H}$, $\sigma(\mathcal{H})_k = (\mathcal{H})_{k+m}$; in other words, σ is a shift by m columns. The shift applied to a submatrix of $\mathcal{H}$ consisting of several columns is applied to each column separately.

Lemma 4.41. Silverman realization algorithm. *Let* rank $\mathcal{H} = n$. *Find an* $n \times n$ *submatrix* Φ *of* $\mathcal{H}$ *that has full rank. Construct the following matrices:*

(i) $\sigma\Phi \in \mathbb{R}^{n\times n}$ *is the submatrix of* $\mathcal{H}$ *having the rows with the same index as those of* Φ *and the columns obtained by shifting each individual column of* Φ *by one block column (i.e., m columns).*

(ii) $\Gamma \in \mathbb{R}^{n\times m}$ *is composed of the same rows as* Φ*; its columns are the first m columns of* $\mathcal{H}$*.*

(iii) $\Lambda \in \mathbb{R}^{p\times n}$ *is composed of the same columns as* Φ*; its rows are the first p rows of* $\mathcal{H}$*.*

The triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$*, where* $\mathbf{C} = \Lambda$, $\mathbf{A} = \Phi^{-1}\sigma\Phi$, *and* $\mathbf{B} = \Phi^{-1}\Gamma$, *is a realization of dimension n of the given sequence of Markov parameters.*

Proof. By assumption there exist $n = \text{rank}\mathcal{H}$ columns of $\mathcal{H}$ that span its column space. Denote these columns by Φ_∞; note that the columns making up Φ_∞ need not be consecutive columns of $\mathcal{H}$. We denote by σ the column right-shift operator. Let $\sigma\Phi_\infty$ denote the n columns of $\mathcal{H}$ obtained by shifting those of Φ_∞ by one block column, i.e., by m individual columns; let Γ_∞ denote the first m columns of $\mathcal{H}$. Since the columns of Φ_∞ form a basis for the space spanned by the columns of $\mathcal{H}$, there exist *unique* matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$ such that

$$\sigma\Phi_\infty = \Phi_\infty\mathbf{A}, \tag{4.66}$$

$$\Gamma_\infty = \Phi_\infty\mathbf{B}. \tag{4.67}$$

Finally, define $\mathbf{C}$ as the first block row, i.e., the first p individual rows, of Φ_∞:

$$\mathbf{C} = (\Phi_\infty)_1. \tag{4.68}$$

For this proof $(\mathbf{M})_k$, $k > 0$, denotes the kth block row of the matrix $\mathbf{M}$. Recall that the first block element of Γ_∞ is $\mathbf{h}_1$, i.e., using our notation $(\Gamma_\infty)_1 = \mathbf{h}_1$. Thus (4.67), together with (4.68), implies

$$\mathbf{h}_1 = (\Gamma_\infty)_1 = (\Phi_\infty\mathbf{B})_1 = (\Phi_\infty)_1\mathbf{B} = \mathbf{CB}.$$

For the next Markov parameter, notice that

$$\mathbf{h}_2 = (\sigma\Gamma_\infty)_1 = (\Gamma_\infty)_2.$$

Thus making use of (4.66), we have

$$h_2 = (\sigma\Gamma_\infty)_1 = (\sigma\Phi_\infty\mathbf{B})_1 = (\Phi_\infty\mathbf{AB})_1 = (\Phi_\infty)_1\mathbf{AB} = \mathbf{CAB}.$$

For the kth Markov parameter, combining (4.67), (4.66), and (4.68), we obtain

$$\mathbf{h}_k = (\sigma^{k-1}\Gamma_\infty)_1 = (\sigma^{k-1}\Phi_\infty\mathbf{B})_1 = (\Phi_\infty\mathbf{A}^{k-1}\mathbf{B})_1 = (\Phi_\infty)_1\mathbf{A}^{k-1}\mathbf{B} = \mathbf{CA}^{k-1}\mathbf{B}.$$

Thus $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ is indeed a realization of dimension n. □

The state dimension of a realization cannot be less than n; indeed, if such a realization exists, the rank of $\mathcal{H}$ will be less than n, which is a contradiction to the assumption that the rank of $\mathcal{H}$ is equal to n. Thus a realization of Σ whose dimension equals rank $\mathcal{H}$ is called a *minimal realization*; notice that the Silverman algorithm constructs minimal realizations. In this context the following holds true.

Lemma 4.42. *A realization of Σ is minimal if and only if it is reachable and observable.*

Proof. Let $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ be some realization of $\mathbf{h}_n$, $n > 0$. Since $\mathcal{H} = \mathcal{O}\mathcal{R}$,

$$\text{rank}\mathcal{H} \leq \min\{\text{rank}\mathcal{O}, \text{rank}\mathcal{R}\} \leq \dim(\mathbf{A}).$$

Let $(\hat{\mathbf{C}}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ be a reachable and observable realization. Since $\mathcal{H} = \hat{\mathcal{O}}\hat{\mathcal{R}}$, and each of the matrices $\hat{\mathcal{O}}$, $\hat{\mathcal{R}}$ contains a nonsingular matrix of size $\hat{\mathbf{A}}$, we conclude that $\dim(\hat{\mathbf{A}}) \leq \text{rank}\mathcal{H}$, which concludes the proof. □

We are now left with the proof of part (3) of the main theorem, namely, that minimal realizations are equivalent. We provide the proof only for a special case; the proof of the general case follows along similar lines.

Outline of proof. *SISO case (i.e., $p = m = 1$).* Let $(\mathbf{C}_i, \mathbf{A}_i, \mathbf{B}_i)$, $i = 1, 2$, be minimal realizations of Σ. We will show the existence of a transformation $\mathbf{T}$, $\det \mathbf{T} \neq 0$ such that (4.24) holds. From Lemma 4.39 we conclude that

$$\mathcal{H}_{n,n} = \mathcal{O}_n^1\mathcal{R}_n^1 = \mathcal{O}_n^2\mathcal{R}_n^2, \tag{4.69}$$

where the superscript is used to distinguish between the two different realizations. Furthermore, the same lemma also implies

$$\mathcal{H}_{n,n+1} = \mathcal{O}_n^1[\mathbf{B}_1 \quad \mathbf{A}_1\mathcal{R}_n^1] = \mathcal{O}_n^2[\mathbf{B}_2 \quad \mathbf{A}_2\mathcal{R}_n^2],$$

which in turn yields

$$\mathcal{O}_n^1\mathbf{A}_1\mathcal{R}_n^1 = \mathcal{O}_n^2\mathbf{A}_2\mathcal{R}_n^2. \tag{4.70}$$

Because of minimality, the following determinants are nonzero: $\det \mathcal{O}_n^i \neq 0$, $\det \mathcal{R}_n^i \neq 0$, $i = 1, 2$. We now define

$$\mathbf{T} = (\mathcal{O}_n^1)^{-1}\mathcal{O}_n^2 = \mathcal{R}_n^1(\mathcal{R}_n^2)^{-1}.$$

Equation (4.69) implies $\mathbf{C}_1 = \mathbf{C}_2\mathbf{T}^{-1}$ and $\mathbf{B}_1 = \mathbf{T}\mathbf{B}_2$, while (4.70) implies $\mathbf{A}_1 = \mathbf{T}\mathbf{A}_2\mathbf{T}^{-1}$. □

Example 4.43. We now investigate the realization problem for the *Fibonacci sequence* given in Example 4.37,

$$\Sigma_3 = \{1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots\},$$

which is constructed according to the rule $\mathbf{h}_1 = 1$, $\mathbf{h}_2 = 2$, and

$$\mathbf{h}_{k+2} = \mathbf{h}_{k+1} + \mathbf{h}_k, \qquad k > 0.$$

The Hankel matrix (4.63) becomes

$$\mathcal{H} = \begin{pmatrix} 1 & 2 & 3 & 5 & 8 & 13 & \cdots \\ 2 & 3 & 5 & 8 & 13 & 21 & \cdots \\ 3 & 5 & 8 & 13 & 21 & 34 & \cdots \\ 5 & 8 & 13 & 21 & 34 & 55 & \cdots \\ 8 & 13 & 21 & 34 & 55 & 89 & \cdots \\ 13 & 21 & 34 & 55 & 89 & 144 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

It readily follows from the law of construction of the sequence that the rank of the Hankel matrix is two. Φ is chosen so that it contains rows 2 and 4 and columns 2 and 5 of $\mathcal{H}$:

$$\Phi = \begin{pmatrix} 3 & 13 \\ 8 & 34 \end{pmatrix} \Rightarrow \Phi^{-1} = \begin{pmatrix} -17 & \frac{13}{2} \\ 4 & -\frac{3}{2} \end{pmatrix}.$$

The remaining matrices are now

$$\sigma\Phi = \begin{pmatrix} 5 & 21 \\ 13 & 55 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 2 \\ 5 \end{pmatrix}, \quad \text{and } \Lambda = (2 \quad 8).$$

It follows that

$$\mathbf{A} = \Phi^{-1}\sigma\Phi = \begin{pmatrix} -1/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix}, \quad \mathbf{B} = \Phi^{-1}\Gamma = \begin{pmatrix} -3/2 \\ 1/2 \end{pmatrix}, \quad \text{and } \mathbf{C} = (2 \quad 8).$$

Furthermore,

$$\mathbf{H}_3(s) = \sum_{k>0} \mathbf{h}_k s^{-k} = \frac{s+1}{s^2 - s - 1}.$$

Concerning the remaining four sequences of Example 4.37, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are realizable, while the last two, namely, $\boldsymbol{\Sigma}_4$ and $\boldsymbol{\Sigma}_5$, are not realizable. In particular,

$$\boldsymbol{\Sigma}_1 = \left(\begin{array}{c|c} 1 & 1 \\ \hline 1 & 0 \end{array}\right), \quad \mathbf{H}_1(s) = \frac{1}{s-1}, \quad \text{and } \boldsymbol{\Sigma}_2 = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ -1 & 2 & 1 \\ \hline 0 & 1 & 0 \end{array}\right], \quad \mathbf{H}_2(s) = \frac{s}{(s-1)^2}.$$

In the last case, $\mathbf{H}_5(s) = e^{s^{-1}} - 1$, which is not rational and hence has no finite-dimensional realization. The fact that $\boldsymbol{\Sigma}_5$ is not realizable follows also from the fact that the determinant of the associated Hankel matrix $\mathcal{H}_{i,j} = \frac{1}{i+j-1}$, of size n, also known as the *Hankel determinant*, is nonzero; it has been shown in [5], namely, that

$$\det \mathcal{H}_n = \prod_{i=2}^{n} \prod_{j=2}^{i} \frac{(j-1)^2}{(2j-1)(2j-2)^2(2j-3)},$$

which implies that the Hankel determinant for $n = 4, 5, 6$ is of the order 10^{-7}, 10^{-17}, and 10^{-43}, respectively.

4.4.2 Realization of proper rational matrix functions

Given is a $p \times m$ matrix $\mathbf{H}(s)$ with proper rational entries, i.e., entries whose numerator degree is no larger than the denominator degree. Consider first the scalar case, i.e., $p = m = 1$. We can write

$$\mathbf{H}(s) = \mathbf{D} + \frac{\mathbf{p}(s)}{\mathbf{q}(s)},$$

where $\mathbf{D}$ is a constant in $\mathbb{R}$ and $\mathbf{p}$, $\mathbf{q}$ are polynomials in s,

$$\begin{aligned} \mathbf{p}(s) &= p_0 + p_1 s + \cdots + p_{\nu-1} s^{\nu-1}, & p_i &\in \mathbb{R}, \\ \mathbf{q}(s) &= q_0 + q_1 s + \cdots + q_{\nu-1} s^{\nu-1} + s^{\nu}, & q_i &\in \mathbb{R}. \end{aligned}$$

In terms of these coefficients p_i and q_i, we can write down a realization of $\mathbf{H}(s)$ as follows:

$$\Sigma_{\mathbf{H}} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{ccccc|c} 0 & 1 & 0 & & 0 & 0 \\ 0 & 0 & 1 & & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & & 1 & 0 \\ -q_0 & -q_1 & -q_2 & \cdots & -q_{\nu-1} & 1 \\ \hline p_0 & p_1 & p_2 & \cdots & p_{\nu-1} & \mathbf{D} \end{array}\right) \in \mathbb{R}^{(\nu+1)\times(\nu+1)} \qquad (4.71)$$

It can be shown that $\Sigma_{\mathbf{H}}$ is indeed a realization of $\mathbf{H}$, i.e.,

$$\mathbf{H}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I}_{\nu} - \mathbf{A})^{-1}\mathbf{B}.$$

This realization is reachable but not necessarily observable; this means that the rank of the associated Hankel matrix is at most ν. The realization is in addition observable if the polynomials $\mathbf{p}$ and $\mathbf{q}$ are coprime. Thus (4.71) is *minimal* if $\mathbf{p}$ and $\mathbf{q}$ are coprime. In this case the rank of the associated Hankel matrix $\mathcal{H}$ is precisely ν.

In the general case, we can write

$$\mathbf{H}(s) = \mathbf{D} + \frac{1}{\mathbf{q}(s)}\mathbf{P}(s),$$

where $\mathbf{q}$ is a scalar polynomial that is the least common multiple of the denominators of the entries of $\mathbf{H}$ and $\mathbf{P}$ is a polynomial matrix of size $p \times m$:

$$\begin{aligned} \mathbf{P}(s) &= \mathbf{P}_0 + \mathbf{P}_1 s + \cdots + \mathbf{P}_{\nu-1} s^{\nu-1}, & \mathbf{P}_i &\in \mathbb{R}^{p\times m}, \\ \mathbf{q}(s) &= q_0 + q_1 s + \cdots + q_{\nu-1} s^{\nu-1} + s^{\nu}, & q_i &\in \mathbb{R}. \end{aligned}$$

The construction given above provides a realization:

$$\Sigma_{\mathbf{H}} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{ccccc|c} \mathbf{0}_m & \mathbf{I}_m & \mathbf{0}_m & \cdots & \mathbf{0}_m & \mathbf{0}_m \\ \mathbf{0}_m & \mathbf{0}_m & \mathbf{I}_m & \cdots & \mathbf{0}_m & \mathbf{0}_m \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_m & \mathbf{0}_m & \mathbf{0}_m & & \mathbf{I}_m & \mathbf{0}_m \\ -q_0\mathbf{I}_m & -q_1\mathbf{I}_m & -q_2\mathbf{I}_m & \cdots & -q_{\nu-1}\mathbf{I}_m & \mathbf{I}_m \\ \hline \mathbf{P}_0 & \mathbf{P}_1 & \mathbf{P}_2 & \cdots & \mathbf{P}_{\nu-1} & \mathbf{D} \end{array}\right),$$

where $\mathbf{0}_m$ is a square zero matrix of size m, $\mathbf{I}_m$ is the identity matrix of the same size, and, consequently, the state of this realization has size νm. Unlike the scalar case, however, the realization $\Sigma_{\mathbf{H}}$ need not be minimal. One way to obtain a minimal realization is by applying the reachable observable-canonical decomposition given in Lemma 4.25. An alternative way is to apply the Silverman algorithm; in this case, $\mathbf{H}$ has to be expanded into a formal power series,

$$\mathbf{H}(s) = \mathbf{h}_0 + \mathbf{h}_1 s^{-1} + \mathbf{h}_2 s^{-2} + \cdots + \mathbf{h}_t s^{-t} + \cdots .$$

The Markov parameters can be computed using the following relationship. Given the polynomial q as above, let

$$q^{(k)}(s) = s^{\nu-k} + q_{n-1}s^{\nu-k-1} + \cdots + q_{k+1}s + q_k, \qquad k = 1, \ldots, \nu, \tag{4.72}$$

denote its ν *pseudo-derivative* polynomials. It follows that the numerator polynomial $\mathbf{P}(s)$ is related to the Markov parameters $\mathbf{h}_k$ and the denominator polynomial q as follows:

$$\mathbf{P}(s) = \mathbf{h}_1 q^{(1)}(s) + \mathbf{h}_2 q^{(2)}(s) + \cdots + \mathbf{h}_{\nu-1} q^{(\nu-1)}(s) + \mathbf{h}_\nu q^{(\nu)}(s). \tag{4.73}$$

This can be verified by direct calculation. Alternatively, assume that $\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, and let $q(s)$ denote the characteristic polynomial of $\mathbf{A}$. Then

$$\operatorname{adj}(s\mathbf{I} - \mathbf{A}) = q^{(\nu)}(s)\mathbf{A}^{\nu-1} + q^{(\nu-1)}(s)\mathbf{A}^{\nu-2} + \cdots q^{(2)}(s)\mathbf{A}^1 + q^{(1)}(s)\mathbf{I}. \tag{4.74}$$

The result (4.73) follows by noting that $\mathbf{P}(s) = \mathbf{C}\operatorname{adj}(s\mathbf{I} - \mathbf{A})\,\mathbf{B}$.

Since $\mathbf{H}$ is rational, the rank of the ensuing *Hankel* matrix associated with the sequence of Markov parameters $\mathbf{h}_k$, $k > 0$, is guaranteed to have *finite rank*. In particular, the following upper bound holds:

$$\operatorname{rank}\mathcal{H} \leq \min\{\nu m, \nu p\}.$$

An important attribute of a rational matrix function is its *McMillan degree*. For proper rational matrix functions $\mathbf{H}$, the McMillan degree turns out to equal the rank of the associated Hankel matrix $\mathcal{H}$; in other words, the McMillan degree in this case is equal to the dimension of any minimal realization of $\mathbf{H}$.

4.4.3 Symmetric systems and symmetric realizations

A system Σ is called symmetric if its Markov parameters are symmetric, $\mathbf{h}_k = \mathbf{h}_k^*$, $k \geq 0$.[3] In other words, Σ is symmetric if $\mathbf{h}_0 = \mathbf{h}_0^*$ and the associated Hankel matrix (4.63) is symmetric, $\mathcal{H} = \mathcal{H}^*$.

Definition 4.44. *A realization is called* symmetric *if* $\mathbf{D} = \mathbf{D}^*$ *and there exists a symmetric matrix* $\Psi = \Psi^*$ *such that*

$$\mathbf{A}\Psi = \Psi\mathbf{A}^*, \;\; \mathbf{B} = \Psi\mathbf{C}^*. \tag{4.75}$$

[3] Recall that if a matrix is real, the superscript $(\cdot)^*$ denotes simple transposition.

It follows that every symmetric system has a symmetric realization.

Lemma 4.45. *A reachable and observable system Σ is symmetric if and only if it possesses a symmetric realization.*

Proof. A moment's reflection shows that if Σ has a symmetric realization, it is symmetric. Conversely, let the system be symmetric; this together with the factorization (4.65) implies

$$\mathcal{H} = \mathcal{O}(\mathbf{C}, \mathbf{A})\mathcal{R}(\mathbf{A}, \mathbf{B}) = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)\mathcal{O}^*(\mathbf{A}^*, \mathbf{C}^*) = \mathcal{H}^*.$$

Thus, since the column span of $\mathcal{H}$ and $\mathcal{H}^*$ are the same, there exists a matrix $\Psi \in \mathbb{R}^{n\times n}$ such that $\mathcal{O}(\mathbf{C}, \mathbf{A})\Psi = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)$; hence $\mathbf{C}\Psi = \mathbf{B}^*$. Furthermore, $\mathcal{O}(\mathbf{C}, \mathbf{A})\mathbf{A}\Psi = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)\mathbf{A}^* = \mathcal{O}(\mathbf{C}, \mathbf{A})\Psi\mathbf{A}^*$. Since $\mathcal{O}$ has full column rank, the equality $\mathbf{A}\Psi = \Psi\mathbf{A}^*$ follows. It remains to show that Ψ is symmetric. Notice that $\mathcal{O}(\mathbf{C}, \mathbf{A})\Psi\mathcal{O}^*(\mathbf{A}^*, \mathbf{C}^*) = \mathcal{R}^*(\mathbf{B}^*, \mathbf{A}^*)\mathcal{O}^*(\mathbf{A}^*, \mathbf{C}^*) = \mathcal{H}$ is symmetric. Again, since $\mathcal{O}$ has full column rank, it has an $n \times n$ nonsingular submatrix, composed of the rows with index $I = \{i_1, \ldots, i_n\}$, which we denote by $\mathcal{O}_I$; thus $\mathcal{O}_I\Psi\mathcal{O}_I^* = \mathcal{H}_{I,I}$, where the latter is the submatrix of the Hankel matrix composed of those rows and columns indexed by I. Thus $\Psi = [\mathcal{O}_I]^{-1}\mathcal{H}_{I,I}\left[\mathcal{O}_I^*\right]^{-1}$. Since $\mathcal{H}_{I,I}$ is symmetric, this proves the symmetry of Ψ. The proof is thus complete. □

4.4.4 The partial realization problem

This problem was studied in [193]. Recursive solutions were provided in [12] and [11]. Recall section 4.4 and in particular Definition 4.36. The realization problem with partial data is defined as follows.

Definition 4.46. *Given the finite sequence of $p \times m$ matrices $\mathbf{h}_k$, $k = 1, \ldots, r$, the* partial realization problem *consists of finding a positive integer n and constant matrices* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *such that*

$$\mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B};\ \mathbf{C} \in \mathbb{R}^{p\times n},\ \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{B} \in \mathbb{R}^{n\times m}, \qquad k = 1, 2, \ldots, N.$$

The triple $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ *is then called a* partial realization *of the sequence* $\mathbf{h}_k$.

Because of Lemma 4.38, a finite sequence of matrices is always realizable. As a consequence, the set of problems arising consists of

(a) *minimality*: given the sequence $\mathbf{h}_k$, $k = 1, \ldots, r$, find the smallest positive integer n for which the partial realization problem is solvable.

(b) *parametrization of solutions*: parametrize all minimal and other solutions.

(c) *recursive construction*: recursive construction of solutions.

Similarly to the realization problem, the partial realization problem can be studied by means of the partially defined Hankel matrix:

$$\mathcal{H}_r = \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & & \mathbf{h}_r \\ \mathbf{h}_2 & & \cdots & \mathbf{h}_r & ? \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ & \mathbf{h}_r & \cdots & ? & ? \\ \mathbf{h}_r & ? & \cdots & ? & ? \end{pmatrix} \in \mathbb{R}^{rp \times rm},$$

where ? denote unknown matrices defining the continuation of the given finite sequence $\mathbf{h}_k$, $k = 1, \ldots, N$.

The *rank* of the partially defined Hankel matrix $\mathcal{H}_k$ is defined as the size of the largest nonsingular submatrix of $\mathcal{H}_k$, independently of the unknown parameters "?". It then follows that the dimension of any partial realization $\mathbf{\Sigma}$ satisfies

$$\dim \mathbf{\Sigma} \geq \operatorname{rank} \mathcal{H}_k = n.$$

Furthermore, there always exists a partial realization of dimension n, which is a minimal partial realization. Once the rank of $\mathcal{H}_k$ is determined, Silverman's algorithm (see Lemma 4.41) can be used to construct such a realization. We illustrate this procedure by means of a simple example.

Example 4.47. Consider the scalar (i.e., $m = p = 1$) sequence $\mathbf{\Sigma} = (1, 1, 1, 2)$; the corresponding Hankel matrix $\mathcal{H}_4$ and its first three submatrices are

$$\mathcal{H}_4 = \begin{pmatrix} 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & a \\ 1 & 2 & a & b \\ 2 & a & b & c \end{pmatrix}, \quad \mathcal{H}_3 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & a \end{pmatrix}, \quad \mathcal{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathcal{H}_1 = (1),$$

where a, b, c denote the unknown continuation of the original sequence. The determinants of these matrices are

$$\det \mathcal{H}_4 = -a^3 + 4a^2 - 8a + 8 + 2ab - 3b - c, \quad \det \mathcal{H}_3 = -1, \quad \det \mathcal{H}_2 = 0, \quad \det \mathcal{H}_1 = 1.$$

It follows that

$$\operatorname{rank} \mathcal{H}_4 = 3.$$

By Lemma 4.41, we choose $\Phi = \mathcal{H}_3$, $\Gamma = \Lambda^* = (1 \;\; 1 \;\; 1)^*$, which implies

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & a^2 - 4a + 8 - b \\ 1 & 0 & -a^2 + 3a - 4 + b \\ 0 & 1 & a - 2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{C} = (1 \;\; 1 \;\; 1).$$

Hence, there are multiple minimal partial realizations of $\mathbf{\Sigma}$. Indeed, the above expressions provide a parametrization of all minimal solutions; the parameters are $a, b \in \mathbb{R}$. Finally, we note that the value of c is uniquely determined by a, b. In this case, for realizations of minimal degree 3, we must have $c = -a^3 + 5a^2 - 12a + 2ab - 4b + 16$.

4.5 The rational interpolation problem*

The realization problem aims at constructing a linear system in internal (state-space) form, from some or all Markov parameters $\mathbf{h}_k$. As already noted, the Markov parameters constitute information about the transfer function $\mathbf{H}(s)$ obtained by expanding it around infinity $\mathbf{H}(s) = \sum_{k\geq 0} \mathbf{h}_k s^{-k}$.

In this section we will address the problem of constructing an external or internal description of a system based on finitely many samples of the transfer function taken at finite points in the complex plane. This problem is known as *rational interpolation*. For simplicity we will assume that the systems in question have a single input and a single output and thus their transfer function is scalar rational. We distinguish between two approaches.

The first approach, presented in subsection 4.5.2, has as a main tool the so-called Löwner matrix. The Löwner matrix encodes the information about the admissible complexity of the solutions as a simple function of its rank. The computation of the solutions can be carried out both in the *external* (transfer function) and in the *internal* (state-space) frameworks. This approach to rational interpolation leads to a generalization of the classical, system theoretic, concept of *realization* of linear dynamical systems.

The second approach, presented in subsection 4.5.3, is known as the *generating system approach* and involves the construction of a polynomial or rational matrix, such that any polynomial or rational combination of its rows yields a solution of the problem at hand. This construction has a system theoretic interpretation as a *cascade interconnection* of two systems, one of which can be chosen freely. This method leads to *recursive solutions* in a natural way by expanding, namely, the cascade interconnection just mentioned. The solutions can be classified according to properties like *complexity*, *norm boundedness*, or *positive realness*.

4.5.1 Problem formulation*

Consider the array of pairs of points

$$\mathbb{P} = \{(s_i; \phi_{ij}) : j = 0, 1, \ldots, \ell_i - 1;\ i = 1, \ldots, k,\ s_i \neq s_j,\ i \neq j\}, \tag{4.76}$$

where we will assume that there are N given pieces of data, that is, $\sum_{i=1}^{k} \ell_i = N$. We are looking for *all* rational functions,

$$\phi(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)},\quad \gcd(\mathbf{n}, \mathbf{d}) = 1, \tag{4.77}$$

where $\mathbf{n}(s)$, $\mathbf{d}(s)$ are *coprime* polynomials, that is, their *greatest common divisor* is a (nonzero) constant, which *interpolate* the points of the array $\mathbb{P}$, i.e.,

$$\left.\frac{d^j\phi(s)}{ds^j}\right|_{s=s_i} = \phi_{ij},\qquad j = 0, 1, \ldots, \ell_i - 1,\ i = 1, \ldots, k. \tag{4.78}$$

In other words, the jth derivative of $\phi(s)$ evaluated at $s = s_i$ is equal to ϕ_{ij}. We distinguish two special cases: (a) the distinct point interpolation problem,

$$\mathbb{P} = \{(s_i; \phi_i) : i = 1, \ldots, N,\ s_i \neq s_j,\ i \neq j\}, \tag{4.79}$$

and (b) the single multiple point interpolation problem,

$$\mathbb{P} = \{(s_0; \phi_{0j}) : \; j = 0, 1, \ldots, N-1\}, \tag{4.80}$$

where the value of the function and derivatives thereof are provided only at $s = s_0$.

Solution of the unconstrained problem

The *Lagrange interpolating polynomial* associated with $\mathbb{P}$ is the unique polynomial of degree less than N which interpolates the points of this array. In the distinct point case (4.79), it is

$$\ell(s) = \sum_{j=1}^{N} \phi_j \prod_{i \neq j} \frac{s - s_i}{s_j - s_i}. \tag{4.81}$$

A similar formula holds for the general case. A parametrization of *all* solutions to (4.77), (4.78) can be given in terms of $\ell(s)$ as follows:

$$\phi(s) = \ell(s) + \mathbf{r}(s)\Pi_{i=1}^{N}(s - s_i), \tag{4.82}$$

where the parameter $\mathbf{r}(s)$ is an arbitrary rational function with no poles at the s_i. Most often, however, one is interested in parametrizing all solutions to the interpolation problem (4.78) which satisfy additional constraints. In such cases, this formula, although general, provides little insight.

Constrained interpolation problems

The first parameter of interest is the *complexity* or *degree* of rational interpolants (4.77). It is defined as

$$\deg \phi = \max\{\deg \mathbf{n}, \; \deg \mathbf{d}\}$$

and is sometimes referred to as the *McMillan degree* of the rational function ϕ. The following problems arise.

Problem (A): ***Parametrization of interpolants by complexity.***

(a) Find the *admissible* degrees of complexity, i.e., those positive integers π for which there exist solutions $\phi(s)$ to the interpolation problem (4.77), (4.78), with $\deg \phi = \pi$.

(b) Given an admissible degree π, construct *all* corresponding solutions.

Another constraint of interest is *bounded realness*, that is, finding interpolants which have poles in the left half of the complex plane (called *stable*) and whose magnitude on the imaginary axis is less than some given positive number μ.

Problem (B): Nevanlinna–Pick. ***Parametrization of interpolants by norm.***

(a) Do there exist bounded real interpolating functions?

(b) If so, what is the minimum norm and how can such interpolating functions be constructed?

A third constraint of interest is *positive realness* of interpolants. A function $\phi : \mathbb{C} \to \mathbb{C}$ is positive real (p.r.) if it maps the closed right half of the complex plane onto itself:

$$s \in \mathbb{C} : \mathcal{R}e(s) \geq 0 \mapsto \phi(s) : \mathcal{R}e(\phi(s)) \geq 0 \text{ for } s \text{ not a pole of } \phi.$$

Thus given the array of points $\mathbb{P}$, the following problem arises.

Problem (C): ***Parametrization of positive real interpolants.***

(a) Does there exist a p.r. interpolating function?

(b) If so, give a procedure to construct such interpolating functions.

In sections 4.5.2 and 4.5.3, we will investigate the constrained interpolation Problem (A) for the special case of *distinct* interpolating points. At the end of section 4.5.3, a short summary of the solution of the other two, that is, Problems (B) and (C), will be provided.

4.5.2 The Löwner matrix approach to rational interpolation*

The idea behind this approach to rational interpolation is to use a formula similar to (4.81) which would be valid for rational functions. Before introducing this formula, we partition the array $\mathbb{P}$ given by (4.79) into two disjoint subarrays $\mathbb{J}$ and $\mathbb{I}$ as follows:

$$\mathbb{J} = \{(s_i, \phi_i) : i = 1, \ldots, r\}, \quad \mathbb{I} = \{(\hat{s}_i, \hat{\phi}_i) : i = 1, \ldots, p\},$$

where for simplicity of notation some of the points have being redefined as follows: $\hat{s}_i = s_{r+i}, \hat{\phi}_i = \phi_{r+i}, i = 1, \ldots, p, p+r = N$. Consider $\phi(s)$ defined by the following equation:

$$\sum_{i=1}^{r} \gamma_i \frac{\phi(s) - \phi_i}{s - s_i} = 0, \qquad \gamma_i \neq 0,\ i = 1, \ldots, r,\ r \leq N.$$

Solving for $\phi(s)$ we obtain

$$\phi(s) = \frac{\sum_{j=1}^{r} \phi_j \gamma_j \Pi_{i \neq j}(s - s_i)}{\sum_{j=1}^{r} \gamma_j \Pi_{i \neq j}(s - s_i)}, \qquad \gamma_j \neq 0. \tag{4.83}$$

Clearly, the above formula, which can be regarded as the rational equivalent of the Lagrange formula, interpolates the first r points of the array $\mathbb{P}$, i.e., the points of the array $\mathbb{J}$. For $\phi(s)$ to interpolate the points of the array $\mathbb{I}$, the coefficients γ_i have to satisfy the following equation:

$$\boldsymbol{L}\gamma = \boldsymbol{0},$$

where

$$\boldsymbol{L} = \begin{bmatrix} \frac{\hat{\phi}_1 - \phi_1}{\hat{s}_1 - s_1} & \cdots & \frac{\hat{\phi}_1 - \phi_r}{\hat{s}_1 - s_r} \\ \vdots & & \vdots \\ \frac{\hat{\phi}_p - \phi_1}{\hat{s}_p - s_1} & \cdots & \frac{\hat{\phi}_p - \phi_r}{\hat{s}_p - s_r} \end{bmatrix} \in \mathbb{R}^{p \times r}, \quad \gamma = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_r \end{bmatrix} \in \mathbb{R}^r. \tag{4.84}$$

$\boldsymbol{L}$ is called the *Löwner matrix*, defined by means of the *row array* $\mathbb{I}$ and the *column array* $\mathbb{J}$. As it turns out, $\boldsymbol{L}$ is the main tool of this approach to the rational interpolation problem.

Remark 4.5.1. As shown by Antoulas and Anderson [25], the (generalized) Löwner matrix associated with the array $\mathbb{P}$ consisting of *one multiple point* (4.80) has *Hankel* structure. In particular,

$$\boldsymbol{L} = \begin{bmatrix} \frac{\phi^{(1)}(s_0)}{1!} & \frac{\phi^{(2)}(s_0)}{2!} & \frac{\phi^{(3)}(s_0)}{3!} & \cdots & \frac{\phi^{(k)}(s_0)}{k!} \\ \frac{\phi^{(2)}(s_0)}{2!} & \frac{\phi^{(3)}(s_0)}{3!} & \frac{\phi^{(4)}(s_0)}{4!} & \cdots & \frac{\phi^{(k+1)}(s_0)}{(k+1)!} \\ \frac{\phi^{(3)}(s_0)}{3!} & \frac{\phi^{(4)}(s_0)}{4!} & \frac{\phi^{(5)}(s_0)}{5!} & \cdots & \frac{\phi^{(k+2)}(s_0)}{(k+2)!} \\ \vdots & \vdots & \vdots & \iddots & \vdots \\ \frac{\phi^{(k)}(s_0)}{k!} & \frac{\phi^{(k+1)}(s_0)}{(k+1)!} & \frac{\phi^{(k+2)}(s_0)}{(k+2)!} & \cdots & \frac{\phi^{(2k)}(s_0)}{(2k)!} \end{bmatrix}.$$

This shows that the Löwner matrix is the right tool for generalizing realization theory to rational interpolation.

From rational function to Löwner matrix

The key result in connection with the Löwner matrix is the following.

Lemma 4.48. *Consider the array of points $\mathbb{P}$ defined by* (4.76), *consisting of samples taken from a given rational function $\phi(s)$, together with a partition into subarrays $\mathbb{J}$, $\mathbb{I}$ as defined in the beginning of the subsection. Let $\mathbf{L}$ be any $p \times r$ Löwner matrix with $p, r \geq \deg \phi$. It follows that* $\operatorname{rank} \boldsymbol{L} = \deg \phi$.

Corollary 4.49. *Under the assumptions of the lemma, any square Löwner submatrix of $\boldsymbol{L}$ of size* $\deg \phi$ *is nonsingular.*

In what follows, given $\mathbf{A} \in \mathbb{R}^{\pi\times\pi}$, $\mathbf{b}, \mathbf{c}^* \in \mathbb{R}^{\pi}$, the following matrices will be of interest:

$$\mathcal{R}_r = [(s_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \ \cdots \ (s_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}] \in \mathbb{R}^{\pi\times r}, \tag{4.85}$$

$$\mathcal{O}_p = [(\hat{s}_1\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{c}^* \ \cdots \ (\hat{s}_p\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{c}^*]^* \in \mathbb{R}^{p\times\pi}. \tag{4.86}$$

As will be shown subsequently in (4.87), the Löwner matrix factors in a product of $\mathcal{O}_p$ times $\mathcal{R}_r$. Therefore, in analogy with the realization problem (where the Hankel matrix factors in a product of an observability times a reachability matrix), we will call $\mathcal{O}_p$ the *generalized observability matrix* and $\mathcal{R}_r$ the *generalized reachability matrix* associated with the underlying interpolation problem.

Proposition 4.50. *Let* $(\mathbf{A}, \mathbf{b})$ *be a reachable pair, where* $\mathbf{A}$ *is a square matrix and* $\mathbf{b}$ *is a vector; in addition, let s_i, $i = 1, \ldots, r$, be scalars that are not eigenvalues of* $\mathbf{A}$. *It follows that the generalized reachability matrix defined by* (4.85) *has rank equal to the size of* $\mathbf{A}$, *provided that $r \geq$* size $(\mathbf{A})$.

For a proof of this proposition see Antoulas and Anderson [24]. Based on this proof, we can now provide a proof of Lemma 4.48.

Proof. We distinguish two cases. **(a)** $\phi(s)$ is *proper rational.* According to the results in section 4.4.2, there exists a minimal quadruple $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$ of dimension π such that

$$\phi(s) = d + \mathbf{c}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

This expression implies

$$\begin{aligned}\hat{\phi}_i - \phi_j &= \mathbf{c}(\hat{s}_i\mathbf{I} - \mathbf{A})^{-1}[(s_j\mathbf{I} - \mathbf{A}) - (\hat{s}_i\mathbf{I} - \mathbf{A})](s_j\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \\ &= \mathbf{c}(\hat{s}_i\mathbf{I} - \mathbf{A})^{-1}[s_j - \hat{s}_i](s_j\mathbf{I} - \mathbf{A})^{-1}\mathbf{b},\end{aligned}$$

and hence

$$[\boldsymbol{L}]_{i,j} = \frac{\hat{\phi}_i - \phi_j}{\hat{s}_i - s_j} = -\mathbf{c}(\hat{s}_i\mathbf{I} - \mathbf{A})^{-1}(s_j\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

Consequently, $\boldsymbol{L}$ can be factorized as follows:

$$\boldsymbol{L} = -\mathcal{O}_p\mathcal{R}_r, \tag{4.87}$$

where $\mathcal{R}_r$ and $\mathcal{O}_p$ are the generalized reachability, observability matrices defined by (4.85), (4.86), respectively. Because of the proposition given above, the rank of both $\mathcal{O}_p$ and $\mathcal{R}_r$ is π. This implies that the rank of their product $\boldsymbol{L}$ is also π. This completes the proof when $\phi(s)$ is proper.

(b) $\phi(s)$ is *not proper rational.* In this case, by means of a bilinear transformation

$$s \mapsto \frac{\alpha s + \beta}{s + \gamma}, \qquad \alpha\gamma - \beta \neq 0,$$

for almost all α, β, γ, the rational function

$$\tilde{\phi}(s) = \phi\left(\frac{\alpha s + \beta}{s + \gamma}\right)$$

will be proper. The Löwner matrices $\boldsymbol{L}$, $\tilde{\boldsymbol{L}}$ attached to ϕ, $\tilde{\phi}$, respectively, are related as follows:

$$(\alpha\gamma - \beta)\tilde{\boldsymbol{L}} = \operatorname{diag}(\alpha - \hat{s}_i)\,\boldsymbol{L}\operatorname{diag}(\alpha - s_i).$$

The parameter α can be chosen so that $\operatorname{diag}(\alpha - \hat{s}_i)$ and $\operatorname{diag}(\alpha - s_j)$ are nonsingular, which implies the desired result. This concludes the proof of the lemma. □

From Löwner matrix to rational function

Given the array of points $\mathbb{P}$ defined by (4.76), we are now ready to tackle the interpolation problem (4.77), (4.78) and, in particular, solve the two problems (a) and (b) of Problem (A). The following definition is needed first.

Definition 4.51. **(a)** *The rank of the array* $\mathbb{P}$ *is*

$$\operatorname{rank} \mathbb{P} = \max_{L} \{\operatorname{rank} L\} = q,$$

where the maximum is taken over all possible Löwner matrices which can be formed from $\mathbb{P}$.

(b) *We will call a Löwner matrix almost square if it has at most one more row than column or vice versa, with the sum of the number of rows and columns being equal to* N.

The following is a consequence of Lemma 4.48.

Proposition 4.52. *The rank of all Löwner matrices having at least* q *rows and* q *columns is equal to* q. *Consequently, almost square Löwner matrices have rank* q.

Let $q = \operatorname{rank} \mathbb{P}$, and assume that $2q < N$. For any Löwner matrix with $\operatorname{rank} \boldsymbol{L} = q$, there exists a column vector $\gamma \neq \mathbf{0}$ of appropriate dimension, say, $r + 1$, satisfying

$$\boldsymbol{L}\gamma = \mathbf{0} \text{ or } \gamma^* \boldsymbol{L} = \mathbf{0}. \tag{4.88}$$

In this case, we can attach to $\boldsymbol{L}$ a rational function denoted by

$$\phi_L(s) = \frac{\mathbf{n}_L(s)}{\mathbf{d}_L(s)} \tag{4.89}$$

using formula (4.83), i.e.,

$$\mathbf{n}_L(s) = \sum_{j=1}^{r+1} \gamma_j \phi_j \prod_{i \neq j} (s - s_i), \quad \mathbf{d}_L(s) = \sum_{j=1}^{r+1} \gamma_j \prod_{i \neq j} (s - s_i). \tag{4.90}$$

The rational function $\phi_L(s)$ just defined has the following properties.

Lemma 4.53. **(a)** $\deg \phi_L \leq r \leq q < N$. **(b)** *There is a unique* ϕ_L *attached to all* $\boldsymbol{L}$ *and* γ *satisfying* (4.88) *as long as* $\operatorname{rank} \boldsymbol{L} = q$. **(c)** *The numerator, denominator polynomials* $\mathbf{n}_L$, $\mathbf{d}_L$ *have* $q - \deg \phi_L$ *common factors of the form* $(s - s_i)$. **(d)** ϕ_L *interpolates exactly* $N - q + \deg \phi_L$ *points of the array* $\mathbb{P}$.

The proof of this result can be found in Antoulas and Anderson [25]. As a consequence of the above lemma and Lemma 4.48, we obtain the next corollary.

Corollary 4.54. ϕ_L *interpolates all given points if and only if* $\deg \phi_L = q$ *if and only if all* $q \times q$ *Löwner matrices which can be formed from the data array* $\mathbb{P}$ *are nonsingular.*

We are now ready to state, from Antoulas and Anderson [25], the main result.

Theorem 4.55. *Given the array of N points* $\mathbb{P}$, *let* $\operatorname{rank} \mathbb{P} = q$.

(a) *If* $2q < N$, *and all square Löwner matrices of size* q *which can be formed from* $\mathbb{P}$ *are nonsingular, there is a unique interpolating function of minimal degree denoted by* $\phi^{min}(s)$, *and* $\deg \phi^{min} = q$.

(b) *Otherwise,* $\phi^{min}(s)$ *is not unique and* $\deg \phi^{min} = N - q$.

The first part of the theorem follows from the previous corollary. The second part can be justified as follows. Part (b) of the proposition above says that as long as $\boldsymbol{L}$ has rank q, there is a unique rational function ϕ_L attached to it. Consequently, for $\boldsymbol{L}$ to yield a different rational function ϕ_L defined by (4.89), (4.90), it will have to *lose* rank. This occurs when $\boldsymbol{L}$ has at most $q - 1$ rows. In this case, its rank is $q - 1$ and there exists a column vector γ such that $\boldsymbol{L}\gamma = \mathbf{0}$. Since $\boldsymbol{L}$ has $N - q + 1$ columns, the degree of the attached ϕ_L will generically (i.e., for almost all γ) be $N - q$. It readily follows that for almost all γ, ϕ_L will interpolate all the points of the array $\mathbb{P}$. This argument shows that there can *never* exist interpolating functions of degree between q and $N - q$. The admissible degree problem can now be solved in terms of the rank of the array $\mathbb{P}$.

Corollary 4.56. *Under the assumptions of the main theorem, if* $\deg \phi^{min} = q$*, the admissible degrees are q, and all integers greater than or equal to $N - q$, while if* $\deg \phi^{min} = N - q$*, the admissible degrees are all integers greater than or equal to $N - q$.*

Remark 4.5.2. **(i)** If $2q = N$, the only solution γ of (4.88) is $\gamma = \mathbf{0}$. Hence, ϕ_L, defined by (4.89), (4.90) does not exist, and part (b) of the above theorem applies.

(ii) To distinguish between case (a) and case (b) of this theorem, we need only to check the nonsingularity of $2q + 1$ Löwner matrices. Construct from $\mathbb{P}$ *any* Löwner matrix of size $q \times (q + 1)$, with row, column sets denoted by $\mathbb{I}_q$, $\mathbb{J}_q$, and call it $\boldsymbol{L}_q$. The Löwner matrix $\boldsymbol{L}_q^*$ of size $(q + 1) \times q$ is now constructed. Its row set $\mathbb{I}_q$ contains the points of the row set $\mathbb{I}_q$ together with the last point of the column set $\mathbb{J}_q$; moreover, its column set $\mathbb{J}_q^*$ contains the points of the column set $\mathbb{J}_q$ with the exception of the last one. The $2q + 1$ Löwner matrices which need to be checked are the $q \times q$ submatrices of $\boldsymbol{L}_q$ and $\boldsymbol{L}_q^*$.

The construction of interpolating functions

Given an admissible degree, we will discuss in this section the construction of all corresponding interpolating functions. Two construction methods will be presented: the first is based on an external (input-output) framework, while the second is based on a state-space framework.

Given the array $\mathbb{P}$, let π be an admissible degree. For the polynomial construction we need to form from $\mathbb{P}$ *any* Löwner matrix having $\pi + 1$ columns,

$$\boldsymbol{L} \in \mathbb{R}^{(N-\pi-1)\times(\pi+1)},$$

and determine a parametrization of all γ such that

$$\boldsymbol{L}\gamma = \mathbf{0}.$$

A parametrization of all interpolating functions of degree π is then

$$\phi_L(s) = \frac{\mathbf{n}_L(s)}{\mathbf{d}_L(s)},$$

where the numerator and denominator polynomials are defined by (4.90). If $\pi \geq N - q$, we have to make sure that there are no common factors between the numerator and the denominator of ϕ_L; this is the case for almost all γ. More precisely, the $2\pi + 1 - N$ (scalar) parameters which parametrize all γ have to avoid the hypersurfaces defined by the equations

$$\mathbf{d}_L(s_i) = 0, \qquad i = 1, \ldots, N.$$

Since we can always make sure that γ depends affinely on these parameters, we are actually dealing with hyperplanes. For details and examples, see Antoulas and Anderson [25].

For use below, notice that ϕ_L will be proper rational if and only if the leading coefficient of $\mathbf{d}_L$ is different from zero; i.e., from the second formula (4.90), we must have

$$\gamma_{\pi_1} + \cdots + \gamma_{\pi_{\pi+1}} \neq 0.$$

For the *state-space* construction of interpolating functions of admissible degree π, we need a Löwner matrix of size $\pi \times (\pi + 1)$:

$$\bar{L} \in \mathbb{R}^{\pi \times (\pi+1)}.$$

Thus, in case $\pi \geq N - q$, we need an array $\bar{\mathbb{P}}$ which contains besides the original N points of the array $\mathbb{P}$, another $2\pi + 1 - N$ points, chosen arbitrarily but subject to the nonsingularity condition given in part (a) of the main theorem (see also the remark at the end of the previous section). Let $\bar{\gamma} \in \mathbb{R}^{\pi+1}$ be such that

$$\bar{L}\bar{\gamma} = \mathbf{0}.$$

If $\bar{\gamma}_{\pi_1} + \cdots + \bar{\gamma}_{\pi_{\pi+1}} \neq 0$, the underlying interpolating function is proper. Otherwise, we need to perform a bilinear transformation which will ensure the properness of the function under construction. (See the proof of Lemma 4.48.) Once the properness condition is guaranteed, the state-space construction proceeds by defining the following two $\pi \times \pi$ matrices:

$$\mathbf{Q} = \bar{L} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \\ -1 & -1 & \cdots & -1 \end{bmatrix}, \quad \sigma\mathbf{Q} = \bar{L} \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s \\ -s_{\pi+1} & -s_{\pi+1} & \cdots & -s_{\pi+1} \end{bmatrix} \in \mathbb{R}^{(\pi+1)\times\pi},$$

where s_i, $i = 1, \ldots, \pi + 1$, are the points which define the column array of $\bar{L}$. Let the quadruple of constant matrices $(\mathbf{A}, \mathbf{b}, \mathbf{c}, d)$ be defined as follows:

$$\left.\begin{array}{rcl} \mathbf{A} & = & (\sigma\mathbf{Q})\mathbf{Q}^{-1} \\ \mathbf{b} & = & (s_1\mathbf{I} - \mathbf{A})[\bar{L}]_{(:,1)} \\ \mathbf{c} & = & [(s_1\mathbf{I} - \mathbf{A})]_{(1,:)} \\ d & = & y_i - \mathbf{c}(s_i\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \end{array}\right\} \tag{4.91}$$

for any s_i, where $[\mathbf{M}]_{(:,1)}$ denotes the first column of $\mathbf{M}$, while $[\mathbf{M}]_{(1,:)}$ denotes the first row. It can be shown that the above quadruple is a minimal realization of the desired interpolating function $\phi(s)$ of degree π:

$$\phi(s) = d + \mathbf{c}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}. \tag{4.92}$$

The steps involved in proving the above result are as follows. First, because of the properness of the underlying function, the matrix $\mathbf{Q}$ is nonsingular. Next, we need to show that none of the s_i's are an eigenvalue of $\mathbf{A}$, that is, $(s_i\mathbf{I} - \mathbf{A})$ is invertible. Finally, we need to show that the rational function given by (4.92) is indeed an interpolating function of the prescribed degree π. These steps can be found in Anderson and Antoulas [24].

Remark 4.5.3. **(i)** In the realization problem the *shift* is defined in terms of the associated Hankel matrix, as the operation that assigns to the ith column the $(i+1)$st column. It follows that $\mathbf{A}$ is determined by this shift. For the more general interpolation problem, formula (4.91) shows that

$$\mathbf{AQ} = \sigma\mathbf{Q}.$$

If we define the shift operation in this case as assigning to the ith column of the Löwner matrix, s_i times itself, then $\sigma\mathbf{Q}$ is indeed the shifted version of $\mathbf{Q}$, and, consequently, $\mathbf{A}$ is again determined by the shift.

(ii) The theory, presented above, has been worked out for the multiple-point as well as for more general interpolation problems; see [25] and [24].

(iii) It is readily checked that the classical system theoretic problem of realization can be interpreted as a rational interpolation problem where all the data are provided at a single point. Our theory has generalized the theory of *realization* to the theory of *interpolation*.

All missing proofs, as well as other details and examples, can be found in [25] and [24]. Some of the results discussed can also be found in [50].

4.5.3 Generating system approach to rational interpolation*

This method for dealing with the rational interpolation problem is based on the factorization of a rational matrix expressed in terms of the data (4.76). It leads to a parametrization of all interpolants that solve Problems (A), (B), and (C) in section 4.5.1. Through the years, solutions to various special cases of the general rational interpolation problem have been worked out in what amounts to a generating system approach. For example, more than three-quarters of a century ago, Problem (B) was solved using this approach. Actually, the generating system was constructed recursively.

This section will make use of certain elementary concepts and results concerning polynomial and rational matrices, for example, invariant factors of polynomial matrices, left coprime polynomial matrices, row-reduced polynomial matrices, unimodular polynomial matrices, and Bezout equations. See section 6.3 of the book by Kailath [191] for an exposition of these concepts and the underlying theory.

To keep the exposition simple, only the distinct point and the single multiple point interpolation problems will be considered. The tools presented, however, are applicable in the general case.

The data in terms of time series

The interpolation array $\mathbb{P}$ defined by (4.76) can be interpreted in terms of time functions. To the distinct point array $\mathbb{P}$ defined by (4.79) we associate the following exponential time series (i.e., functions of time):

$$\mathbb{D} = \left\{ \mathbf{w}_k(t) = \begin{pmatrix} \mathbf{u}_k(t) \\ -\mathbf{y}_k(t) \end{pmatrix},\ \mathbf{u}_k(t) = e^{s_k t},\ \mathbf{y}_k(t) = \phi_k e^{s_k t},\ t \geq 0,\ k = 1, \dots, N \right\}. \tag{4.93}$$

We will also consider the (unilateral) Laplace transform of these time series; in particular, we will consider the $2 \times N$ matrix of rational entries whose kth column is the transform of $\mathbf{w}_k(t)$:

$$\mathbf{W}(s) = [\mathbf{W}_1(s) \ \cdots \ \mathbf{W}_N(s)], \ \text{ where } \mathbf{W}_k(s) = \frac{1}{s - s_k} \begin{pmatrix} 1 \\ -\phi_k \end{pmatrix}, \qquad k = 1, \dots, N. \tag{4.94}$$

It is easy to see that $\mathbf{W}(s)$ has a realization,

$$\mathbf{W}(s) = \mathbf{C}(s\mathbf{I}_N - \mathbf{A})^{-1}, \tag{4.95}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ -\phi_1 & -\phi_2 & \cdots & -\phi_N \end{bmatrix} \in \mathbb{C}^{2\times N}, \ \mathbf{A} = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_N \end{bmatrix} \in \mathbb{C}^{N\times N}. \tag{4.96}$$

Since the interpolation points are distinct, i.e., $s_i \neq s_j, i \neq j$, the pair $(\mathbf{C}, \mathbf{A})$ is observable.

For the single multiple-point interpolation array $\mathbb{P}$ defined by (4.80), a similar construction holds. Let $\mathbf{p}_k$ be the vector-valued polynomial function

$$\mathbf{p}_k(t) = \begin{pmatrix} 1 \\ -\phi_0 \end{pmatrix} \frac{t^{k-1}}{(k-1)!} + \cdots + \begin{pmatrix} 0 \\ -\frac{\phi_{0j}}{j!} \end{pmatrix} \frac{t^{k-j-1}}{(k-j-1)!} + \cdots + \begin{pmatrix} 0 \\ -\frac{\phi_{0,k-1}}{(k-1)!} \end{pmatrix},\ k = 1, \dots, N.$$

The time series in this case are polynomial-exponential:

$$\mathbb{D} = \left\{ \mathbf{w}_k(t) = \begin{pmatrix} \mathbf{u}_k(t) \\ -\mathbf{y}_k(t) \end{pmatrix} = e^{s_0 t}\, \mathbf{p}_k(t),\ t \geq 0,\ k = 1, \dots, N \right\}. \tag{4.97}$$

A straightforward calculation yields the following realization for the (unilateral) Laplace transform of this set of time series $\mathbf{W}(s) = [\mathbf{W}_1(s) \cdots \mathbf{W}_N(s)] = \mathbf{C}(s\mathbf{I}_N - \mathbf{A})^{-1}$, where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_0 & -\frac{\phi_{01}}{1!} & -\frac{\phi_{02}}{2!} & \cdots & -\frac{\phi_{0,N-1}}{(N-1)!} \end{bmatrix} \in \mathbb{C}^{2\times N},$$

$$\mathbf{A} = \begin{bmatrix} s_0 & 1 & & & \\ & s_0 & 1 & & \\ & & \ddots & \ddots & \\ & & & s_0 & 1 \\ & & & & s_0 \end{bmatrix} \in \mathbb{C}^{N\times N}.$$

Again by construction, the pair $(\mathbf{C}, \mathbf{A})$ is observable.

The *realization problem* discussed in section 4.4 can be expressed in terms of the rational interpolation problem. Given the scalar Markov parameters $h_0, h_1, \dots, h_{N-1}$, we seek to determine all rational functions $\phi(s)$ whose behavior at infinity (i.e., formal power series) is

$$\phi(s) = h_0 + h_1 s^{-1} + \cdots + h_{N-1} s^{-N+1} + \cdots.$$

By introducing s^{-1} as the new variable, the behavior at infinity is transformed into the behavior at zero and the Markov parameters become *moments*:

$$\tilde{\phi}(s) = \phi(s^{-1}) = h_0 + h_1 s + \cdots + h_{N-1} s^{N-1} + \cdots.$$

Consequently, $h_k = \frac{1}{k!}\frac{d^k\tilde{\phi}}{dt^k}\big|_{s=0}$, i.e., the realization problem is equivalent to a rational interpolation problem where all the data are provided at zero. From the above considerations, the corresponding time series $\mathbf{W} = [\mathbf{W}_1 \cdots \mathbf{W}_N]$ can be expressed in terms of (4.95), where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -h_0 & -h_1 & -h_2 & \cdots & -h_{N-1} \end{bmatrix} \in \mathbb{R}^{2\times N}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{N\times N}. \tag{4.98}$$

Interpolation in terms of the time series data

Given the polynomial $\mathbf{r}(s) = \sum_{i=0}^{k} r_i s^i$, $r_i \in \mathbb{R}$, we will denote by $\mathbf{r}(\frac{d}{dt})$ the constant coefficient differential operator

$$\mathbf{r}\left(\frac{d}{dt}\right) = r_0 + r_1\frac{d}{dt} + r_2\frac{d^2}{dt^2} + \cdots + r_k\frac{d^k}{dt^k}.$$

The following is a characterization of rational interpolants in terms of the time series in both the time and the frequency domains.

Proposition 4.57. *With the notation introduced above, consider the rational function* $\phi(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}$*, where* $\mathbf{n}$*,* $\mathbf{d}$ *are coprime. This rational function interpolates the points of the array* $\mathbb{P}$ *defined by* (4.79) *if and only if one of the following equivalent conditions holds:*

$$\left[\mathbf{n}\left(\frac{d}{dt}\right) \quad \mathbf{d}\left(\frac{d}{dt}\right)\right] \mathbf{w}_k(t) = 0, \qquad t \geq 0, \tag{4.99}$$

$$[\mathbf{n}(s) \quad \mathbf{d}(s)]\mathbf{W}_k(s) = \mathbf{r}_k(s) \tag{4.100}$$

for $k = 1, \dots, N$*, where* $\mathbf{r}_k(s)$ *is a polynomial.*

Equation (4.99) provides a *time domain characterization*, while (4.100) provides a *frequency domain characterization* of rational interpolants.

Proof. We will give the proof only for the distinct point interpolation problem. From (4.99), given the definition of the time series $\mathbf{w}_k$, follows $(\mathbf{n}(s_k) - \phi_k \mathbf{d}(s_k))\, e^{s_k t} = 0$; since this must hold for all $t \geq 0$, the necessity and sufficiency of the interpolation conditions $\frac{\mathbf{n}(s_k)}{\mathbf{d}(s_k)} = \phi_k$ follow for all k. If we take the unilateral Laplace transform of (4.99), we obtain (4.100), and in particular $\frac{\mathbf{n}(s)-\phi_k \mathbf{d}(s)}{s-s_k} = \mathbf{r}_k(s)$, where $\mathbf{r}_k(s)$ is a polynomial resulting from initial conditions of $\mathbf{y}_k$ at $t = 0^-$. Thus the expression on the left-hand side is a polynomial if and only if $\phi(s_k) = \phi_k$ for all k. □

The solution of Problem (A)

From the frequency domain representation (4.95) of the data, we construct the pair of polynomial matrices $\Xi(s)$, $\Theta(s)$ of size $2 \times N$, 2×2, respectively, such that $\det \Theta(s) \neq 0$, and

$$\mathbf{W}(s) = \Theta(s)^{-1}\Xi(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}. \qquad (4.101)$$

The above operation consists of computing a *left* polynomial denominator $\Theta(s)$ for the data $\mathbf{W}(s)$. Moreover, the polynomial matrices $\Theta(s)$, $\Xi(s)$ must be *left coprime*, that is, every nonsingular polynomial matrix $\mathbf{L}(s)$ such that $\mathbf{P} = \mathbf{L}\tilde{\mathbf{P}}$ and $\mathbf{Q} = \mathbf{L}\tilde{\mathbf{Q}}$, for appropriate polynomial matrices $\tilde{\mathbf{P}}$, $\tilde{\mathbf{Q}}$, is unimodular, that is, its determinant is a nonzero constant. A consequence of left coprimeness is the existence of polynomial matrices $\mathbf{P}(s)$, $\mathbf{Q}(s)$ of size 2×2, $N \times 2$, respectively, such that the so-called *Bezout equation* is satisfied:

$$\Theta(s)\mathbf{P}(s) + \Xi(s)\mathbf{Q}(s) = \mathbf{I}_2.$$

A Θ constructed this way has the following properties.

Proposition 4.58. *The matrix $\Theta(s)$ constructed above satisfies the following:* **(a)** *its invariant factors are* 1 *and* $\chi(s) = \Pi_i(s - s_i)$; **(b)** *its* $(1, 2)$, $(2, 2)$ *entries* $\theta_{12}(s)$, $\theta_{22}(s)$ *are coprime.*

Proof. Because of the observability of the pair $\mathbf{C}, \mathbf{A}$ and the coprimeness of Ξ, Θ, the polynomial matrices $s\mathbf{I} - \mathbf{A}$ and $\Theta(s)$ have a *single* nonunity invariant factor that is the *same*, namely, the characteristic polynomial of $\mathbf{A}$: $\chi(s) = \Pi_{i=1}^N(s - s_i)$. (See, e.g., [8] or Chapter 6 of [191].) Therefore, after a possible normalization by a (nonzero) constant,

$$\det \Theta(s) = \chi(s). \qquad (4.102)$$

Let θ_{ij} denote the (i, j)th entry of Θ. The ith column of (4.101) yields the equation

$$(s - s_i)[\Xi(s)]_{(:,i)} = \begin{pmatrix} \theta_{11}(s) & \theta_{12}(s) \\ \theta_{21}(s) & \theta_{22}(s) \end{pmatrix} \begin{pmatrix} 1 \\ -\phi_i \end{pmatrix} = \begin{pmatrix} \theta_{11}(s) - \phi_i\theta_{12}(s) \\ \theta_{21}(s) - \phi_i\theta_{22}(s) \end{pmatrix},$$
$$i = 1, \ldots, N.$$

Evaluating this expression at $s = s_i$, we obtain $\theta_{11}(s_i) = \phi_i\theta_{12}(s_i)$, $\theta_{21}(s_i) = \phi_i\theta_{22}(s_i)$, $i = 1, \ldots, N$. Because of (4.102), if θ_{12}, θ_{22} were not coprime, their greatest common divisor would have to be a product of terms $(s - s_i)$, where the s_i are the interpolation points. Therefore, by the latter equation, all four entries of Θ would have the same common factor. This, however, contradicts the fact that one of the two invariant factors of Θ is equal to 1. The desired coprimeness is thus established. □

Lemma 4.59. *The rational function* $\phi(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}$, *with* $\mathbf{n}, \mathbf{d}$ *coprime, is an interpolant for the array* $\mathbb{P}$ *if and only if there exist coprime polynomials* $\mathbf{a}(s), \mathbf{b}(s)$ *such that*

$$[\mathbf{n}(s) \quad \mathbf{d}(s)] = [\mathbf{a}(s) \quad \mathbf{b}(s)]\,\Theta(s) \tag{4.103}$$

and

$$\mathbf{a}(s_i)\theta_{12}(s_i) + \mathbf{b}(s_i)\theta_{22}(s_i) \neq 0, \qquad i = 1, \dots, N. \tag{4.104}$$

Proof. If the numerator and the denominator of ϕ satisfy (4.103), there holds

$$[\mathbf{n}(s) \quad \mathbf{d}(s)]\mathbf{W}(s) = [\mathbf{a}(s) \quad \mathbf{b}(s)]\Xi(s).$$

The latter expression is polynomial, and hence by Proposition 4.57 ϕ is an interpolant. Also, $\mathbf{a}$, $\mathbf{b}$ are coprime; otherwise $\mathbf{n}$, $\mathbf{d}$ would not be coprime, which is a contradiction. Finally, we notice that $\mathbf{d} = \mathbf{a}\theta_{12} + \mathbf{b}\theta_{22}$, which implies that conditions (4.104) must be satisfied. This is possible due to part (b) of the proposition above.

Conversely, let ϕ be an interpolant. According to Proposition 4.57, $[\mathbf{n}(s) \quad \mathbf{d}(s)]\mathbf{W}(s)$ is a polynomial row vector. From the Bezout equation, it follows that $\mathbf{P}(s) + \mathbf{W}(s)\mathbf{Q}(s) = [\Theta(s)]^{-1}$. Multiplying this relationship on the left by the row vector $[\mathbf{n}(s) \quad \mathbf{d}(s)]$, we conclude that $[\mathbf{n}(s) \quad \mathbf{d}(s)][\Theta(s)]^{-1}$ must be a polynomial row vector, i.e., there exist polynomials $\mathbf{a}$, $\mathbf{b}$ such that (4.103) holds. Furthermore, the coprimeness of $\mathbf{n}$, $\mathbf{d}$ implies the coprimeness of $\mathbf{a}$, $\mathbf{b}$. □

As shown above, *all* interpolants $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ can be parametrized by means of (4.103), where the parameter $\Gamma = \frac{\mathbf{a}}{\mathbf{b}}$ is an arbitrary rational function subject to constraints (4.104). We can interpret Θ as a two-port with inputs $\mathbf{u}$, $\hat{\mathbf{u}}$, and outputs $\mathbf{y}$, $\hat{\mathbf{y}}$ (see Figure 4.4):

$$\begin{pmatrix} \hat{\mathbf{y}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{y} \end{pmatrix}$$

Γ can be seen as relating $\hat{\mathbf{u}}$ and $\hat{\mathbf{y}}$ in the following manner: $\mathbf{b}(\frac{d}{dt})\hat{\mathbf{u}} = \mathbf{a}(\frac{d}{dt})\hat{\mathbf{y}}$. Then the parametrization of *all* solutions ϕ can be interpreted as a linear system described by the linear, constant coefficient, differential equation $\mathbf{d}(\frac{d}{dt})\mathbf{y} = \mathbf{n}(\frac{d}{dt})\mathbf{u}$. This system, in turn, can be represented by means of a feedback interconnection between Θ and Γ, where Θ is

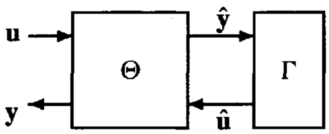


Figure 4.4. *Feedback interpretation of the parametrization of all solutions of the rational interpolation problem.*

fixed and Γ is arbitrary, subject to (4.104). As a consequence of the above interpretation, Θ is called the generating system or generating matrix of the *rational interpolation problem* at hand. Furthermore, (4.103) shows that every interpolant can be expressed in terms of the linear fractional representation,

$$\phi(s) = \frac{\mathbf{a}(s)\theta_{11}(s) + \mathbf{b}(s)\theta_{21}(s)}{\mathbf{a}(s)\theta_{12}(s) + \mathbf{b}(s)\theta_{22}(s)}.$$

Remark 4.5.4. The coprimeness constraint on $\mathbf{n}$, $\mathbf{d}$ has been expressed equivalently as a coprimeness constraint on the parameter polynomials $\mathbf{a}$, $\mathbf{b}$ together with constraints (4.104). The former is a nonlinear constraint in the space of coefficients of $\mathbf{a}$, $\mathbf{b}$; it is automatically satisfied in the case of minimal interpolants discussed below. Constraints (4.104) are linear in the coefficients of $\mathbf{a}$, $\mathbf{b}$. Examples will be worked out later.

To tackle Problem (A), the concept of a *row reduced* generating matrix is needed. Let ν_i be the degree of the ith row of Θ. The row-wise highest coefficient matrix $[\Theta]_{\mathrm{hr}}$ is a 2×2 constant matrix, whose (i, j)th entry is the coefficient of the term s^{ν_i} of the polynomial $\theta_{i,j}(s)$. We will call Θ *row reduced* if $[\Theta]_{\mathrm{hr}}$ is nonsingular. Notice that the row degrees of Θ and the degree of its determinant, which by (4.102) is N, satisfy $\nu_1 + \nu_2 \geq N$, in general. An equivalent characterization of row reducedness is that the row degrees of Θ satisfy $\nu_1 + \nu_2 = N$.

The matrix Θ in (4.101) is unique, up to left multiplication with a unimodular matrix (which is a polynomial matrix with constant nonzero determinant). We use this freedom to transform Θ into *row reduced* form. For simplicity we will use the same symbol, namely, Θ, to denote the row reduced version of this matrix. Let the corresponding row degrees be

$$\kappa_1 = \deg\ (\theta_{11}(s) \quad \theta_{12}(s)) \leq \kappa_2 = \deg\ (\theta_{21}(s) \quad \theta_{22}(s)), \qquad \kappa_1 + \kappa_2 = N.$$

The row degrees of row reduced polynomial matrices have two important properties. First, although the row reduced version of any polynomial matrix is nonunique, the corresponding row degrees are *unique*. Second, because of the so-called predictable-degree property of row reduced polynomial matrices (see, e.g., Chapter 6 of [191]), the degree of $\mathbf{r}(s)\Theta(s)$, with Θ row reduced and $\mathbf{r}$ some polynomial row vector with coprime entries, either can be κ_1 or be greater than or equal to κ_2.

Construction of a row reduced Θ using $\mathcal{O}(\mathbf{C}, \mathbf{A})$. We will now show how a row reduced generating matrix can be constructed directly from the observability matrix $\mathcal{O}(\mathbf{C}, \mathbf{A})$. The procedure involves the determination of two linear dependencies among the rows of this observability matrix, which leads to the two *observability indices* κ_i, $i = 1, 2$, of the pair $(\mathbf{C}, \mathbf{A})$. (For details, see [8] or [191].)

Let $\mathbf{c}_i$ denote the ith row of $\mathbf{C}$, $i = 1, 2$. For simplicity, we will assume that working from top to bottom of the observability matrix, $\mathbf{c}_2\mathbf{A}^{\kappa_1}$ is the first row of $\mathcal{O}$ that is linearly dependent on the preceding ones, i.e., $\mathbf{c}_i\mathbf{A}^j$, $i = 1, 2$, $j \leq \kappa_1$. Then, because of observability, the next row to be linearly dependent on the previous ones will be $\mathbf{c}_1\mathbf{A}^{\kappa_2}$, where $\kappa_1 < \kappa_2$, and $\kappa_1 + \kappa_2 = N$:

$$\begin{aligned}
\mathbf{c}_2\mathbf{A}^{\kappa_1} &= \textstyle\sum_{i=0}^{\kappa_1} \alpha_i \mathbf{c}_1 \mathbf{A}^i &+& \textstyle\sum_{j=0}^{\kappa_1 - 1} \beta_j \mathbf{c}_2 \mathbf{A}^j, \\
\mathbf{c}_1\mathbf{A}^{\kappa_2} &= \textstyle\sum_{i=0}^{\kappa_2 - 1} \gamma_i \mathbf{c}_1 \mathbf{A}^i &+& \textstyle\sum_{j=0}^{\kappa_1} \delta_j \mathbf{c}_2 \mathbf{A}^j.
\end{aligned}$$

It follows that Θ can be read off of the above relationships:

$$\Theta(s) = \begin{pmatrix} -\sum_{i=0}^{\kappa_1}\alpha_i s^i & s^{\kappa_1} - \sum_{i=0}^{\kappa_1-1}\beta_j s^j \\ s^{\kappa_2} - \sum_{i=0}^{\kappa_2-1}\gamma_i s^i & -\sum_{i=0}^{\kappa_1}\delta_j s^j \end{pmatrix}. \tag{4.105}$$

Clearly, $\det\,[\Theta]_{\mathrm{hr}} = -1$, which implies that Θ is row reduced.

Combining the preceding lemma with the above considerations, we obtain the main result, which provides the solution of Problem (A). This result was proved in Antoulas and Willems [15] as well as Antoulas et al. [22].

Theorem 4.60. *Consider* Θ *defined by* (4.101), *which is row reduced, with row degrees* $\kappa_1 \le \kappa_2$.

(i) *If* $\kappa_1 < \kappa_2$ *and* θ_{11}, θ_{21} *are coprime,*

$$\phi^{min}(s) = \frac{\theta_{11}(s)}{\theta_{12}(s)}, \qquad \delta(\phi^{min}) = \kappa_1,$$

is the unique minimal interpolant. Furthermore, there are no interpolants of complexity between κ_1 *and* κ_2.

(ii) *Otherwise, there is a family of interpolating functions of minimal complexity which can be parametrized as follows:*

$$\phi^{min}(s) = \frac{\theta_{21}(s) + \mathbf{a}(s)\theta_{11}(s)}{\theta_{22}(s) + \mathbf{a}(s)\theta_{12}(s)}, \qquad \delta(\phi^{min}) = \kappa_2 = N - \kappa_1,$$

where the polynomial $\mathbf{a}(s)$ *satisfies*

$$\deg \mathbf{a} = \kappa_2 - \kappa_1, \quad \theta_{22}(s_i) + \mathbf{a}(s_i)\theta_{12}(s_i) \neq 0, \qquad i = 1, \dots, N.$$

(iii) *In both cases* (i) *and* (ii), *there are families of interpolants* $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ *of every degree* $\kappa \ge \kappa_2$, *satisfying* (4.103), *where* $\deg \mathbf{a} = \kappa - \kappa_1$, $\deg \mathbf{b} = \kappa - \kappa_2$, *and* $\mathbf{a}$, $\mathbf{b}$ *are coprime.*

Corollary 4.61. Proper rational and polynomial interpolants. *The interpolants above are proper rational provided that* $\mathbf{a}$, $\mathbf{b}$ *satisfy* $[\mathbf{b}(s)\theta_{22}(s) + \mathbf{a}(s)\theta_{12}(s)]_h \neq 0$, *where* $[\mathbf{r}]_h$ *is used to denote the coefficient of the highest power of the polynomial* $\mathbf{r}$. *All polynomial interpolants* $\phi(s) = \mathbf{n}(s)$ *are given by* $-\mathbf{n}(s) = \mathbf{b}(s)\theta_{21}(s) + \mathbf{a}(s)\theta_{11}(s)$, *where the polynomials* $\mathbf{a}$, $\mathbf{b}$ *satisfy the Bezout equation* $\mathbf{b}(s)\theta_{22}(s) + \mathbf{a}(\mathbf{s})\theta_{12}(s) = 1$.

One polynomial interpolant is the *Lagrange interpolating polynomial* $\ell(s)$ given in the distinct point case by (4.81). It is also worth mentioning that a generating matrix (which is *not* row reduced) can be written in terms of $\ell(s)$ and $\chi(s)$ given by (4.102):

$$\Theta(s) = \begin{pmatrix} \chi(s) & 0 \\ \ell(s) & 1 \end{pmatrix}.$$

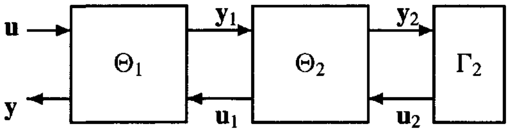


Figure 4.5. *Feedback interpretation of recursive rational interpolation.*

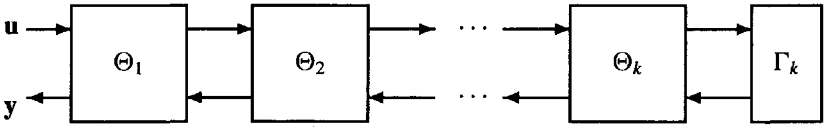


Figure 4.6. *A general cascade decomposition of systems.*

The solution of the unconstrained interpolation problem (4.82) can be obtained in this framework by means of polynomial linear combinations of the rows of this particular generating matrix.

Recursive interpolation

The fact that in the parametrization of all interpolants via the generating system, Γ is arbitrary—except for the avoidance conditions (4.104)—yields, at least conceptually, the solution to the recursive interpolation problem with no additional effort. In particular, if $\mathbb{D} = \mathbb{D}_1 \cup \mathbb{D}_2$, we define Θ_1 as a generating system for the data set $\mathbb{D}_1$ and Θ_2 as a generating system for the modified data set $\Theta_1(\frac{d}{dt})\mathbb{D}_2$. Then the cascade $\Theta = \Theta_2\Theta_1$ of the two generating systems Θ_1 and Θ_2 provides a generating system for $\mathbb{D}$. More generally, Figures 4.5 and 4.6 give a pictorial representation of the solution to the recursive interpolation problem. The *cascade interconnection* is associated with the Euclidean algorithm, the Schur algorithm, the Nevanlinna algorithm, the Berlekamp–Massey algorithm, Darlington synthesis, continued fractions, and Hankel-norm approximation.

Problem (A) was solved recursively in [12]. Earlier, the recursive realization problem with a degree constraint was solved in [9]. The main result of these two papers, which is shown in Figures 4.6 and 4.7, is that a recursive update of the solution of the realization or interpolation problems corresponds to attaching an appropriately defined component to a cascade interconnection of systems.

The scalar case. We conclude this account on recursiveness by making this cascade interconnection explicit for SISO systems. Let the transfer function be $\mathbf{H}(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)}$, assumed for simplicity, strictly proper. As shown by Kalman [193], recursive realization corresponds to the decomposition of $\mathbf{H}(s)$ is a *continued fraction*:

$$\mathbf{H}(s) = \cfrac{1}{\mathbf{d}_1(s) + \cfrac{1}{\mathbf{d}_2(s) + \cfrac{1}{\mathbf{d}_3(s) + \cfrac{1}{\ddots + \cfrac{1}{\mathbf{d}_N(s)}}}}}, \tag{4.106}$$

where $\mathbf{d}_i$ are polynomials of degree κ_i and $\sum_i \kappa_i = \deg \mathbf{d} = n$. In the generating system framework, we have

$$\Theta = \prod_{i=1}^{N} \Theta_i, \quad \text{where } \Theta_i = \begin{pmatrix} 0 & 1 \\ 1 & -\mathbf{d}_i \end{pmatrix}.$$

This cascade decomposition can be simplified in this case, as shown in Figure 4.7.

Furthermore, a state-space realization follows from this decomposition. For details, see [146] and [10]. Here we will illustrate the generic case only, that is, the case where $\kappa_i = 1$ for all i, and hence $N = n$. By appropriate scaling, we will assume that the polynomials are monic (coefficient of highest degree is one); (4.106) becomes

$$\mathbf{H}(s) = \cfrac{\beta_1}{s + \alpha_1 + \cfrac{\beta_2}{s + \alpha_2 + \cfrac{\beta_3}{\ddots + \cfrac{\beta_n}{s + \alpha_n}}}}. \tag{4.107}$$

The following triple is a minimal realization of $\mathbf{H}(s)$, whenever it has a generic decomposition of the form (4.107):

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccccc|c} -\alpha_1 & \beta_2 & 0 & \cdots & 0 & 0 & 1 \\ -1 & -\alpha_2 & \beta_3 & \cdots & 0 & 0 & 0 \\ \vdots & & \ddots & \ddots & & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & -\alpha_{n-1} & \beta_n & 0 \\ 0 & 0 & \cdots & 0 & -1 & -\alpha_n & 0 \\ \hline \beta_1 & 0 & \cdots & 0 & 0 & 0 & \end{array}\right). \tag{4.108}$$

Notice that $\mathbf{A}$ is in *tridiagonal* form, while $\mathbf{B}$ and $\mathbf{C}^*$ are multiples of the first canonical unit vector in $\mathbb{R}^n$. To summarize, we have seen important connections between the following topics:

(a) realization/interpolation,

(b) cascade/feedback interconnection,

(c) linear fractions,

(d) continued fractions,

(e) tridiagonal state space realizations.

Thus partial realization consists of truncating the tail of the continued fraction or, equivalently, of the cascade decomposition of the system or of the tridiagonal state space realization. These issues will play a role in Chapter 10, where an iterative method, the so-called Lanczos procedure, of constructing this tridiagonal realization will be of central importance.

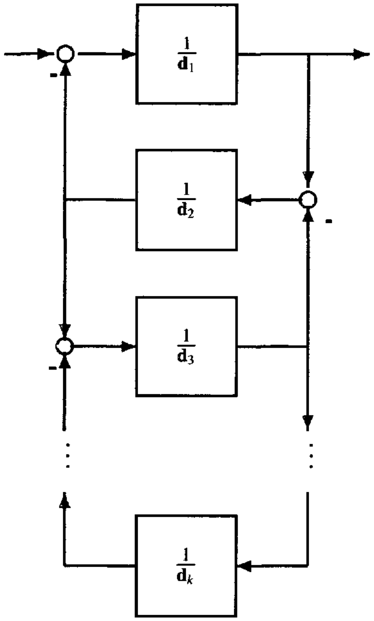


Figure 4.7. *Cascade (feedback) decomposition of scalar systems.*

The solution of Problems (B) and (C)

Given the array $\mathbb{P}$ defined by (4.79)—again we restrict our attention to the distinct point case—together with $\mu > 0$, we wish to find out whether there exist interpolants that are stable (poles in the left half of the complex plane) with magnitude bounded by μ on the imaginary axis. The tool for investigating the existence issue is the *Nevanlinna–Pick* matrix

$$\Pi_\mu = \begin{bmatrix} \frac{\mu^2-\bar{\phi}_1\phi_1}{\bar{s}_1+s_1} & \cdots & \frac{\mu^2-\bar{\phi}_1\phi_N}{\bar{s}_1+s_N} \\ \vdots & \ddots & \vdots \\ \frac{\mu^2-\bar{\phi}_N\phi_1}{\bar{s}_N+s_1} & \cdots & \frac{\mu^2-\bar{\phi}_N\phi_N}{\bar{s}_N+s_N} \end{bmatrix},$$

where $\bar{(\cdot)}$ denotes complex conjugation. A solution exists if and only if this matrix is positive (semi-) definite $\Pi_\mu \geq 0$. Write $\Pi_\mu = \mu^2\Pi_1 - \Pi_2$, where $\Pi_1 > 0$, $\Pi_2 > 0$. Let μ_i^2 be the eigenvalues of $\Pi_1^{-1}\Pi_2$, with μ_1^2 the largest. As long as $\mu > \mu_1$, $\Pi_\mu > 0$, for $\mu = \mu_1$ it becomes semidefinite and for $\mu < \mu_1$ it is indefinite. Thus the smallest norm for which there exist solutions to Problem (B) is the square root of the largest eigenvalue of $\Pi_1^{-1}\Pi_2$.

In [26] it was shown that the Nevanlinna–Pick interpolation problem can be transformed into an interpolation problem *without norm constraint*. This is achieved by adding the so-called mirror image interpolation points to the original data. In terms of trajectories, the mirror image set $\hat{\mathbb{D}}$ of $\mathbb{D}$, defined by (4.93), is

$$\mathbb{D} = \left\{ \begin{pmatrix} 1 \\ -\phi_i \end{pmatrix} e^{s_i t},\ i = 1, \ldots, N \right\} \text{ and } \hat{\mathbb{D}} = \left\{ \begin{pmatrix} -\bar{\phi}_i \\ 1 \end{pmatrix} e^{-\bar{s}_i t},\ i = 1, \ldots, N \right\}.$$

The augmented data set is thus $\mathbb{D}_{aug} = \mathbb{D} \cup \hat{\mathbb{D}}$, and the corresponding pair of matrices is

$$\mathbf{C}_{aug} = (\mathbf{C} \ \ \mathbf{J}\bar{\mathbf{C}}), \ \mathbf{A}_{aug} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\bar{\mathbf{A}} \end{pmatrix}, \text{ where } \mathbf{J} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and $\bar{(\cdot)}$ applied to a matrix denotes complex conjugation (without transposition). We now construct left coprime polynomial matrices Θ_{aug}, Ξ_{aug} such that $\mathbf{C}_{aug}(s\mathbf{I} - \mathbf{A}_{aug})^{-1} = [\Theta_{aug}(s)]^{-1}\Xi_{aug}(s)$. The main result is that the generating system for the data set $\mathbb{D}_{aug}$ is the generating system that solves Problem (B), provided that the parameters $\mathbf{a}$, $\mathbf{b}$ are appropriately restricted. For simplicity of notation, let the entries of Θ_{aug} be denoted by θ_{ij}. The following result can be proved using the results in [26].

Theorem 4.62. Classification by norm. *Let Θ_{aug} be as defined above. The interpolants $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ of $\mathbb{D}$ with norm of ϕ less than or equal to μ on the imaginary axis are given, if they exist, by $\phi = \frac{\mathbf{a}\theta_{11}+\mathbf{b}\theta_{21}}{\mathbf{a}\theta_{12}+\mathbf{b}\theta_{22}}$, where the magnitude of $\frac{\mathbf{a}}{\mathbf{b}}$ on the imaginary axis must be at most μ.*

The above result achieves an *algebraization* of the Nevanlinna–Pick interpolation problem. For related treatments of Problem (B), see [271] and [195]. See also the book by Ball, Gohberg, and Rodman [38]. We conclude this part by mentioning that Problem (C), that is, the problem of positive real interpolation, can also be turned into an unconstrained interpolation problem by adding an appropriately defined mirror-image set of points. For details, see [27], [234], [309].

Concluding remarks and generalizations

The problem of *rational interpolation* has a long history. It was only recently recognized, however, as a problem that generalizes the realization problem. One can distinguish two approaches: state-space and polynomial. The generalization of the state-space framework from the realization to the rational interpolation problem is due to Antoulas and Anderson [25], [23], [26] and Anderson and Antoulas [24]. Therein, the Löwner matrix replaces and generalizes the Hankel matrix as the main tool. The generating system or polynomial approach to rational interpolation with the complexity (McMillan degree) as constraint was put forward in Antoulas and Willems [15] and Antoulas et al. [22].

The above results can be generalized considerably. Consider the array consisting of the distinct interpolation data s_i, V_i, Y_i of size 1×1, $r_i \times p$, $r_i \times m$, respectively, satisfying $s_i \neq s_j$, $i \neq j$, and rank $V_i = r_i \leq p$, of $i = 1, \ldots, N$. The *left tangential* or *left directional interpolation problem* consists of finding all $p \times m$ rational matrices $\Phi(s)$ satisfying $V_i\Phi(s_i) = Y_i$, $i = 1, \ldots, N$, keeping track of their complexity, norm boundedness, or positive realness at the same time. In this case, the generating matrix Θ is a square of size $p + m$, and there are $p + m$ (observability) indices that enter the picture. The *right tangential* or *right directional interpolation problem*, as well as the *bitangential* or *bidirectional interpolation problem*, can be defined similarly. The solution of Problem (A) in all its matrix and tangential versions has been given in the generating system framework in [22]. For a general account on the generating system approach to rational interpolation, see [38].

Examples

Example 4.63. Consider the data array containing four pairs:

$$\mathbb{P} = \{(0,0),\ (1,3),\ (2,4),\ (1/2,2)\}.$$

According to (4.96),

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & -3 & -4 & -2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 2 & \\ & & & \frac{1}{2} \end{pmatrix}.$$

Following the construction leading to (4.105), we need the observability matrix of the (**C**, **A**) pair,

$$\mathcal{O}_4 = \left(\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 0 & -3 & -4 & -2 \\ \hline 0 & 1 & 2 & \frac{1}{2} \\ 0 & -3 & -8 & -1 \\ \hline 0 & 1 & 4 & \frac{1}{4} \\ 0 & -3 & -16 & -\frac{1}{2} \\ \hline 0 & 1 & 8 & \frac{1}{8} \\ 0 & -3 & -32 & -\frac{1}{4} \end{array}\right).$$

Examining the rows from top to bottom, the first linear dependence occurring in $\mathcal{O}_4$ is that of the fourth row on the previous ones:

$$\mathbf{c}_2\mathbf{A} + 6\mathbf{c}_1\mathbf{A} + \mathbf{c}_2 = 0.$$

It follows that $\kappa_1 = 1$. Therefore, $\kappa_2 = N - \kappa_1 = 3$. This means that the next linear dependence is that of the seventh row on the previous ones:

$$2\mathbf{c}_1\mathbf{A}^3 - 9\mathbf{c}_1\mathbf{A}^2 + 4\mathbf{c}_1\mathbf{A} - 2\mathbf{c}_2\mathbf{A} + \mathbf{c}_2 = 0.$$

According to formula (4.105), a row reduced generating matrix Θ is therefore

$$\Theta(s) = \begin{pmatrix} 6s & s+1 \\ s(s-4)(2s-1) & -(2s-1) \end{pmatrix}.$$

By (i) of the theorem, since $6s$ and $s+1$ are coprime polynomials, there is a unique minimal interpolant with McMillan degree 1, namely,

$$\phi^{min}(s) = \frac{6s}{s+1}.$$

Furthermore, there are no interpolants of degree 2. The next family of interpolants has McMillan degree 3. It can be parametrized in terms of the second-degree polynomial $\mathbf{p}(s) = p_0 + p_1 s + p_2 s^2$, as follows:

$$\phi(s) = \frac{s(s-4)(2s-1) + (p_2 s^2 + p_1 s + p_0)6s}{-(2s-1) + (p_2 s^2 + p_1 s + p_0)(s+1)}.$$

The coefficients of $\mathbf{p}$ must satisfy the constraints (4.104) in $\mathbb{R}^3$, which in this case turn out to be

$$\begin{aligned}
&p_0 + 1 \neq 0,\\
&2p_2 + 2p_1 + 2p_0 - 1 \neq 0,\\
&4p_2 + 2p_1 + p_0 - 1 \neq 0,\\
&p_2 + 2p_1 + 4p_0 \neq 0.
\end{aligned}$$

By letting $p_2 = p_1 = 0$ and $p_0 = 2$ in the above family of interpolants, we obtain the Lagrange interpolating polynomial for the data array $\mathbb{P}$:

$$\ell(s) = \frac{s}{3}(2s^2 - 9s + 16).$$

The next family of interpolants has McMillan degree 4. It can be parametrized in terms of a third-degree polynomial $\mathbf{p}$ and a first-degree polynomial $\mathbf{q}$, as follows:

$$\phi(s) = \frac{(p_3 s^3 + p_2 s^2 + p_1 s + p_0)6s + (q_1 s + q_0)s(s-4)(2s-1)}{(p_3 s^3 + p_2 s^2 + p_1 s + p_0)(s+1) - (q_1 s + q_0)(2s-1)}.$$

Firstly, $\mathbf{p}$, $\mathbf{q}$ must be coprime, i.e.,

$$p_3 q_0^3 - p_2 q_0^2 q_1 + p_1 q_0 q_1^2 - p_0 q_1^3 \neq 0.$$

Then, constraints (4.104) must also be satisfied, that is, the free parameters must avoid the following hyperplanes in $\mathbb{R}^6$:

$$\begin{aligned}
&p_0 + q_0 \neq 0,\\
&2p_3 + 2p_2 + 2p_1 + 2p_0 - q_1 - q_0 \neq 0,\\
&8p_3 + 4p_2 + 2p_1 + p_0 - 2q_1 - q_0 \neq 0,\\
&p_3 + 2p_2 + 4p_1 + 8p_0 \neq 0.
\end{aligned}$$

Example 4.64. *Continuation of Example* 4.63: *Recursive construction of interpolants.* The time series associated with $\mathbb{P}$ are

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 1 \\ -3 \end{pmatrix} e^t, \quad \mathbf{w}_3 = \begin{pmatrix} 1 \\ -4 \end{pmatrix} e^{2t}, \quad \mathbf{w}_4 = \begin{pmatrix} 1 \\ -2 \end{pmatrix} e^{\frac{t}{2}}.$$

Following Figure 4.6, we will now construct the generating systems Θ_i, $i = 1, 2, 3, 4$, satisfying $\Theta = \Theta_4 \Theta_3 \Theta_2 \Theta_1$; according to (4.102), since $\det \Pi_{i=1}^4 \Theta(s) = \Pi_{i=1}^4 (s - s_i)$, there must hold $\det \Theta_i = s - s_i$, $i = 1, 2, 3, 4$. The generating system that annihilates $\mathbf{w}_1$ is thus

$$\Theta_1(s) = \begin{pmatrix} s & 0 \\ 0 & 1 \end{pmatrix},$$

since the first error time series, defined as $\mathbf{e}_1 = \Theta_1(\frac{d}{dt})\mathbf{w}_1$, is zero: $\mathbf{e}_1 = \mathbf{0}$. The second error time series is defined similarly, namely, $\mathbf{e}_2 = \Theta_1(\frac{d}{dt})\mathbf{w}_2$, and we have $\mathbf{e}_2 = \mathbf{w}_2$; thus

$$\Theta_2(s) = \begin{pmatrix} 3 & 1 \\ 0 & s-1 \end{pmatrix}.$$

The first error time series remains zero, $\Theta_2\Theta_1(\frac{d}{dt})\mathbf{w}_1 = 0$, and $\Theta_2(\frac{d}{dt})\mathbf{e}_2 = \Theta_2\Theta_1(\frac{d}{dt})\mathbf{w}_2 = \mathbf{0}$; the third error time series is $\Theta_2\Theta_1(\frac{d}{dt})\mathbf{w}_3 = \begin{pmatrix} 2 \\ -4 \end{pmatrix} e^{2t}$, which implies

$$\Theta_3(s) = \begin{pmatrix} s-2 & 0 \\ 2 & 1 \end{pmatrix} \Rightarrow \Theta_3\Theta_2\Theta_1\left(\frac{d}{dt}\right)\mathbf{w}_i = \mathbf{0}, \qquad i = 1, 2, 3,$$

$$\text{while } \mathbf{e}_4 = \Theta_3\Theta_2\Theta_1\left(\frac{d}{dt}\right)\mathbf{w}_4 = \begin{pmatrix} \frac{3}{4} \\ 0 \end{pmatrix} e^{\frac{t}{2}}.$$

Finally

$$\Theta_4(s) = \begin{pmatrix} 2s-1 & 0 \\ 0 & 1 \end{pmatrix} \Rightarrow \Theta(s) = \Theta_4(s)\Theta_3(s)\Theta_2(s)\Theta_1(s)$$
$$= \begin{pmatrix} 3s(s-2)(2s-1) & (s-2)(2s-1) \\ 6s & s+1 \end{pmatrix}.$$

Although this generating matrix is not the same as the one obtained in the previous example, it differs only by left multiplication with a unimodular matrix. In particular, $\mathbf{U}\Theta_4\Theta_3\Theta_2\Theta_1$, where $\mathbf{U}(s) = \begin{pmatrix} 0 & 1 \\ \frac{1}{3} & \frac{1-2s}{3} \end{pmatrix}$ is equal to the generating matrix obtained in the previous example.

Example 4.65. *Realization Example* 4.47 *revisited.* Following the construction leading to (4.105), we need the observability matrix of the $(\mathbf{C}, \mathbf{A})$ pair defined by (4.98):

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 & -2 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The associated observability matrix is $\mathcal{O}_5$:

$$\mathcal{O}_5(\mathbf{C}, \mathbf{A}) = \left(\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 & -2 \\ \hline 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ \hline 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \\ \hline 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The first row that is linearly dependent on the preceding ones is the sixth; the next is the seventh:

$$\mathbf{c}_1\mathbf{A}^2 - \mathbf{c}_2\mathbf{A}^2 + \mathbf{c}_2\mathbf{A} = 0, \quad \mathbf{c}_1\mathbf{A}^3 + \mathbf{c}_1\mathbf{A}^2 - \mathbf{c}_1\mathbf{A} + 2\mathbf{c}_2\mathbf{A} - \mathbf{c}_2 = 0.$$

The corresponding generating system is

$$\Theta(s) = \begin{pmatrix} s^2 & -s^2 + s \\ s^3 + s^2 - s & 2s - 1 \end{pmatrix}.$$

According to the theory, the minimal interpolant has degree 3, and all minimal interpolants form a two-parameter family obtained by multiplying Θ on the left by $[\alpha s + \beta \quad 1]$:

$$\tilde{\phi}^{min}(s) = -\frac{(\alpha s + \beta)s^2 + (s^3 + s^2 - s)}{(\alpha s + \beta)(s^2 - s) - (2s - 1)}, \qquad \alpha,\ \beta \in \mathbb{R}.$$

According to (4.104), the parameters α, β must be such that the denominator of the above expression is nonzero for $s = 0$; therefore, since the value of the denominator for $s = 0$ is 1, these two parameters are *free*. And to obtain matching of the Markov parameters, we replace s by s^{-1}:

$$\phi^{min}(s) = \tilde{\phi}^{min}(s^{-1}) = -\frac{(\alpha + \beta s) + (1 + s - s^2)}{(\alpha + \beta s)(1 - s) - 2s^2 + s^3} = \frac{s^2 - (\beta + 1)s - (\alpha + 1)}{s^3 - (\beta + 2)s^2 + (\beta - \alpha)s + \alpha}.$$

It is readily checked that the power series expansion of ϕ around infinity is

$$\phi^{min}(s) = s^{-1} + s^{-2} + s^{-3} + 2s^{-4} + (\beta + 4)s^{-5} + (\beta^2 + 4\beta + \alpha + 8)s^{-6} + \cdots .$$

4.6 Chapter summary

The purpose of this chapter was to familiarize the reader with fundamental concepts from system theory. Three topics were discussed, namely, the external description, the internal description, and the realization/interpolation problem. The last section on the rational interpolation problem can be omitted on first reading. It forms a natural extension and generalization of the realization problem, which is interesting in its own right. Many aspects discussed in this section, however, will turn out to be important in what follows.

The section on the external description states that a linear time-invariant system is an operator (map) which assigns inputs to outputs. In particular, it is the convolution operator defined in terms of its kernel, which is the impulse response of the system. In the frequency domain, the external description is given in terms of the transfer function, whose series expansion around infinity yields the Markov parameters.

The internal description introduces, besides the external variables (that is, the input and the output), an internal variable called the state. The internal description thus consists of a set of equations describing how the input affects the state and another set describing how the output is obtained from the input and the state. The former set contains differential or difference equations, which are first order in the state (only the first derivative or shift of the state is involved), and it describes the dynamics of the system. The latter set contains algebraic equations (no derivatives or shifts of the state are allowed). The internal description is thus completely determined by means of the quadruple of matrices $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, defined by (4.13).

To get a better understanding of a dynamical system, two *structural concepts* are introduced, namely, those of *reachability* and *observability*. The former is used to investigate the extent to which the states can be influenced by manipulating the input. The latter is used to assess the influence of the state on the output. The *reachability* and *observability* matrices are used to quantitatively answer these questions. Besides these matrices, the so-called gramians can be used as well. Their advantage lies in the fact that they are square symmetric and semidefinite. Furthermore, if the assumption of stability is made, the *infinite gramians* arise, which can be computed by solving appropriately defined linear matrix equations, the so-called Lyapunov equations. These equations play a fundamental role in the computation of approximants to a given system. They will be the subject of detailed study in Chapter 6.

Eliminating the state from the internal description is straightforward and leads to the external description. The converse, however, that is deducing an internal description from the external description is nontrivial. It involves the *construction of state*, and if the data consist of *all* Markov parameters, it is known as the *realization problem*. Conditions for solvability and ways of constructing solutions (if they exist) are discussed in section 4.4. If the data consist of a partial set of Markov parameters, the existence of solutions being no longer an issue, the parametrization of all (minimal complexity) solutions becomes important; this is discussed briefly.

The final section on rational interpolation provides a generalization of the realization results in several ways. First, the input-output data need not be confined to Markov parameters; instead, samples of the transfer function at arbitrary points in the complex plane are allowed. Second, both state-space and polynomial ways of constructing all solutions of a given complexity are discussed. Finally it is pointed out that the machinery which was set up (generating system method) can be used to solve the problems of constructing systems with special properties, like bounded real or positive real transfer function.

Chapter 5

Linear Dynamical Systems: Part 2

To assess the quality of a given approximation, we must be able to measure "sizes" of dynamical systems as well as the distance between two of them. In the first part of this chapter, we discuss various notions of norms of linear dynamical systems. The second part offers a brief review of system stability and an introduction to the related concept of system dissipativity. The chapter concludes with the discussion of $\mathcal{L}_2$-systems, which play a role in Hankel-norm approximation. For more details on the material in the sections that follow, see [96], [116], [71], [149], [370], [360], [361], [288], and, in addition, [119], [177], and [197].

5.1 Time and frequency domain spaces and norms

Consider a linear space $\mathbb{X}$ over $\mathbb{R}$, not necessarily finite-dimensional. Let a norm ν be defined on $\mathbb{X}$, satisfying the three axioms (3.1). $\mathbb{X}$ is then called a *normed space*. In such spaces the concept of *convergence* can be defined as follows. We say that a sequence $\mathbf{x}_k$, $k = 1, 2, \ldots$, *converges* to $\mathbf{x}_*$ if the sequence of real numbers $\nu(\mathbf{x}_k - \mathbf{x}_*) = \| \mathbf{x}_k - \mathbf{x}_* \|$ converges to zero. A sequence $\mathbf{x}_k$, $k = 1, 2, \ldots$, is a *Cauchy sequence* if for all $\epsilon > 0$ there exists an integer m such that $\| \mathbf{x}_i - \mathbf{x}_j \| < \epsilon$ for all indices $i, j > m$. If every Cauchy sequence converges, then $\mathbb{X}$ is called *complete*.

5.1.1 Banach and Hilbert spaces

A Banach space is a normed linear space $\mathbb{X}$ that is complete. A subspace $\mathbb{Y}$ of $\mathbb{X}$ is *closed* if every sequence in $\mathbb{Y}$ that converges in $\mathbb{X}$ has its limit in $\mathbb{Y}$. If $\mathbb{X}$ is finite-dimensional, every subspace is closed. This does not hold if $\mathbb{X}$ is infinite-dimensional.

We now turn our attention to Hilbert spaces. These spaces have more structure than Banach spaces. The additional structure results from the existence of an *inner product*. The inner product is a function from the Cartesian product $\mathbb{X} \times \mathbb{X}$ to $\mathbb{R}$:

$$\langle \cdot, \cdot \rangle : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}, \quad (\mathbf{x}, \mathbf{y}) \longmapsto \langle \mathbf{x}, \mathbf{y} \rangle \in \mathbb{R}. \tag{5.1}$$

This map must satisfy the following four properties:

$$\begin{aligned}
&\langle \mathbf{x}, \mathbf{x}\rangle \geq 0,\\
&\langle \mathbf{x}, \mathbf{x}\rangle = 0 \;\Leftrightarrow\; x = 0,\\
&\text{for fixed } \mathbf{x}_* \in \mathbb{X},\ \langle \mathbf{x}_*, \mathbf{y}\rangle \text{ is a linear function},\\
&\langle \mathbf{x}, \mathbf{y}\rangle^* = \langle \mathbf{y}, \mathbf{x}\rangle.
\end{aligned}$$

This inner product *induces* a norm on $\mathbb{X}$, namely,

$$\mathbf{x} \longmapsto \langle \mathbf{x}, \mathbf{x}\rangle = \| \mathbf{x} \|^2 .$$

Linear spaces $\mathbb{X}$ that are endowed with an *inner product* and are *complete* are called Hilbert spaces. We close this introductory section by defining the space $\mathbb{Y}^\perp$ of a subspace $\mathbb{Y} \subset \mathbb{X}$:

$$\mathbb{Y}^\perp = \{\mathbf{x} \in \mathbb{X} : \langle \mathbf{x}, \mathbf{y}\rangle = 0,\ \forall \mathbf{y} \in \mathbb{Y}\}.$$

The space $\mathbb{Y}^\perp$ is always closed; if $\mathbb{Y}$ is also closed, then $\mathbb{Y}^\perp$ is called its *orthogonal complement*; in this case, $\mathbb{X}$ can be decomposed in a direct sum,

$$\mathbb{X} = \mathbb{Y} \oplus \mathbb{Y}^\perp.$$

5.1.2 The Lebesgue spaces ℓ_p and $\mathcal{L}_p$

In this section, we define the p-norms of infinite sequences and functions. Consider the vector-valued sequences $\mathbf{f} : \mathcal{I} \to \mathbb{R}^n$, where $\mathcal{I} = \mathbb{Z}, \mathcal{I} = \mathbb{Z}_+$, or $\mathcal{I} = \mathbb{Z}_-$. The Hölder p-norms of these sequences are defined as

$$\| \mathbf{f} \|_p = \begin{cases} \left(\sum_{t\in\mathcal{I}} \| \mathbf{f}(t) \|_p^p\right)^{\frac{1}{p}}, & 1 \leq p < \infty,\\[2ex] \sup_{t\in\mathcal{I}} \| \mathbf{f}(t) \|_p, & p = \infty. \end{cases} \tag{5.2}$$

The corresponding ℓ_p spaces are

$$\ell_p^n(\mathcal{I}) = \{\mathbf{f} : \mathcal{I} \to \mathbb{R}^n,\ \| \mathbf{f} \|_p < \infty\}, \qquad 1 \leq p \leq \infty,$$

where, for instance, $\mathcal{I} = \mathbb{Z}, \mathcal{I} = \mathbb{Z}_+$, or $\mathcal{I} = \mathbb{Z}_-$. Consider now continuous-time vector-valued functions $\mathbf{f} : \mathcal{I} \to \mathbb{R}^n$, where $\mathcal{I} = \mathbb{R}, \mathcal{I} = \mathbb{R}_+, \mathcal{I} = \mathbb{R}_-$, or $\mathcal{I} = [a, b]$; the p-norms are defined as

$$\| \mathbf{f} \|_p = \begin{cases} \left(\int_{t\in\mathcal{I}} \| \mathbf{f}(t) \|_p^p \, dt\right)^{\frac{1}{p}}, & 1 \leq p < \infty,\\[2ex] \sup_{t\in\mathcal{I}} \| \mathbf{f}(t) \|_p, & p = \infty. \end{cases} \tag{5.3}$$

The corresponding $\mathcal{L}_p$ spaces are

$$\mathcal{L}_p^n(\mathcal{I}) = \{\mathbf{f} : \mathcal{I} \to \mathbb{R}^n,\ \| \mathbf{f} \|_p < \infty\}, \qquad 1 \leq p \leq \infty,$$

where, for instance, $\mathcal{I} = \mathbb{R}, \mathcal{I} = \mathbb{R}_+, \mathcal{I} = \mathbb{R}_-$, or the finite interval $\mathcal{I} = [a, b]$.

The spaces defined above are sometimes referred to as *time domain* ℓ_p, $\mathcal{L}_p$ spaces since their elements are functions of a real variable, which is mostly *time*. In what follows, *frequency domain* ℓ_p, $\mathcal{L}_p$ spaces will also be introduced.

5.1.3 The Hardy spaces h_p and $\mathcal{H}_p$

We now turn our attention to spaces of functions of a *complex variable*. Since this variable can be interpreted as *complex frequency*, we will refer to these spaces as *frequency domain* spaces.

h_p spaces

Let $\mathcal{D}$, $\partial\mathcal{D}$, $\bar{\mathcal{D}} \subset \mathbb{C}$ denote the (open) unit disc, the unit circle, and the complement of the closed unit disc, respectively. Consider the matrix-valued function $\mathbf{F} : \mathbb{C} \to \mathbb{C}^{q\times r}$, which is analytic in $\bar{\mathcal{D}}$. Its p-norm is defined as follows:

$$\| \mathbf{F} \|_{h_p} = \left(\frac{1}{2\pi} \sup_{|r|>1} \int_0^{2\pi} \| \mathbf{F}(re^{i\theta}) \|_p^p \, d\theta \right)^{\frac{1}{p}} \quad \text{for } p \in [1, \infty)$$

$$\text{and } \| \mathbf{F} \|_{h_\infty} = \sup_{z\in\bar{\mathcal{D}}} \| \mathbf{F}(z) \|_p \quad \text{for } p = \infty.$$

We choose $\| \mathbf{F}(z_0) \|_p$ to be the Schatten p-norm of $\mathbf{F}$ evaluated at $z = z_0$; the Schatten norms are defined in (3.5). However, there are other possible choices (see section 5.6). The resulting h_p spaces are defined as follows:

$$h_p^{q\times r} = h_p^{q\times r}(\bar{\mathcal{D}}) = \{\mathbf{F} : \mathbb{C} \to \mathbb{C}^{q\times r} : \ \| \mathbf{F} \|_{h_p} < \infty\}.$$

In a similar way, one can define the spaces $h_p^{q\times r}(\mathcal{D})$. The following special cases are worth noting:

$$\| \mathbf{F} \|_{h_2} = \left(\frac{1}{2\pi} \sup_{|r|>1} \int_0^{2\pi} \text{trace } \left[\mathbf{F}^*(re^{-i\theta})\mathbf{F}(re^{i\theta})\right] d\theta \right)^{\frac{1}{2}}, \tag{5.4}$$

where $(\cdot)^*$ denotes complex conjugation and transposition; furthermore

$$\| \mathbf{F} \|_{h_\infty} = \sup_{z\in\bar{\mathcal{D}}} \sigma_{max} (\mathbf{F}(z)) \, . \tag{5.5}$$

$\mathcal{H}_p$ spaces

Let $\mathbb{C}_+ \subset \mathbb{C}$ denote the (open) right half of the complex plane: $s = x + iy \in \mathbb{C}$, $x > 0$. Consider the $q \times r$ complex-valued functions $\mathbf{F}$, which are analytic in $\mathbb{C}_+$. Then

$$\| \mathbf{F} \|_{\mathcal{H}_p} = \left(\sup_{x>0} \int_{-\infty}^{\infty} \| \mathbf{F}(x + iy) \|_p^p \, dy \right)^{\frac{1}{p}} \quad \text{for } p \in [1, \infty)$$

$$\text{and } \| \mathbf{F} \|_{\mathcal{H}_\infty} = \sup_{z\in\mathbb{C}_+} \| \mathbf{F}(z) \|_p \quad \text{for } p = \infty.$$

Again, $\| \mathbf{F}(s_0) \|_p$ is chosen to be the Schatten p-norm (see (3.5)) of $\mathbf{F}$ evaluated at $s = s_0$. The resulting $\mathcal{H}_p$ spaces are defined analogously to the h_p spaces:

$$\mathcal{H}_p^{q\times r} = \mathcal{H}_p^{q\times r}(\mathbb{C}_+) = \{\mathbf{F} : \mathbb{C} \to \mathbb{C}^{q\times r} : \ \| \mathbf{F} \|_{\mathcal{H}_p} < \infty\}.$$

The Hardy spaces $\mathcal{H}_p^{q\times r}(\mathbb{C}_-)$ can be defined in a similar way. Two special cases are worth noting:

$$\| \mathbf{F} \|_{\mathcal{H}_2} = \left(\sup_{x>0} \int_{-\infty}^{\infty} \text{trace } \left[\mathbf{F}^*(x-iy)\mathbf{F}(x+iy) \right] dy \right)^{\frac{1}{2}}, \tag{5.6}$$

where $(\cdot)^*$ denotes complex conjugation and transposition, and

$$\| \mathbf{F} \|_{\mathcal{H}_\infty} = \sup_{s\in\mathbb{C}_+} \sigma_{max}\left(\mathbf{F}(s)\right). \tag{5.7}$$

The search for the suprema in the formulas above can be simplified by making use of the ***maximum modulus theorem***, which states that a function $\mathbf{f}$ continuous inside a domain $D \subset \mathbb{C}$ as well as on its boundary ∂D and analytic inside D attains its maximum on the *boundary* ∂D of D. Thus (5.4), (5.5), (5.6), (5.7) become

$$\| \mathbf{F} \|_{h_2} = \left(\frac{1}{2\pi} \int_0^{2\pi} \text{trace } \left[\mathbf{F}^*(e^{-i\theta})\mathbf{F}(e^{i\theta}) \right] d\theta \right)^{\frac{1}{2}}, \quad \| \mathbf{F} \|_{h_\infty} = \sup_{\theta\in[0,2\pi]} \sigma_{max}\left(\mathbf{F}(e^{i\theta})\right) \tag{5.8}$$

and

$$\| \mathbf{F} \|_{\mathcal{H}_2} = \left(\int_{-\infty}^{\infty} \text{trace } \left[\mathbf{F}^*(-iy)\mathbf{F}(iy) \right] dy \right)^{\frac{1}{2}}, \quad \| \mathbf{F} \|_{\mathcal{H}_\infty} = \sup_{y\in\mathbb{R}} \sigma_{max}\left(\mathbf{F}(iy)\right). \tag{5.9}$$

Frequency domain $\mathcal{L}_p$ spaces

If the function $\mathbf{F} : \mathbb{C} \to \mathbb{C}^{q\times r}$ has no singularities (poles) on the imaginary axis but is not necessarily analytic in either the left or the right half of the complex plane, $\mathcal{H}_p$ norms are not defined. Instead, the *frequency domain* $\mathcal{L}_p$ norms of $\mathbf{F}$ are defined as follows:

$$\| \mathbf{F} \|_{\mathcal{L}_p} = \left(\sup_{y\in\mathbb{R}} \int_{-\infty}^{\infty} \| \mathbf{F}(iy) \|_p^p \, dy \right)^{\frac{1}{p}} \text{ for } p \in [1,\infty) \text{ and } \| \mathbf{F} \|_{\mathcal{L}_\infty} = \sup_{y\in\mathbb{R}} \sigma_{max}(\mathbf{F}(iy)). \tag{5.10}$$

The corresponding *frequency domain* $\mathcal{L}_p$ spaces are

$$\mathcal{L}_p(i\mathbb{R}) = \{ \mathbf{F} : \mathbb{C} \to \mathbb{C}^{q\times r} : \| \mathbf{F} \|_{\mathcal{L}_p} < \infty \}. \tag{5.11}$$

In a similar way, one may define the *frequency domain* ℓ_p spaces.

5.1.4 The Hilbert spaces ℓ_2 and $\mathcal{L}_2$

The (real) spaces $\ell_p(\mathcal{I})$ and $\mathcal{L}_p(\mathcal{I})$ are Banach spaces; $\ell_2(\mathcal{I})$ and $\mathcal{L}_2(\mathcal{I})$, however, can be given the structure of Hilbert spaces, where the inner product is defined as follows:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\ell_2} = \sum_{t\in\mathcal{I}} \mathbf{x}^*(t)\mathbf{y}(t), \quad \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}_2} = \frac{1}{2\pi} \int_{\mathcal{I}} \mathbf{x}^*(t)\mathbf{y}(t) dt. \tag{5.12}$$

For $\mathcal{I} = \mathbb{Z}$ and $\mathcal{I} = \mathbb{R}$, respectively, elements (vectors or matrices) with entries in $\ell_2(\mathbb{Z})$ and $\mathcal{L}_2(\mathbb{R})$ have a *transform* defined as follows:

$$\mathbf{f} \longmapsto \mathbf{F}(\xi) = \begin{cases} \sum_{-\infty}^{\infty} \mathbf{f}(t)\xi^{-t} & \text{for discrete-time functions,} \\ \int_{-\infty}^{\infty} \mathbf{f}(t)e^{-\xi t}dt & \text{for continuous-time functions.} \end{cases}$$

Thus if the domain of $\mathbf{f}$ is discrete, $\mathbf{F}(e^{i\theta}) = \mathcal{F}(\mathbf{f})(\theta)$ is the *Fourier transform* of $\mathbf{f}$ and belongs to $\mathcal{L}_2[0, 2\pi]$; analogously, if the domain of $\mathbf{f}$ is continuous, $\mathbf{F}(i\omega) = \mathcal{F}(\mathbf{f})(\omega)$ is the *Fourier transform* of $\mathbf{f}$ belonging to the space $\mathcal{L}_2(i\mathbb{R})$ defined by (5.11). Furthermore, the following bijective correspondences hold:

$$\begin{aligned} \ell_2(\mathbb{Z}) = \ell_2(\mathbb{Z}_-) \oplus \ell_2(\mathbb{Z}_+) &\longrightarrow \mathcal{L}_2[0, 2\pi] = h_2(\mathcal{D}) \oplus h_2(\bar{\mathcal{D}}),. \\ \mathcal{L}_2(\mathbb{R}) = \mathcal{L}_2(\mathbb{R}_-) \oplus \mathcal{L}_2(\mathbb{R}_+) &\longrightarrow \mathcal{L}_2(i\mathbb{R}) = \mathcal{H}_2(\mathbb{C}_-) \oplus \mathcal{H}_2(\mathbb{C}_+). \end{aligned}$$

For simplicity, the above relationships are shown for spaces containing scalar-valued functions. They are, however, equally valid for the corresponding spaces containing matrix-valued functions.

Two results connect the spaces introduced above. We state only the continuous-time versions. The first has the names of *Parseval*, *Plancherel*, and *Paley-Wiener* attached to it.

Proposition 5.1. *The Fourier transform $\mathcal{F}$ is a Hilbert space isomorphism between $\mathcal{L}_2(\mathbb{R})$ and $\mathcal{L}_2(i\mathbb{R})$. In addition, it is an isometry, that is, it preserves distances. The Fourier transform maps $\mathcal{L}_2(\mathbb{R}_+)$, $\mathcal{L}_2(\mathbb{R}_-)$ onto $\mathcal{H}_2(\mathbb{C}_+)$, $\mathcal{H}_2(\mathbb{C}_-)$, respectively.*

The second result asserts that the product of an $\mathcal{L}_\infty$ function and a frequency domain $\mathcal{L}_2$ function is a frequency domain $\mathcal{L}_2$ function. It also asserts that the $\mathcal{L}_\infty$ and $\mathcal{H}_\infty$ norms can be viewed as *induced norms*. Recall that if $(\mathbb{X}, \alpha)$ and $(\mathbb{Y}, \beta)$ are two normed spaces with norms α, β, respectively, just as in the finite-dimensional case, the *α, β-induced norm* of an operator $\mathbf{T}$ with domain $\mathbb{X}$ and codomain $\mathbb{Y}$ is $\| \mathbf{T} \|_{\alpha,\beta} = \sup_{\mathbf{x}\neq\mathbf{0}} \| \mathbf{Tx} \|_\beta / \| \mathbf{x} \|_\alpha$.

Proposition 5.2. *Let $\mathbf{F} \in \mathcal{L}_\infty$ and $\mathbf{G} \in \mathcal{L}_2(i\mathbb{R})$. The product $\mathbf{FG} \in \mathcal{L}_2(i\mathbb{R})$. In addition, the $\mathcal{L}_\infty$ norm can be viewed as an induced norm in the frequency domain space $\mathcal{L}_2(i\mathbb{R})$:*

$$\| \mathbf{F} \|_{\mathcal{L}_\infty} = \| \mathbf{F} \|_{\mathcal{L}_2\text{-ind}} = \sup_{\mathbf{X}\neq\mathbf{0}} \frac{\| \mathbf{FX} \|_{\mathcal{L}_2}}{\| \mathbf{X} \|_{\mathcal{L}_2}}.$$

If $\mathbf{F} \in \mathcal{H}_\infty$ and $\mathbf{G} \in \mathcal{H}_2(\mathbb{C}_+)$, then $\mathbf{FG} \in \mathcal{H}_2(\mathbb{C}_+)$. Furthermore, the $\mathcal{H}_\infty$ norm can be viewed as an induced norm in the frequency domain space $\mathcal{H}_2$, as well as in the time domain space $\mathcal{L}_2$:

$$\| \mathbf{F} \|_{\mathcal{H}_\infty} = \| \mathbf{F} \|_{\mathcal{H}_2\text{-ind}} = \sup_{\mathbf{X}\neq\mathbf{0}} \frac{\| \mathbf{FX} \|_{\mathcal{H}_2}}{\| \mathbf{X} \|_{\mathcal{H}_2}} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\| \mathbf{f} * \mathbf{x} \|_{\mathcal{L}_2}}{\| \mathbf{x} \|_{\mathcal{L}_2}} = \| \mathbf{f} \|_{\mathcal{L}_2\text{-ind}},$$

where $\mathbf{x} \in \mathcal{L}_2(\mathbb{R}_+)$.

5.2 The spectrum and singular values of the convolution operator

Given a linear system Σ, recall the definition of the convolution operator $\mathcal{S}$ given by (4.4) and (4.5) for the discrete-time and the continuous-time case, respectively.

Spectrum and singular values of $\mathcal{S}$

Consider first the case of SISO ($m = p = 1$) discrete-time systems. The convolution operator $\mathcal{S}$ is also known as the Laurent operator. To this operator we associate the *transfer function* $\mathbf{H}(z) = \sum_{t=-\infty}^{\infty} \mathbf{h}_t z^{-t}$, also known as the *symbol* of $\mathcal{S}$. A classical result, which can be found, e.g., in [70], asserts that $\mathcal{S}$ is a bounded operator in $\ell_2(\mathbb{Z})$ if and only if $\mathbf{H}$ belongs to the (frequency domain) space $\ell_\infty(\partial\mathcal{D})$, that is, $\|\mathbf{H}\|_{\ell_\infty} < \infty$. The *spectrum* of this operator is composed of all complex numbers λ such that the *resolvent* $\mathcal{S} - \lambda\mathbf{I}$ is not invertible; furthermore, if there exists $\mathbf{v} \in \ell_2(\mathbb{Z})$ such that $(\mathcal{S} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$, $(\lambda, \mathbf{v})$ is an *eigenvalue, eigenvector* pair of $\mathcal{S}$.

Denoting the convolution operator by $\mathcal{S}(\mathbf{H})$, to stress the connection with $\mathbf{H}$, we have in this case

$$\|\mathcal{S}(\mathbf{H})\|_2 = \|\mathbf{H}\|_{\ell_\infty}.$$

Furthermore, it follows that $[\mathcal{S}(\mathbf{H})]^* = \mathcal{S}(\mathbf{H}^*)$ and $\mathcal{S}(\mathbf{H}_1)\mathcal{S}(\mathbf{H}_2) = \mathcal{S}(\mathbf{H}_1\mathbf{H}_2)$. Thus

$$\mathcal{S}(\mathbf{H})[\mathcal{S}(\mathbf{H})]^* = \mathcal{S}(\mathbf{H}\mathbf{H}^*) = \mathcal{S}(\mathbf{H}^*\mathbf{H}) = [\mathcal{S}(\mathbf{H})]^*\mathcal{S}(\mathbf{H}),$$

which shows that $\mathcal{S}$ is a *normal* operator. This implies that if λ belongs to the spectrum of $\mathcal{S}$, $|\lambda|$ is a singular value of $\mathcal{S}$. Otto Toeplitz in 1911 showed that the *spectrum* of $\mathcal{S}$ is

$$\Lambda(\mathcal{S}) = \{\mathbf{H}(z) : \ | z |= 1\}.$$

If we plot the elements of this set in the complex plane, we obtain a closed curve, which is sometimes referred to as the *Nyquist plot* of the underlying system. Due to the normality of $\mathcal{S}$, the set of singular values of $\mathcal{S}$ is given by absolute values of the elements of its spectrum:

$$\Sigma(\mathcal{S}) = \{ | \mathbf{H}(z) |: \ | z |= 1\}.$$

Therefore, the largest entry of this set is the ℓ_∞ norm of $\mathbf{H}$. Notice that $\mathcal{S}$ is not invertible whenever $0 \in \Lambda(\mathcal{S})$; if $\mathbf{H}$ is rational, this means that $\mathcal{S}$ is not invertible whenever $\mathbf{H}$ has a zero on the unit circle.

Closely related to $\mathcal{S}$ is the *Toeplitz operator* $\mathcal{T} : \ell_2(\mathbb{Z}_+) \to \ell_2(\mathbb{Z}_+)$, $\mathbf{y} = \mathcal{T}(\mathbf{u})$, where $\mathbf{y}(t) = \sum_{k=0}^{\infty} \mathbf{h}_{t-k}\mathbf{u}(k)$, $t \geq 0$. It can be shown (see [70]) that $\|\mathcal{T}\|_2 = \|\mathbf{H}\|_{h_\infty}$ and that the spectrum $\Lambda(\mathcal{T})$ contains $\Lambda(\mathcal{S})$, together with all points enclosed by this curve which have nonzero winding number.

The spectrum $\Lambda(\mathcal{S})$ can be approximated by considering a finite portion of the impulse response, namely, $\mathbf{h}_k$, where k runs over any consecutive set of N integers. For simplicity, let $k = 0, 1, \ldots, N-1$. The corresponding input and output sequences are finite: $\mathbf{u}_k$, $\mathbf{y}_k$, $k = 0, 1, \ldots, N-1$. In this case, $\mathbf{y} = \mathbf{h} \circledast \mathbf{u}$, where $\circledast$ denotes *periodic convolution*, which is the usual convolution but with the sum taken over N consecutive values of $\mathbf{u}$, and outside the interval $[0, N-1]$, the functions are assumed to be periodic: $\mathbf{y}(t) = \sum_{k=0}^{N-1} \mathbf{h}_{t-k}\mathbf{u}(k)$, $t = 0, 1, \dots, N-1$. In matrix form we have

$$\begin{pmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(N-2) \\ \mathbf{y}(N-1) \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{h}_0 & \mathbf{h}_{N-1} & \cdots & \mathbf{h}_2 & \mathbf{h}_1 \\ \mathbf{h}_1 & \mathbf{h}_0 & \ddots & & \mathbf{h}_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{h}_{N-2} & & \ddots & \mathbf{h}_0 & \mathbf{h}_{N-1} \\ \mathbf{h}_{N-1} & \mathbf{h}_{N-2} & \cdots & \mathbf{h}_1 & \mathbf{h}_0 \end{pmatrix}}_{\mathcal{S}_N} \begin{pmatrix} \mathbf{u}(0) \\ \mathbf{u}(1) \\ \vdots \\ \mathbf{u}(N-2) \\ \mathbf{u}(N-1) \end{pmatrix} \Rightarrow \mathbf{y}_N = \mathcal{S}_N \mathbf{u}_N.$$

The $N \times N$ matrix $\mathcal{S}_N$ is called a *circulant*; it is a special Toeplitz matrix constructed by means of cyclic permutation of its first column or row. Circulant matrices are normal; they are diagonalized by means of the discrete Fourier transform (DFT) matrix $(\Phi_N)_{k,\ell=1,\dots,N} = \frac{1}{\sqrt{N}} w_N^{(k-1)(\ell-1)}$, where $w_N = e^{-i\frac{2\pi}{N}}$ is the principal Nth root of unity. The eigenvalues of $\mathcal{S}_N$ are $\bar{\mathbf{H}}(w^k)$, $k = 0, 1, \dots, N-1$, where $\bar{\mathbf{H}}(z) = \sum_{k=0}^{N-1} \mathbf{h}_k z^{-k}$ can be considered an approximant of $\mathbf{H}$. Some examples will illustrate these issues.

Example 5.3. We will first compute the singular values of the convolution operator $\mathcal{S}$ associated with the discrete-time system:

$$y(k+1) - ay(k) = bu(k).$$

The impulse response of this system is $h_0 = 0$, $h_k = ba^{k-1}$, $k > 0$. Hence from the above discussion it follows that

$$\mathcal{S} = b \left(\begin{array}{ccc|ccc} \ddots & \vdots & \vdots & \vdots & \vdots & \\ \cdots & 0 & 0 & 0 & 0 & \cdots \\ \cdots & 1 & 0 & 0 & 0 & \cdots \\ \hline \cdots & a & 1 & 0 & 0 & \cdots \\ \cdots & a^2 & a & 1 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right), \quad \mathcal{S}_N = b \begin{pmatrix} 0 & a^{N-2} & \cdots & a & 1 \\ 1 & 0 & \cdots & a^2 & a \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a^{N-3} & a^{N-4} & \cdots & 0 & a^{N-2} \\ a^{N-2} & a^{N-3} & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{N \times N}$$

The error between the eigenvalues of $\mathcal{S}_N$ and the corresponding elements of the spectrum of $\mathcal{S}$ are of the order of 10^{-2}, 10^{-3} for $N = 6$, $N = 10$, respectively; a similar property holds for the singular values of $\mathcal{S}$ and $\mathcal{S}_N$. The spectrum (Nyquist plot) of this system consists of a circle centered at $(\frac{|ab|}{1-a^2}, 0)$ of radius $\frac{|b|}{1-a^2}$. Thus the largest singular value of $\mathcal{S}$ is $\frac{|b|}{1-|a|}$, while the smallest is $\frac{|b|}{1+|a|}$, which implies that $\mathcal{S}$ is invertible.

As a second example, we consider a discrete-time third-order system known as the *Butterworth filter*. Its transfer function is

$$\mathbf{H}(z) = \frac{1}{6} \frac{(z+1)^3}{z(z^2 + \frac{1}{3})}.$$

The first few Markov parameters are $\mathbf{h}_0 = \frac{1}{6}$, $\mathbf{h}_1 = \frac{1}{2}$, $\mathbf{h}_2 = \frac{4}{9}$, $\mathbf{h}_3 = 0$, $\mathbf{h}_4 = \frac{-4}{27}$, $\mathbf{h}_5 = 0$, $\mathbf{h}_6 = \frac{4}{81}$, $\mathbf{h}_7 = 0$, $\mathbf{h}_8 = \frac{-4}{243}$. The Nyquist plot (i.e., the spectrum of $\mathcal{S}$) is shown in Figure 5.1,

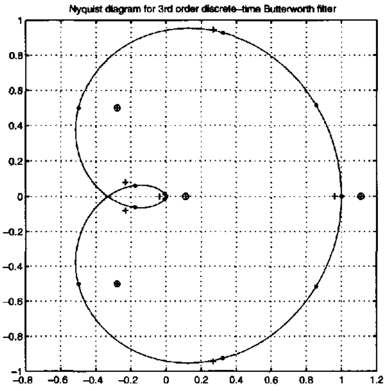


Figure 5.1. *Spectrum of the convolution operator $\mathcal{S}$ compared with the eigenvalues of the finite approximants $\mathcal{S}_N$, $N = 4$ (circles), $N = 6$ (crossed circles), $N = 10$ (dots).*

together with the eigenvalues of $\mathcal{S}_N$, for $N = 4$ (circles) $N = 6$ (crossed circles) and $N = 10$ (dots); the error in the latter case is of the order 10^{-3}.

The singular values of $\mathcal{S}$ are $\Sigma(\mathcal{S}) = [0, 1]$, which implies that $\mathcal{S}$ is not invertible. Finally, the spectrum of the associated Toeplitz operator $\Lambda(\mathcal{T})$ is composed of $\Lambda(\mathcal{S})$ and all points in the interior of this curve.

Generalization to continuous-time systems

For SISO continuous-time systems, similar results hold. Without proof, we state the following facts. The convolution operator defined by (4.5) is bounded in $\mathcal{L}_2(\mathbb{R})$, provided that the associated transfer function $\mathbf{H}(s)$ belongs to (the frequency domain space) $\mathcal{L}_\infty(i\mathbb{R})$, that is, $\mathbf{H}$ is bounded on the imaginary axis. Again, the spectrum of $\mathcal{S}$ is composed of all complex numbers such that the resolvent $\mathcal{S} - \lambda\mathbf{I}$ is not invertible. If the associated $\mathbf{H}$ is rational, this condition is equivalent to the absence of zeros of $\mathbf{H}$ on the imaginary axis.

Just as in the discrete-time case, denoting the convolution operator by $\mathcal{S}(\mathbf{H})$, to stress its connection with $\mathbf{H}$, we have $\|\mathcal{S}(\mathbf{H})\|_2 = \|\mathbf{H}\|_{\mathcal{L}_\infty}$. Furthermore, similar arguments show that $\mathcal{S}$ is normal. The parallel of the Toeplitz result in the continuous-time case is that the spectrum of $\mathcal{S}$ is

$$\Lambda(\mathcal{S}) = \{\mathbf{H}(i\omega) : \omega \in \mathbb{R}\}.$$

This set in the complex plane yields a closed curve known as the *Nyquist plot*. Therefore, the set of singular values of $\mathcal{S}$ is given by

$$\Sigma(\mathcal{S}) = \{\mid \mathbf{H}(i\omega) \mid : \omega \in \mathbb{R}\}.$$

The largest entry of this set is the $\mathcal{L}_\infty$ norm of $\mathbf{H}$.

Generalization to MIMO systems

The spectrum of the convolution operator associated with a MIMO system is defined for $m = p$ (same number of input and output channels). In this case, it is composed of the union of the m curves which constitute the m eigenvalues of the transfer function $\mathbf{H}(i\omega)$. For continuous-time systems, for instance, this gives

$$\Lambda(\mathcal{S}) = \{\, \lambda_k(\mathbf{H}(i\omega)) \,:\; \omega \in \mathbb{R},\; k = 1, \ldots, m \,\}.$$

The convolution operator for MIMO systems is in general not normal. Its singular values (which are defined even if $m \neq p$) are those of $\mathbf{H}(i\omega)$:

$$\Sigma(\mathcal{S}) = \{\, \sigma_k(\mathbf{H}(i\omega)) \,:\; \omega \in \mathbb{R},\; k = 1, \ldots, \min(m, p) \,\}.$$

Therefore,

$$\|\mathcal{S}\|_2 = \sup_{\omega \in \mathbb{R}} \sigma_1(\mathbf{H}(i\omega)),$$

where, as before, σ_1 denotes the largest singular value. Similar statements can be made for discrete-time MIMO systems.

The adjoint of the convolution operator

A detailed study of the singular values of the convolution operator $\mathcal{S}$ requires the use of the adjoint operator. For completeness, this issue is briefly addressed next, although it will not be used in the sequel.

By definition, given the inner product $\langle \cdot, \cdot \rangle$, the adjoint of $\mathcal{S}$, denoted by $\mathcal{S}^*$, is the unique operator satisfying

$$\langle \mathbf{y}, \mathcal{S}\mathbf{u} \rangle = \langle \mathcal{S}^*\mathbf{y}, \mathbf{u} \rangle \tag{5.13}$$

for all $\mathbf{y}$, $\mathbf{u}$ in the appropriate spaces.

By (4.5), $\mathcal{S}$ is an integral operator with kernel $\mathbf{h}(\cdot)$, which by means of state-space data can be expressed as follows: $\mathbf{h}(\cdot) = \mathbf{D}\delta(\cdot) + \mathbf{C}e^{\mathbf{A}(\cdot)}\mathbf{B}$, $t \geq 0$. The former part, namely, $\mathbf{D}$, is an *instantaneous* and not dynamical action; therefore, its adjoint is given by $\mathbf{D}^*$. The adjoint due to the latter part $\mathbf{h}_a(\cdot) = \mathbf{C}e^{\mathbf{A}(\cdot)}\mathbf{B}$, $t > 0$, which denotes dynamical action, can be computed as follows:

$$\begin{aligned}\langle \mathbf{y}, \mathcal{S}\mathbf{u} \rangle &= \int_{-\infty}^{\infty} \mathbf{y}^*(t) \left[\int_{-\infty}^{\infty} \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\, d\tau \right] dt \\ &= \int_{-\infty}^{\infty} \mathbf{u}^*(\tau) \left[\int_{-\infty}^{\infty} \mathbf{B}^* e^{\mathbf{A}^*(t-\tau)}\mathbf{C}^*\mathbf{y}(t)\, dt \right] d\tau = \langle \mathbf{u}, \mathcal{S}^*\mathbf{y} \rangle.\end{aligned}$$

Thus from (5.13), it follows that

$$\mathcal{S}^* : \mathcal{L}^p(\mathbb{R}) \longrightarrow \mathcal{L}^m(\mathbb{R}),\ \mathbf{y} \longmapsto \mathbf{u} = \mathcal{S}^*(\mathbf{y}),\ \text{where } \mathbf{u}(t) = -\int_{+\infty}^{-\infty} \mathbf{B}^* e^{-\mathbf{A}^*(t-\tau)}\mathbf{C}^*\mathbf{y}(\tau) d\tau, \tag{5.14}$$

with time t running backward, that is, from $+\infty$ to $-\infty$. It follows that given $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, its *adjoint* with respect to the usual inner product in $\mathcal{L}_2(\mathbb{R})$ can be defined as

$$\mathbf{\Sigma}^* = \left(\begin{array}{c|c} -\mathbf{A}^* & -\mathbf{C}^* \\ \hline \mathbf{B}^* & \mathbf{D}^* \end{array}\right). \tag{5.15}$$

Furthermore, if $\mathbf{H}_\Sigma(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, the transfer function of the adjoint system Σ^* is

$$\mathbf{H}_{\Sigma^*}(s) = \mathbf{H}^*(-s) = \mathbf{D}^* - \mathbf{B}^*(s\mathbf{I} + \mathbf{A}^*)^{-1}\mathbf{C}^*.$$

Remark 5.2.1. **(a)** According to the computations above, the adjoint can also be defined by replacing $-\mathbf{C}^*$, $\mathbf{B}^*$ with $\mathbf{C}^*$, $-\mathbf{B}^*$, respectively.

(b) Given the state space representation of the original system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, $\mathbf{y} = \mathbf{C}\mathbf{u} + \mathbf{D}\mathbf{u}$ and one for the adjoint $\dot{\mathbf{p}} = -\mathbf{A}^*\mathbf{p} - \mathbf{C}^*\tilde{\mathbf{y}}$, $\tilde{\mathbf{u}} = \mathbf{B}^*\tilde{\mathbf{y}} + \mathbf{D}^*\tilde{\mathbf{y}}$, notice that $\mathbf{u}$, $\tilde{\mathbf{u}}$ and $\mathbf{y}$, $\tilde{\mathbf{y}}$ are, respectively, elements of the same space. A straightforward calculation shows that the following relationship holds:

$$\frac{d}{dt}\mathbf{p}^*(t)\mathbf{x}(t) = \tilde{\mathbf{u}}^*(t)\mathbf{u}(t) - \tilde{\mathbf{y}}^*(t)\mathbf{y}(t).$$

In other words, for fixed t, the derivative of the usual inner product of the state of the original system and that of the adjoint system is equal to the difference of the inner products between $\mathbf{u}$, $\tilde{\mathbf{u}}$ and $\mathbf{y}$, $\tilde{\mathbf{y}}$. The following general result can be shown. Two systems with input $\mathbf{u}$, $\tilde{\mathbf{y}}$ and output $\mathbf{y}$, $\tilde{\mathbf{u}}$, respectively, are adjoints of each other if and only if they admit minimal state space representations such that the states $\mathbf{x}$, $\mathbf{p}$ satisfy the above instantaneous condition. For details, see [287], [120].

(c) According to the above considerations, the adjoint of the autonomous system $\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t)$ is $\frac{d}{dt}\mathbf{p}(t) = -\mathbf{A}^*\mathbf{p}(t)$. It follows that $\frac{d}{dt}\left[\mathbf{p}^*(t)\mathbf{x}(t)\right] = 0$, which means that $\mathbf{p}^*(t)\mathbf{x}(t) = \mathbf{p}^*(t_0)\mathbf{x}(t_0)$ for all $t \geq t_0$.

(d) One way to define the adjoint Σ^* of a discrete-time system Σ is by means of the transfer function. Given $\mathbf{H}(z) = \mathbf{D} + \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, the transfer function of the adjoint system is

$$\mathbf{H}_{\Sigma^*}(z) = \mathbf{H}^*(z^{-1}) = \mathbf{D}^* + \mathbf{B}^*(z^{-1}\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*.$$

A simple calculation shows that if $\mathbf{A}$ is invertible, the adjoint system has the following realization:

$$\Sigma^* = \left[\begin{array}{c|c} (\mathbf{A}^*)^{-1} & -(\mathbf{A}^*)^{-1}\mathbf{C}^* \\ \hline \mathbf{B}^*(\mathbf{A}^*)^{-1} & \mathbf{D}^* - \mathbf{B}^*\mathbf{C}^* \end{array}\right].$$

5.3 Computation of the 2-induced or $\mathcal{H}_\infty$-norm

According to (5.9), if Σ is stable, i.e., the eigenvalues of $\mathbf{A}$ have negative real parts, its 2-norm is

$$\boxed{\|\Sigma\|_2 = \|\mathcal{S}\|_{2\text{-ind}} = \|\mathbf{H}\|_{\mathcal{H}_\infty} = \sup_\omega \sigma_{max}(\mathbf{H}(i\omega)).} \tag{5.16}$$

The last equality follows from Proposition 5.2, which asserts that the 2-induced norm and the $\mathcal{H}_\infty$-norm of a stable Σ are the same. Consider the rational function

$$\Phi_\gamma(s) = \mathbf{I}_m - \frac{1}{\gamma^2}\mathbf{H}_\Sigma^*(-s)\mathbf{H}_\Sigma(s).$$

If the $\mathcal{H}_\infty$-norm of Σ is less than $\gamma > 0$, there is no real ω, such that $\Phi_\gamma(i\omega)$ is zero. Thus the $\mathcal{H}_\infty$-norm of Σ is less than γ if and only if $\Phi_\gamma^{-1}(s)$ has no pure imaginary poles.

Proposition 5.4. *Given* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, *let* $\mathbf{\Sigma}_{\Phi_\gamma} = \left(\begin{array}{c|c} \tilde{\mathbf{A}}(\gamma) & \tilde{\mathbf{B}}(\gamma) \\ \hline \tilde{\mathbf{C}}(\gamma) & \tilde{\mathbf{D}}(\gamma) \end{array}\right)$ *be a realization of* Φ_γ^{-1}. *The* $\mathcal{H}_\infty$*-norm of* $\mathbf{\Sigma}$ *is less than* γ *if and only if* $\tilde{\mathbf{A}}(\gamma)$ *has no eigenvalues on the imaginary axis. If* $\mathbf{D} = \mathbf{0}$,

$$\tilde{\mathbf{A}}(\gamma) = \begin{bmatrix} \mathbf{A} & \frac{1}{\gamma}\mathbf{B}\mathbf{B}^* \\ -\frac{1}{\gamma}\mathbf{C}^*\mathbf{C} & -\mathbf{A}^* \end{bmatrix} \in \mathbb{R}^{2n\times 2n}. \tag{5.17}$$

If $\mathbf{D} \neq \mathbf{0}$ *and* $\gamma^2\mathbf{I} - \mathbf{D}\mathbf{D}^*$ *is nonsingular,*

$$\tilde{\mathbf{A}}(\gamma) = \begin{bmatrix} \mathbf{F} & \mathbf{G} \\ -\mathbf{H} & -\mathbf{F}^* \end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \tag{5.18}$$

where $\mathbf{F} = \mathbf{A} + \gamma\mathbf{B}\mathbf{D}^*(\gamma^2\mathbf{I} - \mathbf{D}\mathbf{D}^*)^{-1}\mathbf{C}$, $\mathbf{G} = \gamma\mathbf{B}(\gamma^2\mathbf{I} - \mathbf{D}^*\mathbf{D})^{-1}\mathbf{B}^*$, $\mathbf{H} = \gamma\mathbf{C}^*(\gamma^2\mathbf{I} - \mathbf{D}\mathbf{D}^*)^{-1}\mathbf{C}$.

The proof of this proposition is left as an exercise (see Problem 30). Notice that $\tilde{\mathbf{A}}$ satisfies

$$\left(\tilde{\mathbf{A}}\mathbf{J}\right)^* = \tilde{\mathbf{A}}\mathbf{J}, \quad \text{where } \mathbf{J} = \begin{pmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}.$$

Such matrices are called *Hamiltonian matrices*. Therefore, the computation of the $\mathcal{H}_\infty$-norm requires the (repeated) computation of the eigenvalues of structured—in this case, Hamiltonian—matrices.

Using Proposition 5.4 to compute the $\mathcal{H}_\infty$-norm

The above fact can be used to compute the $\mathcal{H}_\infty$-norm of $\mathbf{\Sigma}$ to any given accuracy by means of the *bisection algorithm*. We proceed as follows. For a sufficiently large $\bar{\gamma}$, the Hamiltonian defined by (5.17) or (5.18) has no pure imaginary eigenvalues, while for a sufficiently small $\underline{\gamma}$, it does. The algorithm used to find an approximation of the $\mathcal{H}_\infty$-norm consists in bisecting the interval $[\underline{\gamma}, \bar{\gamma}]$: let $\tilde{\gamma} = (\underline{\gamma} + \bar{\gamma})/2$. If $\tilde{\mathbf{A}}(\tilde{\gamma})$ has imaginary eigenvalues, then the interval above is substituted by the interval, where $\underline{\gamma} = \tilde{\gamma}$; otherwise by $\bar{\gamma} = \tilde{\gamma}$. Both of these intervals now have half the length of the previous interval. The procedure continues until the difference $\bar{\gamma} - \underline{\gamma}$ is sufficiently small.

Example 5.5. Consider the continuous-time third-order Butterworth filter. The state-space matrices are

$$\mathbf{A} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \ \mathbf{B} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \ \mathbf{C} = [0 \ 0 \ 1], \ \mathbf{D} = 0,$$

which implies that $\Phi_\gamma(s) = 1 + \frac{1}{\gamma^2}\frac{1}{s^6-1}$ and the Hamiltonian

$$\tilde{\mathbf{A}}(\gamma) = \left[\begin{array}{ccc|ccc} -1 & 0 & 0 & \frac{1}{\gamma} & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & -\frac{1}{\gamma} & 0 & 1 & 0 \end{array}\right].$$

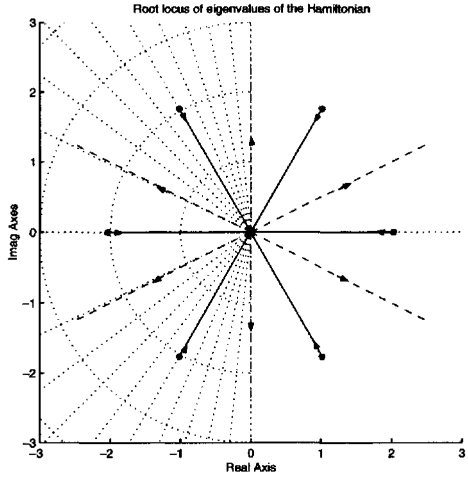


Figure 5.2. *Locus of the eigenvalues of the Hamiltonian as a function of* γ.

The locus of the eigenvalues of the Hamiltonian as a function of the parameter $\gamma \geq 0$ is shown in Figure 5.2. For $\gamma = \infty$, the six eigenvalues lie on the unit circle at the locations indicated. As γ decreases, the eigenvalues move in straight lines toward the origin; they meet at the origin for $\gamma = 1$. For smaller values of γ, the eigenvalues move again in straight lines away from the origin, with two of them traversing the imaginary axis in the positive, negative, direction, respectively. As γ approaches 0, the six eigenvalues go to infinity. Therefore, the $\mathcal{H}_\infty$-norm (which is the 2-norm of the associated convolution operator $\mathcal{S}$) is equal to 1. It is worth examining the sensitivity of the eigenvalues as a function of γ, especially close to the $\mathcal{H}_\infty$-norm. For $\gamma = 1$, the Hamiltonian $\tilde{\mathbf{A}}(1)$ has six zero eigenvalues, while the dimension of its null space is one. Therefore, $\tilde{\mathbf{A}}(1)$ is similar to a Jordan block of size six. Therefore, as detailed in Example 10.5, a perturbation in γ of the order 10^{-8} will result in a perturbation of the eigenvalues of the order 10^{-3}. Thus at the bifurcation point of the locus, we have very high sensitivity. To remedy this situation, a procedure for computing the infinity norm which bypasses this problem has been proposed in [75].

5.4 The Hankel operator and its spectra

Recall section 4.1 and in particular the definition of the convolution operator $\mathcal{S}$ (4.5). In this section, we first define the Hankel operator $\mathcal{H}$ which is obtained from $\mathcal{S}$ by restricting its domain and codomain. Then both the eigenvalues and the singular values of $\mathcal{H}$ are determined. A general reference on Hankel operators is [261].

Given a linear, time invariant, not necessarily causal, discrete-time system, the convolution operator $\mathcal{S}$ induces an operator of interest in the theory of linear systems—the Hankel operator $\mathcal{H}$, defined as follows:

$$\mathcal{H} : \ell_2^m(\mathbb{Z}_-) \longrightarrow \ell_2^p(\mathbb{Z}_+), \ \mathbf{u}_- \longmapsto \mathbf{y}_+, \tag{5.19}$$

$$\text{where } \ \mathbf{y}_+(t) = \mathcal{H}(\mathbf{u}_-)(t) = \sum_{k=-\infty}^{-1} \mathbf{h}_{t-k}\, \mathbf{u}_-(k), \qquad t \geq 0.$$

The matrix representation of $\mathcal{H}$ is given by the lower-left block of the matrix representation (4.4) of $\mathcal{S}$:

$$\begin{pmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \mathbf{y}(2) \\ \vdots \end{pmatrix} = \underbrace{\begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \mathbf{h}_3 & \cdots \\ \mathbf{h}_2 & \mathbf{h}_3 & \mathbf{h}_4 & \cdots \\ \mathbf{h}_3 & \mathbf{h}_4 & \mathbf{h}_5 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}}_{\mathcal{H}} \begin{pmatrix} \mathbf{u}(-1) \\ \mathbf{u}(-2) \\ \mathbf{u}(-3) \\ \vdots \end{pmatrix}.$$

Thus $\mathcal{H}$ turns out to be the same matrix as the Hankel matrix defined by (4.63). Since the Hankel operator of $\mathbf{\Sigma}$ maps *past* inputs into *future* outputs, we have

$$\mathbf{y}(k) = \sum_{j=-\infty}^{-1} \mathbf{C}\mathbf{A}^{(k-j+1)}\mathbf{B}\mathbf{u}(j) = \mathbf{C}\mathbf{A}^{k} \sum_{j=1}^{\infty} \mathbf{A}^{j-1}\mathbf{B}\mathbf{u}(-j),$$

which implies

$$\mathbf{y}(k) = \mathbf{C}\mathbf{A}^{k}\mathbf{x}(0), \ \mathbf{x}(0) = \sum_{j=1}^{\infty} \mathbf{A}^{j-1}\mathbf{B}\mathbf{u}(-j).$$

Briefly, the *properties* of this operator are that, because of (4.65), it has finite rank (at most n) and that the rank equals n if and only if $\mathbf{\Sigma}$ is reachable and observable. Finally, as shown in the next section, $\mathcal{H}$ has a finite set of nonzero singular values. Thus, because $\mathcal{H}$ has finite rank, its eigenvalues and singular values can be computed by means of finite-dimensional quantities, namely, the three gramians.

If $\mathcal{H}$ has domain $\ell_2^m(\mathbb{Z}_-)$ and codomain $\ell_2^p(\mathbb{Z}_+)$, its ℓ_2-induced norm is

$$\| \mathcal{H} \|_{\ell_2\text{-ind}} = \sigma_{max}(\mathcal{H}) = \| \mathbf{\Sigma} \|_H \ \le \ \| \mathbf{\Sigma} \|_{\ell_\infty} .$$

The quantity $\| \mathbf{\Sigma} \|_H$ is called the Hankel-norm of the system $\mathbf{\Sigma}$. If in addition the system is stable, by combining the discrete-time versions of Propositions 5.1 and 5.2, it follows that the ℓ_2-induced norm of $\mathcal{S}$ is equal to the h_∞-Schatten norm (3.6) of its transform, which is the transfer function $\mathbf{H}_{\mathbf{\Sigma}}$. In this case $\|\mathbf{\Sigma}\|_{\ell_\infty}$ above can be replaced by $\|\mathbf{\Sigma}\|_{h_\infty}$, and we often refer to this quantity as the h_∞-norm of $\mathbf{\Sigma}$.

Given a linear, time invariant, continuous-time, not necessarily causal system, similarly to the discrete-time case, the convolution operator $\mathcal{S}$ induces a Hankel operator $\mathcal{H}$, defined as follows:

$$\mathcal{H}: \mathcal{L}_2^m(\mathbb{R}_-) \longrightarrow \mathcal{L}_2^p(\mathbb{R}_+), \ \mathbf{u}_- \longmapsto \mathbf{y}_+, \tag{5.20}$$

$$\text{where} \ \ \mathbf{y}_+(t) = \mathcal{H}(\mathbf{u}_-)(t) = \int_{-\infty}^{0} \mathbf{h}(t-\tau)\mathbf{u}_-(\tau)\,d\tau, \qquad t \ge 0.$$

Sometimes this integral is written in the form $\mathbf{y}_+(t) = \int_t^\infty \mathbf{h}(\tau)\mathbf{u}_-(t-\tau)d\tau$. If $\mathcal{H}$ has domain $\mathcal{L}_2^m(\mathbb{R}_-)$ and codomain $\mathcal{L}_2^p(\mathbb{R}_+)$, its $\mathcal{L}_2$-induced norm is

$$\| \mathcal{H} \|_{\mathcal{L}_2\text{-ind}} = \sigma_{max}(\mathcal{H}) = \| \mathbf{\Sigma} \|_H \ \le \ \| \hat{\mathbf{\Sigma}} \|_{\mathcal{L}_\infty}. \tag{5.21}$$

The quantity $\| \mathbf{\Sigma} \|_H$ is called the Hankel-norm of the system $\mathbf{\Sigma}$. As in the discrete-time case, if the system is stable, by combining Propositions 5.1 and 5.2 it follows that the

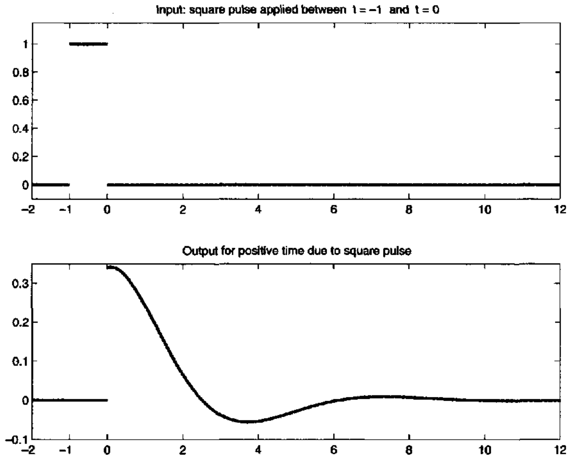


Figure 5.3. *The action of the Hankel operator: past inputs are mapped into future outputs. Top: square pulse input applied between $t = -1$ and $t = 0$. Bottom: resulting output considered for $t > 0$.*

$\mathcal{L}_2$-induced norm of the system defined by the kernel $\mathbf{h}$ is equal to the $\mathcal{H}_\infty$-Schatten norm of its transform, which is the transfer function $\mathbf{H}_\Sigma$, and we often refer to this quantity as the $\mathcal{H}_\infty$-norm of Σ. Figure 5.3 illustrates the action of the Hankel operator. A square pulse of unit amplitude is applied from $t = -1$ until $t = 0$ (upper plot), to the second-order system with transfer function $\mathbf{H}(s) = \frac{1}{s^2+s+1}$; the resulting response for $t > 0$ is depicted (lower plot).

Remark 5.4.1. *The Hankel operator and the Hankel matrix.* We have defined three objects that carry the name *Hankel* and are denoted, by abuse of notation, with the same symbol $\mathcal{H}$. The first is the Hankel *matrix* defined by (4.63) for both discrete- and continuous-time systems; the second is the Hankel *operator* (5.19), defined for *discrete-time* systems; and the third is the Hankel *operator* for *continuous-time* systems defined by (5.20). It turns out that the Hankel matrix is the matrix representation (in the canonical basis) of the Hankel operator for discrete-time systems. Furthermore, the Hankel matrix and the continuous-time Hankel operator are *not* related; for instance, whenever defined, the eigenvalues of the former are not the same as those of the latter.

5.4.1 Computation of the eigenvalues of $\mathcal{H}$

We begin by computing the eigenvalues of the Hankel operator; this problem makes sense only for square systems $m = p$. The first result, Lemma 5.6, asserts that, in this case, the (nonzero) eigenvalues of the Hankel operator are equal to the eigenvalues of the *cross gramian.* Recall that this concept was defined in section 4.3.2 for discrete-time systems, in which case the statement about the equality of eigenvalues follows in a straightforward way provided Σ is stable. Here we prove this result for continuous-time systems. The *cross* *gramian* $\mathcal{X}$ for continuous-time square systems $\mathbf{\Sigma}$ ($m = p$) is defined as the solution $\mathcal{X}$ of the Sylvester equation

$$\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{B}\mathbf{C} = \mathbf{0}. \tag{4.59}$$

If $\mathbf{A}$ is stable, it is well known that the solution of this equation can be written as

$$\mathcal{X} = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{C}e^{\mathbf{A}t}\,dt. \tag{4.60}$$

The next lemma states the desired property of the cross gramian.

Lemma 5.6. *For square minimal stable systems* $\mathbf{\Sigma}$, *the nonzero eigenvalues of the Hankel operator* $\mathcal{H}$ *are equal to the eigenvalues of the cross gramian* $\mathcal{X}$.

Proof. Recall that the Hankel operator maps past inputs into future outputs, namely,

$$\mathcal{H} : \mathbf{u}_- \to \mathbf{y}_+, \quad \mathbf{y}_+(t) = \mathcal{H}(\mathbf{u}_-)(t) = \int_{-\infty}^0 \mathbf{h}(t-\tau)\mathbf{u}(\tau)d\tau, \qquad t \geq 0,$$

where $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$, $t \geq 0$, is the impulse response of $\mathbf{\Sigma}$. The eigenvalue problem of $\mathcal{H}$ for square systems is $\mathcal{H}(\mathbf{u}_-) = \lambda\mathbf{y}_+$, where $\mathbf{y}_+(t) = \mathbf{u}_-(-t)$. Let the function $\mathbf{u}_-$ be an eigenfunction of $\mathcal{H}$. Then

$$\int_{-\infty}^0 \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}_-(\tau)d\tau = \lambda\mathbf{u}_-(-t) \Rightarrow \mathbf{C}e^{\mathbf{A}t}\int_{-\infty}^0 e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}_-(\tau)d\tau = \lambda\mathbf{u}_-(-t),$$

$$\int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{C}e^{\mathbf{A}t}dt \int_{-\infty}^0 e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}_-(\tau)d\tau = \lambda\int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{u}_-(-t)dt.$$

The first integral is equal to the cross gramian $\mathcal{X}$, and the second and the third are equal to the same constant vector, say, $\mathbf{v} \in \mathbb{R}^n$. We thus have $\mathcal{X}\mathbf{v} = \lambda\mathbf{v}$, which shows that if λ is a nonzero eigenvalue of $\mathcal{H}$, it is also an eigenvalue of $\mathcal{X}$. Conversely, let $(\lambda, \mathbf{v})$ be an eigenpair of $\mathcal{X}$, i.e., $\mathcal{X}\mathbf{v} = \lambda\mathbf{v}$:

$$\left[\int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\mathbf{C}e^{\mathbf{A}\tau}d\tau\right]\mathbf{v} = \lambda\,\mathbf{v} \Rightarrow \mathbf{C}e^{\mathbf{A}t}\int_{-\infty}^0 \left[e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{C}e^{-\mathbf{A}\tau}\right]\mathbf{v}\,d\tau = \lambda\,\mathbf{C}e^{\mathbf{A}t}\,\mathbf{v}$$

$$\Rightarrow \int_{-\infty}^0 \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}\underbrace{\mathbf{C}e^{-\mathbf{A}\tau}\mathbf{v}}_{\tilde{\mathbf{u}}(\tau)}\,d\tau = \lambda\mathbf{C}e^{\mathbf{A}t}\,\mathbf{v} = \lambda\tilde{\mathbf{u}}(-t) \Rightarrow \mathcal{H}(\tilde{\mathbf{u}})(t) = \lambda\tilde{\mathbf{u}}(-t), \quad t \geq 0.$$

Therefore $\tilde{\mathbf{u}}$ is an eigenfunction of $\mathcal{H}$. The proof is thus complete. □

5.4.2 Computation of the singular values of $\mathcal{H}$

Next, it is shown (Lemma 5.8) that the (nonzero) *singular values* of the Hankel operator are equal to the square roots of the eigenvalues of the product of the reachability and observability gramians. First, we need a definition.

Definition 5.7. *The* Hankel singular values *of the stable system* Σ, *denoted by*

$$\sigma_1(\Sigma) > \cdots > \sigma_q(\Sigma) \text{ each with multiplicity } m_i, \qquad i = 1, \dots, q, \ \sum_{i=1}^{q} m_i = n, \tag{5.22}$$

are the singular values of $\mathcal{H}_\Sigma$ *defined by* (5.19), (5.20). *The* Hankel-norm *of* Σ *is the largest Hankel singular value:*

$$\|\Sigma\|_H = \sigma_1(\Sigma).$$

The Hankel operator of a not necessarily stable system Σ *is defined as the Hankel operator of its stable and causal part* Σ_+: $\mathcal{H}_\Sigma = \mathcal{H}_{\Sigma_+}$.

The discrete-time case

We start by computing the singular values of the Hankel operator $\mathcal{H}$ defined by (5.19). Because of the factorization of $\mathcal{H}$ given in Lemma 4.39, we have $\mathcal{H} = \mathcal{OR}$, where $\mathcal{O}$, $\mathcal{R}$ are the infinite observability, reachability, matrices, respectively. Thus $\mathcal{H}^*\mathcal{H} = \mathcal{R}^*\mathcal{O}^*\mathcal{OR}$. Let $\mathbf{v}$ be an eigenvector of $\mathcal{H}^*\mathcal{H}$ corresponding to the eigenvalue σ^2; then

$$\mathcal{R}^*\mathcal{O}^*\mathcal{OR}\mathbf{v} = \sigma^2\mathbf{v} \Rightarrow \underbrace{\mathcal{RR}^*}_{\mathcal{P}}\underbrace{\mathcal{O}^*\mathcal{O}}_{\mathcal{Q}}\mathcal{R}\mathbf{v} = \sigma^2\mathcal{R}\mathbf{v}.$$

Recall that by (4.47), (4.48), $\mathcal{RR}^* = \mathcal{P}$ is the reachability and $\mathcal{O}^*\mathcal{O} = \mathcal{Q}$ is the observability gramian of the system. This implies that if $\mathbf{v}$ is an eigenvector of $\mathcal{H}^*\mathcal{H}$, then $\mathcal{R}\mathbf{v}$ is an eigenvector of the product of the two gramians $\mathcal{PQ}$. Conversely, let $\mathcal{PQ}\mathbf{w} = \sigma^2\mathbf{w}$; this implies

$$\mathcal{PQ}\mathbf{w} = \sigma^2\mathbf{w} \Rightarrow \mathcal{OPQ}\mathbf{w} = \sigma^2\mathcal{O}\mathbf{w} \Rightarrow \underbrace{\mathcal{OR}}_{\mathcal{H}}\underbrace{\mathcal{R}^*\mathcal{O}^*}_{\mathcal{H}^*}\mathcal{O}\mathbf{w} = \sigma^2\mathcal{O}\mathbf{w},$$

and therefore σ^2 is an eigenvalue of $\mathcal{HH}^*$ with corresponding eigenvector $\mathcal{O}\mathbf{w}$. We conclude that σ is a singular value of the Hankel operator $\mathcal{H}$ if and only if σ^2 is an eigenvalue of the product of gramians $\mathcal{PQ}$.

The continuous-time case

To compute the singular values of $\mathcal{H}$, we need its *adjoint* $\mathcal{H}^*$. Recall (5.20). For continuous-time systems, the adjoint is defined as follows:

$$\mathcal{H}^* : \mathcal{L}^p(\mathbb{R}_+) \longrightarrow \mathcal{L}^m(\mathbb{R}_-), \ \mathbf{y}_+ \longmapsto \mathbf{u}_- = \mathcal{H}^*(\mathbf{y}_+), \tag{5.23}$$

$$\text{where } \mathbf{u}_-(t) = \int_0^\infty \mathbf{h}^*(-t+\tau)\mathbf{y}(\tau)d\tau, \qquad t \leq 0.$$

In what follows, it is assumed that the underlying system is finite-dimensional, i.e., the rank of the Hankel matrix derived from the corresponding Markov parameters $\mathbf{h}_t$ of Σ is finite. Consequently, by subsection 4.4, there exists a triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ such that (4.20) holds:

$$\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}, \qquad t > 0.$$

It follows that

$$(\mathcal{H}\mathbf{u})(t) = \int_{-\infty}^{0} \mathbf{h}(t-\tau)\mathbf{u}(\tau)d\tau = \mathbf{C}e^{\mathbf{A}t}\underbrace{\int_{-\infty}^{0} e^{-\mathbf{A}\tau}\mathbf{B}u(\tau)d\tau}_{=\mathbf{x}_i} = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}_i, \qquad \mathbf{x}_i \in \mathbb{R}^n.$$

Moreover,

$$(\mathcal{H}^*\mathcal{H}\mathbf{u})(t) = \mathbf{B}^*e^{-\mathbf{A}^*t}\left(\int_0^{\infty} e^{\mathbf{A}^*\sigma}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}\sigma}d\sigma\right)\mathbf{x}_i = \mathbf{B}^*e^{-\mathbf{A}^*t}\mathcal{Q}\mathbf{x}_i,$$

where the expression in parentheses is $\mathcal{Q}$, the infinite observability gramian defined by (4.44). The requirement for $\mathbf{u}$ to be an eigenfunction of $\mathcal{H}^*\mathcal{H}$ is that this last expression be equal to $\sigma_i^2\mathbf{u}(t)$. This implies

$$\mathbf{u}(t) = \frac{1}{\sigma_i^2}\mathbf{B}^*e^{-\mathbf{A}^*t}\mathcal{Q}\mathbf{x}_i, \qquad t \leq 0.$$

Substituting $\mathbf{u}$ in the expression for $\mathbf{x}_i$, we obtain

$$\frac{1}{\sigma_i^2}\left(\int_{-\infty}^{0} e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{B}^*e^{-\mathbf{A}^*\tau}d\tau\right)\mathcal{Q}\mathbf{x}_i = \mathbf{x}_i.$$

Recall the definition (4.43) of the infinite reachability gramian; this equation becomes

$$\mathcal{P}\mathcal{Q}\mathbf{x}_i = \sigma_i^2\mathbf{x}_i.$$

Conversely, let $\mathcal{P}\mathcal{Q}\mathbf{x}_i = \sigma_i\mathbf{x}_i$. We will show that the function $\mathbf{y}_+(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}_i$, $t \geq 0$, is an eigenfunction of $\mathcal{H}\mathcal{H}^*$. To this end we multiply the above expression by $\mathbf{y}_+$ on both sides. Making use of the definition of the gramians $\mathcal{P}$ and $\mathcal{Q}$, the resulting expression is

$$\int_{-\infty}^{0} \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}\left[\int_0^{\infty} \mathbf{B}^*e^{\mathbf{A}^*(-\tau+\sigma)}\mathbf{C}^*\left(\mathbf{C}e^{\mathbf{A}\sigma}\mathbf{x}_i\right)d\sigma\right]d\tau = \sigma_i^2\,\mathbf{C}e^{\mathbf{A}t}\mathbf{x}_i,$$

which immediately implies the desired $\mathcal{H}\mathcal{H}^*\mathbf{y}_+ = \sigma_i^2\,\mathbf{y}_+$.

We conclude that the (nonzero) singular values of the Hankel operator $\mathcal{H}$ are the eigenvalues of the product of the infinite gramians $\mathcal{P}$ and $\mathcal{Q}$ of the system. Therefore $\mathcal{H}$, in contrast to $\mathcal{S}$, has a discrete set of singular values. It can be shown that (4.43) holds for discrete-time systems, where $\mathcal{P}$, $\mathcal{Q}$ are the infinite gramians obtained by solving the discrete Lyapunov equations (4.49), (4.50), respectively. In summary we have the following lemma.

Lemma 5.8. *Given a reachable, observable, and stable discrete- or continuous-time system $\mathbf{\Sigma}$ of dimension n, the* Hankel singular values *of $\mathbf{\Sigma}$ are equal to the positive square roots of the eigenvalues of the product of gramians $\mathcal{P}\mathcal{Q}$,*

$$\sigma_k(\mathbf{\Sigma}) = \sqrt{\lambda_k(\mathcal{P}\mathcal{Q})}, \qquad k = 1, \ldots, n. \tag{5.24}$$

Remark 5.4.2. **(a)** For *discrete-time systems*, $\sigma_k(\Sigma) = \sigma_k(\mathcal{H})$; i.e., the singular values of the system are the singular values of the (block) Hankel matrix defined by (4.63). For *continuous-time systems*, however, $\sigma_k(\Sigma)$ are the singular values of a continuous-time *Hankel operator*. They are not equal to the singular values of the associated matrix of Markov parameters.

(b) It should be noticed that following Proposition 4.35, the Hankel singular values of a continuous-time stable system and those of a discrete-time stable system related by means of the bilinear transformation $s = \frac{z-1}{z+1}$ are the *same*.

(c) If the system is not symmetric, the Hankel singular values σ_i and the singular values π_i of $\mathcal{X}$ satisfy the following majorization inequalities: $\sum_{i=1}^{k} \sigma_i \geq \sum_{i=1}^{k} \pi_i$ and $\sum_{i=1}^{k} \pi_{n-i+1} \geq \sum_{i=1}^{k} \sigma_{n-i+1}$, $i = 1, \ldots, n$. If the Hankel operator is symmetric, the following majorization inequalities hold: $\sum_{i=1}^{k} \pi_i$, $\sum_{i=1}^{k} \sigma_i \leq \sum_{i=1}^{k} \frac{\lambda_i(\mathcal{P})+\lambda_i(\mathcal{Q})}{2}$. Thus in the balanced basis (see Chapter 7), $\sum_{i=1}^{k} \sigma_i \geq \sum_{i=1}^{k} \pi_i$.

The Hankel singular values and the cross gramian

If we are dealing with a symmetric and stable system Σ, its Hankel singular values can be computed using only one gramian, namely, the cross gramian $\mathcal{X}$ defined in (4.59), (4.60).

Recall that according to Lemma 4.45, a symmetric system always possesses a realization that satisfies (4.75). Substituting these relations in equation (4.59), we obtain

$$\mathbf{A}\mathcal{X}+\mathcal{X}\mathbf{A}+\mathbf{B}\mathbf{C}=0 \;\Rightarrow\; \mathbf{A}\mathcal{X}\Psi+\mathcal{X}\mathbf{A}\Psi+\mathbf{B}\mathbf{C}\Psi=0 \;\Rightarrow\; \mathbf{A}\,[\mathcal{X}\Psi]+[\mathcal{X}\Psi]\,\mathbf{A}^*+\mathbf{B}\mathbf{B}^*=0.$$

Since the Lyapunov equation (4.45) has a unique solution, this implies that the reachability and the cross gramians are related in this case: $\mathcal{P} = \mathcal{X}\Psi$. Using a similar argument involving (4.46), we obtain $\mathcal{Q} = \Psi^{-1}\mathcal{X}$; we conclude that

$$\mathcal{P} = \mathcal{X}\Psi,\ \mathcal{Q} = \Psi^{-1}\mathcal{X} \;\Rightarrow\; \mathcal{X}^2 = \mathcal{P}\mathcal{Q}. \tag{5.25}$$

We have thus proved the following result, which also follows as a corollary of Lemma 5.6.

Proposition 5.9. *Given a symmetric minimal and stable system* $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$, *the reachability, observability, and cross gramians satisfy* (5.25)*. It follows that the eigenvalues of* $\mathcal{X}$ *are the eigenvalues of the associated Hankel operator* $\mathcal{H}$, *and the absolute values of these eigenvalues are the Hankel singular values of* $\mathcal{H}$.

Remark 5.4.3. Consider a square system $\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$, that is, $m = p$, and the associated cross gramian $\mathcal{X}$. Since the eigenvalues of $\mathcal{X}$ are the same as the eigenvalues of the Hankel operator $\mathcal{H}$, it follows that trace $\mathcal{H}$ = trace $\mathcal{X}$. Furthermore, according to [250] there holds $2\,\text{trace}\ \mathcal{H} = -\text{trace}\ \left(\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)$. We conclude, therefore, that twice the trace of the cross gramian is equal to minus the trace of the transfer function of Σ evaluated as $s = 0$. This fact can be seen directly by taking the trace of $\mathcal{X} + \mathbf{A}^{-1}\mathcal{X}\mathbf{A} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{C}$; the trace of the left-hand side is namely equal to $2\,\text{trace}\ \mathcal{X}$, while that of the right-hand side is equal to the desired $-\text{trace}\ \left(\mathbf{C}\mathbf{A}^{-1}\mathbf{B}\right)$.

5.4.3 The Hilbert–Schmidt norm

An operator $\mathcal{T} : \mathbb{X} \to \mathbb{Y}$, where $\mathbb{X}$, $\mathbb{Y}$ are Hilbert spaces, is a *Hilbert–Schmidt* operator if there exists a complete orthonormal sequence $\{\mathbf{w}_n\} \in \mathbb{X}$ such that $\sum_{n>0} \|\mathcal{T}(\mathbf{w}_n)\| < \infty$. This property can be readily checked for integral operators $\mathcal{T}$. The integral operator

$$\mathcal{T}(\mathbf{w})(x) = \int_a^b \mathbf{k}(x, y)\mathbf{w}(y)dy, \qquad x \in [c, d],$$

is a *Hilbert–Schmidt* operator if its kernel $\mathbf{k}$ is square integrable in both variables, if the expression

$$\kappa^2 = \text{trace} \left[\int_a^b \int_c^d \mathbf{k}^*(x, y)\mathbf{k}(x, y)dxdy\right],$$

is finite. In this case, κ is the Hilbert–Schmidt norm of $\mathcal{I}$. Following [365], such operators are compact and hence have a discrete spectrum, where each eigenvalue has finite multiplicity, and the only possible accumulation point is zero.

It readily follows that the convolution operator $\mathcal{S}$ defined above is not Hilbert–Schmidt, while the Hankel operator $\mathcal{H}$ is. In particular,

$$\kappa^2 = \text{trace} \left[\int_0^\infty \int_{-\infty}^0 \mathbf{h}^*(t-\tau)\mathbf{h}(t-\tau)\, d\tau\, dt\right].$$

Assuming that the system is stable, this expression is equal to

$$\begin{aligned}
\kappa^2 &= \text{trace} \int_0^\infty \int_{-\infty}^0 \mathbf{B}^* e^{\mathbf{A}^*(t-\tau)}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}\, d\tau\, dt \\
&= \text{trace} \int_0^\infty \mathbf{B}^* e^{\mathbf{A}^* t} \left[\int_{-\infty}^0 e^{-\mathbf{A}^*\tau}\mathbf{B}^*\mathbf{B}e^{-\mathbf{A}\tau}d\tau\right] e^{\mathbf{A}t}\mathbf{B}\, dt \\
&= \text{trace} \int_0^\infty \mathbf{B}^* e^{\mathbf{A}^* t}\, \mathcal{Q}\, e^{\mathbf{A}t}\mathbf{B}\, dt = \text{trace} \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^* t}\, \mathcal{Q}\, dt \\
&= \text{trace}\ [\mathcal{P}\mathcal{Q}] = \sigma_1^2 + \cdots + \sigma_n^2,
\end{aligned}$$

where σ_i are the Hankel singular values of the system $\mathbf{\Sigma}$.

The Hilbert–Schmidt norm for SISO linear systems was given an interpretation in [165]. The result asserts the following.

Proposition 5.10. *Given a SISO stable system, $\pi\kappa^2$ is equal to the area of the Nyquist plot of the associated transfer function, multiplicities included.*

An illustration of this result is given in Example 5.5.2. See also Problem 58.

Example 5.11. For the discrete-time system $y(k+1) - ay(k) = bu(k)$, $|a| < 1$, discussed earlier, the Hankel operator is

$$\mathcal{H} = b \begin{pmatrix} 1 & a & a^2 & \cdots \\ a & a^2 & a^3 & \cdots \\ a^2 & a^3 & a^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

This operator has a single nonzero singular value, which turns out to be

$$\sigma_1(\mathcal{H}) = \frac{|b|}{1-a^2}.$$

In this case, since $\mathcal{H}$ is symmetric, $\lambda_1(\mathcal{H}) = \pm\sigma_1(\mathcal{H})$. Furthermore, the Hilbert–Schmidt norm (which is equal to the trace of $\mathcal{H}$) is $\frac{|b|}{1-a^2}$.

5.4.4 The Hankel singular values of two related operators

Given a system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, consider the operator

$$\mathcal{K}: \mathcal{L}_2[0,T] \to \mathcal{L}_2[0,T],\ \mathcal{K}(\mathbf{u}) = \int_0^T \mathbf{h}(t-\tau)\mathbf{u}(\tau)d\tau.$$

The singular values of this operator can be determined as follows. Recall the definition of the Hamiltonian $\tilde{\mathbf{A}}(\gamma)$ defined by (5.17). It is known (e.g., [368], [363]) that $\gamma > 0$ is a singular value of $\mathcal{K}$ if and only if the determinant of the (2, 2) block of the matrix exponential of $\tilde{\mathbf{A}}(\gamma)T$ is zero:

$$\det[\exp(\tilde{\mathbf{A}}(\gamma)T)]_{22} = 0. \tag{5.26}$$

This problem arises as the computation of the optimal sensitivity for a pure delay:

$$\gamma_{\max} = \inf_{\mathbf{q}\in H_\infty} \|\hat{\mathbf{h}} - e^{-Ts}\mathbf{q}\|_\infty.$$

This was first solved in the above form in [368], with conjecture for the general case

$$\gamma_{\max} = \inf_{\mathbf{q}\in H_\infty} \|\hat{\mathbf{h}} - \mathbf{m}(s)\mathbf{q}\|_\infty,$$

where $\mathbf{m}$ is a general (real) inner function. The conjecture amounts to replacing $\exp(\tilde{\mathbf{A}}(\gamma)T)$ by $\mathbf{m}(-\tilde{\mathbf{A}}(\gamma))$. The conjecture was resolved first by [232] and later streamlined by [302], [122], [363] in a more generalized context; in particular, [363] gave a completely basis-free proof. For the generalization to the case $\mathbf{D} \neq 0$, see [302]. We also note that the result played a crucial role in the solution of the sampled-data $\mathcal{H}_\infty$ control problem [39].

A second instance of the existence of an exact formula for the Hankel singular values is given in [253]. This paper shows that the Hankel singular values of a system whose transfer function is given by the product of a scalar inner (stable and all-pass) function and of a rational MIMO function can be determined using an appropriately defined Hamiltonian.

Here are some details. Let $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{0} \end{array}\right)$ be given with transfer function $\mathbf{H}_\Sigma(s)$, together with $\phi(s)$, which is scalar and inner. The Hankel singular values of the system whose transfer function is $\phi(s)\mathbf{H}_\Sigma(s)$ are the solutions of the following transcendental equation $\Delta(\gamma) = 0$, where

$$\Delta(\gamma) = \det\left[\begin{pmatrix} \mathbf{I} & -\frac{1}{\gamma}\mathcal{P} \\ \mathbf{0} & 0 \end{pmatrix} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ -\frac{1}{\gamma}\mathcal{Q} & \mathbf{I} \end{pmatrix}\phi(-\tilde{\mathbf{A}}(\gamma))\right],$$

$\mathcal{P}$, $\mathcal{Q}$ are the gramians of Σ, and $\tilde{\mathbf{A}}(\gamma)$ is the associated Hamiltonian defined by (5.17).

Example 5.12 (see [254]). The transfer function from u to y for the delay differential equation

$$\dot{y}(t) = -y(t) + cy(t-1) + c\dot{y}(t-1) - cu(t) + u(t-1), \qquad |c| < 1,$$

is

$$\psi(s) = \phi(s)\mathbf{H}(s), \text{ where } \phi(s) = \frac{e^{-s} - c}{1 - ce^{-s}}, \quad \mathbf{H}(s) = \frac{1}{s+1}.$$

The singular values of the associated Hankel operator are characterized by $\Delta(\gamma) = 0$, where

$$\mathbf{F} = \begin{pmatrix} -1 & \frac{1}{\gamma} \\ -\frac{1}{\gamma} & 1 \end{pmatrix}, \quad \mathcal{P} = \frac{1}{2}, \quad \mathcal{Q} = \frac{1}{2}.$$

For this example $\phi(-\mathbf{F}) = \left(e^{\mathbf{F}} - c\mathbf{I}\right)\left(\mathbf{I} - ce^{\mathbf{F}}\right)^{-1}$. The bisection search on the interval $(0.07, 1)$ is employed for $c = 0.75$, which yields the Hankel singular values shown in Figure 5.4. The same figure shows also the determinant as a function of γ.

γ_1	0.9228
γ_2	0.4465
γ_3	0.1615
γ_4	0.1470
γ_5	0.0807
γ_6	0.0771
$\vdots$	$\vdots$

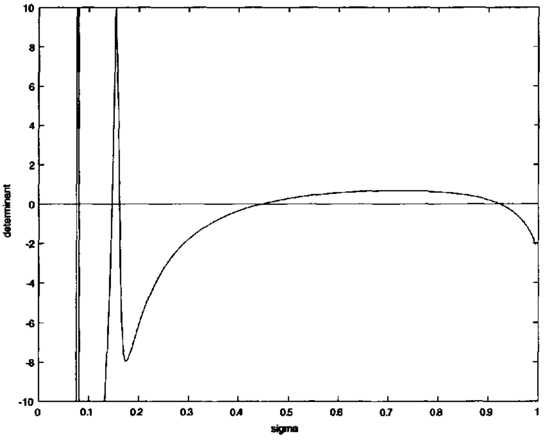


Figure 5.4. *The first six Hankel singular values (left). Plot of the determinant versus $0 \le \gamma \le 1$ (right). The zero crossings are the Hankel singular values.*

Since $\tilde{\mathbf{A}}(\gamma)$ is 2×2, it is possible to compute the determinant in terms of γ. By straightforward calculation, $\phi(\tilde{\mathbf{A}}(\gamma))$ is equal to

$$\frac{1}{\Omega}\begin{bmatrix} \left(1+c^2\right)\cos\omega - \frac{1}{\omega}\left(1-c^2\right)\sin\omega - 2c & \frac{1}{\gamma\omega}\left(1-c^2\right)\sin\omega \\ -\frac{1}{\gamma\omega}\left(1-c^2\right)\sin\omega & \left(1+c^2\right)\cos\omega + \frac{1}{\omega}\left(1-c^2\right)\sin\omega - 2c \end{bmatrix},$$

where $\Omega = 1 - 2c\cos\omega + c^2$ and $\omega = \frac{\sqrt{1-\gamma^2}}{\gamma}$; thus

$$\Delta = \frac{1}{\Omega}\left\{\frac{(1+c^2)}{4}\left(\omega^2 - 3\right)\cos\omega - \frac{(1-c^2)}{4}\frac{(1-3\omega^2)}{\omega}\sin\omega + \frac{c}{2}\left(3-\omega^2\right)\right\}.$$

The Newton iteration can be applied, which confirms the singular values in the table in Figure 5.4.

5.5 The $\mathcal{H}_2$-norm

5.5.1 A formula based on the gramians

The $\mathcal{H}_2$-*norm* of the continuous-time system $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ is defined as the $\mathcal{L}_2$-norm of the impulse response (in the time domain):

$$\| \boldsymbol{\Sigma} \|_{\mathcal{H}_2} = \| \mathbf{h} \|_{\mathcal{L}_2} . \tag{5.27}$$

This is similar for discrete-time systems. Therefore, this norm is bounded only if $\mathbf{D} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ is stable, i.e., the eigenvalues of $\mathbf{A}$ are in the open left half of the complex plane; in this case, there holds (5.9):

$$\| \boldsymbol{\Sigma} \|_{\mathcal{H}_2}^2 = \int_0^\infty \text{trace } [\mathbf{h}^*(t)\mathbf{h}(t)]\, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{trace } [\mathbf{H}^*(-i\omega)\mathbf{H}(i\omega)]\, d\omega,$$

where the second equality is a consequence of Parseval's theorem. Thus using (4.44) we obtain

$$\| \boldsymbol{\Sigma} \|_{\mathcal{H}_2}^2 = \int_0^\infty \text{trace } [\mathbf{B}^* e^{\mathbf{A}^* t}\mathbf{C}^*\mathbf{C}e^{\mathbf{A}t}\mathbf{B}] dt = \text{trace } [\mathbf{B}^*\mathcal{Q}\mathbf{B}].$$

Furthermore, since trace $[\mathbf{h}^*(t)\mathbf{h}(t)] = \text{trace}\,[\mathbf{h}(t)\mathbf{h}^*(t)]$, using (4.43), this last expression is also equal to trace $[\mathbf{C}\mathcal{P}\mathbf{C}^*]$; therefore

$$\boxed{\| \boldsymbol{\Sigma} \|_{\mathcal{H}_2} = \sqrt{\text{trace } [\mathbf{B}^*\mathcal{Q}\mathbf{B}]} = \sqrt{\text{trace } [\mathbf{C}\mathcal{P}\mathbf{C}^*]}\,.} \tag{5.28}$$

A question that arises is whether the $\mathcal{H}_2$-norm is *induced*. The question reduces to whether the Frobenius norm is induced. In [85] it is shown that the Frobenius norm is *not* induced (see also [99]). Furthermore, in [86] it is shown that the largest eigenvalue $\sqrt{\lambda_{\max}(\mathbf{C}\mathcal{P}\mathbf{C}^*)}$ is an induced norm. For details on the definition of this (nonequi-) induced norm, see section 5.6. Consequently, the $\mathcal{H}_2$-norm is an induced norm in only the single-input or the single-output ($m = 1$ or $p = 1$) case. This norm can be interpreted as the maximum amplitude of the output which results from finite energy input signals.

5.5.2 A formula based on the EVD of A

Continuous-time SISO

Given is a stable SISO system $\boldsymbol{\Sigma}$ with transfer function $\mathbf{H}(s)$. We first assume that all poles λ_i are distinct: $\lambda_i \neq \lambda_j$, $i = 1, \ldots, n$. Let c_i be the corresponding residues: $c_i = \mathbf{H}(s)(s - \lambda_i)|_{s=\lambda_i}$, $i = 1, \ldots, n$. Thus $\mathbf{H}(s) = \sum_{i=1}^n \frac{c_i}{s-\lambda_i}$. We will use the notation $\mathbf{H}(s)^* = \mathbf{H}^*(-s)$; if the coefficients of $\mathbf{H}$ are real, $\mathbf{H}(s)^* = \mathbf{H}(-s)$. The following result holds:

$$\boxed{\|\boldsymbol{\Sigma}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^n c_i\ \mathbf{H}(s)^*\Big|_{s=\lambda_i^*} = \sum_{i=1}^n c_i\ \mathbf{H}(-\lambda_i^*).} \tag{5.29}$$

Proof. The following realization for **H** holds:

$$\mathbf{A} = \operatorname{diag}[\lambda_1, \dots, \lambda_n],\ \mathbf{B} = [1 \ \cdots \ 1]^*,\ \mathbf{C} = [c_1 \ \cdots \ c_n].$$

Therefore, the reachability gramian is

$$\mathcal{P}_{i,j} = \frac{1}{\lambda_i + \lambda_j^*}, \qquad i, j = 1, \dots, n.$$

Thus the square of the $\mathcal{H}_2$-norm is $\mathbf{C}\mathcal{P}\mathbf{C}^*$, and the desired result follows. □

We now consider the case where **H** has a single real pole λ of multiplicity n:

$$\mathbf{H}(s) = \frac{c_n}{(s-\lambda)^n} + \cdots + \frac{c_1}{(s-\lambda)}.$$

In this case, a minimal realization of **H** is given by $\mathbf{A} = \lambda\mathbf{I}_n + \mathbf{J}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{J}_n$ is the matrix with ones on the superdiagonal and zeros elsewhere, $\mathbf{B} = [0 \ \cdots 0 \ \ 1]^*$ and $\mathbf{C} = [c_n \ \cdots \ c_1]$. Then

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_2}^2 = \sum_{k=1}^{n} c_k \frac{(-1)^{k-1}}{(k-1)!} \frac{d^{k-1}}{ds^{k-1}} \mathbf{H}(s)\bigg|_{s=-\lambda} = \sum_{k=1}^{n} \frac{c_k}{(k-1)!} \frac{d^{k-1}}{ds^{k-1}} \mathbf{H}(-s)\bigg|_{s=\lambda}.$$

Continuous-time MIMO

Assume that the matrix **A** is diagonalizable (i.e., multiplicities are allowed but the algebraic and geometric multiplicity of eigenvalues is the same). Thus let in the eigenvector basis:

$$\mathbf{A} = \operatorname{diag}(\lambda_1\mathbf{I}_1, \dots, \lambda_k\mathbf{I}_k),\ \ \mathbf{B} = [\mathbf{B}_1^* \ \mathbf{B}_2^* \ \cdots \ \mathbf{B}_k^*]^*,\ \ \mathbf{C} = [\mathbf{C}_1 \ \mathbf{C}_2 \ \cdots \ \mathbf{C}_k],$$

where $\mathbf{I}_j$ is the identity matrix, and $\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$. Then the $\mathcal{H}_2$-norm is given by the expression

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^{k} \operatorname{trace}\ \left[\mathbf{H}^*(-\lambda_i^*)\mathbf{B}_i^*\mathbf{C}_i^*\right].$$

Discrete-time SISO

In this case, we will consider systems whose transfer functions are strictly proper. The constant or nonproper part can be easily taken care of. Thus, let $\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \sum_{i=1}^{n} \frac{c_i}{z-\lambda_i}$. The $\mathcal{H}_2$-norm of this system is

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^{n} c_i \left[\frac{1}{z}\mathbf{H}\left(\frac{1}{z}\right)\right]\bigg|_{z=\lambda_i^*}.$$

We now consider the case where **H** has a single real pole λ of multiplicity n:

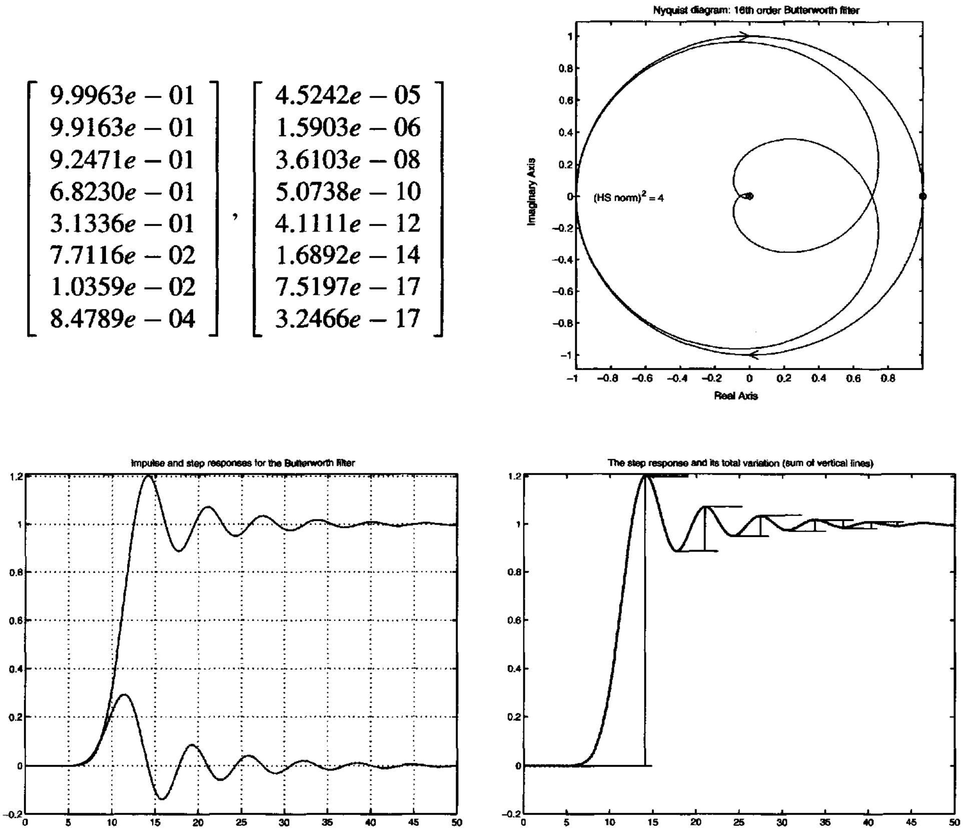
$$\mathbf{H}(z) = \frac{c_n}{(z-\lambda)^n} + \cdots + \frac{c_1}{(z-\lambda)}.$$

In this case (as before), a minimal realization of $\mathbf{H}$ is $\mathbf{A} = \lambda\mathbf{I}_n + \mathbf{J}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{J}_n$ is the matrix with ones on the superdiagonal and zeros elsewhere, $\mathbf{B} = [0 \cdots 0 \;\; 1]^*$ and $\mathbf{C} = [c_n \; \cdots \; c_1]$. Then

$$\|\boldsymbol{\Sigma}\|_{\mathcal{H}_2}^2 = \sum_{k=1}^{n} \frac{c_k}{(k-1)!} \frac{d^{k-1}}{ds^{k-1}} \left[\frac{1}{z} \mathbf{H}\left(\frac{1}{z}\right) \right] \Bigg|_{s=\lambda^*} .$$

An example

We consider a 16th-order, continuous-time, low-pass Butterworth filter. This is a linear dynamical system with low-pass frequency characteristics; in MATLAB it can be generated by using the command `butter(16,1,'s')`, where `16` is the order, `1` is the cutoff frequency, and `'s'` indicates that it is a continuous-time filter. Figure 5.5 shows the Hankel singular values, the impulse response, the step response, and the Nyquist diagram of this system.



$$\begin{bmatrix} 9.9963e-01 \\ 9.9163e-01 \\ 9.2471e-01 \\ 6.8230e-01 \\ 3.1336e-01 \\ 7.7116e-02 \\ 1.0359e-02 \\ 8.4789e-04 \end{bmatrix}, \quad \begin{bmatrix} 4.5242e-05 \\ 1.5903e-06 \\ 3.6103e-08 \\ 5.0738e-10 \\ 4.1111e-12 \\ 1.6892e-14 \\ 7.5197e-17 \\ 3.2466e-17 \end{bmatrix}$$

Figure 5.5. *Sixteenth-order Butterworth filter. Top left: Hankel singular values. Top right: Nyquist plot. Bottom left: impulse response and step response. Bottom right: total variation of step response.*

The various norms are as follows. The $\mathcal{H}_\infty$-norm is equal to 1; the $\mathcal{H}_2$-norm is 0.5646; since the system is SISO, this norm is equal to the 2, ∞-induced norm of the convolution operator. In other words, it is the maximum amplitude obtained by using unit energy (2-norm) input signals. The ∞, ∞-induced norm, also known as the *peak-to-peak gain*, is 2.1314; this is equal to the peak amplitude of the output obtained for some input signal of unit amplitude. This number is also equal to the *total variation* of the step response, i.e., the sum of all peak-to-valley variations of the step response. In this case, the 2-induced norm of the convolution and the Hankel operators is the same. The Frobenius norm of the Hankel operator (also known as the Hilbert–Schmidt norm) is equal to 4. Notice that the Nyquist diagram winds three times around 0, drawing almost perfect circles; this gives an area of 3π for the Nyquist diagram. The remaining area is actually equal to one circle; thus the total area is estimated to 4π, which verifies the result mentioned earlier.

5.6 Further induced norms of $\mathcal{S}$ and $\mathcal{H}^*$

Consider a vector-valued function of time $\mathbf{f} : \mathcal{I} \to \mathbb{R}^n$, where $\mathcal{I} = (a, b) \subset \mathbb{R}$. For a fixed $t = t_0$, $\mathbf{f}(t_0)$ is a vector in $\mathbb{R}^n$, and we can define its norm in one of the different ways indicated in (3.2):

$$\phi(t_0) = \|\mathbf{f}(t_0)\|_q.$$

ϕ in turn is a scalar (and nonnegative) function of time t, and we can define its p norm as follows:

$$\|\phi\|_p = \left[\int_a^b [\phi(t)]^p\, dt\right]^{\frac{1}{p}}.$$

Combining these two formulas, we get the (p, q) norm of $\mathbf{f}$, where p is the *temporal norm*, while q is the *spatial norm*. To distinguish this from the cases discussed, we will use a different notation:

$$|||\, \mathbf{f}\, |||_{(p,q)} = \left[\int_a^b \|\mathbf{f}(t)\|_q^p\, dt\right]^{\frac{1}{p}}. \tag{5.30}$$

Recall the Hölder inequality (3.3), valid for constant vectors. The generalized *Hölder inequality* valid for vector-valued time signals becomes

$$|\langle \mathbf{f}, \mathbf{g}\rangle| \le |||\, \mathbf{f}\, |||_{(p,q)} \cdot |||\, \mathbf{g}\, |||_{(\bar{p},\bar{q})}, \qquad p^{-1} + \bar{p}^{-1} = 1, \;\; q^{-1} + \bar{q}^{-1} = 1.$$

For a proof, see [86]. The (p, q) norm of a *discrete-time* vector-valued function can be defined similarly; if, namely, $\mathcal{I} \subset \mathbb{Z}$, the time integrals in the above expressions are replaced by (discrete) time sums. With these generalized definitions, (5.2) and (5.3) are (p, p) norms, that is, the spatial and temporal norms are taken to be p-norms. We can now define the corresponding (p, q) *Lebesgue spaces*:

$$\mathcal{L}_{(p,q)}(\mathcal{I}, \mathbb{R}^n) = \{\mathbf{f} : \mathcal{I} \to \mathbb{R}^n, \;\; |||\, \mathbf{f}\, |||_{(p,q)} < \infty\}. \tag{5.31}$$

Consider the map $\mathcal{T} : \mathcal{L}_{(p,q)} \to \mathcal{L}_{(r,s)}$; the associated *induced operator norm* is

$$|||\, \mathcal{T}\, |||_{(p,q)}^{(r,s)} = \sup_{\mathbf{u}\neq 0} \frac{|||\, \mathbf{y}\, |||_{(r,s)}}{|||\, \mathbf{u}\, |||_{(p,q)}}. \tag{5.32}$$

These induced operator norms are called *equi-induced* if $p = q = r = s$. Otherwise, they are called *mixed-induced norms*. Such norms are used extensively in control theory. Roughly speaking, given a system with *control* and *disturbance* inputs $\mathbf{v}, \mathbf{u}$, *to-be-controlled* and *measured* outputs $\mathbf{w}, \mathbf{y}$, the problem consists of finding a feedback control law mapping $\mathbf{y}$ to $\mathbf{u}$ such that the (convolution operator of the) closed-loop system $\mathcal{S}_{\text{cl}}$ from $\mathbf{v}$ to $\mathbf{w}$ is minimized in some appropriate sense. In the $\mathcal{H}_\infty$ control problem, $p = q = r = s = 2$, and the largest system energy gain from the disturbance to the to-be-controlled outputs is minimized. If instead of energy, one is interested in maximum amplitude minimization, $p = q = r = s = \infty$, and the resulting problem is the $\mathcal{L}_1$ optimal control problem. If $p = q = 2$ and $r = 2, \infty$ and $s = \infty$, we obtain the generalized $\mathcal{H}_2$ control problem. See [86] for a more complete description of such problems and appropriate references.

We conclude this section by quoting some results from [86] on induced norms of the convolution operator. The following notation is used. Given a matrix $\mathbf{M} \geq 0$, $\lambda_{\max}(\mathbf{M})$ denotes the largest eigenvalue of $\mathbf{M}$, while $\delta_{\max}(\mathbf{M})$ denotes the largest diagonal entry of $\mathbf{M}$.

Proposition 5.13. *Let* $\mathcal{S} : \mathcal{L}_{(p,q)}(\mathbb{R}_+, \mathbb{R}^m) \to \mathcal{L}_{(r,s)}(\mathbb{R}_+, \mathbb{R}^p)$, *and let* $\mathcal{H} : \mathcal{L}_{(p,q)}(\mathbb{R}_-, \mathbb{R}^m) \to \mathcal{L}_{(r,s)}(\mathbb{R}_+, \mathbb{R}^p)$. *The following four items are equi-induced norms in time and in space, in the domain and the range,* $p = q = r = s$:

- $\|\Sigma\|_2 = ||| \mathcal{S} |||_{(2,2)}^{(2,2)} = \sup_{\mathbf{u}\neq\mathbf{0}} \left[\frac{\int_0^\infty (\mathbf{y}_1^2(t)+\cdots+\mathbf{y}_p^2(t))\,dt}{\int_0^\infty (\mathbf{u}_1^2(t)+\cdots+\mathbf{u}_m^2(t))\,dt}\right]^{\frac{1}{2}} = \sup_{\omega\in\mathbb{R}} \sigma_{\max}(\mathbf{H}(i\omega))$,
- $\|\Sigma\|_H = ||| \mathcal{H} |||_{(2,2)}^{(2,2)} = \sup_{\mathbf{u}_-\neq\mathbf{0}} \left[\frac{\int_0^\infty (\|\mathbf{y}_+\|^2)\,dt}{\int_0^\infty \|\mathbf{u}_-\|^2\,dt}\right]^{\frac{1}{2}} = [\lambda_{\max}(\mathcal{P}\mathcal{Q})]^{\frac{1}{2}}$,
- $\|\Sigma\|_\infty = ||| \mathcal{S} |||_{(\infty,\infty)}^{(\infty,\infty)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\sup_i \max|\mathbf{y}_i|}{\sup_i \max|\mathbf{u}_i|} = \max_i \sum_j \eta_{ij}$, *where* $\eta_{ij} = \int_0^\infty |\mathbf{h}_{ij}(t)|\,dt$,
- $\|\Sigma\|_1 = ||| \mathcal{S} |||_{(1,1)}^{(1,1)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\int_0^\infty (|\mathbf{y}_1|+\cdots+|\mathbf{y}_p|)\,dt}{\int_0^\infty (|\mathbf{u}_1(t)|+\cdots+|\mathbf{u}_m(t)|)\,dt} = \max_j \sum_i \eta_{ij}$, *where* $\eta_{ij} = \int_0^\infty |\mathbf{h}_{ij}(t)|\,dt$.

The following are mixed-induced norms; they are equi-induced in time and in space, in the domain $p = q$, *and separately in time and in space in the range,* $r = s$:

- $\|\Sigma\|_{1,2} = ||| \mathcal{S} |||_{(1,1)}^{(2,2)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\left[\int_0^\infty (\mathbf{y}_1^2(t)+\cdots+\mathbf{y}_p^2(t))\,dt\right]^{\frac{1}{2}}}{\int_0^\infty (|\mathbf{u}_1(t)|+\cdots+|\mathbf{u}_m(t)|)\,dt} = [\delta_{\max}(\mathbf{B}^*\mathcal{Q}\mathbf{B})]^{\frac{1}{2}}$,
- $\|\Sigma\|_{2,\infty} = ||| \mathcal{S} |||_{(2,2)}^{(\infty,\infty)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\sup_{t\geq 0} \max_i |\mathbf{y}_i|}{\left[\int_0^\infty (\mathbf{u}_1^2(t)+\cdots+\mathbf{u}_m^2(t))\,dt\right]^{\frac{1}{2}}} = [\delta_{\max}(\mathbf{C}\mathcal{P}\mathbf{C}^*)]^{\frac{1}{2}}$.

The following are further mixed induced norms:

- $||| \mathcal{S} |||_{(1,2)}^{(2,2)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\left[\int_0^\infty (\mathbf{y}_1^2(t)+\cdots+\mathbf{y}_p^2(t))\,dt\right]^{\frac{1}{2}}}{\int_0^\infty \sqrt{\mathbf{u}_1^2(t)+\cdots+\mathbf{u}_m^2(t)}\,dt} = [\lambda_{\max}(\mathbf{B}^*\mathcal{Q}\mathbf{B})]^{\frac{1}{2}}$,
- $||| \mathcal{S} |||_{(2,2)}^{(\infty,2)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\sup_{t\geq 0} \sqrt{\mathbf{y}_1^2(t)+\cdots+\mathbf{y}_p^2(t)}}{\left[\int_0^\infty (\mathbf{u}_1^2(t)+\cdots+\mathbf{u}_m^2(t))\,dt\right]^{\frac{1}{2}}} = [\lambda_{\max}(\mathbf{C}\mathcal{P}\mathbf{C}^*)]^{\frac{1}{2}}$,
- $||| \mathcal{S} |||_{(1,2)}^{(\infty,2)} = \sup_{\mathbf{u}\neq\mathbf{0}} \frac{\sup_{t\geq 0} \sqrt{\mathbf{y}_1^2(t)+\cdots+\mathbf{y}_p^2(t)}}{\int_0^\infty \sqrt{\mathbf{u}_1^2(t)+\cdots+\mathbf{u}_m^2(t)}\,dt} = \sup_{t\geq 0} \sigma_{\max}(\mathbf{h}(t))$.

More generally, $||| \, \mathcal{S} \, |||^{(\infty,s)}_{(\infty,q)}$ *for* $q, s \in [1, \infty]$ *are* $\mathcal{L}_1$ *operator norms. Finally, the induced norms* $||| \cdot |||^{(2,2)}_{(1,q)}$, $||| \cdot |||^{(\infty,s)}_{(2,2)}$, $||| \cdot |||^{(\infty,s)}_{(1,q)}$, $q, s \in [1, \infty]$, *of the convolution operator* $\mathcal{S}$ *and of the Hankel operator* $\mathcal{H}$ *are equal.*

Notice that for SISO systems, the $\mathcal{L}_1$-induced and the $\mathcal{L}_\infty$-induced norms of the convolution operator $\mathcal{S}$ are the same and are equal to $\int_0^\infty \mid \mathbf{h} \mid dt$.

5.7 Summary of norms

Time-domain Lebesgue norms. Consider a vector valued function of time: $\mathbf{w} : \mathbb{R} \to \mathbb{R}^n$. The $\mathcal{L}_1$-, $\mathcal{L}_2$-, and $\mathcal{L}_\infty$-norms of $\mathbf{w}$ are defined as follows. It is assumed that both the *spatial* and the *temporal* norms are the same:

- $||| \, \mathbf{w} \, |||_{(1,1)} = \|\mathbf{w}\|_{\mathcal{L}_1} = \int_0^\infty \|\mathbf{w}(t)\| \, dt$, where $\|\mathbf{w}(t)\| = |\mathbf{w}_1(t)| + \cdots + |\mathbf{w}_n(t)|$.
- $||| \, \mathbf{w} \, |||_{(2,2)} = \|\mathbf{w}\|_{\mathcal{L}_2} = \sqrt{\int_0^\infty \|\mathbf{w}(t)\|^2 \, dt}$, where $\|\mathbf{w}(t)\|^2 = |\mathbf{w}_1(t)|^2 + \cdots + |\mathbf{w}_n(t)|^2$.
- $||| \, \mathbf{w} \, |||_{(\infty,\infty)} = \|\mathbf{w}\|_{\mathcal{L}_\infty} = \max_i \sup_t |\mathbf{w}_i(t)|$.

Frequency-domain Lebesgue norms. Consider a matrix valued function of $s = i\omega$: $\mathbf{H} : i\mathbb{R} \to \mathbb{C}^{p \times m}$ with $p \leq m$. The $\mathcal{L}_1$-, $\mathcal{L}_2$-, and $\mathcal{L}_\infty$-norms of $\mathbf{H}$ are defined as above. Notice that the *spatial* and the *frequency* norms are the same:

- $\|\mathbf{H}\|_{\mathcal{L}_1} = \int_0^\infty \|\mathbf{H}(i\omega)\| \, d\omega$, where $\|\mathbf{H}(i\omega)\| = \sigma_1(\mathbf{H}(i\omega)) + \cdots + \sigma_p(\mathbf{H}(i\omega))$.
- $\|\mathbf{H}\|_{\mathcal{L}_2} = \sqrt{\int_0^\infty \|\mathbf{H}(i\omega)\|^2 \, d\omega}$, where $\|\mathbf{H}(i\omega)\|^2 = \sigma_1^2(\mathbf{H}(i\omega)) + \cdots + \sigma_p^2(\mathbf{H}(i\omega))$.
- $\|\mathbf{H}\|_{\mathcal{L}_\infty} = \sup_\omega \|\mathbf{H}(i\omega)\|$, where $\|\mathbf{H}(i\omega)\| = \sigma_1(\mathbf{H}(i\omega))$.

Frequency-domain Hardy norms. Consider a matrix valued function $\mathbf{H} : \mathbb{C} \to \mathbb{C}^{p \times m}$, with $p \leq m$, of the complex variable $s = \sigma + i\omega \in \mathbb{C}$, that is, of the two real variables σ and ω. We assume that $\mathbf{H}$ is analytic in the closed right-half plane (RHP) (i.e., analytic for $\mathcal{R}e(s) \geq 0$). The $\mathcal{H}_1$-, $\mathcal{H}_2$-, and $\mathcal{H}_\infty$-norms of $\mathbf{H}$ are defined as follows:

- $\|\mathbf{H}\|_{\mathcal{H}_1} = \sup_{\sigma > 0} \int_0^\infty \|\mathbf{H}(\sigma + i\omega)\| \, d\omega$, where $\|\mathbf{H}(s)\| = \sigma_1(\mathbf{H}(s)) + \cdots + \sigma_p(\mathbf{H}(s))$.
- $\|\mathbf{H}\|_{\mathcal{H}_2} = \sqrt{\sup_{\sigma > 0} \int_0^\infty \|\mathbf{H}(\sigma + i\omega)\|^2 \, d\omega}$, where $\|\mathbf{H}(s)\|^2 = \sigma_1^2(\mathbf{H}(s)) + \cdots + \sigma_p^2(\mathbf{H}(s))$.
- $\|\mathbf{H}\|_{\mathcal{H}_\infty} = \sup_{\sigma > 0} [sup_\omega \|\mathbf{H}(\sigma + i\omega)\|]$, where $\|\mathbf{H}(s)\| = \sigma_1(\mathbf{H}(s))$.

As a consequence of the maximum modulus theorem, the Hardy norms are equal to the corresponding frequency domain Lebesgue norms:

$$\|\mathbf{H}\|_{\mathcal{H}_1} = \|\mathbf{H}\|_{\mathcal{L}_1}, \quad \|\mathbf{H}\|_{\mathcal{H}_2} = \|\mathbf{H}\|_{\mathcal{L}_2}, \quad \|\mathbf{H}\|_{\mathcal{H}_\infty} = \|\mathbf{H}\|_{\mathcal{L}_\infty}.$$

Induced and noninduced norms of $\boldsymbol{\Sigma}$. The expressions for the most common induced and noninduced norms of $\boldsymbol{\Sigma}$ are summarized. They involve the norms of the convolution operator $\mathcal{S}$: $\mathbf{u} \mapsto \mathbf{y} = \mathcal{S}(\mathbf{u}) = \mathbf{h} * \mathbf{u}$ and of the Hankel operator $\mathcal{H}$: $\mathbf{u}_- \mapsto \mathbf{y}_+ = \mathcal{H}(\mathbf{u}_-)$. We note that in case of induced norms, the *spatial* and the *temporal* norms are taken to be the same.

$r = s = 2$ $p = q = 2$	$\|\boldsymbol{\Sigma}\|_2 =$	$\|\mathcal{S}\|_{2\text{-ind}} = \sup_\omega \sigma_1(\mathbf{H}(i\omega))$ $= \|\mathbf{H}(s)\|_{\mathcal{H}_\infty}$	$\mathcal{L}_2$-norm of $\mathcal{S}$: sup *of freq. resp.*
$r = s = 2$ $p = q = 2$	$\|\boldsymbol{\Sigma}\|_H =$	$\|\mathcal{H}\|_{2,2\text{-ind}} = \|\boldsymbol{\Sigma}\|_{\mathrm{H}} = \sigma_1(\mathcal{H})$ $= \sqrt{\lambda_{\max}(\mathcal{PQ})}$	$\mathcal{L}_2$-norm of $\mathcal{H}$: *Hankel-norm*
$r = s = \infty$ $p = q = \infty$	$\|\boldsymbol{\Sigma}\|_\infty =$	$\|\mathcal{S}\|_{\infty,\infty\text{-ind}} = \max_i \sum_j \eta_{ij}$, $\eta_{ij} = \int_0^\infty \|\mathbf{h}_{ij}(t)\|\, dt$	$\mathcal{L}_1$-norm of $\mathcal{S}$: *peak row gain*
$r = s = 1$ $p = q = 1$	$\|\boldsymbol{\Sigma}\|_1 =$	$\|\mathcal{S}\|_{1,1\text{-ind}} = \max_j \sum_i \eta_{ij}$, $\eta_{ij} = \int_0^\infty \|\mathbf{h}_{ij}(t)\|\, dt$	$\mathcal{L}_\infty$-norm of $\mathcal{S}$: *peak column gain*
	$\|\boldsymbol{\Sigma}\|_{\mathcal{H}_2} =$	$\sqrt{\text{trace}\ [\mathbf{C}^*\mathcal{P}\mathbf{C}]}$ $= \sqrt{\text{trace}\ [\mathbf{B}\mathcal{Q}\mathbf{B}^*]}$	*RMS gain*

Let σ_i, $i = 1, \ldots, q$ be the distinct Hankel singular values of $\boldsymbol{\Sigma}$ (see (5.22)). The various induced norms of $\mathcal{S}$ and the Hankel singular values satisfy the following relationships:

- $\|\Sigma\|_H \leq \|\Sigma\|_2 \leq \|\Sigma\|_\infty$,
- $\sigma_1 \leq \|\Sigma\|_2 \leq 2(\sigma_1 + \cdots + \sigma_q)$,
- $\|\Sigma\|_\infty \leq 2(\sigma_1 + \cdots + \sigma_q)$,
- $\|\Sigma\|_\infty \leq 2n\, \|\Sigma\|_2$.

5.8 System stability

Having introduced norms for dynamical systems, we are now in a position to discuss the important concept of *stability*, followed by a brief account on the more general concept of *dissipativity*.

5.8.1 Review of system stability

Consider the *autonomous* or *closed* system (i.e., system having no *external* influences):

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t), \qquad \mathbf{A} \in \mathbb{R}^{n\times n}. \tag{5.33}$$

This system is called *stable* if all solution trajectories $\mathbf{x}$ are bounded for positive time: $\mathbf{x}(t)$ for $t > 0$ are bounded. The system is called *asymptotically stable* if all solution trajectories go to zero as time tends to infinity: $\mathbf{x}(t) \to 0$ for $t \to \infty$. Furthermore, the matrix $\mathbf{A}$ is called *Hurwitz* if its eigenvalues have negative real parts (belong to the left half of the complex plane (LHP)). The following holds.

Theorem 5.14. *The autonomous system* (5.33) *is*

- **asymptotically stable** *if and only if all eigenvalues of* $\mathbf{A}$ *have negative real parts, that is,* $\mathbf{A}$ *is Hurwitz.*
- **stable** *if and only if all eigenvalues of* $\mathbf{A}$ *have nonpositive real parts, and, in addition, all pure imaginary eigenvalues have multiplicity one.*

In what follows we identify *stability* with *asymptotic stability*. The *discrete-time* analogue of (5.33) is

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t), \qquad \mathbf{A} \in \mathbb{R}^{n \times n}. \tag{5.34}$$

This system is (asymptotically) stable if the eigenvalues of $\mathbf{A}$ have norm at most one: ($|\lambda(\mathbf{A})| < 1$) $|\lambda(\mathbf{A})| \leq 1$; in the latter case any eigenvalues on the unit circle must have multiplicity one. Here $\mathbf{A}$ is called a *Schur* matrix. The eigenvalues of $\mathbf{A}$ are often referred to as the *poles* of (5.33), (5.34), respectively. Thus we have the well-known statement:

$$\boxed{\textbf{stability} \ \Leftrightarrow \ \textbf{poles in LHP/unit disc.}}$$

We now turn our attention to *forced* systems $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, that is, to systems with inputs and outputs:

$$\mathbf{\Sigma}: \ \frac{d}{dt}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t),$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{C} \in \mathbb{R}^{p \times n}$, $\mathbf{D} \in \mathbb{R}^{p \times m}$. This system is called *internally stable* if $\mathbf{u}(t) = 0$, for $t > 0$, implies $\mathbf{x}(t) \to 0$, for $t \to \infty$. In other words, internal stability of a forced system is equivalent to *zero-input* stability.

Let us now consider the input-output (i.e., the convolution) representation of forced systems: $\mathbf{y} = \mathbf{h} * \mathbf{u}$, where $\mathbf{h}$ is the *impulse response* of the system, assumed causal ($\mathbf{h}(t) = 0$, $t < 0$). We want to address the question of what is meant by stability of such a system, in particular, what properties of $\mathbf{h}$ guarantee stability of the system; moreover, we also want to ask the following question: if the system with impulse response $\mathbf{h}$ is realizable, what is the relation between the stability of the external representation and that of the realization?

The system $\mathbf{\Sigma}$ described by the convolution integral $\mathbf{y}(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\tau)\mathbf{u}(\tau)d\tau$ is *bounded-input, bounded-output* (BIBO) *stable* if any bounded input $\mathbf{u}$ results in a bounded output $\mathbf{y}$:

$$\mathbf{u} \in \mathcal{L}_{\infty}^{m}(\mathbb{R}) \ \Rightarrow \ \mathbf{y} \in \mathcal{L}_{\infty}^{p}(\mathbb{R}).$$

To avoid ambiguity, we stress that both the spatial and the temporal norms are the infinity norms (in the domain and the codomain). The following result holds.

Theorem 5.15. *The system described by the convolution integral given above is BIBO stable if and only if the* $\mathcal{L}_{\infty}$*-induced norm of the convolution operator is finite.*

According to the previous section, this means that the (i, j)th entry $\mathbf{h}_{ij}$ of the $p \times m$ impulse response $\mathbf{h}$ must be absolutely integrable,

$$\int_{0}^{\infty} |\ \mathbf{h}_{ij}(t)\ |\ dt < \infty, \qquad i = 1, \ldots, p, \ j = 1, \ldots, m.$$

Closely related to BIBO stability is $\mathcal{L}_2$ stability. The convolution system $\mathbf{\Sigma}$ mentioned above is said to be $\mathcal{L}_2$ *input-output stable* if all square integrable inputs $\mathbf{u} \in \mathcal{L}_2^m(\mathbb{R})$ produce square integrable outputs $\mathbf{y} \in \mathcal{L}_2^p(\mathbb{R})$.

Theorem 5.16. *The convolution system $\mathbf{\Sigma}$ is $\mathcal{L}_2$ input-output stable if and only if the $\mathcal{L}_2$-induced norm of Σ is finite, which is, in turn, equivalent to the finiteness of the $\mathcal{H}_\infty$-norm of the associated transfer function* $\mathbf{H}$.

In a similar way, one can define $\mathcal{L}_p$ *input-output* stability. In this respect, the following holds.

Proposition 5.17. *The finiteness of the $\mathcal{L}_1$-norm of the impulse response* $\mathbf{h}$ *implies the $\mathcal{L}_p$ input-output stability of Σ for $1 \leq p \leq \infty$.*

We conclude this section with a result that relates internal and input-output stability. For this it is assumed that $\mathbf{h}$ is *realizable*, that is, there exist $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{p\times n}$, with n finite, such that $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$. The following holds.

Theorem 5.18. *Given a system $\mathbf{\Sigma}$ and a finite-dimensional realization $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{0} \end{array}\right)$ thereof, the following statements are equivalent:*

1. $\|\mathbf{\Sigma}\|_1 < \infty$.
2. $\|\mathbf{\Sigma}\|_2 < \infty$.
3. There exists a realization of $\mathbf{\Sigma}$, with $\mathbf{A}$ Hurwitz.
4. Every minimal realization of $\mathbf{\Sigma}$ has $\mathbf{A}$ Hurwitz.

This theorem asserts that *internal stability* implies *external stability* and that, conversely, *external stability* together with *minimality* implies *internal stability*. For instance, the system

$$\frac{d}{dt}\mathbf{x}(t) = \begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}\mathbf{x}(t) + \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}\mathbf{u}(t), \qquad \mathbf{y}(t) = \begin{pmatrix} 0 & 2 \end{pmatrix}\mathbf{x}(t),$$

with impulse response $\mathbf{h}(t) = e^{-t}$, $t \geq 0$, while BIBO stable, is not asymptotically stable.

5.8.2 Lyapunov stability

Consider the (autonomous) system described by

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t)), \qquad \mathbf{x}(t) \in \mathbb{R}^n.$$

The function $\Theta : \mathbb{R}^n \to \mathbb{R}$ is said to be a *Lyapunov function* for the above system if for all solutions $\mathbf{x} : \mathbb{R} \to \mathbb{R}^n$, there holds

$$\boxed{\frac{d}{dt}\Theta(\mathbf{x}(t)) \leq 0.} \tag{5.35}$$

Let $\nabla\Theta(\mathbf{x})$ be the gradient of Θ, which is a row vector whose ith entry is $\frac{\partial\Theta}{\partial \mathbf{x}_i}$. With this notation $\frac{d}{dt}\Theta(\mathbf{x}(t)) = \nabla\Theta(\mathbf{x}(t))\cdot\mathbf{f}(\mathbf{x}(t))$, and hence the above inequality can also be written as

$$\nabla\Theta\cdot\mathbf{f}\leq 0.$$

In the case of *linear systems* (5.33) and *quadratic* Lyapunov functions

$$\Theta(\mathbf{x}) = \mathbf{x}^*\mathbf{P}\mathbf{x},\ \ \mathbf{P}=\mathbf{P}^*\in\mathbb{R}^{n\times n},$$

it follows that

$$\frac{d}{dt}\Theta = \left[\frac{d}{dt}\mathbf{x}^*\right]\mathbf{P}\mathbf{x}+\mathbf{x}^*\mathbf{P}\left[\frac{d}{dt}\mathbf{x}\right] = \mathbf{x}^*\underbrace{\left[\mathbf{A}^*\mathbf{P}+\mathbf{P}\mathbf{A}\right]}_{\mathbf{Q}}\mathbf{x} = \mathbf{x}^*\mathbf{Q}\mathbf{x}.$$

Usually $\mathbf{Q}=\mathbf{Q}^*$ is given. Thus the problem of constructing a Lyapunov function amounts to solving for $\mathbf{P}$ the *Lyapunov equation*:

$$\mathbf{A}^*\mathbf{P}+\mathbf{P}\mathbf{A}=\mathbf{Q}. \tag{5.36}$$

The *discrete-time* analogue of this equation is

$$\mathbf{A}^*\mathbf{P}\mathbf{A}+\mathbf{P}=\mathbf{Q}, \tag{5.37}$$

known as the *discrete-time Lyapunov* or *Stein equation*. The next result summarizes some properties that are discussed in section 6.

Theorem 5.19. *If* $\mathbf{A}$ *is Hurwitz, then for all* $\mathbf{Q}\in\mathbb{R}^{n\times n}$ *there exists a unique* $\mathbf{P}\in\mathbb{R}^{n\times n}$ *that satisfies* (5.36). *Moreover, if* $\mathbf{Q}=\mathbf{Q}^*$, *then* $\mathbf{P}=\mathbf{P}^*$; *and if* $\mathbf{Q}=\mathbf{Q}^*\leq 0$ *and* $(\mathbf{A},\mathbf{Q})$ *is observable, then* $\mathbf{P}=\mathbf{P}^*>0$. *Conversely, if* $\mathbf{Q}=\mathbf{Q}^*\leq 0$ *and* $\mathbf{P}=\mathbf{P}^*>0$ *satisfy* (5.36), *then* (5.33) *is stable. If in addition the pair* $(\mathbf{A},\mathbf{Q})$ *is observable, then* (5.33) *is asymptotically stable.*

Example 5.20. (a) Consider a mass-spring-damper system attached to a wall. Let $\mathbf{x}$, $\dot{\mathbf{x}}$ be the position and velocity of the mass. The following equation holds: $m\ddot{\mathbf{x}}+c\dot{\mathbf{x}}+k\mathbf{x}=0$. This can be written as a first-order system,

$$\dot{\mathbf{w}}=\mathbf{A}\mathbf{w},\ \text{ where }\ \mathbf{A}=\begin{pmatrix}0 & 1\\ -\frac{k}{m} & -\frac{c}{m}\end{pmatrix}\ \text{ and }\ \mathbf{w}=\begin{pmatrix}\mathbf{x}\\ \dot{\mathbf{x}}\end{pmatrix}.$$

We postulate the quadratic Lyapunov function $\mathbf{V}=\frac{1}{2}\mathbf{w}^*\mathbf{P}\mathbf{w}$, where $\mathbf{P}=\begin{pmatrix}k & 0\\ 0 & m\end{pmatrix}$; that is, $\mathbf{V}=\frac{1}{2}k\mathbf{x}^2+\frac{1}{2}m\dot{\mathbf{x}}^2$ is the sum of the potential energy stored in the spring plus the kinetic energy of the mass, and hence always positive, whenever the system is not at rest. According to the above considerations, to prove stability we have to compute $\mathbf{A}^*\mathbf{P}+\mathbf{P}\mathbf{A}=\mathbf{Q}$. It turns out that $\mathbf{Q}=-\begin{pmatrix}0 & 0\\ 0 & c\end{pmatrix}\leq\mathbf{0}$. The negative semidefiniteness of $\mathbf{Q}$ proves stability; asymptotic stability follows from the additional fact that the pair $(\mathbf{A},\mathbf{Q})$ is observable.

(b) We now consider the system in Example 4.3 with $\mathbf{u}=0$ (i.e., the input terminal is short-circuited). The differential equation describing this system is $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, where

$\mathbf{x}^* = [\mathbf{x}_1, \ \mathbf{x}_2]$, $\mathbf{x}_1$ is the current through the inductor, and $\mathbf{x}_2$ is the voltage across the capacitor. Eliminating $\mathbf{x}_2$, we get $L\ddot{\mathbf{x}}_1 + R\dot{\mathbf{x}}_1 + \frac{1}{C}\mathbf{x}_1 = 0$. As the Lyapunov function, we will take the energy stored in this circuit, namely, in the inductor and the capacitor $\mathbf{V} = \frac{1}{2}L\mathbf{x}_1^2 + \frac{1}{2}C\mathbf{x}_2^2$, that is, $\mathbf{V} = \mathbf{x}^*\mathbf{P}\mathbf{x}$, $\mathbf{P} = \frac{1}{2}\begin{pmatrix} L & 0 \\ 0 & C \end{pmatrix}$. Thus $\dot{\mathbf{V}} = \mathbf{x}^*\mathbf{Q}\mathbf{x}$, where $\mathbf{Q} = \mathbf{A}^*\mathbf{P} + \mathbf{P}\mathbf{A} = -\begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}0$. Stability follows from the negative semidefiniteness of $\mathbf{Q}$ and asymptotic stability from the additional fact that $(\mathbf{A}, \mathbf{Q})$ is an observable pair.

5.8.3 $\mathcal{L}_2$-systems and norms of unstable systems

We conclude this section with a brief account on $\mathcal{L}_2$-systems defined on the whole real line $\mathbb{R}$. These are systems that produce $\mathcal{L}_2$ trajectories on $\mathbb{R}$ when fed with trajectories that are $\mathcal{L}_2$ on $\mathbb{R}$. When all poles of the system are in the left half of the complex plane, this property is automatic. However, if the poles are in both the left as well as the right half of the complex plane, but not on the imaginary axis, by giving up *causality* we can define an $\mathcal{L}_2$-system. Then, *all-pass* $\mathcal{L}_2$-systems are defined and characterized in terms of the transfer function and a state representation. Here are the details.

Given is the linear, continuous-time system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ defined for all time $t \in \mathbb{R}$; it is assumed that $\mathbf{\Sigma}$ is at rest in the distant past: $\mathbf{x}(-\infty) = 0$. Its impulse response and transfer function are $\mathbf{h}_\Sigma(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \mathbf{D}\delta(t)$, $t \geq 0$ $\mathbf{h}_\Sigma = 0$, $t < 0$, $\mathbf{H}_\Sigma(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$. The impulse response $\mathbf{h}_\Sigma$ as defined above is not square integrable on $\mathbb{R}$ unless the eigenvalues of $\mathbf{A}$ lie in the left half of the complex plane. Therefore, the associated convolution operator $\mathcal{S}_\Sigma$ defined by (4.5) with $\mathbf{h}_\Sigma$ as defined above does *not* map $\mathcal{L}_2$ inputs $\mathbf{u}$ into $\mathcal{L}_2$ outputs $\mathbf{y}$ unless $\mathbf{A}$ is stable. The system can nevertheless be interpreted as an $\mathcal{L}_2$-system on the whole real line. The formulas below hold for continuous-time systems; the corresponding ones for discrete-time systems follow in a similar way and are omitted.

Let $\mathbf{A}$ have no eigenvalues on the imaginary axis, that is, $\mathcal{R}e\lambda_i(\mathbf{A}) \neq 0$. After basis change, $\mathbf{A}$ can be written in block diagonal form:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_+ & \\ & \mathbf{A}_- \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_+ \\ \mathbf{B}_- \end{pmatrix}, \quad \mathbf{C} = (\mathbf{C}_+ \ \ \mathbf{C}_-),$$

where $\mathcal{R}e\lambda_i(\mathbf{A}_+) < 0$ and $\mathcal{R}e\lambda_i(\mathbf{A}_-) > 0$. We now redefine the impulse response $\mathbf{h}$ as follows:

$$\mathbf{h}_{\mathcal{L}_2}(t) = \begin{cases} \mathbf{h}_+(t) = \mathbf{C}_+e^{\mathbf{A}_+t}\mathbf{B}_+, & t > 0, \\ \mathbf{h}_-(t) = \mathbf{C}_-e^{\mathbf{A}_-t}\mathbf{B}_-, & t \leq 0. \end{cases}$$

The associated system operator $\mathcal{S}_{\mathcal{L}_2}$ defines a map from $\mathcal{L}_2^m(\mathbb{R})$ to $\mathcal{L}_2^p(\mathbb{R})$, and the associated Hankel operator $\mathcal{H}_{\mathcal{L}_2}$ defines a map from $\mathcal{L}_2^m(\mathbb{R}_-)$ to $\mathcal{L}_2^p(\mathbb{R}_+)$; they are defined as follows:

$$\mathcal{S}_{\mathcal{L}_2} : \mathbf{u} \longmapsto \mathbf{y} = \mathcal{S}_{\mathcal{L}_2}\mathbf{u}, \ \ \mathbf{y}(t) = (\mathcal{S}_{\mathcal{L}_2}\mathbf{u})(t) = \mathbf{D}\mathbf{u}(t) + \int_{-\infty}^{\infty} \mathbf{h}_{\mathcal{L}_2}(t-\tau)\mathbf{u}(\tau)d\tau,$$

$$\mathcal{H}_{\mathcal{L}_2} : \mathbf{u}_- \longmapsto \mathbf{y}_+ = \mathcal{H}_{\mathcal{L}_2}\mathbf{u}_-, \ \ \mathbf{y}_+(t) = (\mathcal{H}_{\mathcal{L}_2}\mathbf{u}_-)(t) = \int_{-\infty}^{0} \mathbf{h}_+(t-\tau)\mathbf{u}(\tau)d\tau.$$

An important property of the $\mathcal{L}_2$-system is that it has the same transfer function as the original system: $\mathbf{H}_{\mathcal{L}_2}(s) = \mathbf{H}(s)$.

The above considerations allow the definition of the 2-norm and of the Hankel-norm of unstable systems. The former is done by means of $\mathbf{h}_{\mathcal{L}_2}$ and the latter by means of $\mathbf{h}_+$. Norms of unstable systems play an important role in the theory of Hankel-norm approximation presented in Chapter 8. Here is the precise statement.

Norms of unstable systems. Given an unstable system with no poles on the imaginary axis (unit circle), its 2-norm is defined as the 2-induced norm of the convolution operator of the associated $\mathcal{L}_2$- (ℓ_2) system. This is equal to the supremum of the largest singular value of the transfer function on the imaginary axis (the unit circle) and is know as its $\mathcal{L}_\infty$- (ℓ_∞-) norm.

$$\|\boldsymbol{\Sigma}\|_2 = \|\mathcal{S}_{\mathcal{L}_2}\|_{2-\text{ind}} = \sup_{\omega} \sigma_{\max}(\mathbf{H}(i\omega)) = \|\mathbf{H}(s)\|_{\mathcal{L}_\infty}. \tag{5.38}$$

The Hankel-norm of an unstable $\boldsymbol{\Sigma}$ is equal to the Hankel-norm of its stable subsystem $\boldsymbol{\Sigma}_+$,

$$\|\boldsymbol{\Sigma}\|_H = \|\boldsymbol{\Sigma}_+\|_H, \tag{5.39}$$

and there holds $\| \ \boldsymbol{\Sigma} \ \|_2 \geq \| \ \boldsymbol{\Sigma} \ \|_H$.

Remark 5.8.1. The steps presented in Proposition 5.4 for the computation of the $\mathcal{H}_\infty$-norm of a stable system by means of the Hamiltonian matrix $\tilde{\mathbf{A}}(\gamma)$ can also be applied if the eigenvalues of $\mathbf{A}$ are not necessarily in the left half of the complex plane. In this case, the result of the computation provides the $\mathcal{L}_\infty$-norm of the transfer function $\mathbf{H}$, i.e., according to (5.38), the 2-norm of the unstable system $\boldsymbol{\Sigma}$.

Example 5.21. We will consider a simple second-order discrete-time system with one stable and one unstable pole:

$$\boldsymbol{\Sigma}: \ \mathbf{y}(k+2) - (a+b)\mathbf{y}(k+1) + ab\mathbf{y}(k) = (a-b)\mathbf{u}(k+1), \qquad |a| < 1, \ |b| > 1.$$

The transfer function of $\boldsymbol{\Sigma}$ is

$$\mathbf{H}(z) = \frac{(a-b)z}{(z-a)(z-b)} = \frac{a}{z+a} - \frac{b}{z+b}.$$

The impulse response of this system is $\mathbf{h}(0) = 0$, $\mathbf{h}(k) = a^k - b^k$, $k > 0$. The $\mathcal{H}_2$-norm of the system is the ℓ_2-norm of $\mathbf{h}$ which, since $|b| > 1$, is infinite. The 2-induced norm of the convolution operator is the largest singular value of

$$\mathcal{S} = \left(\begin{array}{cccc|ccc}
\ddots & \vdots & \vdots & & \vdots & \vdots & \\
\cdots & 0 & 0 & & 0 & 0 & \cdots \\
\cdots & a-b & 0 & & 0 & 0 & \cdots \\
\hline
\cdots & a^2-b^2 & a-b & & 0 & 0 & \cdots \\
\cdots & a^3-b^3 & a^2-b^2 & & a-b & 0 & \cdots \\
 & \vdots & \vdots & & \vdots & \vdots & \ddots
\end{array}\right),$$

which is also infinite. As indicated above, these norms are defined as the corresponding norms of the associated ℓ_2-system,

$$\boldsymbol{\Sigma}_{\ell_2}: \ \mathbf{y}(k+1) - a\mathbf{y}(k) = a\mathbf{u}(k+1), \ k = 1, 2, \ldots, \ \ \mathbf{y}(k) - b^{-1}\mathbf{y}(k+1) = \mathbf{u}(k),$$
$$k = 0, -1, -2, \ldots.$$

In this case, the $\mathcal{H}_2$-norm is the square root of the quantity $1 + \sum_{j=1}^{\infty}(a^{2j} + b^{-2j}) = \frac{b^2-a^2}{(1-a^2)(b^2-1)}$. Let the circulant matrix associated with the convolution operator, defined earlier, be

$$\mathcal{S}_{2n+1} = \begin{pmatrix} b^{-n} & a^n & \cdots & a & \cdots & b^{-n+2} & b^{-n+1} \\ b^{-n+1} & b^{-n} & \cdots & a^2 & \cdots & b^{-n+3} & b^{-n+2} \\ \vdots & \vdots & \ddots & & & \vdots & \vdots \\ 1 & b^{-1} & & b^{-n} & & a^2 & a \\ \vdots & \vdots & & & \ddots & \vdots & \vdots \\ a^{n-1} & a^{n-2} & \cdots & b & \cdots & b^{-n} & a^n \\ a^n & a^{n-1} & \cdots & 1 & \cdots & b^{-n+1} & b^{-n} \end{pmatrix} \in \mathbb{R}^{(2n+1)\times(2n+1)}.$$

As $n \to \infty$, its eigenvalues approach those of the convolution operator $\mathcal{S}$. Moreover, the singular values of $\mathcal{S}_{2n+1}$ lie between

$$\frac{|a-b|}{(1+|a|)(1+|b|)} \leq \sigma_i(\mathcal{S}_{2n+1}) \leq \frac{|a-b|}{(1-|a|)(|b|-1)}, \qquad i = 1, 2, \ldots, 2n+1.$$

Thus the ℓ_2-induced norm of $\boldsymbol{\Sigma}_{\ell_2}$ is $\frac{|a-b|}{(1-|a|)(|b|-1)}$. This is also the maximum of the transfer function on the unit circle, that is, the ℓ_∞-norm of $\mathbf{H}$. Finally, by definition, the Hankel-norm of $\boldsymbol{\Sigma}$ is the Hankel-norm of its stable part $\boldsymbol{\Sigma}_+$, i.e., $\mathbf{y}(k+1) - a\mathbf{y}(k) = a\mathbf{u}(k+1)$, which is readily computed to be $a/(1-a^2)$.

All-pass $\mathcal{L}_2$-systems

We are now ready to characterize (square) $\mathcal{L}_2$-systems (i.e., $\mathcal{L}_2$-systems having the same number of inputs and outputs) which are all-pass, i.e., their transfer functions are unitary on the $i\omega$-axis. The results are presented only for continuous-time systems which need not be stable.

Definition 5.22. *An $\mathcal{L}_2$-system is all-pass or unitary if* $\| \mathbf{u} \|_2 = \| \mathbf{y} \|_2$ *for all* $(\mathbf{u}, \mathbf{y})$ *satisfying the system equations.*

As before, the transfer function of $\boldsymbol{\Sigma}$ is denoted by $\mathbf{H}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$. The basic statement given below is that a system is all-pass if and only if all its Hankel singular values are equal to one for some appropriate $\mathbf{D}$.

Theorem 5.23. *The following statements are equivalent:*

1. *The $\mathcal{L}_2$-system $\boldsymbol{\Sigma}$ is square $p = m$ and all-pass.*

2. $\mathbf{H}^*(-i\omega)\mathbf{H}(i\omega) = \mathbf{I}_m$.

3. $\exists \mathcal{Q}: \ \mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$,
   $\mathcal{Q}\mathbf{B} + \mathbf{C}^*\mathbf{D} = \mathbf{0}$,
   $\mathbf{D}^*\mathbf{D} = \mathbf{I}_m$.

4. *The solutions $\mathcal{P}$, $\mathcal{Q}$ of the Lyapunov equations* (4.45) *and* (4.46) *satisfy* $\mathcal{P}\mathcal{Q} = \mathbf{I}_n$ *and* $\mathbf{D}^*\mathbf{D} = \mathbf{I}_m$.

Remark 5.8.2. The third condition of the above theorem can be expressed by means of the following single equation:

$$\begin{pmatrix} \mathbf{A}^* \\ \mathbf{B}^* \end{pmatrix}\begin{pmatrix} \mathcal{Q} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathcal{Q} \\ \mathbf{0} \end{pmatrix}\begin{pmatrix} \mathbf{A} & \mathbf{B} \end{pmatrix} + \begin{pmatrix} \mathbf{C}^* & \mathbf{0} \\ \mathbf{D}^* & -\mathbf{I}_m \end{pmatrix}\begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{0} & -\mathbf{I}_m \end{pmatrix} = \mathbf{0},$$

which is a structured Sylvester equation.

Corollary 5.24. **(a)** *The transfer function of an all-pass system can be written as follows:*

$$\mathbf{H}(s) = (\mathbf{I} - \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathcal{Q}^{-1}\mathbf{C}^*)\mathbf{D}.$$

(b) *Given* $\mathbf{A}, \mathbf{B}, \mathbf{C}$, *there exists* $\mathbf{D}$ *such that the system is all-pass if and only if the solutions $\mathcal{P}$, $\mathcal{Q}$ of the Lyapunov equations* (4.45) *and* (4.46) *satisfy* $\mathcal{P}\mathcal{Q} = I$.

(c) *The* 2*-norm of the system, which is the $\mathcal{L}_\infty$-norm of* $\mathbf{H}$, *is* 1.

Example 5.25. Consider the second-order all-pass system described by the transfer function:

$$\mathbf{H}(s) = \frac{(s+a)(s+b)}{(s-a)(s-b)}, \qquad a + b \neq 0.$$

We first assume that $a > 0$ and $b < 0$. In this case, a minimal realization of $\mathbf{H}$ is given as follows:

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{cc|c} a & 0 & c \\ 0 & b & d \\ \hline c & d & 1 \end{array}\right), \quad c^2 = 2a\frac{a+b}{a-b} > 0, \ d^2 = -2b\frac{a+b}{a-b} > 0.$$

It follows that the solutions of the two Lyapunov equations are equal:

$$\mathcal{P} = \mathcal{Q} = -\begin{pmatrix} \frac{c^2}{2a} & \frac{cd}{a+b} \\ \frac{cd}{a+b} & \frac{d^2}{2b} \end{pmatrix} = \begin{pmatrix} \frac{a+b}{a-b} & \frac{2\sqrt{-ab}}{a-b} \\ \frac{2\sqrt{-ab}}{a-b} & -\frac{a+b}{a-b} \end{pmatrix} \quad \Rightarrow \quad \mathcal{P}\mathcal{Q} = I.$$

Notice that the eigenvalues of $\mathcal{P}$ are $\lambda_1 = 1$, $\lambda_2 = -1$, irrespective of the parameters a and b.

Next, we assume that $a > b > 0$. In this case, a minimal realization of $\mathbf{H}$ differs from the one above by a sign:

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{cc|c} a & 0 & c \\ 0 & b & d \\ \hline c & -d & 1 \end{array}\right), \quad c^2 = 2a\frac{a+b}{a-b} > 0, \ d^2 = 2b\frac{a+b}{a-b} > 0.$$

In this case, $\mathcal{P}$ and $\mathcal{Q}$ are no longer equal:

$$\mathcal{P} = -\begin{pmatrix} \frac{c^2}{2a} & \frac{cd}{a+b} \\ \frac{cd}{a+b} & \frac{d^2}{2b} \end{pmatrix} = \begin{pmatrix} \frac{a+b}{a-b} & \frac{2\sqrt{ab}}{a-b} \\ \frac{2\sqrt{ab}}{a-b} & \frac{a+b}{a-b} \end{pmatrix},$$

$$\mathcal{Q} = -\begin{pmatrix} \frac{c^2}{2a} & -\frac{cd}{a+b} \\ -\frac{cd}{a+b} & \frac{d^2}{2b} \end{pmatrix} = \begin{pmatrix} \frac{a+b}{a-b} & -\frac{2\sqrt{ab}}{a-b} \\ -\frac{2\sqrt{ab}}{a-b} & \frac{a+b}{a-b} \end{pmatrix}.$$

Clearly, the product of $\mathcal{P}$ and $\mathcal{Q}$ is the identity matrix. Furthermore, these two matrices have the same set of eigenvalues, namely, $\lambda_1 = \frac{(\sqrt{a}+\sqrt{b})^2}{a-b}$, $\lambda_2 = \frac{1}{\lambda_1}$, whose product is equal to one.

Proof. **1 $\Leftrightarrow$ 2.** First we notice that $\mathbf{y} = \mathcal{S}(\mathbf{u})$ is equivalent to $\mathbf{Y} = \mathbf{HU}$ in the frequency domain. Thus, the following string of equivalences holds:

$$\| \mathbf{Y} \|_2 = \| \mathbf{HU} \|_2 \;\Leftrightarrow\; \langle \mathbf{Y}, \mathbf{Y} \rangle_2 = \langle \mathbf{HU}, \mathbf{HU} \rangle_2 = \langle \mathbf{U}, \mathbf{H}^*\mathbf{HU} \rangle_2 = \langle \mathbf{U}, \mathbf{U} \rangle \;\Leftrightarrow\; \mathbf{H}^*\mathbf{H} = \mathbf{I}.$$

2 $\Rightarrow$ 3. for $m = p$. We make use of the following auxiliary result:

$$(\mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B})^{-1} = \mathbf{D}^{-1}(\mathbf{I} - \mathbf{C}(s\mathbf{I} - \mathbf{A}^{\times})^{-1}\mathbf{B}\mathbf{D}^{-1}), \text{ where } \mathbf{A}^{\times} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}.$$

Hence, from $[\mathbf{H}(s)]^{-1} = \mathbf{H}^*(-s)$, $s = i\omega$, we obtain that

$$\mathbf{D}^{-1}(\mathbf{I} - \mathbf{C}(s\mathbf{I} - \mathbf{A}^{\times})^{-1}\mathbf{B}\mathbf{D}^{-1}) = \mathbf{D}^* + \mathbf{B}^*(s\mathbf{I} + \mathbf{A}^*)^{-1}(-\mathbf{C}^*) \Rightarrow \mathbf{D}^{-1} = \mathbf{D}^*,$$

and there exists $\mathcal{Q}$, $\det \mathcal{Q} \neq 0$, such that (a) $\mathbf{B}^*\mathcal{Q} = -\mathbf{D}^{-1}\mathbf{C}$, (b) $-\mathcal{Q}^{-1}\mathbf{C}^* = \mathbf{B}\mathbf{D}^{-1}$, and (c) $-\mathcal{Q}^{-1}\mathbf{A}^*\mathcal{Q} = \mathbf{A}^{\times} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$.

Thus (a) and (b) imply $\mathcal{Q}\mathbf{B} + \mathbf{C}^*\mathbf{D} = \mathbf{0}$, while (c) implies $\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} - \mathcal{Q}\mathbf{B}\mathbf{D}^{-1}\mathbf{C} = \mathbf{0}$. Combining the last two equations, we obtain $\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$. Since $\mathcal{R}e\, \lambda_i(\mathbf{A}) \neq 0$, the above equation has a *unique* solution $\mathcal{Q}$, which is symmetric since it also satisfies $\mathbf{A}^*\mathcal{Q}^* + \mathcal{Q}^*\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$.

3 $\Rightarrow$ 2. By direct computation we obtain

$$\mathbf{H}^*(v)\mathbf{H}(w) = \mathbf{I} - (v + w)\mathbf{C}(v\mathbf{I} - \mathbf{A})^{-1}\mathcal{Q}^{-1}(w\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*,$$

which implies the desired result for $w = i\omega = -v$.

3 $\Rightarrow$ 4. Let $\mathcal{P}$ satisfy $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$. It follows that $\mathcal{Q}(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)\mathcal{Q} = \mathbf{0}$, and hence $\mathcal{Q}\mathbf{A}(\mathcal{P}\mathcal{Q}) + (\mathcal{Q}\mathcal{P})\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{B}\mathbf{B}^*\mathcal{Q} = \mathbf{0}$. Substituting $\mathcal{Q}\mathbf{B} = -\mathbf{C}\mathbf{D}$ and subtracting from $\mathbf{A}\mathcal{Q} + \mathbf{A}^*\mathcal{Q} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$, we obtain

$$\underbrace{\mathcal{Q}\mathbf{A}}_{\mathbf{R}}\underbrace{(\mathbf{I} - \mathcal{P}\mathcal{Q})}_{\mathbf{X}} + \underbrace{(\mathbf{I} - \mathcal{Q}\mathcal{P})}_{\mathbf{X}^*}\underbrace{\mathbf{A}^*\mathcal{Q}}_{\mathbf{R}^*} = \mathbf{0} \Rightarrow \mathbf{R}\mathbf{X} + \mathbf{X}^*\mathbf{R}^* = \mathbf{0}.$$

Since $\mathcal{R}e\, \lambda_i(\mathbf{A}) \neq 0$, the unique solution is $\mathbf{X} = \mathbf{0}$, which implies the desired $\mathcal{P}\mathcal{Q} = \mathbf{I}$.

4 $\Rightarrow$ 3. This follows similarly. □

5.9 System dissipativity*

We now wish to generalize the concept of Lyapunov stability to *open* systems, that is, systems with *inputs* and *outputs*:

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \ \ \mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{u}).$$

We define a function s called the *supply function* to the system

$$\mathbf{s} : \mathbb{U} \times \mathbb{Y} \to \mathbb{R}, \ \ (\mathbf{u}, \mathbf{y}) \mapsto \mathbf{s}(\mathbf{u}, \mathbf{y}),$$

which represents something like the *power* delivered to the system by the environment through the external variables, namely, the input $\mathbf{u}$ and output $\mathbf{y}$. Dissipativeness now means that the system absorbs supply (energy). More precisely, the system defined above is said to be *dissipative*, with respect to the supply function s, if there exists a nonnegative function $\Theta : \mathbb{X} \to \mathbb{R}$ such that the *dissipation inequality*

$$\boxed{\Theta(\mathbf{x}(t_1)) - \Theta(\mathbf{x}(t_0)) \leq \int_{t_0}^{t_1} \mathbf{s}(\mathbf{u}(t), \mathbf{y}(t))\, dt} \qquad (5.40)$$

holds for all $t_0 \leq t_1$ and all trajectories $(\mathbf{u}, \mathbf{x}, \mathbf{y})$ which satisfy the system equations. Therefore, if $\int_{t_0}^{t_1} \mathbf{s}\, dt > 0$, we will say that *work is done on the system,* while if this integral is negative, we will say that *work is done by the system.* The nonnegative function Θ is called a *storage function.* It is a *generalized energy* function for the dissipative system in question. Thus the above definition says that the change in internal storage, namely, $\Theta(\mathbf{x}(t_1)) - \Theta(\mathbf{x}(t_0))$, can never exceed what is supplied to the system. Finally, if $\Theta(\mathbf{x})$ is differentiable with respect to time, along trajectories $\mathbf{x}$ of the system, the dissipation inequality can be written as

$$\boxed{\frac{d}{dt}\Theta(\mathbf{x}(t)) \leq \mathbf{s}(\mathbf{u}(t), \mathbf{y}(t)).} \qquad (5.41)$$

Thus the *storage function* generalizes the concept of *Lyapunov function* given in (5.35) from autonomous or closed systems to open systems, that is, systems with inputs and outputs. Consequently the concept of *dissipativity* generalizes to open systems the concept of *stability.*

Two *universal storage functions* can be constructed, namely, the *available storage* and the *required supply.* The former is defined as

$$\Theta_{\text{avail}}(\mathbf{x}_0) = \sup \left[- \int_0^{\infty} \mathbf{s}(\mathbf{u}(t), \mathbf{y}(t))\, dt \right], \qquad (5.42)$$

where $(\mathbf{u}, \mathbf{x}, \mathbf{y})$ satisfy the system equations, and in addition $\mathbf{x}(0) = \mathbf{x}_0$, $\mathbf{x}(\infty) = \mathbf{0}$. Therefore, $\Theta_{\text{avail}}(\mathbf{x}_0)$ is obtained by seeking to *maximize the supply extracted from the system starting at a fixed initial condition.* The latter is

$$\Theta_{\text{req}}(\mathbf{x}_0) = \inf \left[\int_{-\infty}^{0} \mathbf{s}(\mathbf{u}(t), \mathbf{y}(t))\, dt \right], \qquad (5.43)$$

where $(\mathbf{u}, \mathbf{x}, \mathbf{y})$ satisfy the system equations, and in addition $\mathbf{x}(-\infty) = \mathbf{0}$, $\mathbf{x}(0) = \mathbf{x}_0$. In this case, $\Theta_{\text{req}}(\mathbf{x}_0)$ is obtained by seeking to *minimize the supply needed to achieve a fixed initial state.*

A result due to Willems [360], [361] states that a system is dissipative if and only if $\Theta_{\text{req}}(\mathbf{x}_0)$ is finite for all $\mathbf{x}_0$; in this case, storage functions form a convex set and, in addition, the following inequality holds:

$$0 \le \Theta_{\text{avail}}(\mathbf{x}_0) \le \Theta(\mathbf{x}_0) \le \Theta_{\text{req}}(\mathbf{x}_0). \tag{5.44}$$

Thus the available storage and the required supply are extremal storage functions.

Some commonly used quadratic supply functions $\mathbf{s} = (\mathbf{y}^*, \mathbf{u}^*)\mathbf{Q}(\mathbf{y}^*, \mathbf{u}^*)^*$ are

$$\mathbf{s}(\mathbf{u}, \mathbf{y}) = \|\mathbf{u}\|_2^2 - \|\mathbf{y}\|_2^2, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix}, \tag{5.45}$$

$$\mathbf{s}(\mathbf{u}, \mathbf{y}) = \mathbf{u}^*\mathbf{y} + \mathbf{y}^*\mathbf{u}, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}, \tag{5.46}$$

$$\mathbf{s}(\mathbf{u}, \mathbf{y}) = \|\mathbf{u}\|_2^2 + \|\mathbf{y}\|_2^2, \quad \mathbf{Q} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}.$$

Notice that if $\mathbf{s} = 0$, the storage function Θ becomes a Lyapunov function and dissipativity is equivalent to stability.

Remark 5.9.1. *Electrical and mechanical systems.* For electrical circuits consisting of passive components, that is, resistors, capacitors, inductors, and ideal transformers, the sum of the product of the *voltage* and *current* at the external ports, which is the power supplied, constitutes a *supply function*: $\sum_k \mathbf{V}_k\mathbf{I}_k$ electrical power. In mechanical systems, the sum of the product of the force and velocity as well as angle and torque of the various particles is a possible *supply function*: $\sum_k((\frac{d}{dt}\mathbf{x}_k)\,\mathbf{F}_k + (\frac{d}{dt}\theta_k)\,\mathbf{T}_k)$ mechanical power.

Linear systems and quadratic supply functions

Consider the linear system $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ with the supply rate $\mathbf{s}$. The central issue now is, given $\mathbf{s}$, to determine whether $\boldsymbol{\Sigma}$ is dissipative with respect to $\mathbf{s}$. This amounts to the construction of a storage function Θ such that the dissipation inequality holds.

Consider the general *quadratic supply function* which is a function of the external variables, namely, the input $\mathbf{u}$ and the output $\mathbf{y}$:

$$\mathbf{s}(\mathbf{u}, \mathbf{y}) = \begin{pmatrix} \mathbf{y}^* & \mathbf{u}^* \end{pmatrix} \underbrace{\begin{pmatrix} \mathbf{Q}_{\mathbf{yy}} & \mathbf{Q}_{\mathbf{yu}} \\ \mathbf{Q}_{\mathbf{uy}} & \mathbf{Q}_{\mathbf{uu}} \end{pmatrix}}_{\mathbf{Q}} \begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix}, \qquad \mathbf{Q} = \mathbf{Q}^* \in \mathbb{R}^{(p+m)\times(p+m)}, \tag{5.47}$$

where no definiteness assumptions on $\mathbf{Q}$ are necessary at this stage. Given that $\mathbf{y} = \mathbf{C}\mathbf{x}+\mathbf{D}\mathbf{u}$, the supply function can also be expressed in terms of the state $\mathbf{x}$ and the input $\mathbf{u}$: $\mathbf{s} = \mathbf{x}^*\mathbf{Q}_{11}\mathbf{x} + \mathbf{x}^*\mathbf{Q}_{12}\mathbf{u} + \mathbf{u}^*\mathbf{Q}_{21}\mathbf{x} + \mathbf{u}^*\mathbf{Q}_{22}\mathbf{u}$, where

$$\hat{\mathbf{Q}} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{C}^* & \mathbf{0} \\ \mathbf{D}^* & \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \mathbf{Q}_{\mathbf{yy}} & \mathbf{Q}_{\mathbf{yu}} \\ \mathbf{Q}_{\mathbf{uy}} & \mathbf{Q}_{\mathbf{uu}} \end{pmatrix} \begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix}$$

inherits the symmetry property from $\mathbf{Q}$. We are now ready to state the following theorem, due to Willems.

Theorem 5.26. *Consider the system* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, *which is assumed reachable and observable, together with the quadratic supply function* $\mathbf{s}$. *The following statements are equivalent:*

(1) $\mathbf{\Sigma}$ *is dissipative with respect to* $\mathbf{s}$.

(2) $\mathbf{\Sigma}$ *admits a quadratic storage function* $\Theta = \mathbf{x}^*\mathbf{X}\mathbf{x}$ *with* $\mathbf{X} = \mathbf{X}^* \geq \mathbf{0}$.

(3) *There exists a positive semidefinite solution* $\mathbf{X}$ *to the following set of linear matrix inequalities (LMIs):*

$$\underbrace{-\begin{pmatrix} \mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} & \mathbf{X}\mathbf{B} \\ \mathbf{B}^*\mathbf{X} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{C}^* & \mathbf{0} \\ \mathbf{D}^* & \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \mathbf{Q}_{yy} & \mathbf{Q}_{yu} \\ \mathbf{Q}_{uy} & \mathbf{Q}_{uu} \end{pmatrix} \begin{pmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix}}_{\Phi(\mathbf{X})} \geq \mathbf{0}.$$

(4) *There exist* $\mathbf{X} = \mathbf{X}^* \geq \mathbf{0}$, $\mathbf{K}$, *and* $\mathbf{L}$ *such that*

$$\boxed{\begin{aligned} &\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} + \mathbf{K}^*\mathbf{K} = \mathbf{Q}_{11}, \\ &\mathbf{X}\mathbf{B} + \mathbf{K}^*\mathbf{L} = \mathbf{Q}_{12}, \\ &\mathbf{L}^*\mathbf{L} = \mathbf{Q}_{22}. \end{aligned}} \tag{5.48}$$

(5) *There exists* $\mathbf{X}_- = \mathbf{X}_-^* \geq \mathbf{0}$ *such that* $\Theta_{\text{avail}} = \mathbf{x}^*\mathbf{X}_-\mathbf{x}$.

(6) *There exists* $\mathbf{X}_+ = \mathbf{X}_+^* \geq \mathbf{0}$ *such that* $\Theta_{\text{req}} = \mathbf{x}^*\mathbf{X}_+\mathbf{x}$.

(7) *The transfer function* $\mathbf{H}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ *satisfies*

$$\begin{pmatrix} \mathbf{H}^*(-i\omega) & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{Q}_{yy} & \mathbf{Q}_{yu} \\ \mathbf{Q}_{uy} & \mathbf{Q}_{uu} \end{pmatrix} \begin{pmatrix} \mathbf{H}(i\omega) \\ \mathbf{I} \end{pmatrix} \geq 0$$

for all $\omega \in \mathbb{R}$, *such that* $i\omega$ *is not an eigenvalue of* $\mathbf{A}$.

Corollary 5.27. *Let* $\mathbf{Q}_{22}$ *be nonsingular and define the quantities*

$$\mathbf{F} = \mathbf{A} - \mathbf{B}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21}, \quad \mathbf{G} = \mathbf{B}\mathbf{Q}_{22}^{-1}\mathbf{B}^*, \quad \mathbf{J} = \left(\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{21} - \mathbf{Q}_{11}\right).$$

Two further conditions are equivalent to the seven listed in the theorem:

(8) *There exists a solution* $\mathbf{X} = \mathbf{X}^* \geq \mathbf{0}$ *to the algebraic Riccati equation*

$$\mathbf{F}^*\mathbf{X} + \mathbf{X}\mathbf{F} + \mathbf{X}\mathbf{G}\mathbf{X} + \mathbf{J} = \mathbf{0}. \tag{ARE}$$

(9) *There exists a solution* $\mathbf{X} = \mathbf{X}^* \geq \mathbf{0}$ *to the Riccati inequality* $\mathbf{F}^*\mathbf{X} + \mathbf{X}\mathbf{F} + \mathbf{X}\mathbf{G}\mathbf{X} + \mathbf{J} \leq 0$.

In case of dissipativity, the quadratic form

$$\mathbf{d} = \begin{pmatrix} \mathbf{x}^* & \mathbf{u}^* \end{pmatrix} \Phi(\mathbf{X}) \begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix} = \|\mathbf{Kx} + \mathbf{Lu}\|^2 \geq 0$$

is positive semidefinite, where the second equality follows from part (3) of the theorem. Thus

$$\mathbf{d} = \mathbf{s} - \frac{d}{dt}\Theta, \quad \Theta = \mathbf{x}^*\mathbf{Xx} \;\Rightarrow\; \|\mathbf{Kx} + \mathbf{Lu}\|^2 = \mathbf{s} - \frac{d}{dt}\mathbf{x}^*\mathbf{Xx} \geq 0.$$

d is a *dissipativity function* for the system Σ with the supply rate s. This relationship says that what gets dissipated at time t, which is always nonnegative, is equal to the supply at time t minus the rate of change of the storage function.

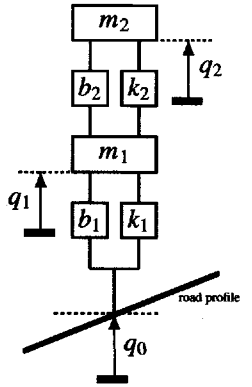
Remark 5.9.2. *There are two refinements.* **(a)** If the dissipation inequality (5.40) (and consequently (5.41)) is satisfied with *equality*, the system Σ is called *conservative* with respect to s. In this case, the inequality in part (3) of Theorem 5.26 is satisfied with equality: $\Phi(\mathbf{X}) = 0$. Consequently, in part (4) of the same theorem $\mathbf{K} = \mathbf{0}$ and $\mathbf{L} = \mathbf{0}$. Thus conditions (5.48) become

$$\mathbf{A}^*\mathbf{X} + \mathbf{XA} = \mathbf{Q}_{11}, \quad \mathbf{XB} = \mathbf{Q}_{12}, \quad \mathbf{Q}_{22} = \mathbf{0}. \tag{5.49}$$

It readily follows that a stable system is *unitary* or *all-pass* if it is conservative with respect to the supply function (5.45). In particular, the conditions in part 3 of Theorem 5.23 coincide with (5.49), where $\mathbf{Q}_{11} = -\mathbf{C}^*\mathbf{C}$, $\mathbf{Q}_{12} = -\mathbf{C}^*\mathbf{D}$, and $\mathbf{Q}_{22} = \mathbf{I} - \mathbf{D}^*\mathbf{D}$. Furthermore, the inequality in part (6) of Theorem 5.26 becomes an equality and, in the all-pass case, becomes identical with part 3 of Theorem 5.23. As a consequence, the Riccati equation is replaced by a Lyapunov equation which has a *unique* solution. Therefore, in the case of conservative systems, there is a unique storage function $\Theta_{\text{avail}} = \Theta = \Theta_{\text{req}}$.

(b) If the dissipation inequality (5.40) (and consequently (5.41)) is *strict*, the system Σ is called *strictly dissipative* with respect to s; as a consequence, all inequalities in Theorem 5.26 are *strict*.

Example 5.28. We consider a simplified model of a car suspension. It is composed of a mass m_1, which models the wheel, together with a spring with constant k_1 and a damper with constant b_1, which model the interaction of the wheel and the road. The car chassis has a mass m_2, and its connection to the wheel is modeled by means of a second spring and damper with constants k_2, b_2, respectively. There is a (control) force f applied to the axle (i.e., to the wheel) which acts on m_1 vertically. The wheel follows the road, which has a profile described by its distance q_0 from a fixed position. Furthermore, the distance of the masses m_1, m_2 from the same fixed position are q_1, q_2, respectively. The equations of motion are as follows. For simplicity time derivatives are denoted by dots:



$$m_2\ddot{q}_2 + b_2(\dot{q}_2 - \dot{q}_1) + k_2(q_2 - q_1) = 0,$$
$$m_1\ddot{q}_1 - b_2(\dot{q}_2 - \dot{q}_1) - k_2(q_2 - q_1) + b_1(\dot{q}_1 - \dot{q}_0) + k_1(q_1 - q_0) = f.$$

Notice that $b_1(\dot{q}_1 - \dot{q}_0) + k_1(q_1 - q_0)$ is the force exerted by the road on the wheel. The inputs to this system are therefore the road profile q_0 and the control force f. The output may be chosen as q_2.

This is a dissipative system, as there are no active components. To formalize this, we introduce the supply function, which consists of the sum of products of force times velocity,

$$\mathbf{s} = f\dot{q}_1 - [b_1(\dot{q}_1 - \dot{q}_0) + k_1(q_1 - q_0)]\,\dot{q}_0.$$

From physical considerations, if follows that a storage function is

$$\Theta = \frac{1}{2}m_1\dot{q}_1^2 + \frac{1}{2}m_2\dot{q}_2^2 + \frac{1}{2}k_1(q_1 - q_0)^2 + \frac{1}{2}k_2(q_2 - q_1)^2,$$

which is the sum of the kinetic energies of the two masses and the sum of the energy stored in the two springs. The dissipation function

$$\mathbf{d} = b_2(\dot{q}_2 - \dot{q}_1)^2 + b_1(\dot{q}_1 - \dot{q}_0)^2$$

is the energy dissipated in the two dashpots. Indeed, a straightforward computation shows that $\mathbf{d} = \mathbf{s} - \dot{\Theta}$. Notice that if the dashpots are absent, i.e., $b_1 = b_2 = 0$, the system is conservative, which means that the energy supplied is stored and no part of it is dissipated as heat. The above storage function is not unique. To characterize all of them, we can apply Theorem 5.26. This is left as an exercise for the reader (see Problem 31).

We will not provide a proof of the above theorem and corollary. We refer instead to the original source, namely, the work of Willems [360], [361]. See also the lecture notes of Scherer and Weiland [288].

Instead, in the sections that follow, we will attempt to clarify two special cases. The first uses the supply function (5.46) and is used to discuss issues related to *passivity* of a given system. The equalities (5.48) provide in this case a state-space characterization of passivity and are known as the *positive real lemma* or as the Kalman–Yakubovich–Popov (*KYP*) lemma. The second case uses the supply function (5.45) and investigates *contractivity* of the system, that is, whether the system has $\mathcal{H}_\infty$-norm less than unity. Equations (5.48) provide a characterization of contractivity and are known as the *bounded real lemma*.

5.9.1 Passivity and the positive real lemma*

In this section, we will consider a special class of systems, namely, *passive* systems. Roughly speaking, these are systems that do not generate energy, that is, the energy dissipated is never greater than the energy generated. This turns out to be closely related to the concept of *positive realness*, which is important in electric networks composed of passive components (resistors R, capacitors C, and inductors L). We will show that such systems are *dissipative* with respect to the *supply rate* $\mathbf{s} = \mathbf{y}^*\mathbf{u} + \mathbf{u}^*\mathbf{y}$ defined in (5.46). For simplicity, we consider only SISO systems. This section follows the notes of [359].

The real rational function $\mathbf{H}(s)$ is *positive real* if it maps the right half of the complex plane $\mathbb{C}$ onto itself:

$$s \in \mathbb{C},\ \ \mathcal{R}e(s) \geq 0 \Rightarrow \mathcal{R}e(\mathbf{H}(s)) \geq 0,\ \ s \text{ not a pole of } \mathbf{H}.$$

Henceforth, let $\mathbf{H} = \frac{\mathbf{n}}{\mathbf{d}}$ with $\mathbf{n}, \mathbf{d}$ coprime polynomials and $\deg(\mathbf{d}) = n$, $\deg(\mathbf{n}) = m \leq n$. Recall first the *partial fraction expansion* of this rational function. Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct zeros of $\mathbf{d}$ (the *poles* of $\mathbf{H}$) and $n_1, n_2, \ldots, n_k$, their multiplicities ($\sum_{i=1}^{k} n_i = n$). We can now express $\mathbf{H}$ as follows:

$$\mathbf{H}(s) = a_0 + \sum_{i=1}^{k} \sum_{\ell=1}^{n_i} \frac{a_{i\ell}}{(s - \lambda_i)^{\ell}},$$

where a_0 and $a_{i\ell}$ are complex numbers that are uniquely determined by $\mathbf{H}$, $a_{in_i} \neq 0$, for $i = 1, 2, \ldots, k$. If a pole, λ_k is simple, that is, $n_k = 1$, a_{i1} is the *residue* of $\mathbf{H}$ at the pole λ_i; in fact, $a_{k1} = (s - \lambda_k)\mathbf{H}(s)|_{s=\lambda_k}$. The following result follows from the definition.

Proposition 5.29. *$\mathbf{H}(s) \neq 0$ is positive real if and only if $\mathbf{H}(s^{-1})$ is positive real if and only if $\mathbf{H}^{-1}(s)$ is positive real.*

Proof. Observe (using continuity) the obvious fact that $\mathbf{H}$ is positive real if and only if $\mathcal{Re}(\mathbf{H}(s)) \geq 0$ for all but a finite number of points s in the right half plane $\mathcal{Re}(s) > 0$. The map $s \mapsto 1/s$ is a bijection for $s \neq 0$, $\mathcal{Re}(s) \geq 0$. Hence $\mathcal{Re}\big(\mathbf{H}(s)\big) \geq 0$ for all but a finite number of points in $\mathcal{Re}(s) \geq 0$ if and only if $\mathcal{Re}(\mathbf{H}(s^{-1})) \geq 0$ for all but a finite number of points in $\mathcal{Re}(s) \geq 0$. Hence the first equivalence is proved. For the second, we notice that $\alpha \neq 0$ satisfies $\mathcal{Re}(\alpha) \geq 0$ if and only if $\mathcal{Re}(\alpha^{-1}) \geq 0$. Hence $\mathcal{Re}\big(\mathbf{H}(s)\big) \geq 0$ for all but a finite number of points s with $\mathcal{Re}(s) \geq 0$ if and only if $\mathcal{Re}\big(\mathbf{H}^{-1}(s)\big) \geq 0$ for all but a finite number of points s, $\mathcal{Re}(s) \geq 0$, and the proof is complete. □

The next result provides the characterization of positive realness in terms of the external representation of $\mathbf{\Sigma}$.

Theorem 5.30. *The following statements are equivalent:*

(a) $\mathbf{H}$ *is positive real.*

(b) $\mathbf{H}$ *satisfies the following conditions:*

(i) $\mathcal{Re}\big(\mathbf{H}(i\omega)\big) \geq 0$ *for all* $\omega \in \mathbb{R}$ *such that* $i\omega$ *is not a pole of* $\mathbf{H}$.

(ii) *The system is stable, that is, all poles* λ *of* $\mathbf{H}$ *lie in the left half of the complex plane:* $\mathcal{Re}(\lambda) \leq 0$.

(iii) *The poles of* $\mathbf{H}$ *on the imaginary axis are simple, and the associated residues are real and positive.*

(iv) *The difference between the degree of the numerator and that of the denominator is at most one:* $\deg(\mathbf{n}) - \deg(\mathbf{d}) \leq 1$. *Furthermore, if there is a pole at infinity, it is, as those on the imaginary axis, simple, and the associated residue is positive:* $\lim_{|s|\to\infty} \frac{\mathbf{H}(s)}{s} \geq 0$.

(c) *All input-output pairs* $(\mathbf{u}, \mathbf{y})$ *of compact support, which are compatible with the system equations,*

$$\mathbf{d}\left(\frac{d}{dt}\right)\mathbf{y} = \mathbf{n}\left(\frac{d}{dt}\right)\mathbf{u}, \tag{5.50}$$

satisfy the dissipation inequality

$$\int_{-\infty}^{0} \mathbf{u}(t)\,\mathbf{y}(t)\,dt \geq 0. \tag{5.51}$$

Proof. **(a) ⇒ (b).** (b)(i) is trivial. To prove (b)(ii) and (iii), examine the partial fraction expansion of $\mathbf{H}$ in a small circle around a right half plane or an imaginary axis pole. To prove (b)(iv), examine the real part of $\mathbf{H}(s)$ in a large circle, centered at 0, where the polynomial part of $\mathbf{H}$ dominates.

(b) ⇒ (c). Consider the partial fraction expansion of $\mathbf{H}$. Part (b) shows that $\mathbf{H}$ may be written as

$$\mathbf{H}(s) = a_0 s + \frac{b_0}{s} + \sum_{k=1}^{r} \frac{a_k s}{s^2 + \omega_k^2} + \mathbf{H}_0(s),$$

with $a_0 \geq 0$, $b_0 \geq 0$, $a_k > 0$, $0 \neq \omega_k \in \mathbb{R}$, $\mathbf{H}_0$ is proper, and all its poles are in the open left half of the complex plane. Note that $\mathcal{R}e(\mathbf{H}_0(i\omega)) = \mathcal{R}e(\mathbf{H}(i\omega))$ for all $\omega \in \mathbb{R}$, and $i\omega$ is not a pole of $\mathbf{H}$. Let $\mathbf{H}_0 = \mathbf{n}_0/\mathbf{d}_0$, with $\mathbf{n}_0, \mathbf{d}_0$ coprime polynomials. Note that $\mathbf{d}_0$ is Hurwitz. Now consider the systems

$$\begin{array}{lrcll}
\Sigma_0: & \mathbf{y}_0 & = & a_0 \frac{d}{dt}\mathbf{u}, & \\
\Sigma_1: & \frac{d}{dt}\mathbf{y}_1 & = & b_0 \mathbf{u}, & \\
\Sigma_{2,i}: & \frac{d^2}{dt^2}\mathbf{y}_{2i} + \omega_i^2 \mathbf{y}_{2i} & = & a_i \frac{d}{dt}\mathbf{u}, & i = 1, 2, \ldots, k, \\
\Sigma_3: & \mathbf{d}_0(\frac{d}{dt})\mathbf{y}_3 & = & \mathbf{n}_0(\frac{d}{dt})\mathbf{u}, & \\
\Sigma: & \mathbf{y} & = & \mathbf{y}_1 + \mathbf{y}_2 + \sum_{i=1}^{k} \mathbf{y}_{2i} + \mathbf{y}_3. &
\end{array}$$

Clearly the pair of input-output functions $(\mathbf{u}, \mathbf{y})$ satisfies (5.50) if and only if $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_{2i}$, $i = 1, 2, \ldots, k$, $\mathbf{y}_3$ satisfy the above equations.

It is easy to see (using observability) that if $(\mathbf{u}, \mathbf{y})$ have compact support, where $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 + \sum_{i=1}^{k} \mathbf{y}_{2i} + \mathbf{y}_3$, the summands $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_{2i}, \mathbf{y}_3$ must have compact support. Hence, to prove that (5.51) holds for Σ, it suffices to prove it for Σ_0, Σ_1, $\Sigma_{2,i}$, $i = 1, 2, \ldots, k$, and Σ_3.

We now prove that for each of these systems, the inequality analogous to (5.51) holds. Consider compact support trajectories satisfying the equations of each of these systems. For Σ_0,

$$\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}_0(t)\,dt = a_0 \int_{-\infty}^{0} \mathbf{u}(t) \left[\frac{d}{dt}\mathbf{u}(t)\right] dt = \frac{1}{2} a_0 \mathbf{u}^2(0) \geq 0.$$

For Σ_1, assuming $b_0 > 0$,

$$\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}_1(t)\,dt = \frac{1}{b_0} \int_{-\infty}^{0} \left[\frac{d}{dt}\mathbf{y}_1(t)\right] \mathbf{y}_1(t)\,dt = \frac{1}{2}\frac{\mathbf{y}_1^2(0)}{b_0} \geq 0.$$

For $\mathbf{\Sigma}_{2i}$ we introduce the auxiliary variable ξ: $\mathbf{u} = \left(\omega_i^2 + \frac{d^2}{dt^2}\right)\xi$, $\mathbf{y}_{2i} = a_i\frac{d}{dt}\xi$. Hence

$$\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}_{2i}(t)\,dt = a_i \int_{-\infty}^{0} \left(\omega_i^2 + \frac{d^2}{dt^2}\right)\xi\,\frac{d}{dt}\xi\,dt = \frac{1}{2}a_i\omega_i^2\xi^2(0) + \frac{1}{2}a_i\left[\frac{d}{dt}\xi(0)\right]^2.$$

Finally, consider $\mathbf{\Sigma}_3$. Using Parseval's theorem, it follows that for $\mathcal{L}_2$ signals, there holds

$$\int_{-\infty}^{\infty} \mathbf{u}(t)\mathbf{y}_3(t)\,dt = \int_{-\infty}^{\infty} \hat{\mathbf{u}}(-i\omega)\mathbf{H}_0(i\omega)\hat{\mathbf{u}}(i\omega)\,dt = \int_{-\infty}^{\infty} \frac{1}{2}\mathcal{R}e(\mathbf{H}_0(i\omega))|\hat{\mathbf{u}}(i\omega)|^2\,dt \geq 0,$$

where $\hat{\mathbf{u}}$ is the Fourier transform of $\mathbf{u}$. This shows the inequality for integrals over the real line $\mathbb{R}$, but we need it for integrals on the negative real line $\mathbb{R}_-$ only. To show that $\int_{-\infty}^{0}\mathbf{u}(t)\mathbf{y}_3(t)\,dt \geq 0$, consider the input $\mathbf{u}'$ which is defined as $\mathbf{u}'(t) = \mathbf{u}(t), t \leq 0, \mathbf{u}'(t) = 0, t > 0$. Let the corresponding output be $\mathbf{y}_3'$. Since $\mathbf{u}' \in \mathcal{L}_2(\mathbb{R}, \mathbb{R})$, and since $\mathbf{d}_0$ is Hurwitz, it follows that $\mathbf{y}_3' \in \mathcal{L}_2(\mathbb{R}, \mathbb{R})$. Hence

$$\int_{-\infty}^{\infty} \mathbf{u}'(t)\mathbf{y}_3'(t)\,dt = \int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}_3(t)\,dt \geq 0.$$

Note that $\mathbf{u}'$ may not be in C^∞, and therefore a smoothness argument is, strictly speaking, required to complete the proof of nonnegativity.

The desired inequality thus follows $\int_{-\infty}^{0}\mathbf{u}(t)\mathbf{y}(t)\,dt \geq 0$, where $\mathbf{y} = \mathbf{y}_0 + \mathbf{y}_1 + \sum_{i=1}^{k}\mathbf{y}_{2i} + \mathbf{y}_3$ for all $(\mathbf{u}, \mathbf{y})$ of compact support.

(c) $\Rightarrow$ (a). We show that by considering exponential trajectories, (c) implies $\mathcal{R}e\big(\mathbf{H}(s)\big) > 0$ for all s with $\mathcal{R}e(s) > 0$, and s is not a pole of $\mathbf{H}$. Consider $\mathbf{u}(t) = e^{st}$ and $\mathbf{y}(t) = \mathbf{H}(s)e^{st}$. Obviously, $(\mathbf{u}, \mathbf{y})$ is an admissible input-output pair for the system. Now by (c),

$$0 \leq \mathcal{R}e\left(\int_{-\infty}^{0} \mathbf{u}^*(t)\mathbf{y}(t)\,dt\right) = \mathcal{R}e\big(\mathbf{H}(s)\big)\int_{-\infty}^{0} e^{\mathcal{R}e(s)t}\,dt.$$

Hence $\mathcal{R}e(\mathbf{H}(s)) \geq 0$. Note that such input-output pairs do not have compact support, and therefore an approximation argument is, strictly speaking, required to deduce the nonnegativity of $\mathcal{R}e(\int_{-\infty}^{0}\mathbf{u}^*(t)\mathbf{y}(t)\,dt)$ from (c). □

Assume without loss of generality that $\mathbf{H} = \frac{\mathbf{n}}{\mathbf{d}}$ is proper; otherwise, exchange the roles of the input and the output, and consider $\mathbf{H}^{-1}$. Let $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ be a minimal (i.e., reachable and observable) state-space representation system of $\mathbf{H}$, with state-space $\mathbb{R}^n$. We now discuss what constraints are imposed by positive realness on state representations.

Theorem 5.31. *The following conditions are equivalent:*

1. $\mathbf{H}$ *is positive real.*

2. $\mathbf{\Sigma}$ *is dissipative with respect to the supply rate* $\mathbf{s} = \mathbf{u}^*\mathbf{y} + \mathbf{y}^*\mathbf{u}$.

3. *There exists* $\mathbf{X} \in \mathbb{R}^{n\times n}$, $\mathbf{X} = \mathbf{X}^* > 0$, *such that the LMI*

$$\begin{bmatrix} \mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} & \mathbf{X}\mathbf{B} - \mathbf{C}^* \\ \mathbf{B}^*\mathbf{X} - \mathbf{C} & -\mathbf{D} - \mathbf{D}^* \end{bmatrix} \leq 0$$

holds.

Proof. **(1) ⇒ (2).** From the previous result, it follows that $\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt \geq 0$ for all $(\mathbf{u}, \mathbf{y})$ having compact support. Define $\Theta_{\mathrm{req}} : \mathbb{R}^n \to \mathbb{R}$ by (5.43), namely,

$$\Theta_{\mathrm{req}}(\mathbf{x}_0) = \inf \left[\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt \right],$$

where the infimum is taken over all $(\mathbf{u}, \mathbf{x}, \mathbf{y})$ of compact support that satisfy the system equations and in addition satisfying $\mathbf{x}(0) = \mathbf{x}_0$. Further, for all such $(\mathbf{u}, \mathbf{x}, \mathbf{y})$, there holds

$$\begin{aligned}\Theta_{\mathrm{req}}\big(\mathbf{x}(t_1)\big) &= \inf \left[\int_{-\infty}^{t_1} \mathbf{u}'(t)\mathbf{y}'(t)dt \right] \leq \int_{t_0}^{t_1} \mathbf{u}(t)\mathbf{y}(t)dt + \inf \left[\int_{-\infty}^{t_0} \mathbf{u}'(t)\mathbf{y}'(t)dt \right] \\ &= \int_{t_0}^{t_1} \mathbf{u}(t)\mathbf{y}(t)dt + \Theta_{\mathrm{req}}\big(\mathbf{x}(t_0)\big),\end{aligned}$$

where the infima are taken over all elements $(\mathbf{u}', \mathbf{x}', \mathbf{y}')$ of compact support, constrained by $\mathbf{x}'(t_1) = \mathbf{x}(t_1)$ for the first and $\mathbf{x}'(t_0) = \mathbf{x}(t_0)$ for the second infimum. This proves that since the dissipation inequality (5.40) is satisfied, Σ defines a dissipative system with storage function Θ_{req}.

(2) ⇒ (3). We first prove that (2) implies $\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt \geq 0$ for all $(\mathbf{u}, \mathbf{y})$ of compact support. Let Θ be the storage function. Then for all $(\mathbf{u}, \mathbf{x}, \mathbf{y})$ of compact support, there holds

$$\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt \geq \Theta\big(\mathbf{x}(0)\big) - \Theta(0) \geq \inf_{\mathbf{x}\in\mathbb{R}^n} \Theta(\mathbf{x}) - \Theta(0).$$

Since Θ is bounded from below, this implies that $\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt$ is bounded from below for all $(\mathbf{u}, \mathbf{y})$. We claim that 0 is a lower bound. This is easily proven by contradiction: assume that there exists $(\mathbf{u}, \mathbf{y})$ of compact support, such that $\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt < 0$. Now use the input $\kappa\mathbf{u}$ and let $\kappa \to \infty$ to obtain a contradiction.

We now prove that $\int_{-\infty}^{0} \mathbf{u}(t)\mathbf{y}(t)\,dt \geq 0$ for all $(\mathbf{u}, \mathbf{y})$ of compact support implies (3). Note that the LMI implies

$$\frac{d}{dt}\mathbf{x}^*\mathbf{X}\mathbf{x} \leq \mathbf{u}^*\mathbf{y}$$

for all $(\mathbf{u}, \mathbf{x}, \mathbf{y})$, or, equivalently, that $\mathbf{x}^*\mathbf{X}\mathbf{x}$ is a storage function.

It hence suffices to prove that $\Theta_{\mathrm{req}}(\mathbf{x})$ is a quadratic function of $\mathbf{x}$. Here is the idea. We will need the convolution operator $\mathcal{S}$ which maps the space of $\mathbf{C}^{\infty}(\mathbb{R}_-, \mathbb{R}^m)$ functions of compact support into itself by $\mathcal{S}(\mathbf{u})(t) = \mathbf{D}\mathbf{u}(t) + \int_{-\infty}^{t} \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau$. We will also need the operator $\mathcal{T}$ which maps the space of $\mathbf{C}^{\infty}(\mathbb{R}_-, \mathbb{R}^m)$ functions of compact support into $\mathbb{R}^n$ by $\mathcal{T}(\mathbf{u}) = \int_{-\infty}^{0} e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}(t)dt$. Θ_{req} can be defined as

$$\Theta_{\mathrm{req}}(\mathbf{x}_0) = \inf \langle \mathbf{u}, \mathcal{S}(\mathbf{u}) \rangle_{\mathcal{L}_2(\mathbb{R}_-, \mathbb{R}^m)},$$

where the infimum is taken over all $\mathbf{u} \in \mathbf{C}^{\infty}(\mathbb{R}_-, \mathbb{R}^m)$ of compact support subject to the constraint $\mathcal{T}(\mathbf{u}) = \mathbf{x}_0$. By assumption, $\langle \mathbf{u}, \mathcal{S}\mathbf{u} \rangle_{\mathcal{L}_2(\mathbb{R}_-, \mathbb{R}^m)} \geq 0$. The fact that the infimum is hence a quadratic function of $\mathbf{x}_0$ readily follows from *first principles*: $\Theta_{\mathrm{req}}(\mathbf{x}_0)$ is the infimum of a quadratic functional with a linear constraint.

(3) ⇒ (1) is trivial. □

From Theorem 5.26, and in particular (5.48) and (**ARE**), we obtain the following result.

Corollary 5.32. (a) Positive Real Lemma. *The minimal system* $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ *is dissipative with respect to the supply rate* $\mathbf{s} = \mathbf{y}^*\mathbf{u} + \mathbf{u}^*\mathbf{y}$ *if and only if there exist* $\mathbf{X} = \mathbf{X}^* \geq 0$, $\mathbf{K}$, *and* $\mathbf{L}$ *such that*

$$\boxed{\begin{aligned}&\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} + \mathbf{K}^*\mathbf{K} = \mathbf{0},\\ &\mathbf{X}\mathbf{B} + \mathbf{K}^*\mathbf{L} = \mathbf{C}^*,\\ &\mathbf{D} + \mathbf{D}^* = \mathbf{L}^*\mathbf{L}.\end{aligned}} \tag{5.52}$$

(b) *Let* $\mathbf{D} + \mathbf{D}^*$ *be nonsingular, and* $\Delta = (\mathbf{D} + \mathbf{D}^*)^{-1}$. *The system* $\mathbf{\Sigma}$ *is positive real if and only if there exists a positive semidefinite solution* $\mathbf{X} = \mathbf{X}^* \geq \mathbf{0}$ *to the algebraic Riccati equation*

$$(\mathbf{A}^* - \mathbf{C}^*\Delta\mathbf{B}^*)\mathbf{X} + \mathbf{X}(\mathbf{A} - \mathbf{B}\Delta\mathbf{C}) + \mathbf{X}\mathbf{B}\Delta\mathbf{B}^*\mathbf{X} + \mathbf{C}^*\Delta\mathbf{C} = \mathbf{0}. \tag{PRARE}$$

We conclude this section with a result concerning positive real rational functions with real or pure imaginary poles.

Corollary 5.33. *Assume that all the roots of both* **d** *and* **n** *are real. Then* **H** *is positive real if the roots of* **d** *and* **n** *are nonpositive and interlacing and the leading coefficients of* **d** *and* **n** *have the same sign. If all the roots of both* **d** *and* **n** *are purely imaginary,* **H** *is positive real if and only if the roots of* **d** *and* **n** *interlace, and the leading coefficients of* **d** *and* **n** *have the same sign.*

Proof. Assume, without loss of generality, that **H** is proper (otherwise, consider $\mathbf{H}^{-1}$). The first assumption on the poles and zeros implies that **H** can be written as $\mathbf{H}(s) = a_0 + \sum_{i=1}^{n} \frac{a_i}{s+b_i}$, with $0 \leq b_0 < b_1 < \cdots < b_n$. The assumptions of the zeros (examine the behavior of **H** on the real axis) and on the leading coefficients imply further that $a_0 \leq 0$, and $a_i > 0$ for $i = 1, 2, \ldots, n$. Obviously, now, **H** is positive real, since each term of the sum is. This can also been seen from the state representation with

$$\mathbf{A} = \operatorname{diag}(-b_1, -b_2, \ldots, -b_n), \quad \mathbf{B}^* = \mathbf{C} = [\sqrt{a_1} \ \ \sqrt{a_2} \ \ \cdots \ \ \sqrt{a_n}], \quad \mathbf{D} = a_0.$$

The LMI is satisfied with $\mathbf{X} = \mathbf{I}_n$.

Using part (b) of Theorem 5.30, we conclude that **H** positive real implies that its partial fraction expansion looks like

$$\mathbf{H}(s) = \frac{a_0}{s} + \sum_{i=1}^{n} \frac{a_i s}{s^2 + \omega_i^2}$$

with $a_0 \geq 0$, $a_i > 0$ for $i = 1, 2, \ldots, n$ and $0 \leq \omega_1 < \omega_2 < \cdots < \omega_n$. By investigating the behavior of the imaginary part of **H** on the imaginary axis, we conclude that between any two poles there must be at least one zero. Since **H** is strictly proper, it follows that there is exactly one zero between any two poles.

To prove the converse, use the interlacing property to conclude that $\mathbf{H}$ admits a partial fraction expansion as above. This implies that $\mathbf{H}$ is positive real, since it is the sum of positive real functions. This can also be seen from the state representation with

$$\mathbf{A} = \operatorname{diag}\left(0, \begin{pmatrix} 0 & \omega_1 \\ -\omega_1 & 0 \end{pmatrix}, \ldots, \begin{pmatrix} 0 & \omega_n \\ -\omega_n & 0 \end{pmatrix}\right),$$

$$\mathbf{B}^* = \mathbf{C} = [\sqrt{a_0}\ \ \sqrt{a_1}\ \ 0\ \ \sqrt{a_2}\ \ 0\ \ \cdots\ \ \sqrt{a_n}\ \ 0],\quad \mathbf{D} = 0.$$

If $a_0 = 0$, delete the first element. Observe that the LMI is satisfied with $\mathbf{X} = \mathbf{I}_n$. □

Some remarks on network synthesis

Theorem 5.34. $\mathbf{H}$ *is the driving point impedance of an electric circuit that consists of an interconnection of a finite number of positive R, positive L, positive C, and transformers if and only if* $\mathbf{H}$ *is positive real.*

The *only if* part of the above theorem is *analysis*. The *if* part is *synthesis*. This result, undoubtedly the most spectacular in electric circuit theory, was proved by Otto Brune in his dissertation [77].

It turns out that the required number of reactances (that is, the number of L and C combined) is equal to $\max\{\deg(\mathbf{d}), \deg(\mathbf{n})\}$ (the *McMillan degree* of $\mathbf{H}$). Brune's synthesis, however, requires ideal transformers. In 1949, Bott and Duffin proved in a half-page (!) paper, which appeared in the Letters to the Editor section of the *Journal of Applied Physics* [69], that transformers are not needed. This problem was a well-known open problem at that time, and Bott solved it as a graduate student working under the direction of Richard Duffin at the Carnegie Institute of Technology.

Transformerless synthesis requires a number of reactances that is larger than $\max\{\deg(\mathbf{d}), \deg(\mathbf{n})\}$. In terms of state representations, this means that we will end up with a state dimension that is larger than the McMillan degree of $\mathbf{H}$. The Bott–Duffin synthesis is a strong case for the importance of nonminimal state representations. However, the price one has to pay to avoid transformers is the exponential increase in the number of reactances.

Problem. *When does* $\mathbf{d}(\frac{d}{dt})\mathbf{V} = \mathbf{n}(\frac{d}{dt})\mathbf{I}$ *describe an electric circuit that consists of an interconnection of a finite number of positive R, positive L, positive C, and transformers?*

The above is an open problem. The answer is more involved than if and only if the transfer function is positive real and the common factors are stable.

An important advance came in the 1960s with the *positive real lemma*, which introduced the state description in network analysis and synthesis. A detailed treatment of these issues can be found in the book by Anderson and Vongpanitlerd [6].

5.9.2 Contractivity and the bounded real lemma*

Consider a stable system $\mathbf{\Sigma}$ (eigenvalues of $\mathbf{A}$ have negative real part) with the quadratic supply rate $\mathbf{s} = \|\mathbf{u}\|^2 - \|\mathbf{y}\|^2$ defined by (5.45). We will now see that the dissipativity of $\mathbf{\Sigma}$ with respect to $\mathbf{s}$ is equivalent to, among other things, the $\mathcal{H}_\infty$ norm of $\mathbf{\Sigma}$ being no greater than one. Systems satisfying this relationship are, for instance, passive electric circuits with input a voltage $\mathbf{V}$ and output some current $\mathbf{I}$, where $\mathbf{u} = \frac{1}{2}(\mathbf{V} + \mathbf{I})$ and $\mathbf{y} = \frac{1}{2}(\mathbf{V} - \mathbf{I})$. The following is a special case of Theorem 5.26.

Lemma 5.35. *Given is a stable and reachable system* $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{0} \end{array}\right)$ *with initial condition* $\mathbf{x}(0) = 0$. *The following statements are equivalent:*

(1) *The* 2*-induced norm of* $\boldsymbol{\Sigma}$ *is at most equal to one:* $\|\boldsymbol{\Sigma}\|_2 \leq 1$.

(2) *There exists a positive semidefinite* $\mathbf{X} \in \mathbb{R}^{n \times n}$ *and a matrix* $\mathbf{K} \in \mathbb{R}^{m \times n}$ *such that*

$$\frac{d}{dt}\mathbf{x}^*\mathbf{X}\mathbf{x} = \mathbf{u}^*\mathbf{u} - \mathbf{y}^*\mathbf{y} - \|\mathbf{u} + \mathbf{K}\mathbf{x}\|^2. \tag{5.53}$$

(3) *There exists a positive semidefinite solution* $\mathbf{X} \geq 0$ *of the Riccati equation*

$$\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} + \mathbf{C}^*\mathbf{C} + \mathbf{X}\mathbf{B}\mathbf{B}^*\mathbf{X} = \mathbf{0}. \tag{5.54}$$

The solution is positive definite $\mathbf{X} > 0$ *if the pair* $(\mathbf{C}, \mathbf{A})$ *is observable.*

(4) **Bounded Real Lemma.** *There exists* $\mathbf{X} \geq \mathbf{0}$ *such that*

$$\left.\begin{array}{l} \mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} + \mathbf{C}^*\mathbf{C} + \mathbf{K}^*\mathbf{K} = \mathbf{0} \\ \mathbf{X}\mathbf{B} + \mathbf{K}^* = \mathbf{0} \end{array}\right\}. \tag{5.55}$$

Proof. (2) → (1). Integrating (5.53) from $t = 0$ to $t = T$, and keeping in mind that $\mathbf{x}(0) = 0$, we obtain

$$\|\mathbf{u}\|^2_{\mathcal{L}_2(0,T)} - \|\mathbf{y}\|^2_{\mathcal{L}_2(0,T)} \geq \mathbf{x}^*(T)\mathbf{X}\mathbf{x}(T) \geq 0 \qquad \forall T > 0, \tag{5.56}$$

where the last inequality follows from the fact that $\mathbf{X}$ is positive semidefinite. Therefore,

$$\frac{\|\mathbf{y}\|^2_{\mathcal{L}_2(0,T)}}{\|\mathbf{u}\|^2_{\mathcal{L}_2(0,T)}} \leq 1 \qquad \forall T > 0. \tag{5.57}$$

Therefore, this property holds as $T \to \infty$; in other words, the $\mathcal{L}_2(0, \infty)$-norm of $\mathbf{y}$ is always less than that of $\mathbf{u}$ and hence the 2-induced norm of the convolution operator is less than or equal to 1.

(1) → (2). The proof of this part requires the construction of a positive semidefinite $\mathbf{X}$ satisfying (5.53). This is done as follows. Consider the available storage,

$$\Theta_{\text{avail}}(\mathbf{x}(0)) = \sup_{\mathbf{u}} \left\{ -\int_0^\infty (\|\mathbf{u}\|^2 - \|\mathbf{y}\|^2)dt \right\},$$

subject to the system equations and $\mathbf{x}(0) = \mathbf{0}$. We will show that it is a quadratic storage function. First notice that $\Theta_{\text{avail}}(\mathbf{x}(0)) \geq 0$ (take $\mathbf{u} = \mathbf{0}$ and $\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}(0)$, $t \geq 0$).

Now with $t_1 \geq 0$, we have

$$\Theta_{\text{avail}}(\mathbf{x}(0)) \geq -\int_0^{t_1} s\,dt - \int_{t_1}^{\infty} s\,dt.$$

Taking the supremum over all trajectories with $\mathbf{x}(t_1)$ fixed, the second term on the right-hand side of the above inequality becomes $\Theta_{\text{avail}}(\mathbf{x}(t_1))$. Therefore, Θ_{avail} is a storage function since it satisfies $\Theta_{\text{avail}}(\mathbf{x}(t_0)) \geq -\int_0^{t_1} s\,dt + \Theta_{\text{avail}}(\mathbf{x}(t_1))$. Furthermore, since the supply function is quadratic, by optimal control the storage function Θ_{avail} (which is the optimal cost) is also quadratic: $\Theta = \mathbf{x}^*\mathbf{X}\mathbf{x}$, where $\mathbf{X} \geq \mathbf{0}$. Statement (2) states furthermore that if a storage function exists, there exists $\mathbf{K}$ such that (5.53) is satisfied. Computing $\frac{d}{dt}(\mathbf{x}^*\mathbf{X}\mathbf{x}) - \mathbf{u}^*\mathbf{u} + \mathbf{y}^*\mathbf{y} + (\mathbf{u}^* + \mathbf{x}^*\mathbf{K}^*)(\mathbf{u} + \mathbf{K}\mathbf{x})$ and collecting terms, we obtain the expression

$$\mathbf{e} = \mathbf{x}^*(\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} + \mathbf{C}^*\mathbf{C} + \mathbf{K}^*\mathbf{K})\mathbf{x} + \mathbf{u}^*(\mathbf{B}^*\mathbf{X} + \mathbf{K})\mathbf{x} + \mathbf{x}^*(\mathbf{X}\mathbf{B} + \mathbf{K})\mathbf{u}.$$

By requiring that $\mathbf{e} = 0$ for all $\mathbf{x}$, $\mathbf{u}$ we obtain $\mathbf{K} = -\mathbf{X}\mathbf{B}$, where $\mathbf{X}$ satisfies the Riccati equation (5.54).

(2) → (3). The left-hand side of (5.53) can be rewritten as $(\frac{d}{dt}\mathbf{x}^*)\mathbf{X}\mathbf{x} + \mathbf{x}^*\mathbf{X}(\frac{d}{dt}\mathbf{x}) = (\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u})^*\mathbf{X}\mathbf{x} + \mathbf{x}^*\mathbf{X}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u})$. Thus, expanding the right-hand side of this equation and collecting terms on the left-hand side, we obtain $\mathbf{e} = 0$. As above, this implies $\mathbf{K} = -\mathbf{B}^*\mathbf{X}$, and substituting in the term which is quadratic in $\mathbf{x}$ we obtain (5.54). The implication (3) → (2) follows by reversing the above steps.

Finally, the equivalence (3) ↔ (4) is straightforward. □

Remark 5.9.3. If $\mathbf{D} \neq \mathbf{0}$, the *bounded real lemma* states that the $\mathcal{H}_\infty$-norm of $\mathbf{\Sigma}$ is at most one if and only if there exists $\mathbf{X} \geq \mathbf{0}$, $\mathbf{K}$, $\mathbf{L}$ such that

$$\boxed{\begin{aligned} &\mathbf{A}^*\mathbf{X} + \mathbf{X}\mathbf{A} + \mathbf{C}^*\mathbf{C} + \mathbf{K}^*\mathbf{K} = \mathbf{0}, \\ &\mathbf{X}\mathbf{B} + \mathbf{C}^*\mathbf{D} + \mathbf{K}^*\mathbf{L} = \mathbf{0}, \\ &\mathbf{I} - \mathbf{D}^*\mathbf{D} = \mathbf{L}^*\mathbf{L}. \end{aligned}} \tag{5.58}$$

This follows from (5.48) with $\mathbf{Q}_{11} = -\mathbf{C}^*\mathbf{C}$, $\mathbf{Q}_{12} = -\mathbf{C}^*\mathbf{D}$, and $\mathbf{Q}_{22} = \mathbf{I} - \mathbf{D}^*\mathbf{D}$.

5.10 Chapter summary

In the first part of this chapter we discussed various norms for vector- and matrix-valued functions of time and frequency. As in the case of constant vectors and matrices, the most useful are the p norms, where $p = 1, 2, \infty$. Norms of linear systems $\mathbf{\Sigma}$ are introduced next; of importance are the 2-induced norm of the *convolution operator* $\mathcal{S}$ and the 2-induced norm of the associated *Hankel operator* $\mathcal{H}$. The former is also known as the ∞-norm of the system because it is the ∞-norm of the transfer function of $\mathbf{\Sigma}$. The latter is known as the Hankel-norm of $\mathbf{\Sigma}$. Important invariants are the singular values of $\mathcal{H}$, known as *Hankel singular values* of $\mathbf{\Sigma}$. Lemma 5.8 shows that they are the square roots of the eigenvalues of the product of the (infinite) reachability and observability gramians. Consequently, their computation involves the solution of two Lyapunov equations. As we will see, these invariants provide a trade-off between desired accuracy and required complexity for a certain type of approximation. Another important norm is the $\mathcal{H}_2$-norm of $\mathbf{\Sigma}$; two different ways of computing it are given.

For vector- or matrix-valued time/frequency signals, one needs to choose a norm for both the spatial dimension and the time/frequency. Although these are usually taken to be the same, this is not necessary. Various so-called mixed-induced norms are discussed in section 5.6. For a summary of this part, see section 5.7. The system norms discussed above assume the stability of $\mathbf{\Sigma}$. Following formulas (5.38) and (5.39), these definitions can be extended to unstable $\mathbf{\Sigma}$.

The second part of the chapter is devoted to stability and dissipativity. The former is a property of the trajectories of an autonomous or free system (they must all remain bounded or decay to zero). The concept of Lyapunov function was also discussed. For open systems, that is, systems with inputs and outputs, the proper generalization of stability is provided by the concept of dissipativity. The idea is that given a supply function (e.g., the power supplied to an electric circuit), only part of the supplied energy can be stored, while the rest is dissipated. Besides the supply function $\mathbf{s}$, also important are the storage function Θ and the dissipation function $\mathbf{d}$, which are related by means of the equation $\mathbf{d} = \mathbf{s} - \frac{d}{dt}\Theta$. The important special cases of passive and contractive systems are discussed in some detail.

Chapter 6

Sylvester and Lyapunov Equations

This chapter is dedicated to the investigation of the *Sylvester equation*, which is a linear matrix equation. A special case is the *Lyapunov equation*. These equations underlie many of the considerations for model reduction. In particular, the Lyapunov equation has a remarkable property, described by the *inertia result*, which is a powerful way of checking stability of approximants. Lyapunov was interested in studying the distribution of the roots of a polynomial equation with respect to the imaginary axis in the complex plane. Of course, one can do this by defining a matrix whose characteristic polynomial is equal to the given one. But Lyapunov wanted to accomplish this using *quadratic forms*. He observed that under mild assumptions, the solution to an appropriate Lyapunov equation is symmetric—and thus defines a quadratic form—and its eigenvalues are distributed in the same way as the roots of the original polynomial. The distribution of the eigenvalues of a symmetric matrix can be determined, using a classical result of Jacobi, by checking the signs of the principal minors of the matrix; for details, see, e.g., Chapter X of [134].

We begin by listing several ways of solving the Sylvester and Lyapunov equations (section 6.1). In section 6.2, the inertia result is stated and proved using two different methods. Next, various algorithms for the numerical solution of such equations are presented. For the semidefinite Lyapunov equation, the most reliable is the square root method, which computes a Cholesky factor of the solution directly, without computing the solution first. The chapter concludes with remarks on the numerical stability of solution algorithms for the Sylvester and Lyapunov equations.

6.1 The Sylvester equation

Given the matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times k}$, and $\mathbf{C} \in \mathbb{R}^{n\times k}$, the matrix equation

$$\boxed{\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{B} = \mathbf{C}} \tag{6.1}$$

in the unknown $\mathbf{X} \in \mathbb{R}^{n\times k}$ is the Sylvester equation. In the case in which $\mathbf{B} = \mathbf{A}^*$ and $\mathbf{C} = \mathbf{Q} = \mathbf{Q}^*$, the resulting equation,

$$\boxed{\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* = \mathbf{Q},} \tag{6.2}$$

in the unknown $\mathcal{P} \in \mathbb{R}^{n\times n}$ is referred to as the Lyapunov equation.

Before discussing methods for solving this equation, we state a condition for the existence of a solution. First we need to define the matrices

$$\mathbf{M}_1 = \begin{pmatrix} \mathbf{A} & -\mathbf{C} \\ \mathbf{0} & -\mathbf{B} \end{pmatrix}, \ \mathbf{M}_2 = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\mathbf{B} \end{pmatrix}, \tag{6.3}$$

both being square with size $(n + k)$. Roth [276] proved the following result.

Proposition 6.1. *Equation* (6.1) *has a solution if and only if the matrices* $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *are similar.*

Proof. Let $\mathbf{X}$ be a solution of (6.1). Then $\mathbf{M}_3 = \begin{pmatrix} \mathbf{I} & \mathbf{X} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ is nonsingular and satisfies $\mathbf{M}_1\mathbf{M}_3 = \mathbf{M}_3\mathbf{M}_2$. Conversely, if $\mathbf{M}_1$ and $\mathbf{M}_2$ are similar, there exists a nonsingular matrix $\mathbf{M}_3 = \begin{pmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{0} & \mathbf{X}_{22} \end{pmatrix}$ such that $\mathbf{M}_1\mathbf{M}_3 = \mathbf{M}_3\mathbf{M}_2$. It follows that $\mathbf{X} = \mathbf{X}_{12}\mathbf{X}_{22}^{-1}$ is a solution of (6.1). □

Several methods for solving this equation have been developed and will be discussed. We will also derive conditions (6.7) and (6.8), which guarantee the existence of a *unique* solution. The methods are as follows:

1. The Kronecker product method,
2. The eigenvalue/complex integration method,
3. The eigenvalue/eigenvector method,
4. The characteristic polynomial method,
5. The invariant subspace method,
6. The sign function method,
7. The infinite sum method, and
8. The square root method (section 6.3.3).

Note that *reachability* and *observability* are intimately related with properties of the Sylvester equation; see, e.g., [97].

6.1.1 The Kronecker product method

The Sylvester equation (6.1) can be analyzed using the *Kronecker product*. For details, see, among others, [219], [220], [181].

Given $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{p\times q}$ and $\mathbf{R} \in \mathbb{R}^{r\times s}$, their *Kronecker product* is defined as follows:

$$\mathbf{P} \otimes \mathbf{R} = \left[p_{ij}\mathbf{R}\right] = \begin{pmatrix} p_{11}\mathbf{R} & \cdots & p_{1q}\mathbf{R} \\ & & \\ p_{p1}\mathbf{R} & \cdots & p_{pq}\mathbf{R} \end{pmatrix} \in \mathbb{R}^{pr\times qs}.$$

We can also write

$$\mathbf{P} \otimes \mathbf{R} = \left[p_1 \otimes \mathbf{R} \cdots p_p \otimes \mathbf{R}\right] = \left[p_1 \otimes r_1 \cdots p_1 \otimes r_m | \cdots | p_p \otimes r_1 \cdots p_p \otimes r_m\right],$$

where p_i, r_i, denotes the ith column of $\mathbf{P}$, $\mathbf{R}$, respectively. The *vector* associated with a matrix $\mathbf{T} \in \mathbb{R}^{q\times r}$ is

$$\mathrm{vec}\,(\mathbf{T}) = (t_{11} \; \cdots \; t_{q1} \; t_{12} \; \cdots \; t_{q2} \; \cdots \; t_{1r} \; \cdots \; t_{qr})^* \in \mathbb{R}^{qr}.$$

It follows that

$$\mathrm{vec}\,(\mathbf{x}\mathbf{y}^*) = [y_i\mathbf{x}] = \mathbf{y} \otimes \mathbf{x} \;\Rightarrow\; \mathrm{vec}\,\left(\mathbf{x} \;\; \hat{\mathbf{x}}\right)\begin{pmatrix}\mathbf{y}^* \\ \hat{\mathbf{y}}^*\end{pmatrix} = \mathbf{y} \otimes \mathbf{x} + \hat{\mathbf{y}} \otimes \hat{\mathbf{x}}.$$

Therefore, if $\mathbf{C}^* = (\mathbf{c}_1 \; \cdots \; \mathbf{c}_k) \in \mathbb{R}^{n\times k}$, we can write $\mathbf{C} = \mathbf{e}_1\mathbf{c}_1^* + \cdots + \mathbf{e}_k\mathbf{c}_k^*$, which implies

$$\mathrm{vec}\,\mathbf{C} = \mathbf{e}_1 \otimes \mathbf{c}_1 + \cdots + \mathbf{e}_k \otimes \mathbf{c}_k \in \mathbb{R}^{nk}.$$

Using the Kronecker product, the left-hand side of (6.1) can be rewritten as $\mathrm{vec}\,(\mathbf{AX} + \mathbf{XB}) = \mathcal{A}_{\mathbf{B}}\,\mathrm{vec}\,(\mathbf{X})$, and thus

$$\mathcal{A}_{\mathbf{B}}\,\mathrm{vec}\,(\mathbf{X}) = \mathrm{vec}\,(\mathbf{C}), \quad \text{where} \quad \mathcal{A}_{\mathbf{B}} = \mathrm{I_k} \otimes \mathbf{A} + \mathbf{B}^* \otimes \mathrm{I_n}. \tag{6.4}$$

Let the EVD of $\mathbf{A}$ and $\mathbf{B}$ be

$$\mathbf{AV} = \mathbf{V}\Lambda, \;\; \mathbf{W}^*\mathbf{B} = \mathbf{M}\mathbf{W}^*, \tag{6.5}$$

respectively. We assume for simplicity that these matrices are diagonalizable:

$$\Lambda = \mathrm{diag}\,(\lambda_i), \;\; \mathbf{M} = \mathrm{diag}\,(\mu_j), \;\; \mathbf{V} = [\mathbf{v}_1 \; \cdots \; \mathbf{v}_n], \;\; \mathbf{W} = [\mathbf{w}_1 \; \cdots \; \mathbf{w}_k]. \tag{6.6}$$

Given the eigenvectors of $\mathbf{A}$ and $\mathbf{B}$, the eigenvectors of $\mathcal{A}_{\mathbf{B}}$ are

$$\begin{aligned}\left[\mathrm{vec}\,(\mathbf{w}_1\mathbf{v}_1^*) \cdots \mathrm{vec}\,(\mathbf{w}_n\mathbf{v}_k^*)\right] &= [\mathbf{w}_1 \otimes \mathbf{v}_1 \cdots \mathbf{w}_1 \otimes \mathbf{v}_k \; \cdots \; \mathbf{w}_n \otimes \mathbf{v}_1 \cdots \mathbf{w}_n \otimes \mathbf{v}_k] \\ &= [\mathbf{w}_1 \otimes \mathbf{V} \; \cdots \; \mathbf{w}_n \otimes \mathbf{V}] = \mathbf{W} \otimes \mathbf{V}.\end{aligned}$$

It will be shown in section 6.1.3 that the eigenvalues of $\mathcal{A}_{\mathbf{B}}$ are

$$\lambda(\mathcal{A}_{\mathbf{B}}) = \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}) \qquad \forall\, i, j.$$

As a consequence, we obtain the following proposition.

Proposition 6.2. *The Sylvester equation* (6.1) *has a unique solution* **X** *if and only if*

$$\lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}) \neq 0 \qquad \forall\, i, j. \tag{6.7}$$

In this case, (6.4) *yields*

$$\mathrm{vec}\,(\mathbf{X}) = \mathcal{A}_{\mathbf{B}}^{-1}\mathrm{vec}\,(\mathbf{C}).$$

Corollary 6.3. *The Lyapunov equation* (6.2) *has a unique solution* $\mathcal{P}$ *if and only if*

$$\lambda_i(\mathbf{A}) + \lambda_j^*(\mathbf{A}) \neq 0 \qquad \forall\, i, j. \tag{6.8}$$

Furthermore, if the solution is unique, it is Hermitian $\mathcal{P} = \mathcal{P}^*$.

6.1.2 The eigenvalue/complex integration method

Consider the Sylvester equation, where the term $z\mathbf{X}$, $z \in \mathbb{C}$, is added and subtracted on both sides:

$$(\mathbf{A} - z\mathbf{I})\mathbf{X} + \mathbf{X}(\mathbf{B} + z\mathbf{I}) = \mathbf{C} \;\Rightarrow\; \mathbf{X}(z\mathbf{I} + \mathbf{B})^{-1} + (\mathbf{A} - z\mathbf{I})^{-1}\mathbf{X} = (\mathbf{A} - z\mathbf{I})^{-1}\mathbf{C}(\mathbf{B} + z\mathbf{I})^{-1}.$$

Assuming that the sets of eigenvalues of **A** and $-\mathbf{B}$ are disjoint, let Γ be a Cauchy contour[4] that contains all the eigenvalues of **A** but none of the eigenvalues of $-\mathbf{B}$. There holds

$$\int_\Gamma (\mathbf{B} + z\mathbf{I})^{-1} dz = 0, \quad \int_\Gamma (\mathbf{A} - z\mathbf{I})^{-1} dz = 2\pi i\, \mathbf{I}_n.$$

The earlier expression, together with these integrals, yields the solution of the Lyapunov equation as a complex integral:

$$\mathbf{X} = \frac{1}{2\pi i}\int_\Gamma (\mathbf{A} - z\mathbf{I})^{-1}\mathbf{C}(\mathbf{B} + z\mathbf{I})^{-1} dz. \tag{6.9}$$

Remark 6.1.1. This method can also be used to obtain a solution of the equation $\mathbf{A}_1\mathbf{X}\mathbf{A}_2 + \mathbf{B}_1\mathbf{X}\mathbf{B}_2 = \mathbf{C}$. Provided that the spectra of the associated pencils $z\mathbf{B}_1 - \mathbf{A}_1$ and $z\mathbf{A}_2 + \mathbf{B}_2$ are disjoint, the solution can be expressed as follows:

$$\mathbf{X} = \frac{1}{2\pi i}\int_\Gamma (z\mathbf{B}_1 - \mathbf{A}_1)^{-1}\mathbf{C}(z\mathbf{A}_2 + \mathbf{B}_2)^{-1}\, dz,$$

where Γ is a Cauchy contour that includes the spectrum of the pencil $z\mathbf{B}_1 - \mathbf{A}_1$ and excludes that of the pencil $z\mathbf{A}_2 + \mathbf{B}_2$.

The special case of Lyapunov equations

If we consider the Lyapunov equation, in other words, $\mathbf{B} = \mathbf{A}^*$, $\mathbf{C} = \mathbf{Q}$, and **A** has eigenvalues in $\mathbb{C}_-$, we can choose the contour Γ above as the imaginary axis, $z = i\omega$,

[4] See (almost) any book on complex analysis for details on complex integration and Cauchy contours.

and formula (6.9) becomes

$$\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{Q}(-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\, d\omega. \tag{6.10}$$

It readily follows by Parseval's theorem of Fourier transforms that $\mathcal{P}$ can be expressed in the time domain as

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

which is the same as (4.43) and (4.51) with $\mathbf{Q}$ in place of $\mathbf{BB}^*$.

We now examine the case where $\mathbf{A}$ has eigenvalues both in $\mathbb{C}_-$ and in $\mathbb{C}_+$ but not on the imaginary axis. Furthermore, we assume that the contour Γ is the imaginary axis. The question we wish to address is, *What equation does the quantity $\hat{\mathcal{P}}$ satisfy*?

$$\hat{\mathcal{P}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{Q}(-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\, d\omega \in \mathbb{R}^{n\times n}. \tag{6.11}$$

First notice that under the assumptions above, this expression is well defined. The question posed is of interest when dealing with Lyapunov equations where $\mathbf{A}$ is neither stable (eigenvalues in $\mathbb{C}_-$) nor antistable (eigenvalues in $\mathbb{C}_+$) but has eigenvalues in both half planes, except on the imaginary axis. In such a case, the solution of the Lyapunov equation does not admit an integral representation in the time domain (such as $\int e^{\mathbf{A}t}\mathbf{Q}e^{\mathbf{A}^*t}\, dt$).

To address this issue, we define the projection Π onto the *stable* eigenspace of $\mathbf{A}$ in $\mathbb{R}^n$; then $\mathbf{I} - \Pi$ is the projection onto the *antistable* eigenspace of $\mathbf{A}$ in $\mathbb{R}^n$. This projection can be computed as follows. Let $\mathbf{T}$ be a transformation such that $\mathbf{T}^{-1}\mathbf{AT} = \text{diag}\,\{\Lambda_-, \Lambda_+\}$, where all the eigenvalues of Λ_- are in $\mathbb{C}_-$ and all the eigenvalues of Λ_+ in $\mathbb{C}_+$. Then $\Pi = \mathbf{T}\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{T}^{-1}$. Notice that

$$\mathbf{A} = (\mathbf{I} - \Pi)\mathbf{A}(\mathbf{I} - \Pi) + \Pi\mathbf{A}\Pi = \mathbf{A}_- + \mathbf{A}_+.$$

The main result of this section is due to Godunov [142, Chapter 10].

Lemma 6.4. *$\hat{\mathcal{P}}$ defined by* (6.11) *solves the following Lyapunov equation:*

$$\mathbf{A}\hat{\mathcal{P}} + \hat{\mathcal{P}}\mathbf{A}^* = \Pi\mathbf{Q}\Pi - (\mathbf{I} - \Pi)\mathbf{Q}(\mathbf{I} - \Pi). \tag{6.12}$$

With the notation $\mathbf{Q}_- = \Pi\mathbf{Q}\Pi$ and $\mathbf{Q}_+ = (\mathbf{I} - \Pi)\mathbf{Q}(\mathbf{I} - \Pi)$, this Lyapunov equation can be considered as the sum of the Lyapunov equations:

$$\mathbf{A}_-\hat{\mathcal{P}}_- + \hat{\mathcal{P}}_-\mathbf{A}_-^* = \mathbf{Q}_-, \quad \mathbf{A}_+\hat{\mathcal{P}}_+ + \hat{\mathcal{P}}_+\mathbf{A}_+^* = -\mathbf{Q}_+.$$

Thus $\hat{\mathcal{P}} = \hat{\mathcal{P}}_- - \hat{\mathcal{P}}_+$ has an integral representation in the time domain, namely,

$$\hat{\mathcal{P}} = \underbrace{\int_0^{\infty} e^{\mathbf{A}_-t}\mathbf{Q}_-e^{\mathbf{A}_-^*t}\, dt}_{\hat{\mathcal{P}}_-} - \underbrace{\int_{-\infty}^{0} e^{\mathbf{A}_+t}\mathbf{Q}_+e^{\mathbf{A}_+^*t}\, dt}_{\hat{\mathcal{P}}_+}.$$

where

$$\mathbf{A}_- = \mathbf{T}\begin{pmatrix} \Lambda_- & \\ & \mathbf{0} \end{pmatrix}\mathbf{T}^{-1} \text{ and } \mathbf{A}_+ = \mathbf{T}\begin{pmatrix} \mathbf{0} & \\ & \Lambda_+ \end{pmatrix}\mathbf{T}^{-1}.$$

6.1.3 The eigenvalue/eigenvector method

Define the linear operator

$$\mathcal{L}:\ \mathbb{R}^{n\times k}\to\mathbb{R}^{n\times k},\quad \text{where } \mathbf{X}\longmapsto\mathcal{L}(\mathbf{X})=\mathbf{AX}+\mathbf{XB}. \tag{6.13}$$

One way to analyze the operator $\mathcal{L}$ is to use the EVD of $\mathbf{A}$ and $\mathbf{B}$.

Recall the EVDs given in (6.5) and (6.6); it follows that $\mathbf{v}_i$, $\mathbf{w}_j^*$ are right, left, eigenvectors of $\mathbf{A}$, $\mathbf{B}$, corresponding to the eigenvalues λ_i, μ_j, respectively. Then, since

$$\mathcal{L}(\mathbf{v}_i\mathbf{w}_j^*)=\mathbf{A}\mathbf{v}_i\mathbf{w}_j^*+\mathbf{v}_i\mathbf{w}_j^*\mathbf{B}=(\lambda_i+\mu_j)\mathbf{v}_i\mathbf{w}_j^*,$$

we conclude that $\lambda_i+\mu_j$ is an eigenvalue of $\mathcal{L}$ corresponding to the eigenvector $\mathbf{v}_i\mathbf{w}_j^*$. Let also the left eigenvectors of $\mathbf{A}$ be

$$\mathbf{V}^{-1}=\hat{\mathbf{V}}=\begin{pmatrix}\hat{\mathbf{v}}_1^*\\ \vdots\\ \hat{\mathbf{v}}_n^*\end{pmatrix}$$

and the right eigenvectors of $\mathbf{B}$ be $\hat{\mathbf{W}}=\mathbf{W}^{-*}=[\hat{\mathbf{w}}_1\cdots\hat{\mathbf{w}}_k]$; it readily follows that $\sum_{i=1}^n\mathbf{v}_i\hat{\mathbf{v}}_i^*=\mathbf{I}_n$ and $\sum_{i=1}^k\hat{\mathbf{w}}_i\mathbf{w}_i^*=\mathbf{I}_k$. Given these eigenvalue decompositions, the solution of the Sylvester equation can be explicitly written.

Theorem 6.5. *With the notation established above, the solution of the Sylvester equation* (6.1) *can be expressed as a* sum *of* rank-one *matrices:*

$$\mathbf{X}=\sum_{i=1}^n\mathbf{v}_i\hat{\mathbf{v}}_i^*\,\mathbf{C}\,(\lambda_i\mathbf{I}_k+\mathbf{B})^{-1}=\sum_{j=1}^k(\mu_j\mathbf{I}_n+\mathbf{A})^{-1}\,\mathbf{C}\,\hat{\mathbf{w}}_j\mathbf{w}_j^*. \tag{6.14}$$

These expressions can also be written as

$$\mathbf{X}=\mathbf{V}\begin{bmatrix}\hat{\mathbf{v}}_1^*\,\mathbf{C}\,(\lambda_1\mathbf{I}_k+\mathbf{B})^{-1}\\ \vdots\\ \hat{\mathbf{v}}_n^*\,\mathbf{C}\,(\lambda_n\mathbf{I}_k+\mathbf{B})^{-1}\end{bmatrix}=\begin{bmatrix}(\mu_1\mathbf{I}_n+\mathbf{A})^{-1}\,\mathbf{C}\,\hat{\mathbf{w}}_1 & \cdots & (\mu_k\mathbf{I}_n+\mathbf{A})^{-1}\,\mathbf{C}\,\hat{\mathbf{w}}_k\end{bmatrix}\mathbf{W}^* \tag{6.15}$$

Proof. From (6.1) follows $(\mathbf{A}-\lambda_i\mathbf{I}_n)\mathbf{X}+\mathbf{X}(\mathbf{B}+\lambda_i\mathbf{I}_k)=\mathbf{C}$, where λ_i is an eigenvalue of $\mathbf{A}$, with corresponding left/right eigenvectors $\hat{\mathbf{v}}_i$, $\mathbf{v}_i$; hence $\hat{\mathbf{v}}_i^*(\mathbf{A}-\lambda_i\mathbf{I}_n)=\mathbf{0}$, which leads to $\hat{\mathbf{v}}_i^*\mathbf{X}=\hat{\mathbf{v}}_i^*\mathbf{C}(\lambda_i\mathbf{I}_k+\mathbf{B})^{-1}$; hence $\mathbf{v}_i\hat{\mathbf{v}}_i^*\mathbf{X}=\mathbf{v}_i\hat{\mathbf{v}}_i^*\mathbf{C}(\lambda_i\mathbf{I}_k+\mathbf{B})^{-1}$. Due to $\sum_{i=1}^n\mathbf{v}_i\hat{\mathbf{v}}_i^*=\mathbf{I}_n$, we have $\sum_{i=1}^n\mathbf{v}_i\hat{\mathbf{v}}_i^*\mathbf{X}=\mathbf{X}$. The first formula follows; the second formula can be shown similarly. □

Note that the formulas above contain a number of ingredients which are the same as those in [151].

Remark 6.1.2. *The Sylvester equation and the Löwner matrix.* We will now present a connection between the solution of the Sylvester equation and the Löwner matrix; consequently a connection is established between the Sylvester equation and rational interpolation.

Let us assume for argument's sake that $\mathbf{C}$ is rank one; it can then be written as $\mathbf{C} = \mathbf{c}_1\mathbf{c}_2^*$, $\mathbf{c}_1 \in \mathbb{R}^n$, $\mathbf{c}_2 \in \mathbb{R}^k$. Then (6.15) can be written as follows:

$$\mathbf{X} = \underbrace{\mathbf{V}\begin{bmatrix} \hat{\mathbf{v}}_1^*\mathbf{c}_1 & & \\ & \ddots & \\ & & \hat{\mathbf{v}}_n^*\mathbf{c}_1 \end{bmatrix}}_{\tilde{\mathbf{V}}} \underbrace{\begin{bmatrix} \mathbf{c}_2^*(\lambda_1\mathbf{I}+\mathbf{B})^{-1} \\ \vdots \\ \mathbf{c}_2^*(\lambda_n\mathbf{I}+\mathbf{B})^{-1} \end{bmatrix}}_{\mathcal{O}(\mathbf{c}_2^*,\mathbf{B})}$$

$$= \underbrace{\begin{bmatrix} (\mu_1\mathbf{I}+\mathbf{A})^{-1}\mathbf{c}_1 & \cdots & (\mu_k\mathbf{I}+\mathbf{A})^{-1}\mathbf{c}_1 \end{bmatrix}}_{\mathcal{R}(\mathbf{A},\mathbf{c}_1)} \underbrace{\begin{bmatrix} \mathbf{c}_2^*\hat{\mathbf{w}}_1 & & \\ & \ddots & \\ & & \mathbf{c}_2^*\hat{\mathbf{w}}_n \end{bmatrix}\mathbf{W}^*}_{\tilde{\mathbf{W}}^*}.$$

Thus $\tilde{\mathbf{V}}$ is the matrix of scaled right eigenvectors of $\mathbf{A}$ while $\tilde{\mathbf{W}}$ is the matrix of scaled left eigenvectors of $\mathbf{B}$. It is interesting to notice that the remaining matrices $\mathcal{O}$ and $\mathcal{R}$ are directly related with rational interpolation. Recall (4.85) and (4.86) as well as (4.87), which show that the Löwner matrix can be factorized as a product of *generalized* reachability and observability matrices. The matrices defined above, namely, $\mathcal{R}(\mathbf{A}, \mathbf{c}_1)$ and $\mathcal{O}(\mathbf{c}_2^*, \mathbf{B})$, are precisely these generalized matrices, $\mathbf{X} = \tilde{\mathbf{V}}\mathcal{O}(\mathbf{c}_2^*, \mathbf{B}) = \mathcal{R}(\mathbf{A}, \mathbf{c}_1)\tilde{\mathbf{W}}^*$, and if $\mathbf{B} = \mathbf{A}$, $\mathbf{X}^2 = -\tilde{\mathbf{V}}\boldsymbol{L}\tilde{\mathbf{W}}^*$. Thus $\mathbf{X}^2$ and $\boldsymbol{L}$ are the same up to congruence.

6.1.4 Characteristic polynomial methods

Two methods make use of characteristic polynomials. The first uses the characteristic polynomial of the operator $\mathcal{L}$ and the second the characteristic polynomial of either $\mathbf{A}$ or $\mathbf{B}$. We begin with the former.

Given a matrix $\Gamma \in \mathbb{R}^{n\times n}$, with characteristic polynomial $\chi_\Gamma(s) = s^n + \gamma_{n-1}s^{n-1} + \cdots + \gamma_1 s + \gamma_0$, the Cayley–Hamilton theorem implies that $\chi_\Gamma(\Gamma) = \mathbf{0}$. If Γ is invertible, that is, $\det\Gamma = \gamma_0 \neq 0$, the inverse is given by

$$\Gamma^{-1} = \frac{-1}{\gamma_0}\left[\Gamma^{n-1} + \gamma_{n-1}\Gamma^{n-2} + \cdots + \gamma_2\Gamma + \gamma_1\mathbf{I}\right].$$

Consequently, the solution of $\Gamma\mathbf{x} = \mathbf{b}$ can be expressed as

$$\mathbf{x} = \Gamma^{-1}\mathbf{b} = \frac{-1}{\gamma_0}\left[\Gamma^{n-1}\mathbf{b} + \gamma_{n-1}\Gamma^{n-2}\mathbf{b} + \cdots + \gamma_2\Gamma\mathbf{b} + \gamma_1\mathbf{b}\right]. \tag{6.16}$$

Therefore, if the coefficients of the characteristic polynomial of Γ are known, an expression for the solution of $\Gamma\mathbf{x} = \mathbf{b}$ can be written, involving powers of Γ.

Next we will apply this method to the Sylvester equation (6.1). Recall the Sylvester operator $\mathcal{L}$ defined by (6.13). Its characteristic polynomial $\chi_\mathcal{L}$ has degree nk, and its roots are $\lambda_i + \mu_j$, $i = 1, \ldots, n$, $j = 1, \ldots, k$, where λ_i, μ_j are the eigenvalues of $\mathbf{A}$, $\mathbf{B}$, respectively. Therefore, the characteristic polynomial of $\mathcal{L}$ is the same as that of $\mathcal{A}_\mathbf{B}$ defined in (6.4). Let

$$\chi_\mathcal{L}(s) = s^{nk} + c_{nk-1}s^{nk-1} + c_{nk-2}s^{nk-2} + \cdots + c_2 s^2 + c_1 s + c_0.$$

Notice that the Sylvester equation can be written as $\mathcal{L}(\mathbf{X}) = \mathbf{C}$; therefore, $\mathbf{X} = \mathcal{L}^{-1}(\mathbf{C})$ (provided that the inverse exists). Cayley–Hamilton implies

$$\chi_{\mathcal{L}}\left(\mathcal{L}(\mathbf{C})\right) = \mathbf{0} \;\Rightarrow\; \mathcal{L}^{-1}(\mathbf{C}) \cdot \chi_{\mathcal{L}}\left(\mathcal{L}(\mathbf{C})\right) = \mathbf{0}.$$

In analogy to (6.16), this yields an expression for the solution of the Sylvester equation (6.1) of the following form:

$$\mathbf{X} = \frac{-1}{c_0}\left[\mathcal{L}^{nk-1}(\mathbf{C}) + c_{nk-1}\mathcal{L}^{nk-2}(\mathbf{C}) + \cdots + c_3\mathcal{L}^2(\mathbf{C}) + c_2\mathcal{L}^1(\mathbf{C}) + c_1\mathcal{L}^0(\mathbf{C})\right]. \quad (6.17)$$

The quantities $\mathcal{L}^j(\mathbf{C})$ can be obtained recursively as follows:

$$\mathcal{L}^0(\mathbf{C}) = \mathbf{C}, \;\; \mathcal{L}^j(\mathbf{C}) = \mathbf{A}\mathcal{L}^{j-1}(\mathbf{C}) + \mathcal{L}^{j-1}(\mathbf{C})\mathbf{B}, \qquad j = 1, \ldots, nk. \quad (6.18)$$

This procedure has been adapted from Peeters and Rapisarda [264], who, using (6.17) and (6.18) as a basis, propose a general recursive method for solving polynomial Lyapunov equations. A similar method for solving Sylvester and Lyapunov equations was proposed by Hanzon and Peeters [166].

The *second method* that falls under the characteristic polynomial approach makes use of the characteristic polynomial of either **A** or **B**. Let these be

$$\alpha(s) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0, \;\; \beta(s) = s^k + \beta_{k-1}s^{k-1} + \cdots + \beta_1 s + \beta_0.$$

We also consider the *pseudoderivative* polynomials $\alpha^{(i)}$, $i = 1, \ldots, n$, and $\beta^{(j)}$, $j = 1, \ldots, k$, defined as in (4.72). Then from $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$ follow the relationships

$$\begin{array}{rcl}
\mathbf{X} - \mathbf{X} & = & \mathbf{0}, \\
\mathbf{AX} + \mathbf{XB} & = & \mathbf{C}, \\
\mathbf{A}^2\mathbf{X} - \mathbf{XB}^2 & = & \mathbf{AC} - \mathbf{CB}, \\
\mathbf{A}^3\mathbf{X} + \mathbf{XB}^3 & = & \mathbf{A}^2\mathbf{C} - \mathbf{ACB} + \mathbf{CB}^2, \\
\mathbf{A}^4\mathbf{X} - \mathbf{XB}^4 & = & \mathbf{A}^3\mathbf{C} - \mathbf{A}^2\mathbf{CB} + \mathbf{ACB}^2 - \mathbf{CB}^3, \\
\vdots & & \vdots
\end{array}$$

We now form a linear combination of the above equations; this linear combination is given by the coefficients of the characteristic polynomial of A. We thus obtain

$$\alpha(\mathbf{A})\mathbf{X} - \mathbf{X}\alpha(-\mathbf{B}) = \alpha^{(1)}(\mathbf{A})\mathbf{C} - \alpha^{(2)}(\mathbf{A})\mathbf{CB} + \alpha^{(3)}(\mathbf{A})\mathbf{CB}^2 - \alpha^{(4)}(\mathbf{A})\mathbf{CB}^3 + \cdots.$$

By the Cayley–Hamilton method, the first term is zero, and therefore we obtain the solution

$$\mathbf{X} = -\left[\sum_{i=1}^{n} \alpha^{(i)}(\mathbf{A})\mathbf{CB}^{i-1}\right][\alpha(-\mathbf{B})]^{-1}, \quad (6.19)$$

where $\alpha(-\mathbf{B})$ is invertible because **A** and $-\mathbf{B}$ are assumed to have no common eigenvalues. In a similar way, making use of the characteristic polynomial of **B**, we obtain the dual formula to (6.19),

$$\mathbf{X} = -[\beta(-\mathbf{A})]^{-1}\left[\sum_{j=1}^{k} \mathbf{A}^{j-1}\mathbf{C}\beta^{(j)}(\mathbf{B})\right]. \quad (6.20)$$

We can go one step further and compute the inverse of the quantity in the second bracket. For this purpose, we need to solve the following polynomial *Bezout* equation for the coefficients of polynomials $\mathbf{p}(-s)$ and $\mathbf{q}(s)$:

$$\alpha(s)\mathbf{p}(-s)+\beta(-s)\mathbf{q}(s)=1.$$

Then, since $\alpha(-\mathbf{B})\mathbf{p}(\mathbf{B})=\mathbf{I}_k$ and $\beta(-\mathbf{A})\mathbf{q}(\mathbf{A})=\mathbf{I}_n$, we obtain the formulas

$$\mathbf{X}=-\left[\sum_{i=1}^{n}\alpha^{(i)}(\mathbf{A})\mathbf{C}\mathbf{B}^{i-1}\right]\mathbf{p}(\mathbf{B})=-\mathbf{q}(\mathbf{A})\left[\sum_{j=1}^{k}\mathbf{A}^{j-1}\mathbf{C}\beta^{(j)}(\mathbf{B})\right]$$

Example 6.6. We now consider the above methods for solving the Sylvester equation (6.1). It is assumed that

$$\mathbf{A}=\begin{pmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\-24&-50&-35&-10\end{pmatrix},\ \mathbf{B}=\begin{pmatrix}-1&-3/2\\-1/2&0\end{pmatrix},\ \mathbf{C}=-\begin{pmatrix}1&1/2\\1/2&1\\3/2&1\\3/2&2\end{pmatrix}$$

(a) The first method uses the Kronecker product:

$$\mathcal{A}_{\mathbf{B}}=(\mathbf{I}_2\otimes\mathbf{A})+(\mathbf{B}^*\otimes\mathbf{I}_4)$$

$$=\left(\begin{array}{cccc|cccc}-1&1&0&0&-\frac{1}{2}&0&0&0\\0&-1&1&0&0&-\frac{1}{2}&0&0\\0&0&-1&1&0&0&-\frac{1}{2}&0\\-24&-50&-35&-11&0&0&0&-\frac{1}{2}\\\hline-\frac{3}{2}&0&0&0&0&1&0&0\\0&-\frac{3}{2}&0&0&0&0&1&0\\0&0&-\frac{3}{2}&0&0&0&0&1\\0&0&0&-\frac{3}{2}&-24&-50&-35&-10\end{array}\right)$$

$$\operatorname{vec}(\mathbf{C})=-\left(\begin{array}{c}1\\\frac{1}{2}\\\frac{3}{2}\\\frac{3}{2}\\\hline\frac{1}{2}\\1\\1\\2\end{array}\right)\Rightarrow\operatorname{vec}(\mathbf{X})=-\mathcal{A}_{\mathbf{B}}^{-1}\operatorname{vec}(\mathbf{C})=\left(\begin{array}{c}1\\0\\0\\-2\\\hline0\\1\\-1\\-1\end{array}\right)\Rightarrow\mathbf{X}=\begin{pmatrix}1&0\\0&1\\0&-1\\-2&-1\end{pmatrix}$$

(b) We will now use the Cayley–Hamilton method applied to the characteristic polynomial of $\mathcal{L}$. We have

$$\chi_{\mathcal{L}}(s)=s^8+24s^7+243s^6+1350s^5+\tfrac{35787}{8}s^4+\tfrac{17937}{2}s^3+\tfrac{167387}{16}s^2+\tfrac{50505}{8}s+\tfrac{363825}{256}.$$

Using formulas (6.18), we successively apply the Sylvester operator defined by (6.13) to $\mathbf{C}$ for $j = 1, \ldots, 7$. Then $\mathbf{X} = \frac{256}{363825}\mathbf{X}'$ is equal to the expression obtained above, where

$$\mathbf{X}' = \mathcal{L}^7(\mathbf{C}) + 24\mathcal{L}^6(\mathbf{C}) + 243\mathcal{L}^5(\mathbf{C}) + 1350\mathcal{L}^4(\mathbf{C}) + \tfrac{35787}{8}\mathcal{L}^3(\mathbf{C}) + \tfrac{17937}{2}\mathcal{L}^2(\mathbf{C})$$
$$+ \tfrac{167387}{16}\mathcal{L}^1(\mathbf{C}) + \tfrac{50505}{8}\mathcal{L}^0(\mathbf{C}).$$

(c) The next method uses complex integration. The expression $\mathbf{F}(s) = (\mathbf{A} - s\mathbf{I}_4)^{-1} \cdot \mathbf{C}(\mathbf{B} + s\mathbf{I}_2)^{-1}$ is

$$\mathbf{F}(s) = \frac{1}{\mathbf{d}(s)} \begin{bmatrix} -(43s^3 + 4s^4 + 178s^2 + 393s + 144) & -(153s^2 + 28s^3 + 423s + 216 + 2s^4) \\ -2(-s + 14s^3 + s^4 + 79s^2 - 12) & -(-96 + 183s^2 + 43s^3 - 32s + 4s^4) \\ -2(-86s^2 - 86s + 34s^3 + 3s^4 - 24) & -(-142s + 24 - 197s^2 + 53s^3 + 4s^4) \\ 2(201s^3 + 271s^2 + 146s + 3s^4 + 24) & (517s^2 + 322s + 120 + 387s^3 - 8s^4) \end{bmatrix},$$

where $\mathbf{d}(s) = (s + 4)(s + 3)(s + 2)(s + 1)(2s + 1)(2s - 3)$. A partial fraction expansion can be obtained in MATLAB by using the command `diff(int(F))`:

$$\mathbf{F}(s) = \frac{1}{s+4}\begin{bmatrix} -\frac{2}{3} & -\frac{2}{3} \\ \frac{8}{3} & \frac{8}{3} \\ -\frac{32}{3} & -\frac{32}{3} \\ \frac{128}{3} & \frac{128}{3} \end{bmatrix} + \frac{1}{s+3}\begin{bmatrix} 3 & 3 \\ -9 & -9 \\ 27 & 27 \\ -81 & -81 \end{bmatrix} + \frac{1}{s+2}\begin{bmatrix} -5 & -5 \\ 10 & 10 \\ -20 & -20 \\ 40 & 40 \end{bmatrix}$$
$$+ \frac{1}{s+1}\begin{bmatrix} \frac{11}{3} & \frac{8}{3} \\ -\frac{11}{3} & -\frac{8}{3} \\ \frac{11}{3} & \frac{8}{3} \\ -\frac{11}{3} & -\frac{8}{3} \end{bmatrix} + \frac{1}{2s+1}\begin{bmatrix} -\frac{1}{2} & \frac{3}{2} \\ \frac{1}{2} & -\frac{3}{2} \\ -\frac{1}{2} & \frac{3}{2} \\ \frac{1}{2} & -\frac{3}{2} \end{bmatrix} + \frac{1}{2s-3}\begin{bmatrix} -\frac{3}{2} & -\frac{3}{2} \\ -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{7}{2} & \frac{7}{2} \end{bmatrix}.$$

To evaluate the integral (6.9), we need the *residues* of $\mathbf{F}(s)$ at $s = -4$, $s = -3$, $s = -2$, and $s = -1$; these are

$$\mathbf{X}_1 = (s+1)\mathbf{F}(s)|_{s=-1} = \frac{1}{3}\begin{bmatrix} 11 & 8 \\ -11 & -8 \\ 11 & 8 \\ -11 & -8 \end{bmatrix}, \quad \mathbf{X}_2 = (s+2)\mathbf{F}(s)|_{s=-2} = \begin{bmatrix} -5 & -5 \\ 10 & 10 \\ -20 & -20 \\ 40 & 40 \end{bmatrix},$$

$$\mathbf{X}_3 = (s+3)\mathbf{F}(s)|_{s=-3} = \begin{bmatrix} 3 & 3 \\ -9 & -9 \\ 27 & 27 \\ -81 & -81 \end{bmatrix}, \quad \mathbf{X}_4 = (s+4)\mathbf{F}(s)|_{s=-4} = \frac{1}{3}\begin{bmatrix} -2 & -2 \\ 8 & 8 \\ -32 & -32 \\ 128 & 128 \end{bmatrix}.$$

It follows that $\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3 + \mathbf{X}_4$. We can also obtain this solution by using the dual expression of (6.9). This involves the complex integration of $\mathbf{G}(s) = (\mathbf{A} +$

$s\mathbf{I}_4)^{-1}\mathbf{C}(\mathbf{B} - s\mathbf{I}_2)^{-1}$. In this case, the Cauchy contour must contain the eigenvalues of **B** but not those of **A**. Therefore, the residues of $\mathbf{G}(s)$ at $s = \frac{1}{2}$ and $s = -\frac{3}{2}$ are needed:

$$\mathbf{G}(s) = \frac{1}{s-1}\begin{bmatrix} -\frac{11}{3} & -\frac{8}{3} \\ \frac{11}{3} & \frac{8}{3} \\ -\frac{11}{3} & -\frac{8}{3} \\ \frac{11}{3} & \frac{8}{3} \end{bmatrix} + \frac{1}{s-2}\begin{bmatrix} 5 & 5 \\ -10 & -10 \\ 20 & 20 \\ -40 & -40 \end{bmatrix} + \frac{1}{s-3}\begin{bmatrix} -3 & -3 \\ 9 & 9 \\ -27 & -27 \\ 81 & 81 \end{bmatrix}$$

$$+\frac{1}{s-4}\begin{bmatrix} \frac{2}{3} & \frac{2}{3} \\ -\frac{8}{3} & -\frac{8}{3} \\ \frac{32}{3} & \frac{32}{3} \\ -\frac{128}{3} & -\frac{128}{3} \end{bmatrix} + \frac{1}{2s+3}\begin{bmatrix} \frac{3}{2} & \frac{3}{2} \\ \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \\ -\frac{7}{2} & -\frac{7}{2} \end{bmatrix} + \frac{1}{2s-1}\begin{bmatrix} \frac{1}{2} & -\frac{3}{2} \\ -\frac{1}{2} & \frac{3}{2} \\ \frac{1}{2} & -\frac{3}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}.$$

In this case $\mathbf{X} = \mathbf{Y}_1 + \mathbf{Y}_2$, where

$$\mathbf{Y}_1 = \left(s - \frac{1}{2}\right)\mathbf{G}(s)\Big|_{s=\frac{1}{2}} = \frac{1}{4}\begin{bmatrix} 1 & -3 \\ -1 & 3 \\ 1 & -3 \\ -1 & 3 \end{bmatrix},\ \mathbf{Y}_2 = \left(s + \frac{3}{2}\right)\mathbf{G}(s)\Big|_{s=-\frac{3}{2}} = \frac{1}{4}\begin{bmatrix} 3 & 3 \\ 1 & 1 \\ -1 & -1 \\ -7 & -7 \end{bmatrix}.$$

(d) The eigenvector method is applied to **B**; we have

$$\mathbf{B} = \mathbf{V}\Lambda\mathbf{V}^{-1},\ \mathbf{V} = \begin{bmatrix} -1 & 3 \\ 1 & 1 \end{bmatrix},\ \Lambda = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{3}{2} \end{bmatrix}.$$

Therefore, according to the right-hand-side formula (6.14), the solution is

$$\mathbf{X} = \underbrace{[(\Lambda(1,1)\mathbf{I}_4 + \mathbf{A})^{-1}\mathbf{C}\mathbf{V}(:,1) \mid (\Lambda(2,2)\mathbf{I}_4 + \mathbf{A})^{-1}\mathbf{C}\mathbf{V}(:,2)]}_{\begin{bmatrix} 1 & -3 \\ -1 & -1 \\ 1 & 1 \\ -1 & 7 \end{bmatrix}} \underbrace{\mathbf{V}^{-1}}_{\begin{bmatrix} -\frac{1}{4} & \frac{3}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix}} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 2 & 1 \end{bmatrix}.$$

(e) Finally, the characteristic polynomial method, and in particular formula (6.20), yields

$$\mathbf{X} = \left(\mathbf{A}^2 - \mathbf{A} - \frac{3}{4}\mathbf{I}_4\right)^{-1}(-\mathbf{A}\mathbf{C} + \mathbf{C}\mathbf{B} + \mathbf{C}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & -1 \\ -2 & -1 \end{bmatrix}.$$

Using the Bezout equation we can express the inverse in the above expression as a polynomial in **A**. This is done as follows. We solve the polynomial equation, which is always possible, since $\mathbf{p}(s)$, $\mathbf{q}(-s)$ are coprime. It turns out that

$$\mathbf{p}(s)\mathbf{a}(s) + \mathbf{q}(-s)\mathbf{b}(s) = 1 \Rightarrow \mathbf{p}(\mathbf{A})\mathbf{a}(\mathbf{A}) + \mathbf{q}(-\mathbf{A})\mathbf{b}(\mathbf{A}) = \mathbf{I}_4 \Rightarrow \mathbf{q}(-\mathbf{A})\mathbf{b}(\mathbf{A}) = \mathbf{I}.$$

This implies that $\mathbf{b}(\mathbf{A})$ is the inverse of $\mathbf{q}(-\mathbf{A})$; the four polynomials are

$$\begin{aligned}
\mathbf{p}(s) &= s^4 + 10s^3 + 35s^2 + 50s + 24,\\
\mathbf{q}(-s) &= s^2 - s - 3/4,\\
\mathbf{a}(s) &= \tfrac{4}{3465}(-64s + 100),\\
\mathbf{b}(s) &= \tfrac{4}{3465}(64s^3 + 604s^2 + 1892s + 2045).
\end{aligned}$$

Thus

$$\mathbf{X} = \frac{4}{3465}(64\mathbf{A}^3 + 604\mathbf{A}^2 + 1892\mathbf{A} + 2045\mathbf{I}_4)(-\mathbf{AC} + \mathbf{CB} + \mathbf{C}).$$

6.1.5 The invariant subspace method

Recall the definition (6.3) of $\mathbf{M}_1$. The method proposed here stems from Roth's condition stated in Proposition 6.1. Suppose that a relationship of the form

$$\underbrace{\begin{bmatrix} \mathbf{A} & -\mathbf{C} \\ \mathbf{0} & -\mathbf{B} \end{bmatrix}}_{\mathbf{M}_1}\underbrace{\begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix}}_{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix}\mathbf{R}, \text{ where } \mathbf{V} \in \mathbb{R}^{(n+k)\times k},\ \mathbf{V}^*\mathbf{V} = \mathbf{I}_k,\ \mathbf{R} \in \mathbb{R}^{k\times k}, \quad (6.21)$$

has been obtained. If $\mathbf{V}_2$ is nonsingular, the solution of the Sylvester equation can be obtained as $\mathbf{X} = \mathbf{V}_1\mathbf{V}_2^{-1}$. Indeed, we obtain from (6.21)

$$\mathbf{A}\underbrace{\mathbf{V}_1\mathbf{V}_2^{-1}}_{\mathbf{X}} - \mathbf{C} = \underbrace{\mathbf{V}_1\mathbf{V}_2^{-1}}_{\mathbf{X}}\underbrace{\mathbf{V}_2\mathbf{R}\mathbf{V}_2^{-1}}_{-\mathbf{B}} = -\mathbf{XB}.$$

Moreover, it turns out that $\mathbf{V}_2$ is nonsingular if and only if $\mathbf{R}$ and $-\mathbf{B}$ are similar (see Problem [7] in Chapter 15). In this case, the columns of $\mathbf{V}$ form a basis for the eigenspace of $\mathbf{M}_1$, which corresponds to the eigenvalues of $-\mathbf{B}$. We summarize this discussion in the next lemma.

Lemma 6.7. *Let* (6.21) *hold where* $\mathbf{V} = [\mathbf{V}_1^*,\ \mathbf{V}_2^*]$ *is the eigenspace of* $\mathbf{M}_1$ *corresponding to the eigenvalues of* $-\mathbf{B}$. *The solution of the Sylvester equation* (6.1) *is given by* $\mathbf{X} = \mathbf{V}_1\mathbf{V}_2^{-1}$.

The desired invariant subspace can be obtained by computing a real partial Schur decomposition of $\mathbf{M}_1$, where $\mathbf{R}$ is quasi-upper triangular with the eigenvalues of $-\mathbf{B}$ appearing on the diagonal.

6.1.6 The sign function method

Consider the matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$ with eigenvalue decomposition $\mathbf{Z} = \mathbf{V}\Lambda\mathbf{V}^{-1}$, $\Lambda = \text{diag}(\Lambda_+, \Lambda_-)$, where Λ_+, Λ_- contain Jordan blocks corresponding to the r, $n-r$ eigenvalues with positive, negative, real parts, respectively. The *sign function* of this matrix is defined as $\mathbf{Z}_\sigma = \mathbf{V}\text{diag}(\mathbf{I}_r, -\mathbf{I}_{n-r})\mathbf{V}^{-1}$. The purpose of this section is to show that under certain conditions, it can be used to solve the Sylvester and the Lyapunov equations.

Given a complex number $z \in \mathbb{C}$ with $\mathcal{R}e(z) \neq 0$, we perform the following iteration, known as the *sign iteration*:

$$z_{n+1} = \frac{1}{2}\left(z_n + \frac{1}{z_n}\right), \qquad z_0 = z.$$

It readily follows that the fixed points of this iterations are ± 1. Therefore, if $\mathcal{R}e(z) > 0$, $\lim_{n\to\infty} z_n = 1$, and if $\mathcal{R}e(z) < 0$, $\lim_{n\to\infty} z_n = -1$. This iteration can also be defined for matrices $\mathbf{Z} \in \mathbb{C}^{r\times r}$: $\mathbf{Z}_{n+1} = (\mathbf{Z}_n + \mathbf{Z}_n^{-1})/2$, $\mathbf{Z}_0 = \mathbf{Z}$. Fixed points in this case are matrices that satisfy $\mathbf{Z}^2 = \mathbf{I}_r$; in other words, matrices that are diagonalizable and their eigenvalues are ± 1. It can be easily shown that if the matrix has eigenvalues ± 1 but is not diagonalizable, convergence to a diagonalizable matrix with eigenvalues ± 1 is achieved in finitely many steps (equal to half the size of the largest Jordan block). An important property of the sign iteration is that it is invariant under similarity, i.e., if the jth iterate of $\mathbf{Z}$ is $\mathbf{Z}_j$, the jth iterate of $\mathbf{V}\mathbf{Z}\mathbf{V}^{-1}$ is $\mathbf{V}\mathbf{Z}_j\mathbf{V}^{-1}$. The following holds.

Proposition 6.8. *Let* $\mathbf{J}$ *be a Jordan block, i.e.,* $\mathbf{J} = \lambda\mathbf{I}_r + \mathbf{N}$, *where* λ *has positive real part and* $\mathbf{N}$ *is the nilpotent matrix with ones above the diagonal and zeros elsewhere. The sign iteration of* $\mathbf{J}$ *converges to* $\mathbf{I}_r$.

Proof. Since the matrix is upper triangular, the elements on the diagonal will converge to 1. Furthermore, this limit, denoted by $\mathbf{J}'$, has Toeplitz structure. Thus each iteration applied to $\mathbf{J}'$ will bring zeros to the successive superdiagonals. Thus $\mathbf{J}'$ (which is upper triangular and Toeplitz with ones on the diagonal) will converge in r steps to the identity matrix. □

Corollary 6.9. *If* $\mathbf{Z} \in \mathbb{C}^{r\times r}$ *and* $\mathcal{R}e\lambda_i(\mathbf{Z}) < 0$ ($\mathcal{R}e\lambda_i(\mathbf{Z}) > 0$), *the sign iteration converges to* $\mathbf{Z}_n \to -\mathbf{I}_r$ ($\mathbf{Z}_n \to +\mathbf{I}_r$), *respectively.*

Proof. Let $\mathbf{Z} = \mathbf{V}\Lambda\mathbf{V}^{-1}$ be the EVD of $\mathbf{Z}$. Since the iteration is not affected by similarity, we need to consider the iterates of the Jordan blocks $\mathbf{J}_i$ of Λ. The proposition implies that each Jordan block converges to $\pm\mathbf{I}$, depending on whether the real part of the corresponding eigenvalue is positive or negative. □

We will now consider a matrix of the following type:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{A} & -\mathbf{C} \\ \mathbf{0} & -\mathbf{B} \end{pmatrix}, \quad \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathcal{R}e\lambda_i(\mathbf{A}) < 0,\ \mathbf{B} \in \mathbb{R}^{k\times k},\ \mathcal{R}e\lambda_i(\mathbf{B}) < 0,\ \mathbf{C} \in \mathbb{R}^{n\times k}.$$

Proposition 6.10. *The iteration* $\mathbf{Z}_{n+1} = (\mathbf{Z}_n + \mathbf{Z}_n^{-1})/2$, $\mathbf{Z}_0 = \mathbf{Z}$, *defined above, converges to*

$$\lim_{j\to\infty} \mathbf{Z}_j = \begin{pmatrix} -\mathbf{I}_n & 2\mathbf{X} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix},$$

where $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{B} = \mathbf{C}$.

Proof. Let $\mathbf{Z}\mathbf{V} = \mathbf{V}\Lambda$ with $\Lambda = \operatorname{diag}(\Lambda_1, \Lambda_2)$ be the eigenvalue decomposition of $\mathbf{Z}$. $\mathbf{V}$ is upper block triangular and

$$\begin{pmatrix} \mathbf{A} & -\mathbf{C} \\ \mathbf{0} & -\mathbf{B} \end{pmatrix}\begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{0} & \mathbf{V}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{0} & \mathbf{V}_{22} \end{pmatrix}\begin{pmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \Lambda_2 \end{pmatrix}.$$

This readily implies that the solution of the Sylvester equation is $\mathbf{V}_{12}\mathbf{V}_{22}^{-1}$. The block triangular structure of $\mathbf{Z}$ is preserved during the iterations, i.e., $\mathbf{Z}_j$ is also block upper triangular. Furthermore, by the proposition above, the limits of the $(1, 1)$ and $(2, 2)$ blocks are $-\mathbf{I}_n$, $\mathbf{I}_k$, respectively. Thus $\mathbf{Z}_j$ as $j \to \infty$ has the form claimed. It remains to show that $\mathbf{X}$ satisfies the Sylvester equation as stated. This follows since $\lim_{j\to\infty} \mathbf{Z}_j\mathbf{V} = \mathbf{V}\lim_{j\to\infty} \Lambda_j$, and therefore

$$\begin{pmatrix} -\mathbf{I}_n & 2\mathbf{X} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix}\begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{0} & \mathbf{V}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{0} & \mathbf{V}_{22} \end{pmatrix}\begin{pmatrix} -\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \end{pmatrix}.$$

This implies $\mathbf{X} = \mathbf{V}_{12}\mathbf{V}_{22}^{-1}$, which is indeed the solution to the Sylvester equation, as claimed. □

Remark 6.1.3. **(a)** The above result shows that the solution of the Sylvester equation where $\mathbf{A}$ and $\mathbf{B}$ have eigenvalues either in the left or the right half planes can be solved by means of the sign iteration.

(b) The above result also shows that given a matrix $\mathbf{Z} = \mathbf{V}\Lambda\mathbf{V}^{-1} \in \mathbb{C}^{n\times n}$, with k eigenvalues in the left half plane, the sign iteration yields the matrix $\mathbf{Z}_\infty = \mathbf{V}\mathrm{diag}(-\mathbf{I}_k, \mathbf{I}_{n-k})\mathbf{V}^{-1}$. Therefore, $\Pi_\pm = \frac{1}{2}(\mathbf{I}_n \pm \mathbf{Z}_\infty)$ yields the spectral projectors of $\mathbf{Z}$ onto its antistable/stable eigenspaces, respectively.

(c) For Lyapunov equations $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* = \mathbf{Q}$, the starting matrix is

$$\mathbf{Z} = \begin{pmatrix} \mathbf{A} & -\mathbf{Q} \\ \mathbf{0} & -\mathbf{A}^* \end{pmatrix}, \quad \mathbf{A} \in \mathbb{R}^{n\times n}, \quad \mathcal{R}e\lambda_i(\mathbf{A}) < 0 \;\Rightarrow\; \mathbf{Z}_j = \begin{pmatrix} \mathbf{A}_j & -\mathbf{Q}_j \\ \mathbf{0} & -\mathbf{A}_j^* \end{pmatrix}.$$

The iterations can also be written as follows:

$$\mathbf{A}_{j+1} = \frac{1}{2}\left(\mathbf{A}_j + \mathbf{A}_j^{-1}\right), \; \mathbf{A}_0 = \mathbf{A}; \quad \mathbf{Q}_{j+1} = \frac{1}{2}\left(\mathbf{Q}_j + \mathbf{A}_j^{-1}\mathbf{Q}_j\mathbf{A}_j^{-*}\right), \; \mathbf{Q}_0 = \mathbf{Q}.$$

The limits of these iterations are $\mathbf{A}_\infty = -\mathbf{I}_n$ and $\mathbf{Q}_\infty = 2\mathcal{P}$, where $\mathcal{P}$ is the solution of the Lyapunov equation.

(d) The convergence of the sign iteration is ultimately *quadratic*. To accelerate convergence in the beginning, one can introduce a scaling constant as follows: $\mathbf{Z}_{n+1} = \frac{1}{2\gamma_n}(\mathbf{Z}_n + \gamma_n^2\mathbf{Z}_n^{-1})$. It has been proposed in [64] that for the solution of the Lyapunov equation this constant be chosen as $\gamma_n = 1/\sqrt[n]{|\mathrm{det}\mathbf{A}_n|}$.

(e) Often, in the solution of the Lyapunov equation the constant term is provided in factored form $\mathbf{Q} = \mathbf{R}\mathbf{R}^*$. As a consequence, we can obtain the $(j + 1)$st iterate in factored form,

$$\mathbf{Q}_{j+1} = \mathbf{R}_{j+1}\mathbf{R}_{j+1}^*, \quad \text{where } \mathbf{R}_{j+1} = \frac{1}{\sqrt{2}}\left[\mathbf{R}_j, \;\; \mathbf{A}_j^{-1}\mathbf{R}_j\right] \;\Rightarrow\; \mathbf{Q}_\infty = \mathbf{R}_\infty\mathbf{R}_\infty^* = 2\mathcal{P}.$$

Thus the solution is obtained in factored form. However, $\mathbf{R}_\infty$ has infinitely many columns, although its rank cannot exceed n. To avoid this, we need to perform at each step a rank revealing RQ factorization $\mathbf{R}_j = \mathbf{T}_j\mathbf{U}_j$, where $\mathbf{T}_j = [\Delta_j^*, \;\mathbf{0}]^*$, where Δ_j is upper triangular and $\mathbf{U}_j\mathbf{U}_j^* = \mathbf{I}_j$. Thus at the jth step $\mathbf{R}_j$ can be replaced by $\mathbf{T}_j$, which has exactly as many columns as the rank of $\mathbf{R}_j$.

(f) The sign function was first used for solving the Lyapunov equation by Roberts [274]. For a recent overview of the sign function, see Kenney and Laub [198].

Benner and coworkers have contributed to the solution of Lyapunov equations by computing full-rank (triangular) factors [52], as well as by computing low-rank factors of rank k, by means of an $O(k^3)$ SVD (k is the maximum numerical rank of the solution) [55]. The use of the sign function in Newton's method for algebraic Riccati equations (AREs) and its application to stochastic truncation are described in [54]. In [52], the sign function is used for the first time to solve Lyapunov equations arising in generalized linear systems (i.e., systems containing both differential and algebraic equations). The computation of (triangular) Cholesky factors with the sign function was also introduced independently in [223]. In [56], [57], the sign function method is used for solving discrete Lyapunov equations; balancing related model reduction techniques for discrete-time systems are also developed. Benner, Quintana-Orti, and Quintana-Orti [58] give a survey of all model reduction algorithms together with detailed performance studies. Portable software for large-scale model reduction on parallel computers based on the sign function has been developed in the references cited above.

Example 6.11. Consider the number $z = 2$. We perform the iteration $z_{n+1} = \frac{1}{2}(z_n + \frac{1}{z_n})$, $z_0 = 2$. The iterates converge to 1. The error $e_n = z_n - 1$ is as follows: $e_1 = 2.5000\,10^{-1}$, $e_2 = 2.5000\,10^{-2}$, $e_3 = 3.0488\,10^{-4}$, $e_4 = 3.6461\,10^{-8}$, $e_5 = 1.0793\,10^{-15}$, $e_6 = 5.8246\,10^{-31}$, $e_7 = 1.6963\,10^{-61}$, $e_8 = 1.4388\,10^{-122}$. This shows that convergence is fast (the exponent of the error doubles at each iteration).

Example 6.12. Let

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -3 & -3 \end{bmatrix}, \; \mathbf{B} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}, \; \mathbf{C} = \begin{bmatrix} 0 & -1 \\ -1 & 2 \\ -2 & -7 \end{bmatrix}.$$

We form the matrix

$$\mathbf{Z} = \left[\begin{array}{c|c} \mathbf{A} & -\mathbf{C} \\ \hline \mathbf{0} & -\mathbf{B} \end{array}\right].$$

Since both **A** and **B** are composed of one Jordan block with eigenvalues -1, the iteration will converge in finitely many steps. Indeed, in two iterations we get

$$\mathbf{Z}_1 = \left[\begin{array}{ccc|cc} -3/2 & -1 & -1/2 & 3 & 3 \\ 1/2 & 0 & 1/2 & 1 & -1 \\ -1/2 & -1 & -3/2 & -1 & 3 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right], \; \mathbf{Z}_2 = \left[\begin{array}{ccc|cc} -1 & 0 & 0 & 2 & 2 \\ 0 & -1 & 0 & 2 & 0 \\ 0 & 0 & -1 & -2 & 2 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Therefore, the solution of the Sylvester equation $\mathbf{AX} + \mathbf{XB} + \mathbf{C} = \mathbf{0}$ is

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

6.1.7 Solution as an infinite sum

Equations of the type $\mathbf{AXB} + \mathbf{C} = \mathbf{X}$, for the unknown matrix $\mathbf{X}$, are known as *discrete-time Sylvester* or *Stein equations*. They are closely related to continuous-time Sylvester equations.

Proposition 6.13. *Given* $\mathbf{K} \in \mathbb{R}^{n\times n}$, $\mathbf{L} \in \mathbb{R}^{m\times m}$, $\mathbf{M} \in \mathbb{R}^{n\times m}$, *where the first two have no eigenvalues equal to* 1, *we define the matrices:*

$$\hat{\mathbf{K}} = (\mathbf{I}+\mathbf{K})(\mathbf{I}-\mathbf{K})^{-1},\ \hat{\mathbf{L}} = (\mathbf{I}+\mathbf{L})(\mathbf{I}-\mathbf{L})^{-1},\ \hat{\mathbf{M}} = 2(\mathbf{I}-\mathbf{K})^{-1}\mathbf{M}(\mathbf{I}-\mathbf{L})^{-1}.$$

It follows that $\mathbf{X} \in \mathbb{R}^{n\times m}$ *satisfies the Sylvester equation,*

$$\mathbf{KX} + \mathbf{XL} = \mathbf{M},$$

if and only if it satisfies the Stein equation,

$$\hat{\mathbf{K}}\mathbf{X}\hat{\mathbf{L}} - \mathbf{X} = \hat{\mathbf{M}}.$$

Furthermore, $\mathcal{R}e(\lambda_i(\mathbf{K})) < 0$ *is equivalent to* $|\lambda_i(\hat{\mathbf{K}})| < 1$, *and* $\mathcal{R}e(\lambda_i(\mathbf{L})) < 0$ *is equivalent to* $|\lambda_i(\hat{\mathbf{L}})| < 1$. *If these conditions are satisfied, there holds*

$$\mathbf{X} = \int_0^\infty e^{\mathbf{K}t}\mathbf{M}e^{\mathbf{L}t}\,dt = \sum_{k\geq 0}\hat{\mathbf{K}}^k\hat{\mathbf{M}}\hat{\mathbf{L}}^{\mathrm{k}}.$$

Thus, according to the above formula, provided that $\mathbf{K}$ and $\mathbf{L}$ have eigenvalues in the left half plane, the solution of the Sylvester equation can be expressed as an *infinite sum.*

6.2 The Lyapunov equation and inertia

In this section, a remarkable property of the Lyapunov equation (6.2), $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* = \mathbf{Q}$, for which $\mathbf{Q}$ is *semidefinite*, $\mathbf{Q} \geq \mathbf{0}$ or $\mathbf{Q} \leq \mathbf{0}$, is analyzed. This property allows for the *count* of the location of the eigenvalues of $\mathbf{A}$ given the count of the location of the eigenvalues of the solution, and vice versa. This is the *inertia* result associated with the Lyapunov equation. We begin with a definition.

Definition 6.14. *Given a square matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$, *let the number of eigenvalues in the left half plane, on the imaginary axis, in the right half plane, be denoted by* $\nu(\mathbf{A})$, $\delta(\mathbf{A})$, $\pi(\mathbf{A})$, *respectively. The triple*

$$\text{in}\,(\mathbf{A}) = (\nu(\mathbf{A}),\ \delta(\mathbf{A}),\ \pi(\mathbf{A}))$$

is called the inertia *of* $\mathbf{A}$.

Matrices with inertia $(n, 0, 0)$ are referred to as *stable*, while matrices with inertia $(0, 0, n)$ are *antistable*. The following result is a consequence of the Courant–Fischer max-min characterization of eigenvalues.

Proposition 6.15. Sylvester's law ef inertia. *Let* $\mathbf{A} = \mathbf{A}^*$ *and* $\mathbf{X}$ *be real with* $\det \mathbf{X} \neq 0$. *Then* $\text{in}(\mathbf{A}) = \text{in}(\mathbf{X}^*\mathbf{A}\mathbf{X})$.

The first fundamental relationship between inertia and the Lyapunov equation is the following.

Lemma 6.16. *Given* $\mathbf{Q} > \mathbf{0}$, *the Lyapunov equation has a unique positive definite solution* $\mathcal{P} > \mathbf{0}$ *if and only if* $\mathbf{A}$ *is an antistable matrix, i.e.,* $\text{in}(\mathbf{A}) = (0, 0, n)$.

Proof. **(a)** We first show that if $\mathcal{P} > \mathbf{0}$, then $\mathbf{A}$ has inertia equal to $(0, 0, n)$. Let $\mathbf{A}^*\mathbf{y}_i = \lambda_i\mathbf{y}_i$, $1 \leq i \leq n$. If we multiply (6.2) by $\mathbf{y}_i^*$ on the left and by $\mathbf{y}_i$ on the right, we obtain

$$(\lambda_i^* + \lambda_i)\mathbf{y}_i^*\mathcal{P}\mathbf{y}_i = \mathbf{y}_i^*\mathbf{Q}\mathbf{y}_i > 0, \qquad 1 \leq i \leq n.$$

By assumption, $\mathbf{y}_i^*\mathcal{P}\mathbf{y}_i > 0$ and $\mathbf{y}_i^*\mathbf{Q}\mathbf{y}_i > 0$ for $1 \leq i \leq n$. We conclude that $(\lambda_i^* + \lambda_i) > 0$, which proves the assertion. **(b)** Conversely, if $\mathbf{A}$ is antistable, we define

$$\mathcal{P} = \int_{-\infty}^{0} e^{\mathbf{A}\tau}\mathbf{Q}e^{\mathbf{A}^*\tau}d\tau,$$

which exists because of the antistability of $\mathbf{A}$. Furthermore, it is positive definite $\mathcal{P} > 0$. Finally, it is a solution of (6.2), since $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* = \int_{-\infty}^{0}[\mathbf{A}e^{\mathbf{A}\tau}\mathbf{Q}e^{\mathbf{A}^*\tau} + e^{\mathbf{A}\tau}\mathbf{Q}e^{\mathbf{A}^*\tau}\mathbf{A}^*]d\tau = \int_{-\infty}^{0} d[e^{\mathbf{A}\tau}\mathbf{Q}e^{\mathbf{A}^*\tau}] = \mathbf{Q}$. □

The following generalization holds [257].

Theorem 6.17. *Let* $\mathbf{A}, \mathcal{P}, \mathbf{Q}$ *satisfy* (6.2) *with* $\mathbf{Q} > 0$. *It follows that* $\text{in}(\mathbf{A}) = \text{in}(\mathcal{P})$.

Proof. Assume that $\delta(\mathbf{A}) \neq 0$; this implies the existence of $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{x}^*\mathbf{A} = i\omega\mathbf{x}^*$. Multiplying both sides of (6.2) by $\mathbf{x}^*$, $\mathbf{x}$, respectively, we obtain $\mathbf{x}^*\mathbf{Q}\mathbf{x} = 0$, which is a contradiction to the positive definiteness of $\mathbf{Q}$. It follows that $\delta(\mathbf{A}) = 0$, i.e., $\text{in}(\mathbf{A}) = (k, 0, r)$ for $r + k = n$ and $k \geq 0$. We may thus assume without loss of generality that $\mathbf{A}$ has the form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}, \quad \text{where} \quad \text{in}(\mathbf{A}_1) = (0,\ 0,\ r) \quad \text{and} \quad \text{in}(\mathbf{A}_2) = (k,\ 0,\ 0).$$

If we partition $\mathcal{P}$ conformally to $\mathbf{A}$, (6.2) implies

$$\mathbf{A}_1\mathcal{P}_{11} + \mathcal{P}_{11}\mathbf{A}_1^* = \mathbf{Q}_{11} > \mathbf{0} \quad \text{and} \quad \mathbf{A}_2\mathcal{P}_{22} + \mathcal{P}_{22}\mathbf{A}_2^* = \mathbf{Q}_{22} > \mathbf{0}.$$

Using the previous lemma, we conclude that $\mathcal{P}_{11} > \mathbf{0}$ while $\mathcal{P}_{22} < \mathbf{0}$.

There remains to show that the matrix $\mathcal{P}$ has the same number of positive eigenvalues as $\mathcal{P}_{11}$ and the same number of negative eigenvalues as $\mathcal{P}_{22}$, i.e., that $\text{in}(\mathcal{P}) = (k, 0, r)$. Toward this goal, notice that we can write

$$\mathcal{P}_{11} = \mathbf{V}_{11}\mathbf{V}_{11}^* \in \mathbb{R}^{r\times r}, \quad \mathcal{P}_{22} = -\mathbf{V}_{22}\mathbf{V}_{22}^* \in \mathbb{R}^{k\times k}, \quad \det \mathbf{V}_{ii} \neq 0, \qquad i = 1, 2.$$

Hence

$$\mathcal{P} = \text{diag}(\mathbf{V}_{11},\ \mathbf{V}_{22})\mathbf{Y}\text{diag}(\mathbf{V}_{11}^*,\ \mathbf{V}_{22}^*),\quad \text{where } \mathbf{Y} = \begin{pmatrix} \mathbf{I}_r & \mathbf{Y}_{12} \\ \mathbf{Y}_{12}^* & -\mathbf{I}_k \end{pmatrix}.$$

Finally, assuming that $r \leq k$, let $\mathbf{Y}_{12} = \mathbf{U}\Sigma\mathbf{W}^*$, where $\mathbf{U}\mathbf{U}^* = \mathbf{I}$, $\mathbf{W}\mathbf{W}^* = I$, and $\Sigma = (\bar{\Sigma}\ \ \mathbf{0})$, where $\mathbf{0}$ is the zero matrix with r rows and k columns, and $\bar{\Sigma} = \text{diag}\{\sigma_1, \ldots, \sigma_r\}$, $\sigma_i \geq \sigma_{i+1} \geq 0$. Then

$$\mathbf{Y} = \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{pmatrix} \Phi \begin{pmatrix} \mathbf{U}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^* \end{pmatrix},\quad \text{where } \Phi = \begin{pmatrix} \mathbf{I}_r & \bar{\Sigma} & \mathbf{0} \\ \bar{\Sigma}^* & -\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}_{k-r} \end{pmatrix}.$$

Clearly the matrix Φ has eigenvalues $\pm\sqrt{1+\sigma_i^2}$, $i = 1, \ldots, r$, and -1 with multiplicity $k - r$, i.e., Φ has r positive and k negative eigenvalues. Hence by the Sylvester law of inertia, both $\mathbf{Y}$ and $\mathcal{P}$ have the same number of positive and negative eigenvalues as Φ. This completes the proof. □

If we relax the positive definiteness of $\mathbf{Q}$ we obtain the following result.

Lemma 6.18. *Let* $\mathbf{A}, \mathcal{P}, \mathbf{Q}$ *satisfy* (6.2) *with* $\mathbf{Q} \geq 0$. *Then* $\delta(\mathbf{A}) = 0$ *implies* $\pi(\mathbf{A}) \geq \pi(\mathcal{P})$ *and* $\nu(\mathbf{A}) \geq \nu(\mathcal{P})$. *Furthermore,* $\delta(\mathcal{P}) = 0$ *implies* $\pi(\mathcal{P}) \geq \pi(\mathbf{A})$ *and* $\nu(\mathcal{P}) \geq \nu(\mathbf{A})$.

Proof. The proof follows by using continuity arguments. First, assume that the quantities $\mathbf{A}, \mathcal{P}, \mathbf{Q}$ satisfy (6.2) with $\mathbf{Q} \geq \mathbf{0}$ and $\delta(\mathbf{A}) = 0$. Let $\mathcal{P}'$ be the solution of $\mathbf{A}\mathcal{P}' + \mathcal{P}'\mathbf{A}^* = \mathbf{I}$; thus $\text{in}(\mathcal{P}') = \text{in}(\mathbf{A})$. Define

$$\mathcal{P}_\epsilon = \mathcal{P} + \epsilon\mathcal{P}',\qquad \epsilon > 0,$$

the solution of $\mathbf{A}\mathcal{P}_\epsilon + \mathcal{P}_\epsilon\mathbf{A}^* = \mathbf{Q} + \epsilon\mathbf{I}$. This implies that

$$\text{in}\,(\mathcal{P}_\epsilon) = \text{in}\,(\mathbf{A}),\qquad \epsilon > 0.$$

Since the above holds for all $\epsilon > 0$ and due to the continuity of eigenvalues as a function of ϵ, the result follows by letting $\epsilon \to 0$. Similarly, if $\delta(\mathcal{P}) = 0$, define

$$\mathbf{A}_\epsilon = \mathbf{A} + \epsilon\mathcal{P}^{-1},\qquad \epsilon > 0.$$

Substituting this into (6.2) we obtain

$$\mathbf{A}_\epsilon\mathcal{P} + \mathcal{P}\mathbf{A}_\epsilon^* = \mathbf{Q} + 2\epsilon\mathbf{I}_n > \mathbf{0},\qquad \epsilon > 0.$$

This implies $\text{in}(\mathcal{P}) = \text{in}(\mathbf{A}_\epsilon)$ for $\epsilon > 0$. Since this expression is positive definite for all $\epsilon > 0$, and again due to the continuity of the eigenvalues as a function of ϵ, the desired result follows by letting $\epsilon \to 0$. □

Theorem 6.19. *Let* $\mathbf{A}, \mathcal{P}, \mathbf{Q}$ *satisfy* (6.2) *with* $\mathbf{Q} \geq 0$. *If the pair* $(\mathbf{A}, \mathbf{Q})$ *is reachable, then*

$$\text{in}\,(\mathbf{A}) = \text{in}\,(\mathcal{P}).$$

Proof. We will show that reachability implies (a) $\delta(\mathbf{A}) = 0$ and (b) $\delta(\mathcal{P}) = 0$. Hence the result follows by applying the lemma above. There remains to show (a) and (b).

Assume that $\delta(\mathbf{A}) \neq 0$; there exists ω and $\mathbf{v}$ such that $\mathbf{A}^*\mathbf{v} = i\omega\mathbf{v}$. Thus $\mathbf{v}^*(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^*)\mathbf{v} = \mathbf{v}^*\mathbf{Q}\mathbf{v}$. The left-hand side is equal to $(-i\omega + i\omega)\mathbf{v}^*\mathcal{P}\mathbf{v} = 0$. Hence the left-hand side is zero; we conclude that $\mathbf{v}^*\mathbf{Q} = \mathbf{0}$. This is a contradiction to the fourth condition of Lemma 4.15; that is, if $\delta(\mathbf{A}) \neq 0$, the pair $(\mathbf{A}, \mathbf{Q})$ is not reachable. This proves (a).

If we assume now that $\delta(\mathcal{P}) \neq 0$, there exists a vector $\mathbf{v} \neq \mathbf{0}$ such that $\mathcal{P}\mathbf{v} = \mathbf{0}$. Hence $\mathbf{v}^*(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^*)\mathbf{v} = \mathbf{v}^*\mathbf{Q}\mathbf{v}$. The left-hand side is zero and therefore $\mathbf{v}^*\mathbf{Q}\mathbf{v} = 0$, which implies $\mathbf{v}^*\mathbf{Q} = \mathbf{0}$. This in turn implies $\mathbf{v}^*(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^*) = \mathbf{v}^*\mathbf{A}\mathcal{P} = \mathbf{v}^*\mathbf{Q} = \mathbf{0}$, i.e., $\mathbf{v}^*\mathbf{A}\mathcal{P} = \mathbf{0}$. Repeating the above procedure for the equation $\mathbf{A}(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^*)\mathbf{A}^* = \mathbf{A}\mathbf{Q}\mathbf{A}^*$, we conclude that $\mathbf{v}^*\mathbf{A}\mathbf{Q} = \mathbf{0}$. Consequently $\delta(\mathcal{P}) \neq 0$ implies $\mathbf{v}^*\mathbf{A}^{k-1}\mathbf{Q} = \mathbf{0}$ for $k > 0$, which means that the pair $(\mathbf{A}, \mathbf{Q})$ is not reachable. This proves (b). □

Remark 6.2.1. An inertia result with respect to the unit circle can also be formulated. Consider the *discrete-time Lyapunov or Stein equation*,

$$\mathbf{F}\mathbf{X}\mathbf{F}^* + \mathbf{R} = \mathbf{X}. \tag{6.22}$$

If the inertia of $\mathbf{F}$ with respect to the unit circle $\partial\mathcal{D}$ is defined as the number of eigenvalues which are outside, on, or inside the unit circle, and denoted by $\text{in}_{\partial\mathcal{D}}$, and the inertia of $\mathbf{X} = \mathbf{X}^*$ is defined as before, we conclude from Proposition 6.13 that if the pair $(\mathbf{F}, \mathbf{R})$ is reachable,

$$\text{in}_{\partial\mathcal{D}}(\mathbf{F}) = \text{in}\,(\mathbf{X}).$$

6.3 Sylvester and Lyapunov equations with triangular coefficients

Consider the Sylvester equation (6.1). There exist orthogonal matrices $\mathbf{U}, \mathbf{V}$ such that $\mathbf{U}\mathbf{A}\mathbf{U}^*$ and $\mathbf{V}\mathbf{B}\mathbf{V}^*$ are in Schur form, that is, upper triangular. Thus if we multiply (6.1) on the left by $\mathbf{U}$ and on the right by $\mathbf{V}^*$, we obtain $(\mathbf{U}\mathbf{A}\mathbf{U}^*)(\mathbf{U}\mathbf{X}\mathbf{V}^*) + (\mathbf{U}\mathbf{X}\mathbf{V}^*)(\mathbf{V}\mathbf{B}\mathbf{V}^*) = (\mathbf{U}\mathbf{C}\mathbf{V}^*)$. For simplicity we will denote the transformed quantities by the same symbol,

$$\mathbf{A} \leftarrow \mathbf{U}\mathbf{A}\mathbf{U}^*,\ \ \mathbf{X} \leftarrow \mathbf{U}\mathbf{X}\mathbf{V}^*,\ \ \mathbf{B} \leftarrow \mathbf{V}\mathbf{B}\mathbf{V}^*,\ \ \mathbf{C} \leftarrow \mathbf{U}\mathbf{C}\mathbf{V}^*,$$

where $\mathbf{A}, \mathbf{B}$ are in Schur (upper triangular) form. In this case we will refer to the Sylvester and Lyapunov equations in *Schur basis*.

6.3.1 The Bartels–Stewart algorithm

With the coefficient matrices triangular, the solution can be obtained columnwise as follows. Let the columns of $\mathbf{X}, \mathbf{C}$ be denoted by $\mathbf{x}_k$, $\mathbf{c}_k$ and the entries of $\mathbf{A}, \mathbf{B}$ by a_{ij}, b_{ij}, respectively.

The first column of $\mathbf{X}$ satisfies $\mathbf{A}\mathbf{x}_1 + \mathbf{x}_1 b_{11} = \mathbf{c}_1$. Thus $\mathbf{x}_1 = (\mathbf{A} + b_{11}\mathbf{I})^{-1}\mathbf{c}_1$. The second column satisfies $\mathbf{A}\mathbf{x}_2 + \mathbf{x}_1 b_{12} + \mathbf{x}_2 b_{22} = \mathbf{c}_2$. Thus $\mathbf{x}_2 = (\mathbf{A} + b_{22}\mathbf{I})^{-1}(\mathbf{c}_2 - b_{12}\mathbf{x}_1)$, which can be computed once $\mathbf{x}_1$ is known. Similarly,

$$(\mathbf{A} + b_{\ell\ell}\mathbf{I})\mathbf{x}_\ell = \mathbf{c}_\ell - \sum_{j=1}^{\ell-1} b_{j\ell}\mathbf{x}_j \;\Rightarrow\; \mathbf{x}_\ell = (\mathbf{A} + b_{\ell\ell}\mathbf{I})^{-1}\left(\mathbf{c}_\ell - \sum_{j=1}^{\ell-1} b_{j\ell}\mathbf{x}_j\right),$$
$$\ell = 1, \ldots, k. \tag{6.23}$$

Thus the fact that $\mathbf{B}$ is in Schur form allows the calculation of the columns of the solution recursively. In addition, the fact that $\mathbf{A}$ is in Schur form implies that $\mathbf{A} + b_{\ell\ell}\mathbf{I}$ is also in Schur form and therefore the inverse is also in Schur form and can be computed explicitly.

Remark 6.3.1. Recall that given the upper triangular $\mathbf{M} = (\mu_{ij})$, $\mu_{ij} = 0$, for $i > j$, $i, j = 1, \ldots, n$, its inverse $\mathbf{N} = (\nu_{ij})$ is also upper triangular, that is, $\nu_{ij} = 0$, $i > j$, $i, j = 1, \ldots, n$, and its nonzero elements are defined recursively as $\nu_{nn} = \frac{1}{\mu_{nn}}$, and, the entries of all rows from $i = k + 1, \ldots, n$ having been computed, the entries of the kth row are

$$\nu_{kk} = \frac{1}{\mu_{kk}}, \quad \nu_{k\ell} = \frac{-1}{\mu_{kk}} \sum_{j=k+1}^{\ell} \mu_{kj}\nu_{j\ell}, \quad \ell = k+1, \ldots, n,$$
$$\text{and} \quad k = n, n-1, \ldots, 2, 1.$$

The above procedure is the *Bartels–Stewart algorithm* [43] in complex arithmetic, because if $\mathbf{A}$ or $\mathbf{B}$ has complex eigenvalues, the solution of the Sylvester equation $\mathbf{X}$ in the (complex) Schur basis will be complex. On the other hand, for the Lyapunov equation, even if $\mathbf{A}$ has complex eigenvalues, $\mathbf{X}$ will be real.

The Bartels–Stewart algorithm in real arithmetic

To avoid complex arithmetic, the *real Schur form* can be used. In this case, if $\mathbf{A}$ or $\mathbf{B}$ has complex eigenvalues, it can be transformed by (real) orthogonal transformation to quasi-upper-triangular form, where the blocks on the diagonal are 1×1 for real and 2×2 for complex eigenvalues. In the latter case, let $b_{\ell+1,\ell} \neq 0$, that is, there is a 2×2 block on the diagonal between rows/columns ℓ and $\ell + 1$. In this case, we must solve simultaneously for $\mathbf{x}_\ell$ and $\mathbf{x}_{\ell+1}$; (6.23) becomes

$$\mathbf{A}[\mathbf{x}_\ell, \; \mathbf{x}_{\ell+1}] + [\mathbf{x}_\ell, \; \mathbf{x}_{\ell+1}] \begin{bmatrix} b_{\ell,\ell} & b_{\ell,\ell+1} \\ b_{\ell+1,\ell} & b_{\ell+1,\ell+1} \end{bmatrix} = \Bigg[\underbrace{\mathbf{c}_\ell - \sum_{j=1}^{\ell-1} b_{j\ell}\,\mathbf{x}_j}_{=\tilde{\mathbf{c}}_\ell}, \; \underbrace{\mathbf{c}_{\ell+1} - \sum_{j=1}^{\ell-1} b_{j,\ell+1}\mathbf{x}_j}_{=\tilde{\mathbf{c}}_{\ell+1}}\Bigg], \tag{6.24}$$

where the entries of the right-hand side of this expression are denoted by $\tilde{\mathbf{c}}_\ell$, $\tilde{\mathbf{c}}_{\ell+1}$, respectively, for simplicity. The above system of equations can be solved using different methods.

Following [311], the above system of equations is equivalent to

$$\left[\mathbf{A}^2 + (b_{\ell,\ell} + b_{\ell+1,\ell+1})\mathbf{A} + (b_{\ell,\ell}b_{\ell+1,\ell+1} - b_{\ell,\ell+1}b_{\ell+1,\ell})\mathbf{I}\right][\mathbf{x}_\ell,\ \mathbf{x}_{\ell+1}] = [\bar{\mathbf{c}}_\ell,\ \bar{\mathbf{c}}_{\ell+1}],$$

where $\bar{\mathbf{c}}_\ell = \mathbf{A}\tilde{\mathbf{c}}_\ell + b_{\ell+1,\ell+1}\tilde{\mathbf{c}}_\ell - b_{\ell+1,\ell}\tilde{\mathbf{c}}_{\ell+1}$, and $\bar{\mathbf{c}}_{\ell+1} = \mathbf{A}\tilde{\mathbf{c}}_{\ell+1} + b_{\ell,\ell}\tilde{\mathbf{c}}_{\ell+1} - b_{\ell,\ell+1}\tilde{\mathbf{c}}_\ell$. This can be solved using the block version system solve routine in LAPACK [7]. We refer to [311] for a comparison between different methods for solving (6.24). Recursive algorithms for the solution of linear matrix equations with triangular coefficients are also treated in [189].

6.3.2 The Lyapunov equation in the Schur basis

For this purpose we assume that in (6.2) the solution $\mathcal{P}$ is symmetric; we also assume that $\mathbf{Q} = -\mathbf{B}\mathbf{B}^*$. Thus we consider the *Lyapunov equation*,

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}, \tag{4.45}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathcal{P} = \mathcal{P}^* \in \mathbb{R}^{n\times n}$. Throughout this section we assume that

$$(\mathbf{A}, \mathbf{B}) \text{ is reachable.} \tag{6.25}$$

Recall the PBH condition for reachability stated in part 4 of Theorem 4.15. As a consequence, both $\delta(\mathbf{A}) = 0$ and $\delta(\mathcal{P}) = 0$, i.e., $\mathbf{A}$ has no eigenvalues on the imaginary axis and both matrices are nonsingular.

To continue our discussion, it will be convenient to assume that $\mathbf{A}$ is in Schur form, i.e., $\mathbf{A}$ is upper triangular. As shown earlier, there is no loss in generality with this assumption as it amounts to a transformation of the Lyapunov equation into an equivalent system using the Schur form basis vectors. Once the system is in Schur form, partition $\mathbf{A}$, $\mathbf{B}$, and $\mathcal{P}$ compatibly is as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix},\ \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix},\ \mathcal{P} = \begin{pmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} \\ \mathcal{P}_{12}^* & \mathcal{P}_{22} \end{pmatrix}, \tag{6.26}$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are upper triangular.

Proposition 6.20. *Assume* $\mathbf{A}$, $\mathbf{B}$, $\mathcal{P}$ *satisfy the Lyapunov equation, with* $(\mathbf{A}, \mathbf{B})$ *reachable, where partitioning* (6.26) *holds. The following statements hold:* **(a)** *The pair* $\mathbf{A}_{22}$, $\mathbf{B}_2$ *is reachable.* **(b)** $\delta(\mathcal{P}_{22}) = 0$, *i.e.,* $\mathcal{P}_{22}$ *is nonsingular.* **(c)** *The pair* $(\mathbf{A}_{11}, \hat{\mathbf{B}}_1)$ *is reachable, where* $\hat{\mathbf{B}}_1 = \mathbf{B}_1 - \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathbf{B}_2$.

Proof. **(a)** Let $\mathbf{z}_2$ be any left eigenvector of $\mathbf{A}_{22}$. Then $\mathbf{z} = [\mathbf{0}, \mathbf{z}_2^*]^*$ is a left eigenvector of $\mathbf{A}$ and the PBH condition implies $\mathbf{0} \neq \mathbf{z}^*\mathbf{B} = \mathbf{z}_2^*\mathbf{B}_2$. This is true for any left eigenvector of $\mathbf{A}_{22}$, and therefore the PBH condition also implies the reachability of $(\mathbf{A}_{22}, \mathbf{B}_2)$. **(b)** Since $\mathbf{A}_{22}\mathcal{P}_{22} + \mathcal{P}_{22}\mathbf{A}_{22}^* + \mathbf{B}_2\mathbf{B}_2^* = \mathbf{0}$, part (b) follows from the fact that $\delta(\mathbf{A}) = \delta(\mathcal{P}) = 0$, stated earlier. **(c)** As a consequence of (b), the Lyapunov equation can be transformed into

$$\hat{\mathbf{A}}\hat{\mathcal{P}} + \hat{\mathcal{P}}\hat{\mathbf{A}}^* + \hat{\mathbf{B}}\hat{\mathbf{B}}^* = \mathbf{0}, \tag{6.27}$$

where $\hat{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1}$, $\hat{\mathbf{B}} = \mathbf{T}\mathbf{B}$, $\hat{\mathcal{P}} = \mathbf{T}\mathcal{P}\mathbf{T}^*$, and

$$\mathbf{T} = \begin{pmatrix} \mathbf{I} & -\mathcal{P}_{12}\mathcal{P}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \hat{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_{11} & \hat{\mathbf{A}}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix}, \hat{\mathbf{B}} = \begin{pmatrix} \hat{\mathbf{B}}_1 \\ \mathbf{B}_2 \end{pmatrix}, \hat{\mathcal{P}} = \begin{pmatrix} \hat{\mathcal{P}}_{11} & \mathbf{0} \\ \mathbf{0} & \mathcal{P}_{22} \end{pmatrix} \tag{6.28}$$

with

$$\begin{aligned} \hat{\mathbf{A}}_{12} &= \mathbf{A}_{12} - \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathbf{A}_{22} + \mathbf{A}_{11}\mathcal{P}_{12}\mathcal{P}_{22}^{-1}, \\ \hat{\mathbf{B}}_1 &= \mathbf{B}_1 - \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathbf{B}_2, \\ \hat{\mathcal{P}}_{11} &= \mathcal{P}_{11} - \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathcal{P}_{12}^*. \end{aligned}$$

From (6.27) and (6.28) follow three equations:

$$\mathbf{A}_{11}\hat{\mathcal{P}}_{11} + \hat{\mathcal{P}}_{11}\mathbf{A}_{11}^* + \hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^* = \mathbf{0}, \ \mathbf{A}_{22}\mathcal{P}_{22} + \mathcal{P}_{22}\mathbf{A}_{22}^* + \mathbf{B}_2\mathbf{B}_2^* = \mathbf{0}, \ \hat{\mathbf{A}}_{12} = -\hat{\mathbf{B}}_1\mathbf{B}_2^*\mathcal{P}_{22}^{-1}. \tag{6.29}$$

Suppose that there is a left eigenvector $\mathbf{z}_1$ of $\mathbf{A}_{11}$ such that $\mathbf{z}_1^*\hat{\mathbf{B}}_1 = \mathbf{0}$. Then $\mathbf{z}_1^*\hat{\mathbf{A}}_{12} = \mathbf{0}$ and it follows that $\mathbf{z} = [\mathbf{z}_1^*, \mathbf{0}]^*$ is a left eigenvector of $\hat{\mathbf{A}}$ such that $\mathbf{z}^*\hat{\mathbf{B}} = \mathbf{z}_1^*\hat{\mathbf{B}}_1 = \mathbf{0}$ in contradiction to the PBH condition. □

Now, we are ready to prove the main theorem, Theorem 6.23, of this section. It is based on Lemma 6.22.

Definition 6.21. *A diagonal matrix is called a* signature *if its diagonal entries consist only of* 1 *or* -1.

Lemma 6.22. *Let* $\mathbf{A}$, $\mathbf{B}$, *and* $\mathcal{P}$ *satisfy the Lyapunov equation with* $(\mathbf{A}, \mathbf{B})$ *reachable. If* $\mathbf{A}$ *is in Schur form, then* $\mathcal{P}$ *can be expressed in factored form,* $\mathcal{P} = \mathbf{U}\mathbf{S}\mathbf{U}^*$, *where* $\mathbf{U}$ *is upper triangular and* $\mathbf{S}$ *is a signature matrix.*

Proof. The proof is given by induction on n, the order of $\mathbf{A}$. The required property clearly holds for $n = 1$. Assume that it holds for Lyapunov equations of order $k < n$, where (6.25) is satisfied. We show that the same property must also hold for Lyapunov equations of order n, satisfying (6.25).

To prove this, we can assume without loss of generality that the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathcal{P}$ (where $\mathbf{A}$ has dimension n) are partitioned as in (6.26), where the $(1, 1)$ block has dimension $k < n$ and the $(2, 2)$ block has dimension $n-k < n$. Due to reachability, we may also assume that these matrices are in the form (6.28) and satisfy the transformed Lyapunov equation (6.27). By Proposition 6.20, both of the pairs $(\mathbf{A}_{11}, \hat{\mathbf{B}}_1)$ and $(\mathbf{A}_{22}, \mathbf{B}_2)$ are reachable and the induction hypothesis can be applied to each of the two reduced-order Lyapunov equations, giving $\hat{\mathcal{P}}_{11} = \mathbf{U}_{11}\mathbf{S}_1\mathbf{U}_{11}^*$ and $\mathcal{P}_{22} = \mathbf{U}_{22}\mathbf{S}_2\mathbf{U}_{22}^*$. Transforming back from (6.27) gives $\mathcal{P} = \mathbf{U}\mathbf{S}\mathbf{U}^*$ with

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ 0 & \mathbf{U}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathcal{P}_{12}\mathcal{P}_{22}^{-1} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & 0 \\ 0 & \mathbf{U}_{22} \end{bmatrix}$$

and $\mathbf{S} = \begin{pmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{pmatrix}$. The induction is thus complete. □

An iterative proof of the inertia result

A consequence of the above lemma is a self-contained proof by induction of the inertia result associated with the Lyapunov equation. It first appeared in [19].

Theorem 6.23. *Assume that* **A**, **B**, *and* $\mathcal{P}$ *satisfy the Lyapunov equation* (4.45) *with* (**A**, **B**) *reachable. Then*

$$\nu(\mathbf{A}) = \pi(\mathcal{P}) \text{ and } \pi(\mathbf{A}) = \nu(\mathcal{P}).$$

Proof. Again, the proof is given by induction on n, the order of **A**. First, assume that **A** is in *Schur form*. The properties stated in the above theorem clearly hold for $n = 1$. Assume they hold for Lyapunov equations of order $k < n$, satisfying (6.25). We show as a consequence that these properties must also hold for Lyapunov equations of order n, satisfying (6.25).

If we partition the matrices as in the proof of Lemma 6.22, it follows that the Lyapunov equations (6.29) are satisfied. Each one of these has size less than n and hence the induction hypothesis applies:

$$\nu(\mathbf{A}_{11}) = \pi(\hat{\mathcal{P}}_{11}),\ \nu(\hat{\mathcal{P}}_{11}) = \pi(\mathbf{A}_{11}) \text{ and } \nu(\mathbf{A}_{22}) = \pi(\mathcal{P}_{22}),\ \nu(\mathcal{P}_{22}) = \pi(\mathbf{A}_{22}).$$

Since **A** is in Schur form, there holds $\nu(\mathbf{A}) = \nu(\mathbf{A}_{11}) + \nu(\mathbf{A}_{22})$, and $\pi(\mathbf{A}) = \pi(\mathbf{A}_{11}) + \pi(\mathbf{A}_{22})$; due to the structure of $\mathcal{P}$, we have $\nu(\mathcal{P}) = \nu(\hat{\mathcal{P}}_{11}) + \nu(\mathcal{P}_{22})$, and $\pi(\mathcal{P}) = \pi(\mathcal{P}_{11}) + \pi(\mathcal{P}_{22})$, completing the induction.

If **A** is *not* in Schur form, the upper triangular **U** in the considerations above is replaced by **QU**, where $\tilde{\mathbf{A}} = \mathbf{Q}^*\mathbf{A}\mathbf{Q}$ is the original matrix (not in Schur form). The solution of the corresponding Lyapunov equation is $\tilde{\mathcal{P}} = (\mathbf{QU})\mathbf{S}(\mathbf{QU})^*$. □

Remark 6.3.2. The considerations laid out in the proof of the theorem above lead to a UL factorization of the solution $\mathcal{P}$ to the Lyapunov equation. If **A** is in Schur form, the factorization $\mathcal{P} = \mathbf{USU}^*$ holds, where **U** is upper triangular and **S** is a signature matrix. The question is, When does the solution $\tilde{\mathcal{P}}$ in the original coordinate system possess such a factorization?

If the principal minors $\det \tilde{\mathcal{P}}(k : n, k : n)$, $k = 1, \ldots, n$, are different from zero, the UL factorization of $\tilde{\mathbf{P}}$ exists; let $\tilde{\mathcal{P}} = \bar{\mathbf{U}}\bar{\mathbf{L}}$, where the diagonal entries of $\bar{\mathbf{U}}$ can be chosen to be positive and those of $\bar{\mathbf{L}}$ can be chosen to have the same magnitude as the corresponding entries of $\bar{\mathbf{U}}$. Since $\tilde{\mathcal{P}}$ is symmetric, there exists a signature matrix $\tilde{\mathbf{S}}$ such that $(\tilde{\mathbf{S}})^{-1}\bar{\mathbf{L}} = \bar{\mathbf{U}}^*$, and the required factorization follows. It should be noted that the nonsingularity of the minors defined above is basis dependent and cannot always be satisfied. This is the case whenever **A** has eigenvalues with both positive and negative real parts. Actually, it is easy to show in this case that there always exists a basis such that $\tilde{\mathcal{P}}$ does *not* have an LU factorization. For example if $n = 2$, let the solution $\mathcal{P}_1$ be diagonal; by basis change the transformed solution $\mathcal{P}_2$,

$$\mathcal{P}_1 = \begin{pmatrix} \alpha & 0 \\ 0 & -\beta \end{pmatrix},\ \alpha,\ \beta > 0 \ \Rightarrow\ \mathcal{P}_2 = \begin{pmatrix} 0 & \sqrt{\alpha\beta} \\ \sqrt{\alpha\beta} & \alpha-\beta \end{pmatrix},$$

does not have an LU factorization. Of course, if $\mathcal{P}$ is positive or negative definite, the result is the Cholesky factorization, which always exists.

The theorem we have just established determines the inertia of $\mathbf{A}$ if a symmetric solution to a Lyapunov equation is available (as it is in our special case). No a priori assumptions on the spectrum of A are required.

Remark 6.3.3. **(a)** Consider the Lyapunov equation (6.2). If the condition $\mathbf{A}+\mathbf{A}^* < \mathbf{0}$ is satisfied, the following relationship holds:

$$\operatorname{trace} \mathcal{P} = -\operatorname{trace}\left[\mathbf{B}^*(\mathbf{A}+\mathbf{A}^*)^{-1}\mathbf{B}\right].$$

This follows from the fact that if $\mathcal{P}\mathbf{x}_i = \lambda_i\mathbf{x}_i$, then $\lambda_i\mathbf{x}_i^*(\mathbf{A}+\mathbf{A}^*)\mathbf{x}_i = \mathbf{x}_i^*\mathbf{B}^*\mathbf{B}\mathbf{x}_i$. By assumption we can write $-(\mathbf{A}+\mathbf{A}^*) = \mathbf{R}^*\mathbf{R}$ for some $\mathbf{R} \in \mathbb{R}^{n\times n}$. Thus with $\mathbf{y}_i = \mathbf{R}\mathbf{x}_i$ we have $\lambda_i\mathbf{y}_i^*\mathbf{y}_i = \mathbf{y}_i^*\mathbf{R}^{-*}\mathbf{B}^*\mathbf{B}\mathbf{R}^{-1}\mathbf{y}_i$, and thus

$$\begin{aligned}\lambda_i \operatorname{trace}\left[\mathbf{y}_i^*\mathbf{y}_i\right] = \lambda_i \operatorname{trace}\left[\mathbf{y}_i\mathbf{y}_i^*\right] &= \operatorname{trace}\left[\mathbf{y}_i^*\mathbf{R}^{-*}\mathbf{B}^*\mathbf{B}\mathbf{R}^{-1}\mathbf{y}_i\right] \\ &= \operatorname{trace}\left[\mathbf{B}\mathbf{R}^{-1}\mathbf{y}_i\mathbf{y}_i^*\mathbf{R}^{-*}\mathbf{B}^*\right].\end{aligned}$$

The desired result follows by summing the above equations for $i = 1, \ldots, n$ and using the fact that $\sum_i \mathbf{y}_i\mathbf{y}_i^* = \mathbf{I}_n$ and $\operatorname{trace} \mathcal{P} = \sum_i \lambda_i$.

(b) If $\mathbf{A}$ is stable, the condition $\bar{\mathbf{A}}+\bar{\mathbf{A}}^* < \mathbf{0}$ can always be satisfied for some $\bar{\mathbf{A}}$ which is similar to $\mathbf{A}$. By assumption, there exists $\mathbf{Q} > \mathbf{0}$ such that $\mathbf{A}\mathbf{Q}+\mathbf{Q}\mathbf{A}^* < \mathbf{0}$; then $\bar{\mathbf{A}}+\bar{\mathbf{A}}^* < \mathbf{0}$, where $\bar{\mathbf{A}} = \mathbf{Q}^{-1/2}\mathbf{A}\mathbf{Q}^{1/2}$.

6.3.3 The square root method for the Lyapunov equation

If $\mathbf{A}$ is stable (that is, all eigenvalues are in the left half plane), the solution of the Lyapunov equation is positive definite $\mathcal{P} > 0$. In this case, in the Schur basis, the solution according to Lemma 6.22 can be written as $\mathcal{P} = \mathbf{U}\mathbf{U}^*$, where $\mathbf{U}$ is upper triangular. It was observed by Hammarling [164] that this square root factor $\mathbf{U}$ can be computed *without explicitly computing* $\mathcal{P}$ first. This is a consequence of the discussion of the preceding section.

Let us assume that $\mathbf{A}$ is in Schur form, that is, (quasi) upper triangular, and partition $\mathbf{A}$, $\mathbf{B}$, $\mathcal{P}$ as in (6.26). The Lyapunov equation can be further transformed into (6.27) using the transformation $\mathbf{T}$ given in (6.28). Thus (assuming stability of $\mathbf{A}$) $\mathcal{P}$ can be expressed in factored form:

$$\mathcal{P} = \underbrace{\begin{pmatrix} \hat{\mathcal{P}}_{11}^{1/2} & \mathcal{P}_{12}\mathcal{P}_{22}^{-1/2} \\ \mathbf{0} & \mathcal{P}_{22}^{1/2} \end{pmatrix}}_{\mathbf{U}} \underbrace{\begin{pmatrix} \hat{\mathcal{P}}_{11}^{1/2} & \mathbf{0} \\ \mathcal{P}_{22}^{-1/2}\mathcal{P}_{12}^* & \mathcal{P}_{22}^{1/2} \end{pmatrix}}_{\mathbf{U}^*} = \mathbf{U}\mathbf{U}^*. \tag{6.30}$$

The problem is now to successively compute the smaller pieces of $\mathcal{P}$, namely, $\mathcal{P}_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, $\mathcal{P}_{12} \in \mathbb{R}^{k\times(n-k)}$, and finally $\hat{\mathcal{P}}_{11} \in \mathbb{R}^{k\times k}$, where $k < n$. To that effect, we have the three equations (6.29), which we rewrite as follows:

$$\left.\begin{aligned} &\mathbf{A}_{22}\mathcal{P}_{22} + \mathcal{P}_{22}\mathbf{A}_{22}^* + \mathbf{B}_2\mathbf{B}_2^* = \mathbf{0}, \\ &\mathbf{A}_{11}\mathcal{P}_{12} - \mathcal{P}_{12}\left[\mathcal{P}_{22}^{-1}\mathbf{A}_{22}\mathcal{P}_{22} + \mathcal{P}_{22}^{-1}\mathbf{B}_2\mathbf{B}_2^*\right] + \mathbf{A}_{12}\mathcal{P}_{22} + \mathbf{B}_1\mathbf{B}_2^* = \mathbf{0}, \\ &\mathbf{A}_{11}\hat{\mathcal{P}}_{11} + \hat{\mathcal{P}}_{11}\mathbf{A}_{11}^* + \hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^* = \mathbf{0}. \end{aligned}\right\} \tag{6.31}$$

We can thus solve the first (Lyapunov) equation for $\mathcal{P}_{22}$; this is an equation of small size, or at least of smaller size than the original one. Then, we solve the second equation, which is a Sylvester equation for $\mathcal{P}_{12}$. Then, given the fact that the pair $(\mathbf{A}_{11}, \hat{\mathbf{B}}_1)$ is reachable according to Proposition 6.20, the problem of solving a Lyapunov equation of size n is reduced to that of solving a Lyapunov equation of smaller size $k < n$. At the same time, we have computed the last $n - k$ rows and columns of the factor $\mathbf{U}$ and, because of triangularity, of the solution $\mathcal{P}$.

If we wish to apply this method in practice, we choose $k = n - 1$; in other words, we compute successively the last row/column of the square root factors. From the first equation (6.31), we obtain

$$\boxed{\mathcal{P}_{22} = -\frac{\mathbf{B}_2\mathbf{B}_2^*}{\mathbf{A}_{22} + \mathbf{A}_{22}^*} \in \mathbb{R}.} \tag{6.32}$$

Then, from the middle equation follows $\left[\mathbf{A}_{11} + \mathbf{A}_{22}^*\mathbf{I}_{n-1}\right]\mathcal{P}_{12} = \mathbf{A}_{12}\frac{\mathbf{B}_2\mathbf{B}_2^*}{\mathbf{A}_{22}+\mathbf{A}_{22}^*} - \mathbf{B}_1\mathbf{B}_2^*$, which implies

$$\boxed{\mathcal{P}_{12} = \left[\mathbf{A}_{11} + \mathbf{A}_{22}^*\mathbf{I}_{n-1}\right]^{-1}\left[\frac{\mathbf{B}_2}{\mathbf{A}_{22} + \mathbf{A}_{22}^*}\mathbf{A}_{12} - \mathbf{B}_1\right]\mathbf{B}_2^* \in \mathbb{R}^{n-1}.} \tag{6.33}$$

Thus, we have determined the last column of the upper triangular square root factor $\mathbf{U}$ in (6.30). There remains to determine the upper triangular square root factor $\hat{\mathcal{P}}_{11}^{1/2}$ which satisfies the Lyapunov equation given by the third equation (6.31). Recall that $\hat{\mathbf{B}}_1 = \mathbf{B}_1 - \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathbf{B}_2$ and, by Proposition 6.20, the pair $(\mathbf{A}_{11}, \hat{\mathbf{B}}_1)$ is reachable. Consequently, the problem is reduced to one of dimension *one less*.

Example 6.24. Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -24 & -50 & -35 & -10 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

The Schur decomposition of $\mathbf{A}$ is $\mathbf{A} = \mathbf{U}\tilde{\mathbf{A}}\mathbf{U}^*$, where

$$\mathbf{U} = \begin{bmatrix} -1 & 11 & 82 & 6 \\ 1 & -7 & 21 & 11 \\ -1 & -1 & -112 & 6 \\ 1 & 17 & -51 & 1 \end{bmatrix} \cdot \operatorname{diag}\left[\frac{1}{2}, \frac{1}{\sqrt{460}}, \frac{1}{\sqrt{22310}}, \frac{1}{\sqrt{194}}\right]$$

and $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}} = \mathbf{U}^*\mathbf{B}$ are

$$\tilde{\mathbf{A}} = \begin{bmatrix} -1 & -\frac{3\sqrt{5}}{\sqrt{23}} & \frac{133\sqrt{5}}{\sqrt{4462}} & -\frac{460}{\sqrt{194}} \\ 0 & -2 & \frac{109}{\sqrt{194}} & -\frac{773\sqrt{10}}{\sqrt{2231}} \\ 0 & 0 & -3 & \frac{49\sqrt{5}}{\sqrt{23}} \\ 0 & 0 & 0 & -4 \end{bmatrix},$$

$$\tilde{\mathbf{B}} = \operatorname{diag}\left[\frac{1}{2}, \frac{1}{\sqrt{460}}, \frac{1}{\sqrt{22310}}, \frac{1}{\sqrt{194}}\right] \begin{pmatrix} 1 \\ 17 \\ -51 \\ 1 \end{pmatrix},$$

$$\mathcal{P} = \begin{bmatrix} \frac{1}{2016} & 0 & -\frac{1}{2520} & 0 \\ 0 & \frac{1}{2520} & 0 & -\frac{1}{504} \\ -\frac{1}{2520} & 0 & \frac{1}{504} & 0 \\ 0 & -\frac{1}{504} & 0 & \frac{151}{2520} \end{bmatrix} = \mathbf{LL}^* = \mathbf{UU}^*, \quad \text{where}$$

$$\mathbf{L} = \begin{bmatrix} \frac{\sqrt{14}}{168} & 0 & 0 & 0 \\ 0 & \frac{\sqrt{70}}{420} & 0 & 0 \\ -\frac{\sqrt{14}}{210} & 0 & \frac{\sqrt{6}}{60} & 0 \\ 0 & -\frac{\sqrt{70}}{84} & 0 & \frac{\sqrt{5}}{10} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \frac{\sqrt{6}}{120} & 0 & -\frac{\sqrt{14}}{420} & 0 \\ 0 & \frac{\sqrt{755}}{1510} & 0 & -\frac{\sqrt{10570}}{12684} \\ 0 & 0 & \frac{\sqrt{14}}{84} & 0 \\ 0 & 0 & 0 & \frac{\sqrt{10570}}{420} \end{bmatrix}.$$

Example 6.25. We wish to solve $\mathbf{F}\mathcal{P} + \mathcal{P}\mathbf{F}^* + \mathbf{GG}^* = \mathbf{0}$, using Hammarling's algorithm, where

$$\mathbf{F} = \frac{1}{2}\begin{bmatrix} -4 & \sqrt{2} & \sqrt{2} & 0 \\ -\sqrt{2} & 0 & 12 & 0 \\ -\sqrt{2} & 0 & 0 & \sqrt{2} \\ 0 & \sqrt{2} & 0 & -4 \end{bmatrix}, \quad \mathbf{G} = \sqrt{2}\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}.$$

First, **F** is transformed into Schur form; the orthogonal matrix that achieves this is

$$\mathbf{W} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ -1 & 1 & 0 & -1 \\ 1 & 0 & 1 & -1 \\ -1 & 0 & 1 & 1 \end{bmatrix} \cdot \operatorname{diag}\begin{bmatrix} \frac{1}{2} & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \frac{1}{2} \end{bmatrix}.$$

Thus

$$\mathbf{A} = \mathbf{WFW}^* = \begin{bmatrix} -1 & 2 & 3 & 4 \\ 0 & -1 & 2 & 3 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \quad \mathbf{B} = \mathbf{WG} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

• *Step* 1. $\mathbf{A} \leftarrow \mathbf{A}, \mathbf{B} \leftarrow \mathbf{B}, n \leftarrow n, k \leftarrow n-1$. Equation (6.32) is solved for $\mathcal{P}_{22}$; then (6.33) is solved for the $n-1$ vector $\mathcal{P}_{12}$. We get $\mathcal{P}_{22} = 1/2, \mathcal{P}_{12} = [21/4 \quad 9/4 \quad 1]^*$. According to (6.30), the last column of $\mathbf{U}$ is $\mathbf{U}(1:3,4) = \sqrt{2}[21/4 \quad 9/4 \quad 1]^*$, $\mathbf{U}(4,4) = \sqrt{2}/2$. We also have $\mathbf{B}_3 = \mathbf{B}(1:3,1) - \mathcal{P}_{12}\frac{\mathbf{B}(4,1)}{\mathcal{P}_{22}} = [-19/2 \quad -7/2 \quad -1]^*$.

• *Step* 2. $\mathbf{A} \leftarrow \mathbf{A}(1:3,1:3), \mathbf{B} \leftarrow \mathbf{B}_3, n \leftarrow n-1, k \leftarrow n-1$. Equation (6.32) is solved for $\mathcal{P}_{22}$; then (6.33) is solved for the $n-2$ vector $\mathcal{P}_{12}$. $\mathcal{P}_{22} = 1/2, \mathcal{P}_{12} = [31/4 \quad 9/4]^*$. According to (6.30), the third column of $\mathbf{U}$ is $\mathbf{U}(1:2,3) = \sqrt{2}[31/4 \quad 9/4]^*$, $\mathbf{U}(3,3) = \sqrt{2}/2$. We also have $\mathbf{B}_2 = \mathbf{B}_3(1:2,1) - \mathcal{P}_{12}\frac{\mathbf{B}_3(3,1)}{\mathcal{P}_{22}} = [6 \quad 1]^*$.

• *Step* 3. $\mathbf{A} \leftarrow \mathbf{A}(1:2,1:2), \mathbf{B} \leftarrow \mathbf{B}_2, n \leftarrow n-1, k \leftarrow n-1$. Equation (6.32) is solved for $\mathcal{P}_{22}$; then (6.33) is solved for the $n-3$ vector $\mathcal{P}_{12}$. $\mathcal{P}_{22} = 1/2, \mathcal{P}_{12} = [7/2]^*$. According to (6.30), the second column of $\mathbf{U}$ is $\mathbf{U}(1:1,2) = \sqrt{2}[7/2]^*$, $\mathbf{U}(2,2) = \sqrt{2}/2$. We also have $\mathbf{B}_3 = \mathbf{B}_2(1:1,1) - \mathcal{P}_{12}\frac{\mathbf{B}_2(2,1)}{\mathcal{P}_{22}} = [-1]^*$.

• *Step* 4. $\mathbf{A} \leftarrow \mathbf{A}(1:1,1:1), \mathbf{B} \leftarrow \mathbf{B}_3, n \leftarrow n-1, k \leftarrow n-1$. Equation (6.32) is solved for $\mathcal{P}_{22}$; there is no equation (6.33) in this case: $\mathcal{P}_{22} = 1/2$, and hence $\mathbf{U}(1,1) = \sqrt{2}/2$.

Putting together the columns computed above, we obtain the upper triangular factor of the solution to the Lyapunov equation in the Schur basis:

$$\mathbf{U} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 7 & \frac{31}{2} & \frac{21}{2} \\ 0 & 1 & \frac{9}{2} & \frac{9}{2} \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

To obtain the square root factor in the original coordinates we have to multiply by $\mathbf{W}^*$:

$$\mathbf{U} \leftarrow \mathbf{W}^*\mathbf{U} = \operatorname{diag}\left[\frac{\sqrt{2}}{4}, \frac{1}{2}, \frac{1}{2}, \frac{\sqrt{2}}{4}\right] \cdot \begin{bmatrix} 1 & 6 & 12 & 7 \\ 1 & 8 & 20 & 15 \\ 0 & 0 & 1 & 3 \\ 1 & 6 & 10 & 5 \end{bmatrix}.$$

Thus the solution $\mathcal{P}$ in the original quantities is

$$\mathcal{P} = \mathbf{U}\mathbf{U}^* = \begin{bmatrix} \frac{115}{4} & \frac{197\sqrt{2}}{4} & \frac{33\sqrt{2}}{8} & 24 \\ \frac{197\sqrt{2}}{4} & \frac{345}{2} & \frac{65}{4} & \frac{81\sqrt{2}}{2} \\ \frac{33\sqrt{2}}{8} & \frac{65}{4} & \frac{5}{2} & \frac{25\sqrt{2}}{8} \\ 24 & \frac{81\sqrt{2}}{2} & \frac{25\sqrt{2}}{8} & \frac{81}{4} \end{bmatrix}.$$

Finally, MATLAB gives the following solution:

```
lyap(F,G*G') =
   2.8750e+001   6.9650e+001   5.8336e+000   2.4000e+001
   6.9650e+001   1.7250e+002   1.6250e+001   5.7276e+001
   5.8336e+000   1.6250e+001   2.5000e+000   4.4194e+000
   2.4000e+001   5.7276e+001   4.4194e+000   2.0250e+001,
```

which up to machine precision is the same as the $\mathcal{P}$ obtained above.

6.3.4 Algorithms for Sylvester and Lyapunov equations

We will now quote two algorithms given in [311]. The first computes the solution of the Sylvester equation in real arithmetic, and the second computes the square root factors of a Lyapunov equation in complex arithmetic.

Algorithm: Solution of the Sylvester equation $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$ in real arithmetic.

```
Input data:   $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{k\times k}$, $\mathbf{C} \in \mathbb{R}^{n\times k}$.
Output data:  $\mathbf{X} \in \mathbb{R}^{n\times n}$.

  1. Compute the real Schur decomposition $\mathbf{A} = \mathbf{QRQ}^*$ with $\mathbf{R}$: quasi-upper triangular.
  2. If $\mathbf{B} = \mathbf{A}$, set $\mathbf{U} = \mathbf{Q}$, $\mathbf{T} = \mathbf{R}$; else if $\mathbf{A} = \mathbf{A}^*$, get $\mathbf{U}$, $\mathbf{T}$ from $\mathbf{Q}$, $\mathbf{R}$ as follows:
       $\mathtt{idx} = [\mathtt{size}(\mathbf{A}, 1) : -1 : 1]$; $\mathbf{U} = \mathbf{Q}(:, \mathtt{idx})$; $\mathbf{T} = \mathbf{R}(\mathtt{idx}, \mathtt{idx})^*$;
     else compute the real Schur decomposition $\mathbf{B} = \mathbf{UTU}^*$ with $\mathbf{T}$ quasi-upper triangular.
  3. $\mathbf{C} \leftarrow \mathbf{Q}^*\mathbf{CQ}$, $\mathbf{R2} \leftarrow \mathbf{R} * \mathbf{R}$, $\mathbf{I} \leftarrow \mathrm{eye}(n)$, $j \leftarrow 1$.
  4. While $(j < k + 1)$
       • if $j < n$ and $\mathbf{T}(j + 1, j) < 10 * \epsilon * \max(|\,\mathbf{T}(j, j)\,|, |\,\mathbf{T}(j + 1, j + 1)\,|)$
           (a) $\mathbf{b} \leftarrow -\mathbf{C}(:, j) - \mathbf{X}(:, 1 : j - 1) * \mathbf{T}(1 : j - 1, j)$;
           (b) solve $(\mathbf{R} + \mathbf{T}(j, j)\mathbf{I})\mathbf{x} = \mathbf{b}$, and set $\mathbf{X}(:, j) \leftarrow \mathbf{x}$;
           (c) $j \leftarrow j + 1$.
       • else
           (a) $t_{11} \leftarrow \mathbf{T}(j, j), t_{12} \leftarrow \mathbf{T}(j, j+1), t_{21} \leftarrow \mathbf{T}(j+1, j), t_{22} \leftarrow \mathbf{T}(j+1, j+1)$;
           (b) $\mathbf{b} \leftarrow -\mathbf{C}(:, j : j + 1) - \mathbf{X}(:, 1 : j - 1) * \mathbf{T}(1 : j - 1, j : j + 1)$;
               $\mathbf{b} \leftarrow [\mathbf{Rb}(:, 1) + t_{22}\mathbf{b}(:, 1) - t_{21}\mathbf{b}(:, 2), \mathbf{Rb}(:, 2) + t_{11}\mathbf{b}(:, 2) - t_{12}\mathbf{b}(:, 1)]$
           (c) block solve the linear equations
               $[\mathbf{R2} + (t_{11} + t_{22})\mathbf{R} + (t_{11}t_{22} - t_{12}t_{21})\mathbf{I}]\mathbf{x} = \mathbf{b}$, and set $\mathbf{X}(:, j : j + 1) \leftarrow \mathbf{x}$;
           (d) $j \leftarrow j + 2$
       • end if
  5. The solution $\mathbf{X}$ in the original basis is $\mathbf{X} \leftarrow \mathbf{QXQ}^*$.
```

The algorithm, up to part (c) of "else," is the *Bartels–Stewart algorithm*, which is coded in MATLAB as `lyap.m`. Thereby, all quantities obtained are real. If **A** or **B** has complex eigenvalues, this is achieved at the expense of introducing 2 × 2 blocks on the diagonal in the Schur form, which makes a step like the block solve necessary.

The following is essentially Hammarling's algorithm for computation of the square root factor of the solution to the Lyapunov equation (in case the latter is positive definite). It is implemented in complex arithmetic. In other words, if **A** has complex eigenvalues and **R** is the complex Schur form, the resulting square root factor in the Schur basis will in general be complex. Of course, in the original basis it will be real. A version of this algorithm in real arithmetic can be derived in a way similar to that employed for the Bartels–Stewart algorithm. For details, see [311]. This algorithm is also implemented in SLICOT.

Algorithm: Computation of the square root factor of the solution to the Lyapunov equation

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$$

in complex arithmetic.

Input data: $\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{U}^* \in \mathbb{R}^{n\times n}$, $\mathbf{R} \in \mathbb{C}^{n\times n}$ upper triangular, $\mathcal{R}e\lambda_j(\mathbf{A}) < 0$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $(\mathbf{A}, \mathbf{B})$ reachable.

Output data: square root factor $\mathbf{T} \in \mathbb{R}^{n\times n}$ of solution $\mathcal{P} = \mathbf{T}\mathbf{T}^* > 0$.

```
• $\mathbf{T} \leftarrow$ zeros(n, n)
• for $j = n : -1 : 2$ do
    1. $\mathbf{b} \leftarrow \mathbf{B}(j,:)$; $\mu \leftarrow \|\mathbf{b}\|_2$; $\mu_1 \leftarrow \sqrt{-2 * \mathrm{real}(\mathbf{R}(j,j))}$
    2. if $\mu \neq 0$
           $\mathbf{b} \leftarrow \mathbf{b}/\mu$; $\mathbf{I} \leftarrow$ eye(j-1)
           temp $\leftarrow \mathbf{B}(1:j-1,:) * \mathbf{b}^* * \mu_1 + \mathbf{R}(1:j-1,j) * \mu/\mu_1$
           solve for $\mathbf{u}$: $(\mathbf{R}(1:j-1,1:j-1) + \mathbf{R}(j,j)^* * \mathbf{I})\mathbf{u} = -$temp
           $\mathbf{B}(1:j-1,:) \leftarrow \mathbf{B}(1:j-1,:) - \mathbf{u} * \mathbf{b} * \mu_1$
       else
           $\mathbf{u} \leftarrow$ zeros(j - 1, 1)
       end if
    3. $\mathbf{T}(j,j) \leftarrow \mu/\mu_1$
    4. $\mathbf{T}(1:j-1,j) \leftarrow \mathbf{u}$
• $\mathbf{T}(1,1) \leftarrow \|\mathbf{B}(1,:)\|_2/\sqrt{-2 * \mathrm{real}(\mathbf{R}(1,1))}$
• Square root factor in original basis $\mathbf{T} \leftarrow \mathbf{U}^*\mathbf{T}$.
```

6.4 Numerical issues: Forward and backward stability*

In applications, due to finite precision and various errors, the Sylvester and the Lyapunov equations can be solved only approximately. The question of how good these solutions are arises therefore. As discussed in section 3.3.2, there are two ways to assess the quality of a solution: *forward and backward stability*. The former is explained as follows. Let $\tilde{\mathbf{X}}$ be an approximate solution of the Sylvester equation (6.1). This means that the *residual*,

$$\mathbf{R} = \mathbf{A}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}\mathbf{B} - \mathbf{C},$$

is different from zero. A small relative residual,

$$\frac{\|\mathbf{R}\|}{\|\tilde{\mathbf{X}}\|},$$

indicates forward stability. It was shown in [43] that

$$\frac{\|\mathbf{R}\|_F}{\|\tilde{\mathbf{X}}\|_F} \leq \mathcal{O}(\epsilon)(\|\mathbf{A}\|_F + \|\mathbf{B}\|_F),$$

where $\mathcal{O}(\epsilon)$ is a polynomial in ϵ (machine precision) of modest degree. Therefore, the solution of the Sylvester equation is *forward stable*.

To address the issue of backward stability, we need to define the *backward error* of an approximate solution $\tilde{\mathbf{X}}$. Following [170], the *normwise* backward error is

$$\eta(\tilde{\mathbf{X}}) = \min\{\epsilon : (\mathbf{A}+\mathbf{E})\tilde{\mathbf{X}}+\tilde{\mathbf{X}}(\mathbf{B}+\mathbf{F}) = \mathbf{C}+\mathbf{G},\ \|\mathbf{E}\|_F \leq \epsilon\alpha,\ \|\mathbf{F}\|_F \leq \epsilon\beta,\ \|\mathbf{G}\|_F \leq \epsilon\gamma\}.$$

The positive constants α, β, γ provide freedom in measuring the perturbations. A frequent choice is

$$\alpha = \|\mathbf{A}\|_F,\ \beta = \|\mathbf{B}\|_F,\ \gamma = \|\mathbf{G}\|_F.$$

Before stating a general result, we discuss an example.

Example 6.26. Given the positive numbers σ_1, σ_2, let

$$\mathbf{A} = \begin{pmatrix} \frac{1}{2\sigma_1} & \frac{1}{\sigma_1+\sigma_2} \\ \frac{1}{\sigma_1+\sigma_2} & \frac{1}{2\sigma_2} \end{pmatrix} = \mathbf{B},\ \mathbf{C} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The exact solution of the Sylvester equation (6.1) is $\mathbf{X} = \operatorname{diag}(\sigma_1, \sigma_2)$. Assume now that the following approximate solution has been obtained instead: $\tilde{\mathbf{X}} = \operatorname{diag}(\sigma_1, \tilde{\sigma}_2)$, where $\sigma_2 - \tilde{\sigma}_2$ is small. The residual in this case is

$$\mathbf{R} = \mathbf{A}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}\mathbf{A} - \mathbf{C} = \begin{pmatrix} 0 & \frac{\sigma_2-\tilde{\sigma}_2}{\sigma_1+\sigma_2} \\ \frac{\sigma_2-\tilde{\sigma}_2}{\sigma_1+\sigma_2} & \frac{\sigma_2-\tilde{\sigma}_2}{\sigma_2} \end{pmatrix}.$$

Assuming that $\mathbf{C}$ is known exactly (i.e., $\mathbf{G} = \mathbf{0}$), to determine the backward error we need to solve $\mathbf{E}\tilde{\mathbf{X}} + \tilde{\mathbf{X}}\mathbf{F} = -\mathbf{R}$ for the perturbations $\mathbf{E}$ and $\mathbf{F}$. To preserve the structure, we will assume that $\mathbf{E} = \mathbf{F}^* = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. It turns out that $a = 0$ and $d = \frac{\tilde{\sigma}_2-\sigma_2}{2\sigma_2\tilde{\sigma}_2}$. Furthermore, the least squares (smallest norm) solution of $\sigma_1 c + \tilde{\sigma}_2 d = \frac{\tilde{\sigma}_2-\sigma_2}{\sigma_1+\sigma_2}$ is $c = \sigma_1(\tilde{\sigma}_2 - \sigma_2)/(\sigma_1 + \sigma_2)\sqrt{\sigma_1^2 + \tilde{\sigma}_2^2}$, $d = \tilde{\sigma}_2(\tilde{\sigma}_2 - \sigma_2)/(\sigma_1 + \sigma_2)\sqrt{\sigma_1^2 + \tilde{\sigma}_2^2}$. Thus, assuming that $\sigma_1 = 1 \gg \sigma_2, \tilde{\sigma}_2$, the 2-norm of the relative residual is approximately $\phi = \frac{|\tilde{\sigma}_2-\sigma_2|}{\sigma_2}$, while that of the perturbation $\mathbf{E}$ is approximately equal to $\frac{\phi}{\tilde{\sigma}_2}$. This shows that the backward error in this case is equal to the forward error divided by the smallest singular value of the approximate solution. Depending on the value of $\tilde{\sigma}_2$, the backward error can therefore be big.

The fact that the conclusion reached in the example holds in general was shown by Higham in [170]. We will give the formula here, assuming that the right-hand side of the equation is known exactly (i.e., $\mathbf{G} = \mathbf{0}$), $n = k$, and the approximate solution $\tilde{\mathbf{X}}$ has full rank n. Let the smallest singular value of $\tilde{\mathbf{X}}$ be σ_n. The following upper bound holds for the backward error:

$$\eta(\tilde{\mathbf{X}}) \leq \mu \cdot \frac{\|\mathbf{R}\|_F}{(\alpha+\beta)\|\tilde{\mathbf{X}}\|_F}, \quad \text{where } \mu = \frac{(\alpha+\beta)}{\sqrt{\alpha^2+\beta^2}} \cdot \frac{\|\tilde{\mathbf{X}}\|_F}{\sigma_n(\tilde{\mathbf{X}})}.$$

Here, μ is an amplification factor that indicates by how much the backward error can exceed the relative residual. Indeed, if the approximate solution is badly conditioned (near rank deficient), the backward error may become arbitrarily large.

Remark 6.4.1. To put the issue of backward error in perspective, we mention that the solution of the linear set of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ is backward stable. In fact, the following can be proved:

$$\min\{\epsilon : (\mathbf{A}+\mathbf{E})\tilde{\mathbf{x}} = \mathbf{b}+\mathbf{f},\ \|\mathbf{E}\|_2 \leq \epsilon\alpha,\ \|\mathbf{f}\|_2 \leq \epsilon\beta\} = \frac{\|\mathbf{r}\|_2}{\alpha\|\tilde{\mathbf{x}}\|_2+\beta}.$$

In this case the backward error is directly linked to the relative residual, namely, $\frac{\|\mathbf{r}\|}{\|\tilde{\mathbf{x}}\|}$.

In contrast, the solution of the system of equations $\mathbf{A}\mathbf{X} = \mathbf{B}$, where $\mathbf{B}$ is a matrix and not just a vector, is not backward stable. If we choose $\mathbf{B} = \mathbf{I}$, this amounts to the computation of the inverse of the matrix $\mathbf{A}$, which, in turn, is the special case of the Sylvester equation where $\mathbf{B} = \mathbf{0}$. Thus even without knowledge of the above formula, we would not expect the solution of the Sylvester equation to be backward stable.

The condition number of the solutions

Another question that comes up when solving Sylvester or Lyapunov equations concerns the properties that the coefficient matrices must satisfy so that the solution is *well-conditioned.*

Before addressing this question, we define the *condition number* of the Sylvester equation around a solution $\mathbf{X}$. Consider

$$(\mathbf{A}+\delta\mathbf{A})(\mathbf{X}+\delta\mathbf{X}) + (\mathbf{X}+\delta\mathbf{X})(\mathbf{B}+\delta\mathbf{B}) = \mathbf{C},$$

where again we assume for simplicity that the right-hand side is known exactly. After retaining only first-order terms, we get

$$\mathbf{A}(\delta\mathbf{X}) + (\delta\mathbf{X})\mathbf{B} = (\delta\mathbf{A})\mathbf{X} + \mathbf{X}(\delta\mathbf{B}) \Rightarrow \mathcal{L}\text{vec}(\delta\mathbf{X}) = [\mathbf{X}^* \otimes \mathbf{I}_n,\ \ \mathbf{I}_k \otimes \mathbf{X}] \begin{bmatrix} \text{vec}(\delta\mathbf{A}) \\ \text{vec}(\delta\mathbf{B}) \end{bmatrix},$$

where $\mathcal{L} = \mathbf{I}_k \otimes \mathbf{A} + \mathbf{B}^* \otimes \mathbf{I}_n$ is the Lyapunov operator. Following [170], we will call the quantity

$$\kappa = \frac{1}{\|\mathbf{X}\|_F} \cdot \|\mathcal{L}^{-1}[\alpha(\mathbf{X}^* \otimes \mathbf{I}_n),\ \ \beta(\mathbf{I}_k \otimes \mathbf{X})]\|_2 \tag{6.34}$$

the *condition number* of the Sylvester equation. Again, we shall attempt to clarify this formula by considering an example.

Example 6.27. Let $\mathbf{A} = \begin{pmatrix} -\frac{1}{2} & -b \\ b & 0 \end{pmatrix} = \mathbf{B}^*$, $b > 0$, and $\mathbf{C} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. The solution of the Sylvester equation (6.1)—which is a Lyapunov equation—is the identity matrix $\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which is *perfectly conditioned*. We will show that depending on b, the associated condition number (6.34) can be arbitrarily large. In fact, for $b = 10^{-1}$, $\kappa = 27.8$; $b = 10^{-2}$, $\kappa = 2.5\,10^3$; $b = 10^{-3}$, $\kappa = 2.5\,10^5$; $b = 10^{-8}$, $\kappa = 2.5\,10^{15}$.

The conclusion from this example is that there does not seem to be any connection between the condition number of the Sylvester and the Lyapunov equations and their solutions. In other words, the derivative of the Lyapunov function can be big even around solutions that are perfectly conditioned.

6.5 Chapter summary

The Sylvester equation is a linear matrix equation. The Lyapunov equation is also a linear matrix equation where the coefficient matrices satisfy symmetry constraints. The solution of these equations is a prominent part of many model reduction schemes. Therefore, in the preceding chapter these equations were studied in some detail.

First we notice the multitude of methods for solving the Sylvester and Lyapunov equations. The Kronecker method reduces their solution to that of a linear system of equations of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{x}$, $\mathbf{b}$ are vectors. The complex integration method uses the properties of Cauchy integrals to compute solutions. Then come the eigenvalue/eigenvector method, followed by two characteristic polynomial methods, an invariant subspace method, and one method using the sign function.

The second section discusses an important property of the Lyapunov equation, known as *inertia*. The inertia of a matrix (with respect to some domain $\mathcal{D}$ in the complex plane) is composed of three nonnegative numbers, namely, the number of eigenvalues inside the domain, on its boundary, and outside the domain. Depending on the application, $\mathcal{D}$ can be taken as the open left half plane, or the open unit disc. The inertia result then asserts that if the right-hand side of (6.2) is (semi) definite, the inertia on the coefficient matrix $\mathbf{A}$ and of the solution $\mathcal{P}$ are related.

In the third section, consequences of transforming the coefficient matrices to triangular form are discussed. The Bartels–Stewart algorithm was the first numerically reliable method for solving the Sylvester equation. The Lyapunov equation with triangular coefficient matrices is studied next. Following a proof of the inertia result which is iterative in the size of the equation, an algorithm is derived that if the solution to the Lyapunov equation is positive (semi) definite, computes a *square root* (i.e., Cholesky) factor of the solution. This is the algorithm first derived by Hammarling. It has the advantage of increased precision and will prove essential for the computation of the Hankel singular values.

The last section addresses the numerical issue of accuracy of a computed solution. As discussed in section 3.3.2, the stability of the corresponding algorithm is important. It is thus shown that although the solution of the Sylvester equation is *forward stable*, it is, in general, *not* backward stable. This is a property shared with the computation of the inverse of a matrix (which can be obtained by solving a special Sylvester equation). Moreover, the conditioning of the Sylvester equation and of its solution have in general no connection.

Part III

SVD-based Approximation Methods

This page intentionally left blank

Chapter 7

Balancing and Balanced Approximations

A central concept in system theory with application to model reduction is that of *balanced representation* of a system Σ. Roughly speaking, the states in such a representation are such that the *degree of reachability* and the *degree of observability* of each state are the same.

Model reduction requires the elimination of some of the state variables from the original or a transformed system representation, a task which can be accomplished easily. Difficulties arise, however, when one wishes (i) to determine whether the reduced system has inherited properties from the original one, for instance, stability, and (ii) to have some idea of what has been eliminated, for instance, when one seeks an estimate of the norm of the error system.

Indeed, if we first transform the system to a *balanced representation* and then eliminate (truncate) some of the state variables, stability is preserved and there is an a priori computable error bound for the error system.

From a mathematical viewpoint, balancing methods consist of the simultaneous diagonalization of appropriate reachability and observability gramians, which are positive definite matrices. In system theory, however, other instances of positive definite matrices are attached to a linear system, notably, solutions to various Riccati equations. Correspondingly, there exist several other types of balancing. Section 7.5 lists some of these methods and examines in some detail *weighted balancing*, which provides a link with moment matching reduction methods.

The concept of balancing is first encountered in the work of Mullis and Roberts [245] on the design of digital filters. The system theoretic significance of this concept was recognized a few years later by Moore [243].

Technical description

Given a stable linear system Σ with impulse response $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$, $t > 0$ ($\mathbf{D}$ is irrelevant in this case and is omitted), let $\mathbf{h}_r(t) = e^{\mathbf{A}t}\mathbf{B}$, $\mathbf{h}_o(t) = \mathbf{C}e^{\mathbf{A}t}$ be the input-to-state and the state-to-output responses of the system, respectively; clearly, $\mathbf{h}(t) = \mathbf{C}\mathbf{h}_r(t) = \mathbf{h}_o(t)\mathbf{B}$.

The *reachability gramian* is then defined as $\mathcal{P} = \int_0^\infty \mathbf{h}_r(t)\mathbf{h}_r^*(t)dt$, while the *observability gramian* is defined as $\mathcal{Q} = \int_0^\infty \mathbf{h}_o^*(t)\mathbf{h}_o(t)dt$. The significance of these quantities stems from the fact that given a state $\mathbf{x}$, the smallest amount of energy needed to steer the system from $\mathbf{0}$ to $\mathbf{x}$ is given by (4.55)

$$\mathcal{E}_r = \mathbf{x}^*\mathcal{P}^{-1}\mathbf{x},$$

while the energy obtained by observing the output of the system with initial condition $\mathbf{x}$ and no excitation function is given by (4.56)

$$\mathcal{E}_o = \mathbf{x}^*\mathcal{Q}\mathbf{x}.$$

Thus, one way to reduce the number of states is to eliminate those which require a large amount of energy $\mathcal{E}_r$ to be reached and/or yield small amounts of observation energy $\mathcal{E}_o$. However, these concepts are *basis dependent*, and therefore for such a scheme to work, one would have to look for a basis in which these two concepts are *equivalent*. Such a basis exists. It is called a *balanced basis*, and in this basis there holds

$$\mathcal{P} = \mathcal{Q} = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n),$$

where the σ_i are the Hankel singular values of $\mathbf{\Sigma}$. Approximation in this basis takes place by truncating the state $\mathbf{x} = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_n)^*$ to $\hat{\mathbf{x}} = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_k)^*$, $k < n$. Although balanced truncation does not seem to be optimal in any norm, it has nevertheless a variational interpretation. The balanced basis is determined, namely, by a nonsingular transformation $\mathbf{T}$ which solves the following minimization problem:

$$\min_{\mathbf{T}} \operatorname{trace}\left[\mathbf{T}\mathcal{P}\mathbf{T}^* + \mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1}\right].$$

The minimum of this expression is twice the sum of the Hankel singular values, $2\sum_{i=1}^n \sigma_i$, and minimizing $\mathbf{T}$ are *balancing transformations* (see Proposition 7.7).

Approximation by balanced truncation preserves stability, and the $\mathcal{H}_\infty$-norm (the maximum of the frequency response) of the error system is bounded above by twice the sum of the neglected singular values $2(\sigma_{k+1} + \ \cdots \ + \sigma_n)$.

This chapter is concerned with balancing and approximation by balanced truncation. Approximation by truncation is defined, and its two main properties, preservation of stability and error bound, are discussed in detail. Subsequently, a canonical form for continuous-time balanced systems is derived, followed by numerical considerations in computing balanced representations. The latter part of the chapter, which can be omitted on first reading, discusses various generalizations of the concept of balancing. In particular, stochastic, bounded real, and positive real balancing are presented, followed by frequency selective balancing. This leads to balancing and balanced truncation for unstable systems.

7.1 The concept of balancing

From Lemma 4.29 it follows that the states that are difficult to reach, i.e., those that require a large amount of energy to reach, are (have a significant component) in the span of the eigenvectors of the reachability gramian $\mathcal{P}$ corresponding to small eigenvalues. Similarly, the states that are difficult to observe, i.e., those that yield small amounts of observation energy, are those that lie (have a significant component) in the span of the eigenvectors of the observability gramian $\mathcal{Q}$ corresponding to small eigenvalues as well. This observation suggests that reduced-order models may be obtained by eliminating those states that are difficult to reach or difficult to observe. However, states that are difficult to reach may not be difficult to observe and vice versa. Here is a simple example illustrating this point.

Example 7.1. Consider the following continuous-time, stable, and minimal system:

$$\mathbf{A} = \begin{pmatrix} 1 & 3 \\ -1 & -2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 0 & 1 \end{pmatrix}.$$

The reachability gramian $\mathcal{P}$ and the observability gramian $\mathcal{Q}$ can be obtained by solving the Lyapunov equations (4.45) and (4.46), respectively:

$$\mathcal{P} = \begin{pmatrix} \frac{5}{2} & -1 \\ -1 & \frac{1}{2} \end{pmatrix}, \quad \mathcal{Q} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

The set of eigenvalues Λ and the corresponding eigenvectors $\mathbf{V}$ are

$$\Lambda_{\mathcal{P}} = \begin{pmatrix} 2.91421 & 0 \\ 0 & 0.08578 \end{pmatrix}, \quad \mathbf{V}_{\mathcal{P}} = \begin{pmatrix} 0.92388 & 0.38268 \\ -0.38268 & 0.92388 \end{pmatrix},$$

$$\Lambda_{\mathcal{Q}} = \begin{pmatrix} 1.30901 & 0 \\ 0 & 0.19098 \end{pmatrix}, \quad \mathbf{V}_{\mathcal{Q}} = \begin{pmatrix} 0.52573 & -0.85865 \\ 0.85865 & 0.52573 \end{pmatrix}.$$

To find how difficult to observe the states are, we compute

$$\mathbf{V}_{\mathcal{Q}}^{*}\mathbf{V}_{\mathcal{P}} = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix},$$

where $\alpha = 0.16018$ and $\beta = 0.98709$. This means that the eigenvector of $\mathcal{P}$ that corresponds to the smallest eigenvalue, i.e., the state that is the most difficult state to reach, gives almost maximum observation energy; conversely, the eigenvector of $\mathcal{P}$ that corresponds to the largest eigenvalue, i.e., the state that is the easiest state to reach, gives almost minimum observation energy.

The above example suggests that if we wish to base a model reduction procedure to the degree to which states are difficult to reach, or difficult to observe, we need to search for a basis in which states that are difficult to reach are *simultaneously* difficult to observe, and vice versa. From these considerations, the following question arises:

> Given a continuous- or discrete-time, stable system $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$, does there exist a basis in the state space in which states that are *difficult to reach* are also *difficult to observe*?

The answer to this question is affirmative. The transformation that achieves this goal is called a *balancing transformation*. Recall that under an equivalence transformation, the gramians are transformed as shown in (4.57):

$$\tilde{\mathcal{P}} = \mathbf{T}\mathcal{P}\mathbf{T}^{*}, \quad \tilde{\mathcal{Q}} = \mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1} \quad \Rightarrow \quad \tilde{\mathcal{P}}\tilde{\mathcal{Q}} = \mathbf{T}(\mathcal{P}\mathcal{Q})\mathbf{T}^{-1}.$$

The problem is to find $\mathbf{T}$, $\det \mathbf{T} \neq 0$, such that the transformed Gramians $\tilde{\mathcal{P}}$, $\tilde{\mathcal{Q}}$ are equal:

$$\tilde{\mathcal{P}} = \tilde{\mathcal{Q}}.$$

This ensures that the states that are difficult to reach are precisely those that are difficult to observe.

Definition 7.2. *The reachable, observable, and stable system $\boldsymbol{\Sigma}$ is* balanced *if $\mathcal{P} = \mathcal{Q}$. $\boldsymbol{\Sigma}$ is* principal-axis balanced *if $\mathcal{P} = \mathcal{Q} = \Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_n)$.*

Furthermore, Lemma 5.8 implies that the quantities σ_i, $i = 1, \ldots, n$, are the *Hankel singular values* of the system $\boldsymbol{\Sigma}$.

Remark 7.1.1. From a linear algebraic point of view, the concept of *balancing* consists of the *simultaneous diagonalization* of two positive (semi) definite matrices.

We can now state the main lemma of this section. For this we need the Cholesky factor $\mathbf{U}$ of $\mathcal{P}$ and the eigenvalue decomposition of $\mathbf{U}^*\mathcal{Q}\mathbf{U}$:

$$\mathcal{P} = \mathbf{U}\mathbf{U}^*, \ \mathbf{U}^*\mathcal{Q}\mathbf{U} = \mathbf{K}\Sigma^2\mathbf{K}^*. \tag{7.1}$$

Lemma 7.3. Balancing transformation. *Given the reachable, observable, and stable system $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ and the corresponding gramians $\mathcal{P}$ and $\mathcal{Q}$, a (principal axis) balancing transformation is given as follows:*

$$\boxed{\mathbf{T} = \Sigma^{1/2}\mathbf{K}^*\mathbf{U}^{-1}} \quad \text{and} \quad \boxed{\mathbf{T}^{-1} = \mathbf{U}\mathbf{K}\Sigma^{-1/2}.} \tag{7.2}$$

Proof. It is readily verified that $\mathbf{T}\mathcal{P}\mathbf{T}^* = \Sigma$ and $\mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1} = \Sigma$. □

If the Hankel singular values are distinct (i.e., have multiplicity one), balancing transformations $\hat{\mathbf{T}}$ are uniquely determined from $\mathbf{T}$ given above, up to multiplication by a *sign* matrix $\mathbf{S}$, i.e., a diagonal matrix with ± 1 on the diagonal: $\hat{\mathbf{T}} = \mathbf{S}\mathbf{T}$. In the general case we have the next corollary.

Corollary 7.4. *Let there be k distinct singular values σ_i, with multiplicities m_i, respectively. Every principal-axis balancing transformation $\hat{\mathbf{T}}$ has the form $\hat{\mathbf{T}} = \mathbf{S}\mathbf{T}$, where $\mathbf{T}$ is defined by* (7.2), *and $\mathbf{S}$ is a block diagonal unitary matrix with an arbitrary $m_i \times m_i$ unitary matrix as ith block, $i = 1, \ldots, k$.*

Proposition 7.5. *Balanced systems have the following property. In continuous time $\| e^{\mathbf{A}t} \|_2 < 1$, $t > 0$. In discrete time $\| \mathbf{A} \|_2 \leq 1$ with strict inequality holding if the Hankel singular values are distinct.*

Example 7.6 (*continuation*). Using (5.24), we obtain the Hankel singular values $\sigma_1 = 0.8090, \sigma_2 = 0.3090$. Define the quantity $\gamma = 0.6687$. Using (7.2), we obtain the (principal-axis) balancing transformation:

$$\mathbf{T} = \gamma \begin{pmatrix} 1 & \frac{1}{4\sigma_1^2} \\ 1 & \frac{1}{4\sigma_2^2} \end{pmatrix}.$$

Thus we have

$$\Sigma_{\text{bal}} = \left(\begin{array}{c|c} \mathbf{TAT}^{-1} & \mathbf{TB} \\ \hline \mathbf{CT}^{-1} & \end{array}\right) = \left(\begin{array}{cc|c} -\frac{\gamma^2}{2\sigma_1} & \frac{\gamma^2}{\sigma_1-\sigma_2} & \gamma \\ \frac{\gamma^2}{\sigma_2-\sigma_1} & -\frac{\gamma^2}{2\sigma_2} & \gamma \\ \hline -\gamma & \gamma & \end{array}\right) = \left(\begin{array}{cc|c} -0.2763 & 0.8944 & 0.6687 \\ -0.8944 & -0.7236 & 0.6687 \\ \hline -0.6687 & 0.6687 & \end{array}\right).$$

Finally, notice that the above form matches the canonical form (7.24) with γ, σ_1, σ_2 as above, and $s_1 = -1, s_2 = 1$.

Variational interpretation

Balancing transformations can be obtained by minimization of an appropriate expression. First recall that the product of the gramians is similar to a positive definite matrix: $\mathcal{P}\mathcal{Q} \sim \mathcal{P}^{1/2}\mathcal{Q}\mathcal{P}^{1/2} = \mathbf{U}\Sigma^2\mathbf{U}^*$. The following holds true.

Proposition 7.7. *Given that* trace $\mathcal{P}\mathcal{Q} =$ trace Σ^2, *we have*

$$\text{trace}\,(\mathcal{P} + \mathcal{Q}) \geq 2 \text{ trace } \Sigma.$$

Therefore, the lower bound is attained in the balanced case: $\mathcal{P} = \mathcal{Q} = \Sigma$.

Proof. The following equalities hold:

$$\begin{aligned} \mathcal{P} + \mathcal{Q} &= \mathcal{P} + \mathcal{P}^{-1/2}\mathbf{U}\Sigma^2\mathbf{U}^*\mathcal{P}^{-1/2} \\ &= (\mathcal{P}^{1/2}\mathbf{U} - \mathcal{P}^{-1/2}\mathbf{U}\Sigma)(\mathbf{U}^*\mathcal{P}^{1/2} - \Sigma\mathbf{U}^*\mathcal{P}^{-1/2}) \\ &\quad + \mathcal{P}^{-1/2}\mathbf{U}\Sigma\mathbf{U}^*\mathcal{P}^{1/2} + \mathcal{P}^{1/2}\mathbf{U}\Sigma\mathbf{U}^*\mathcal{P}^{-1/2}. \end{aligned}$$

Notice that $\mathbf{M} = (\mathcal{P}^{1/2}\mathbf{U} - \mathcal{P}^{-1/2}\mathbf{U}\Sigma)(\mathbf{U}^*\mathcal{P}^{1/2} - \Sigma\mathbf{U}^*\mathcal{P}^{-1/2})$ is a positive (semi) definite expression and hence its trace is positive. Thus we obtain

$$\text{trace}\,(\mathcal{P} + \mathcal{Q}) = \text{trace } \mathbf{M} + 2\,\text{trace } \Sigma \;\Rightarrow\; \text{trace}\,(\mathcal{P} + \mathcal{Q}) \geq 2 \text{ trace } \Sigma.$$

It readily follows that the lower bound is attained in the balanced case. □

7.2 Model reduction by balanced truncation

Let $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ be balanced with gramians equal to $\boldsymbol{\Sigma}$; partition the corresponding matrices as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \; \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}, \; \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \; \mathbf{C} = (\mathbf{C}_1 \;\; \mathbf{C}_2). \tag{7.3}$$

Definition 7.8. *The systems*

$$\boxed{\boldsymbol{\Sigma}_i = \left(\begin{array}{c|c} \mathbf{A}_{ii} & \mathbf{B}_i \\ \hline \mathbf{C}_i & \end{array}\right), \qquad i = 1, 2,} \tag{7.4}$$

are reduced-order systems obtained from $\boldsymbol{\Sigma}$ *by* balanced truncation.

Reduced-order models obtained by balanced truncation have certain *guaranteed* properties. However these properties are slightly different for discrete- and continuous-time systems, and hence we state two theorems. Recall Definition 5.7 of the Hankel singular values and notation (5.22).

Theorem 7.9. Balanced truncation: Continuous-time systems. *Given the reachable, observable, and stable (poles in the open left half plane) continuous-time system* $\boldsymbol{\Sigma}$*, the reduced-order systems* $\boldsymbol{\Sigma}_i$*,* $i = 1, 2$*, obtained by balanced truncation have the following properties:*

1. $\boldsymbol{\Sigma}_i$ *is balanced and has no poles in the open right half plane.*

2. *If* $\lambda_p(\Sigma_1) \neq \lambda_q(\Sigma_2)\ \forall p, q$*,* $\boldsymbol{\Sigma}_i$ *for both* $i = 1$ *and* $i = 2$ *are in addition reachable, and observable, with no poles on the imaginary axis.*

3. *Let the distinct singular values of* $\boldsymbol{\Sigma}$ *be* σ_i *with multiplicities* m_i*,* $i = 1, \ldots, q$*. Let* $\boldsymbol{\Sigma}_1$ *have singular values* σ_i*,* $i = 1, \ldots, k$*, with the multiplicity* m_i*,* $i = 1, \ldots, k$*,* $k < q$*. The* $\mathcal{H}_\infty$*-norm of the difference between the full-order system* $\boldsymbol{\Sigma}$ *and the reduced-order system* $\boldsymbol{\Sigma}_1$ *is upper bounded by twice the sum of the neglected Hankel singular values, multiplicities not included:*

$$\| \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_1 \|_{\mathcal{H}_\infty} \leq 2\,(\sigma_{k+1} + \cdots + \sigma_q). \tag{7.5}$$

Furthermore, equality holds if $\Sigma_2 = \sigma_q \mathbf{I}_{m_q}$*, i.e.,* Σ_2 *is equal to the smallest Hankel singular value of* $\boldsymbol{\Sigma}$*.*

Theorem 7.10. Balanced truncation: Discrete-time systems. *Given the reachable, observable, and stable (poles in open unit disc) discrete-time system* $\boldsymbol{\Sigma}$*, the reduced-order systems* $\boldsymbol{\Sigma}_i$*,* $i = 1, 2$*, obtained by balanced truncation have the following properties:*

1. $\boldsymbol{\Sigma}_i$*,* $i = 1, 2$*, have poles in the closed unit disc; these systems are in general not balanced.*

2. *If* $\lambda_{min}(\Sigma_1) > \lambda_{max}(\Sigma_2)$*,* $\boldsymbol{\Sigma}_1$ *is in addition reachable and observable.*

3. *If* $\boldsymbol{\Sigma}_1$ *has singular values* σ_i*,* $i = 1, \ldots, k$*, with the multiplicity* m_i*,* $i = 1, \ldots, k$*,* $k < q$*, the* h_∞*-norm of the difference between full and reduced-order models is upper bounded by twice the sum of the neglected Hankel singular values multiplicities not included:*

$$\| \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_1 \|_{h_\infty} \leq 2\,(\sigma_{k+1} + \cdots + \sigma_q).$$

The last part of the above theorems says that if the neglected singular values are small, the amplitude Bode plots of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$ are guaranteed to be *close*. Below is an outline of the proof of parts 1 and 2. Notice that in part 3 for both continuous- and discrete-time systems, the multiplicities of the neglected singular values do *not* enter the upper bound. These error bounds are due to Glover [139] and Enns [107].

Remark 7.2.1. *$\mathcal{L}_1$-error bound.* In addition to (7.5) for the $\mathcal{H}_\infty$-norm of the error system, bounds for the $\mathcal{L}_1$-norm of the impulse response of the error system can be derived. One such bound was obtained in [141]. It is valid for infinite-dimensional systems and involves a minimization. An a priori computable bound involving the Hankel singular values and their multiplicities was derived in [218]. When the Hankel singular values have multiplicity one, this bound is as follows:

$$\|\mathbf{h}-\mathbf{h}_k\|_{\mathcal{L}_1} \leq 2\left[4\sum_{\ell=k+1}^{n}\ell\sigma_\ell - 3\sum_{\ell=k+1}^{n}\sigma_\ell\right],$$

where $\mathbf{h}$ is the impulse response of the original system and $\mathbf{h}_k$ is the impulse response of the kth-order approximant obtained by balanced truncation.

Example 7.11 (*continuation*). If we reduce the system discussed earlier by balanced truncation, the $\mathcal{H}_\infty$-norm of the error between $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$ is equal to the theoretical upper bound, namely, $2\sigma_2 = 0.6180$. Furthermore, the $\mathcal{H}_\infty$-norm of the error between $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_2$ is also equal to the theoretical upper bound, which in this case is $2\sigma_1 = 1.618033$.

7.2.1 Proof of the two theorems

Balanced truncation: Continuous-time

Proof. *Part* 1. We work with the subsystem $\mathbf{\Sigma}_1 = (\mathbf{C}_1, \mathbf{A}_{11}, \mathbf{B}_1)$. By construction, the following equations hold:

$$\mathbf{A}_{11}\Sigma_1 + \Sigma_1\mathbf{A}_{11}^* + \mathbf{B}_1\mathbf{B}_1^* = \mathbf{0},\quad \mathbf{A}_{11}^*\Sigma_1 + \Sigma_1\mathbf{A}_{11} + \mathbf{C}_1^*\mathbf{C}_1 = \mathbf{0}.$$

Clearly, the system $\mathbf{\Sigma}_1$ is balanced. To prove stability, since Σ_1 is positive definite, by Lemma 6.18 we conclude that the eigenvalues of $\mathbf{A}_{11}$ are in the left half plane or on the imaginary axis.

Part 2. There remains to show that if Σ_1 and Σ_2 have no diagonal entries in common, then $\mathbf{A}_{11}$ has no eigenvalues on the imaginary axis.

Assume on the contrary that $\mathbf{A}_{11}$ has eigenvalues on the imaginary axis. For simplicity, we assume that $\mathbf{A}_{11}$ has only one complex eigenvalue on the imaginary axis; we also assume that $\sigma_1 = 1$ with multiplicity one. Let $\mathbf{A}_{11}\mathbf{v} = \nu\mathbf{v}$ for $\nu = j\omega$; it follows that $\mathbf{v}^*\mathbf{A}_{11}^* = \nu^*\mathbf{v}^*$. Multiplying the second equation above on the left, right, by $\mathbf{v}^*$, $\mathbf{v}$, respectively, we obtain $\mathbf{C}_1\mathbf{v} = \mathbf{0}$. By multiplying the same equation only on the right by $\mathbf{v}$, we obtain $(\mathbf{A}_{11}^* + \nu\mathbf{I})\Sigma_1\mathbf{v} = \mathbf{0}$. Similarly, by multiplying the first equation on the left, right, by $\mathbf{v}^*\Sigma_1$, $\Sigma_1\mathbf{v}$, respectively, we conclude that $\mathbf{B}_1^*\Sigma_1\mathbf{v} = \mathbf{0}$; on multiplication of the same equation on the right by $\Sigma_1\mathbf{v}$, we obtain the relationship $(\mathbf{A}_{11} - \nu\mathbf{I})\Sigma_1^2\mathbf{v} = \mathbf{0}$. The consequence of this is that $\Sigma_1^2\mathbf{v}$ must be a multiple of $\mathbf{v}$. Due to the assumptions above, we can take $\mathbf{v}$ to be equal to the first unit vector $\mathbf{v} = \mathbf{e}_1$. Next we need to consider the equations

$$\mathbf{A}_{21}\Sigma_1 + \Sigma_2\mathbf{A}_{12}^* + \mathbf{B}_2\mathbf{B}_1^* = \mathbf{0}, \quad \Sigma_2\mathbf{A}_{21} + \mathbf{A}_{12}^*\Sigma_1 + \mathbf{C}_2^*\mathbf{C}_1 = \mathbf{0}.$$

Denote the first column of $\mathbf{A}_{21}$, $\mathbf{A}_{12}^*$ by $\mathbf{a}$, $\mathbf{b}$, respectively. Multiplying these latter equations on the right by $\mathbf{v}$ we have $\mathbf{a} + \Sigma_2\mathbf{b} = \mathbf{0}$ and $\Sigma_2\mathbf{a} + \mathbf{b} = \mathbf{0}$. The eigenvalues of Σ_2 are different from those of Σ_1 and hence different from 1. Therefore, the first column of $\mathbf{A}_{21}$ is zero: $\mathbf{a} = \mathbf{0}$. Therefore, the column vector $(\mathbf{v}^*, \ \mathbf{0})^*$ is an eigenvector of the whole matrix $\mathbf{A}$ corresponding to the eigenvalue ν. This, however, is a contradiction to the reachability of the pair $(\mathbf{A}, \mathbf{B})$. The conclusion is that $\mathbf{A}_{11}$ cannot have eigenvalues on the imaginary axis, which completes the proof. □

Alternative proof of part 2. It is assumed that $\mathbf{A}_{11}$ has eigenvalues both on the imaginary axis and in the LHP. Thus, there exists a block-diagonal transformation $\mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$ such that $\mathbf{T}_{11}\mathbf{A}_{11}\mathbf{T}_{11}^{-1} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_{22} \end{pmatrix}$, where all eigenvalues of $\mathbf{F}_{11}$ are (strictly) in the LHP (i.e., have negative real parts) and the eigenvalues of $\mathbf{F}_{22}$ are on the imaginary axis. Let the quantities $\mathbf{B}$, $\mathbf{C}$, and Σ be transformed as follows:

$$\bar{\mathbf{A}} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{0} & \mathbf{F}_{13} \\ \mathbf{0} & \mathbf{F}_{22} & \mathbf{F}_{23} \\ \mathbf{F}_{31} & \mathbf{F}_{32} & \mathbf{F}_{33} \end{pmatrix}, \quad \bar{\mathbf{B}} = \begin{pmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \mathbf{G}_3 \end{pmatrix}, \quad \bar{\mathbf{C}} = \begin{pmatrix} \mathbf{H}_1 & \mathbf{H}_2 & \mathbf{H}_3 \end{pmatrix},$$

$$\bar{\mathcal{P}} = \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{0} \\ \mathbf{P}_{12}^* & \mathbf{P}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{33} \end{pmatrix}, \quad \bar{\mathcal{Q}} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} & \mathbf{0} \\ \mathbf{Q}_{12}^* & \mathbf{Q}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Q}_{33} \end{pmatrix},$$

where $\mathbf{F}_{22}$ has eigenvalues on the imaginary axis, $\mathbf{F}_{33} = \mathbf{A}_{22}$, $\mathbf{G}_3 = \mathbf{B}_2$, $\mathbf{H}_3 = \mathbf{C}_2$, $\mathbf{P}_{33} = \mathbf{Q}_{33} = \Sigma_2$. The proof now proceeds in three steps:

(a) First, we show that $\mathbf{G}_2 = \mathbf{0}$ and $\mathbf{H}_2 = \mathbf{0}$.

(b) Next, we show that $\mathbf{P}_{12} = \mathbf{0}$, $\mathbf{Q}_{12} = \mathbf{0}$, that is, in the above basis, the gramians are block diagonal.

(c) Finally, provided that $\sigma_k \neq \sigma_{k+1}$, or equivalently $\lambda_i(\mathbf{P}_{22}\mathbf{Q}_{22}) \neq \lambda_j(\mathbf{P}_{33}\mathbf{Q}_{33})$, for all i, j, it follows that $\mathbf{F}_{23} = \mathbf{0}$ and $\mathbf{F}_{32} = \mathbf{0}$.

The consequence of these three relationships is that $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is not reachable; this implication, however, contradicts the original assumption of reachability.

The proof of the above three facts is based on the Lyapunov equations of the transformed triple, namely, $\bar{\mathbf{A}}^*\bar{\mathbf{Q}} + \bar{\mathbf{Q}}\bar{\mathbf{A}} + \bar{\mathbf{C}}^*\bar{\mathbf{C}} = \mathbf{0}$ and $\bar{\mathbf{A}}\bar{\mathbf{P}} + \bar{\mathbf{P}}\bar{\mathbf{A}}^* + \bar{\mathbf{B}}\bar{\mathbf{B}}^* = \mathbf{0}$,

$$\begin{pmatrix} \mathbf{F}_{11} & \mathbf{0} & \mathbf{F}_{13} \\ \mathbf{0} & \mathbf{F}_{22} & \mathbf{F}_{23} \\ \mathbf{F}_{31} & \mathbf{F}_{32} & \mathbf{F}_{33} \end{pmatrix} \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{0} \\ \mathbf{P}_{12}^* & \mathbf{P}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{33} \end{pmatrix} + \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \mathbf{0} \\ \mathbf{P}_{12}^* & \mathbf{P}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{P}_{33} \end{pmatrix} \begin{pmatrix} \mathbf{F}_{11}^* & \mathbf{0} & \mathbf{F}_{31}^* \\ \mathbf{0} & \mathbf{F}_{22}^* & \mathbf{F}_{32}^* \\ \mathbf{F}_{13}^* & \mathbf{F}_{23}^* & \mathbf{F}_{33}^* \end{pmatrix}$$
$$+ \begin{pmatrix} \mathbf{G}_1\mathbf{G}_1^* & \mathbf{G}_1\mathbf{G}_2^* & \mathbf{G}_1\mathbf{G}_3^* \\ \mathbf{G}_2\mathbf{G}_1^* & \mathbf{G}_2\mathbf{G}_2^* & \mathbf{G}_2\mathbf{G}_3^* \\ \mathbf{G}_3\mathbf{G}_1^* & \mathbf{G}_3\mathbf{G}_2^* & \mathbf{G}_3\mathbf{G}_3^* \end{pmatrix} = \mathbf{0},$$

and the associated observability Lyapunov equation.

The (2,2) equation is $\mathbf{F}_{22}\mathbf{P}_{22}+\mathbf{P}_{22}\mathbf{F}_{22}^*+\mathbf{G}_2\mathbf{G}_2^*=\mathbf{0}$. Because $\mathbf{F}_{22}$ has imaginary eigenvalues exclusively, it follows that $\mathbf{G}_2=\mathbf{0}$. (This is left as an exercise; see Problem 5.) Similarly, $\mathbf{H}_2=\mathbf{0}$. This proves (a). The (1,2) equation is $\mathbf{F}_{11}\mathbf{P}_{12}+\mathbf{P}_{12}\mathbf{F}_{22}^*=\mathbf{0}$. Since $\mathbf{F}_{11}$ and $\mathbf{F}_{22}$ have disjoint sets of eigenvalues, we conclude that $\mathbf{P}_{12}=\mathbf{0}$; similarly, $\mathbf{Q}_{12}=\mathbf{0}$, which proves (b). Finally, the (2,3) equation is $\mathbf{F}_{23}\mathbf{P}_{33}+\mathbf{P}_{22}\mathbf{F}_{32}^*=\mathbf{0}$; the observability Lyapunov equation yields $\mathbf{F}_{32}^*\mathbf{Q}_{33}+\mathbf{Q}_{22}\mathbf{F}_{23}=\mathbf{0}$. These two equations yield

$$\mathbf{F}_{23}\mathbf{P}_{33}\mathbf{Q}_{33}+\mathbf{P}_{22}\mathbf{Q}_{22}\mathbf{F}_{23}=\mathbf{0}.$$

Recall that $\mathbf{P}_{33}=\mathbf{Q}_{33}=\Sigma_2$ and the eigenvalues of $\mathbf{P}_{22}\mathbf{Q}_{22}$ are a subset of the eigenvalues of Σ_1^2. Because the singular values are ordered in decreasing order and because of the assumption $\sigma_k\neq\sigma_{k+1}$, the spectra of $\mathbf{P}_{33}\mathbf{Q}_{33}$ and $\mathbf{P}_{22}\mathbf{Q}_{22}$ are disjoint. Therefore, $\mathbf{F}_{23}=\mathbf{0}$; consequently, $\mathbf{F}_{32}=\mathbf{0}$. This proves part (c). As noted earlier, these three parts imply the lack of reachability and observability of the original triple, which is a contradiction to the assumption of minimality. Hence the reduced system is asymptotically stable.

Balanced truncation: Discrete time

Proof. From the Lyapunov equations follows

$$\mathbf{A}_{11}\Sigma_1\mathbf{A}_{11}^*+\mathbf{A}_{12}\Sigma_2\mathbf{A}_{12}^*+\mathbf{B}_1\mathbf{B}_1^*=\Sigma_1.$$

This shows that the truncated subsystem need not be balanced. Let $\mathbf{A}_{11}^*\mathbf{v}=\nu\mathbf{v}$. Multiplying this equation by $\mathbf{v}^*$, $\mathbf{v}$ we obtain

$$(1-|\nu|^2)\mathbf{v}^*\Sigma_1\mathbf{v}=\mathbf{v}^*\mathbf{A}_{12}\Sigma_2\mathbf{A}_{12}^*\mathbf{v}+\mathbf{v}^*\mathbf{B}_1\mathbf{B}_1^*\mathbf{v}\geq 0.$$

Since $\Sigma_1>0$, this immediately implies $|\nu|\leq 1$. There remains to show that the inequality is strict. Assume the contrary, i.e., that $|\nu|=1$. Then $\mathbf{v}^*\mathbf{A}_{12}=\mathbf{0}$. Hence $(\mathbf{v}^*\ \ \mathbf{0})$ is a left eigenvector of $\mathbf{A}$ and at the same time it is in the left kernel of $\mathbf{B}$, which means that $(\mathbf{A},\mathbf{B})$ is not reachable. This is, however, a contradiction, and part 1 of the theorem is proved.

The proof of part 2 is also by contradiction. Assume that the subsystem $\mathbf{\Sigma}_1$ is not reachable. Then there exists a left eigenvector ν of unit length, such that $\mathbf{A}_{11}^*\mathbf{v}=\nu\mathbf{v}$ and $\mathbf{B}_1^*\mathbf{v}=\mathbf{0}$. The equation given at the beginning of the proof implies

$$(1-|\nu|^2)\mathbf{v}^*\Sigma_1\mathbf{v}=\mathbf{v}^*\mathbf{A}_{12}\Sigma_2\mathbf{A}_{12}^*\mathbf{v}.$$

Moreover, $\mathbf{v}^*\Sigma_1\mathbf{v}\geq\sigma_{min}(\Sigma_1)$ and the right-hand side is no bigger than $\|\mathbf{A}_{21}\mathbf{v}\|^2\sigma_{max}(\Sigma_2)$. Furthermore, since $\|\mathbf{A}\|\leq 1$ and $\mathbf{A}_{11}^*\mathbf{v}=\nu\mathbf{v}$, it follows that $\|\mathbf{A}_{21}\mathbf{v}\|^2\leq 1-|\nu|^2$. Combining all these relationships together, we get

$$(1-|\nu|^2)\sigma_{min}(\Sigma_1)\leq(1-|\nu|^2)\sigma_{max}(\Sigma_2)\ \Rightarrow\ \sigma_{min}(\Sigma_1)\leq\sigma_{max}(\Sigma_2)$$

since the system is stable by part 1. This last inequality yields the desired contradiction which proves part 2.

For the proof of part 3, see [173]. □

A proof of the $\mathcal{H}_\infty$ error bound (7.5)

Recall the partitioning of the balanced system (7.3). The accordingly partitioned state is $\mathbf{x} = (\mathbf{x}_1^* \ \mathbf{x}_2^*)^*$. The reduced system is assumed to be Σ_1 defined by (7.4), and its state will be denoted by ξ. The resulting error system is $\boldsymbol{\Sigma}_e = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_1$; it has input $\mathbf{u}$, state $[\mathbf{x}_1^* \ \mathbf{x}_2^* \ \xi^*]^*$, and output $\mathbf{e} = \mathbf{y} - \bar{\mathbf{y}}$. Σ_e has the following state-space realization:

$$\Sigma_e = \left[\begin{array}{c|c} \mathbf{F} & \mathbf{G} \\ \hline \mathbf{H} & \end{array}\right] = \left[\begin{array}{ccc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{0} & \mathbf{B}_1 \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{0} & \mathbf{B}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{C}_2 & -\mathbf{C}_1 & \end{array}\right]. \tag{7.6}$$

Recall that the Lyapunov equations corresponding to the balanced system $\boldsymbol{\Sigma}$ are

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}^* & \mathbf{A}_{21}^* \\ \mathbf{A}_{12}^* & \mathbf{A}_{22}^* \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1^* & \mathbf{B}_2^* \end{bmatrix} = \mathbf{0},$$

$$\begin{bmatrix} \mathbf{A}_{11}^* & \mathbf{A}_{21}^* \\ \mathbf{A}_{12}^* & \mathbf{A}_{22}^* \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1^* \\ \mathbf{C}_2^* \end{bmatrix} \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} = \mathbf{0}.$$

The next step is to make a basis change $\mathbf{T}$ in the state-space:

$$\begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \\ \bar{\mathbf{x}}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 - \xi \\ \mathbf{x}_2 \\ \mathbf{x}_1 + \xi \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{I} & 0 & -\mathbf{I} \\ 0 & \mathbf{I} & 0 \\ \mathbf{I} & 0 & \mathbf{I} \end{bmatrix}}_{\mathbf{T}} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \xi \end{bmatrix}.$$

Therefore, $\boldsymbol{\Sigma}_e = \left[\begin{array}{c|c} \bar{\mathbf{F}} & \bar{\mathbf{G}} \\ \hline \bar{\mathbf{H}} & \end{array}\right]$, where the transformed quantities are

$$\bar{\mathbf{F}} = \mathbf{T}\mathbf{F}\mathbf{T}^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{0} \\ \frac{1}{2}\mathbf{A}_{21} & \mathbf{A}_{22} & \frac{1}{2}\mathbf{A}_{21} \\ \mathbf{0} & \mathbf{A}_{12} & \mathbf{A}_{11} \end{pmatrix}, \quad \bar{\mathbf{G}} = \mathbf{T}\mathbf{G} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_2 \\ 2\mathbf{B}_1 \end{pmatrix},$$

$$\bar{\mathbf{H}} = \mathbf{H}\mathbf{T}^{-1} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{0} \end{pmatrix}.$$

The following is the crucial result.

Lemma 7.12 (*see* [358]). *With the set-up as above, assume that the part that is eliminated is all-pass, that is,* $\Sigma_2 = \sigma\mathbf{I}$, $\sigma \in \mathbb{R}$, $\mathbf{I} \in \mathbb{R}^{r\times r}$. *It follows that the* $\mathcal{H}_\infty$*-norm of the error system is*

$$\|\boldsymbol{\Sigma}_e\|_{\mathcal{H}_\infty} \leq 2\sigma. \tag{7.7}$$

Proof. We show instead that $\|\frac{1}{2\sigma}\boldsymbol{\Sigma}_e\|_{\mathcal{H}_\infty} \leq 1$. The resulting state space realization is

$$\frac{1}{2\sigma}\boldsymbol{\Sigma}_e = \left[\begin{array}{c|c} \bar{\mathbf{F}} & \frac{1}{2\sigma}\bar{\mathbf{G}} \\ \hline \bar{\mathbf{H}} & \end{array}\right].$$

The matrix

$$\mathbf{X} = \begin{pmatrix} \Sigma_1 & & \\ & 2\mathbf{I}\sigma & \\ & & \Sigma_1^{-1}\sigma^2 \end{pmatrix}$$

is clearly positive definite. It can readily be verified that it satisfies the Riccati equation,[5]

$$\bar{\mathbf{F}}^*\mathbf{X} + \mathbf{X}\bar{\mathbf{F}} + \bar{\mathbf{H}}^*\bar{\mathbf{H}} + \frac{1}{(2\sigma)^2}\mathbf{X}\bar{\mathbf{G}}\bar{\mathbf{G}}^*\mathbf{X} = \mathbf{0}.$$

The claim follows from Lemma 5.35, part 3. □

It should be mentioned that (7.7) holds with equality. The above proof leads to a proof of the inequality. The results in section 8.6.4 lead to a proof of the relationship with equality (at least in the SISO case). For the proof of the general case, we proceed as follows. Let $\boldsymbol{\Sigma}$ have r distinct Hankel singular values $\sigma_1, \ldots, \sigma_r$, with multiplicity $m_1, \ldots, m_r$, such that $\sum_i m_i = n$. The following notation is used:

$$\boldsymbol{\Sigma}_{1,k} = \left[\begin{array}{ccc|c} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1k} & \mathbf{B}_1 \\ \mathbf{A}_{21} & \cdots & \mathbf{A}_{2k} & \mathbf{B}_2 \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{A}_{k1} & \cdots & \mathbf{A}_{kk} & \mathbf{B}_k \\ \hline \mathbf{C}_1 & \cdots & \mathbf{C}_k & \end{array}\right] \text{ with } \Sigma_k = \begin{bmatrix} \sigma_1\mathbf{I} & & & \\ & \sigma_2\mathbf{I} & & \\ & & \ddots & \\ & & & \sigma_k\mathbf{I} \end{bmatrix}, \quad k = 1, \ldots, r.$$

Notice that with the above notation, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{1,r}$ and $\boldsymbol{\Sigma}_{1,0}$ is the zero system. We can thus write

$$\begin{aligned} \boldsymbol{\Sigma} &= (\boldsymbol{\Sigma}_{1,r} - \boldsymbol{\Sigma}_{1,r-1}) + (\boldsymbol{\Sigma}_{1,r-1} - \boldsymbol{\Sigma}_{1,r-2}) + \cdots + (\boldsymbol{\Sigma}_{1,2} - \boldsymbol{\Sigma}_{1,1}) + (\boldsymbol{\Sigma}_{1,1} - \boldsymbol{\Sigma}_{1,0}) \\ \Rightarrow \|\boldsymbol{\Sigma}\| &\le \|\boldsymbol{\Sigma}_{1,r} - \boldsymbol{\Sigma}_{1,r-1}\| + \|\boldsymbol{\Sigma}_{1,r-1} - \boldsymbol{\Sigma}_{1,r-2}\| + \cdots \\ &\qquad \cdots + \|\boldsymbol{\Sigma}_{1,2} - \boldsymbol{\Sigma}_{1,1}\| + \|\boldsymbol{\Sigma}_{1,1} - \boldsymbol{\Sigma}_{1,0}\| \\ \Rightarrow \|\boldsymbol{\Sigma}\| &\le 2\sigma_r + 2\sigma_{r-1} + \cdots + 2\sigma_2 + 2\sigma_1. \end{aligned}$$

[5] The following string of equalities holds where for simplicity $\sigma = 1$:

$$\bar{\mathbf{F}}^*\mathbf{X} + \mathbf{X}\bar{\mathbf{F}} + \mathbf{H}^*\mathbf{H} + \tfrac{1}{4}\mathbf{X}\bar{\mathbf{G}}\bar{\mathbf{G}}^*\mathbf{X} = \mathbf{0},$$

$$\begin{aligned} &\begin{pmatrix} \mathbf{A}_{11}^*\Sigma_1 & \mathbf{A}_{21}^* & \mathbf{0} \\ \mathbf{A}_{12}^*\Sigma_1 & 2\mathbf{A}_{22}^* & \mathbf{A}_{12}^*\Sigma_1^{-1} \\ \mathbf{0} & \mathbf{A}_{21}^* & \mathbf{A}_{11}^*\Sigma_1^{-1} \end{pmatrix} + \begin{pmatrix} \Sigma_1\mathbf{A}_{11} & \Sigma_1\mathbf{A}_{12} & \mathbf{0} \\ \mathbf{A}_{21} & 2\mathbf{A}_{22} & \mathbf{A}_{21} \\ \mathbf{0} & \Sigma_1^{-1}\mathbf{A}_{12} & \Sigma_1^{-1}\mathbf{A}_{11} \end{pmatrix} \\ &+ \begin{pmatrix} \mathbf{C}_1^*\mathbf{C}_1 & \mathbf{C}_1^*\mathbf{C}_2 & \mathbf{0} \\ \mathbf{C}_2^*\mathbf{C}_1 & \mathbf{C}_2^*\mathbf{C}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2\mathbf{B}_2^* & \mathbf{B}_2\mathbf{B}_1^*\Sigma_1^{-1} \\ \mathbf{0} & \Sigma_1^{-1}\mathbf{B}_1\mathbf{B}_2^* & \Sigma_1^{-1}\mathbf{B}_1\mathbf{B}_1^*\Sigma_1^{-1} \end{pmatrix} \\ &= \left(\begin{array}{c|c|c} \mathbf{A}_{11}^*\Sigma_1 + \Sigma_1\mathbf{A}_{11} + \mathbf{C}_1^*\mathbf{C}_1 & \mathbf{A}_{21}^* + \Sigma_1\mathbf{A}_{12} + \mathbf{C}_1^*\mathbf{C}_2 & \mathbf{0} \\ \hline \mathbf{A}_{12}^*\Sigma_1 + \mathbf{A}_{21} + \mathbf{C}_2^*\mathbf{C}_1 & 2\mathbf{A}_{22}^* + 2\mathbf{A}_{22} + \mathbf{C}_2^*\mathbf{C}_2 + \mathbf{B}_2\mathbf{B}_2^* & \mathbf{A}_{12}^*\Sigma_1^{-1} + \mathbf{A}_{21} + \mathbf{B}_2\mathbf{B}_1^*\Sigma_1^{-1} \\ \hline \mathbf{0} & \mathbf{A}_{21}^* + \Sigma_1^{-1}\mathbf{A}_{12} + \Sigma_1^{-1}\mathbf{B}_1\mathbf{B}_2^* & \mathbf{A}_{11}^*\Sigma_1^{-1} + \Sigma_1^{-1}\mathbf{A}_{11} + \Sigma_1^{-1}\mathbf{B}_1\mathbf{B}_1^*\Sigma_1^{-1} \end{array}\right). \end{aligned}$$

From the Lyapunov equations given earlier, and keeping in mind that $\Sigma_2 = \mathbf{I}$, it follows that each one of the (block) entries of the above matrix is zero.

The first inequality follows from the triangle inequality and the second from Lemma 7.12, which asserts that $\|\boldsymbol{\Sigma}_{1,k} - \boldsymbol{\Sigma}_{1,k-1}\| \leq 2\sigma_k$ for all $k = 1, \ldots, r$. Thus, clearly, the error bound (7.5) holds.

7.2.2 $\mathcal{H}_2$-norm of the error system for balanced truncation

In this section we compute the $\mathcal{H}_2$-norm of the error system. Then using the expression for the solution of the Sylvester equation (6.1) given in (6.14), we obtain a *computable upper bound* for this $\mathcal{H}_2$-norm (7.10).

Again recall that $\boldsymbol{\Sigma}$ is the balanced partitioned system (7.3), $\boldsymbol{\Sigma}_1$ is the reduced system obtained by balanced truncation 7.4, and $\boldsymbol{\Sigma}_e$ is the resulting error system (7.6). Let $\mathbf{Y}$ be such that

$$\mathbf{A}^*\mathbf{Y} + \mathbf{Y}\mathbf{A}_{11} + \mathbf{C}^*\mathbf{C}_1 = \mathbf{0}, \tag{7.8}$$

which in partitioned form is

$$\begin{pmatrix} \mathbf{A}_{11}^* & \mathbf{A}_{21}^* \\ \mathbf{A}_{12}^* & \mathbf{A}_{22}^* \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \mathbf{A}_{11} + \begin{pmatrix} \mathbf{C}_1^* \\ \mathbf{C}_2^* \end{pmatrix} \mathbf{C}_1 = \mathbf{0}.$$

From (5.28), the $\mathcal{H}_2$-norm of the error system $\boldsymbol{\Sigma}_e$ is the square root of the following expression:

$$\| \boldsymbol{\Sigma}_e \|_{\mathcal{H}_2}^2 = \text{trace} \left[\begin{pmatrix} \mathbf{B}_1^* & \mathbf{B}_2^* & \mathbf{B}_1^* \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & -\mathbf{Y}_1 \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & -\mathbf{Y}_2 \\ -\mathbf{Y}_1^* & -\mathbf{Y}_2^* & \boldsymbol{\Sigma}_1 \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{B}_1 \end{pmatrix} \right].$$

Lemma 7.13. *With the notation established above, the $\mathcal{H}_2$-norm of the error system $\boldsymbol{\Sigma}_e = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_1$ is*

$$\| \boldsymbol{\Sigma}_e \|_{\mathcal{H}_2}^2 = \text{trace} \left[\left(\mathbf{B}_2\mathbf{B}_2^* + 2\,\mathbf{Y}_2\mathbf{A}_{12} \right) \boldsymbol{\Sigma}_2 \right]. \tag{7.9}$$

The first term in the above expression is the $\mathcal{H}_2$-norm of the neglected subsystem of the original system, while the second term is the inner product of the second block row of the gramian with $\mathbf{Y}$. It is this second term that we try to express by means of the original system data below. Next, we derive an upper bound for the $\mathcal{H}_2$ of the error in terms of the $\mathcal{H}_\infty$-norm of the auxiliary system,

$$\boldsymbol{\Sigma}_{\text{aux}} = \left(\begin{array}{cc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \boldsymbol{\Sigma}_2\mathbf{A}_{21} \\ \hline \mathbf{0} & \mathbf{A}_{12}\boldsymbol{\Sigma}_2 & \mathbf{0} \end{array} \right).$$

Notice that the transfer function of $\boldsymbol{\Sigma}_{\text{aux}}$ can be written as

$$\mathbf{H}_{\text{aux}}(s) = \mathbf{A}_{12}\boldsymbol{\Sigma}_2 \left[s\mathbf{I} - \mathbf{A}_{22} - \mathbf{A}_{21} \left(s\mathbf{I} - \mathbf{A}_{11} \right)^{-1} \mathbf{A}_{12} \right]^{-1} \boldsymbol{\Sigma}_2\mathbf{A}_{21}.$$

It is worth noting that this expression is *quadratic* in the neglected part $\boldsymbol{\Sigma}_2$ of the gramian.

Corollary 7.14. *With the notation established above, there holds*

$$\| \boldsymbol{\Sigma}_e \|_{\mathcal{H}_2}^2 \leq \text{trace} \left[\mathbf{B}_2^*\boldsymbol{\Sigma}_2\mathbf{B}_2 \right] + 2\,k\, \| \boldsymbol{\Sigma}_{\text{aux}} \|_{\mathcal{H}_\infty}. \tag{7.10}$$

Remark 7.2.2. The first term in (7.10) is linear in the neglected singular values Σ_2, while the second is quadratic in Σ_2.

Proof of the lemma. The following equalities hold:

$$\begin{aligned}\|\, \boldsymbol{\Sigma}_{\rm e} \,\|^2 &= \text{trace } \left[\mathbf{B}_1^*\Sigma_1\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_1\mathbf{B}_1 + \mathbf{B}_2^*\Sigma_2\mathbf{B}_2 \right.\\ &\qquad \left. - \mathbf{B}_2^*\mathbf{Y}_2\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_1^*\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_2^*\mathbf{B}_2 + \mathbf{B}_1^*\Sigma_1\mathbf{B}_1\right]\\ &= \text{trace } \left[\mathbf{B}_2^*\Sigma_2\mathbf{B}_2\right] + 2 \text{ trace}\,[\,\mathbf{B}_1^*\Sigma_1\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_1\mathbf{B}_1 \underbrace{-\mathbf{B}_2^*\mathbf{Y}_2\mathbf{B}_1}_{(\diamond)}\,].\end{aligned}$$

From the (1,2) entry of the Lyapunov equation for the balanced reachability gramian follows $\mathbf{A}_{12}\Sigma_2 + \Sigma_1\mathbf{A}_{21}^* + \mathbf{B}_1\mathbf{B}_2^* = \mathbf{0}$ and, consequently, $-\left(\mathbf{B}_1\mathbf{B}_2^*\right)\mathbf{Y}_2 = \left(\mathbf{A}_{12}\Sigma_2 + \Sigma_1\mathbf{A}_{21}^*\right)\mathbf{Y}_2$. This implies that trace ($\diamond$) is

$$\text{trace}\,[\,-\mathbf{B}_2^*\mathbf{Y}_2\mathbf{B}_1\,] = \text{trace}\,[\,-\mathbf{B}_1\mathbf{B}_2^*\mathbf{Y}_2\,] = \text{trace}\,[\,\mathbf{A}_{12}\Sigma_2\mathbf{Y}_2 + \Sigma_1\mathbf{A}_{21}^*\mathbf{Y}_2\,].$$

Substituting yields

$$\begin{aligned}\|\, \boldsymbol{\Sigma}_{\rm e} \,\|^2 &= \text{trace } \left[\mathbf{B}_2^*\Sigma_2\mathbf{B}_2\right] + 2 \text{ trace}\,[\,\mathbf{A}_{12}\Sigma_2\mathbf{Y}_2\,]\\ &\quad + 2\, \underbrace{\text{trace}\,[\,\mathbf{B}_1^*\Sigma_1\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_1\mathbf{B}_1 + \Sigma_1\mathbf{A}_{21}^*\mathbf{Y}_2\,]}_{(\triangleleft)}.\end{aligned}$$

We show that $(\triangleleft) = 0$. The (1,1) entry of (7.8) implies $\mathbf{A}_{11}^*\mathbf{Y}_1 + \mathbf{A}_{21}^*\mathbf{Y}_2 + \mathbf{Y}_1\mathbf{A}_{11} + \mathbf{C}_1^*\mathbf{C}_1 = \mathbf{0}$ and, consequently, $\Sigma_1\mathbf{A}_{21}^*\mathbf{Y}_2 = -(\Sigma_1\mathbf{A}_{11}^*\mathbf{Y}_1 + \Sigma_1\mathbf{Y}_1\mathbf{A}_{11} + \Sigma_1\mathbf{C}_1^*\mathbf{C}_1)$. Thus the expression trace $[\,\mathbf{B}_1^*\Sigma_1\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_1\mathbf{B}_1 + \Sigma_1\mathbf{A}_{21}^*\mathbf{Y}_2\,]$ equals

$$\begin{aligned}&\text{trace}\,[\,\mathbf{B}_1^*\Sigma_1\mathbf{B}_1 - \mathbf{B}_1^*\mathbf{Y}_1\mathbf{B}_1\,] - \text{trace}\,[\,\Sigma_1\mathbf{A}_{11}^*\mathbf{Y}_1 + \Sigma_1\mathbf{Y}_1\mathbf{A}_{11} + \Sigma_1\mathbf{C}_1^*\mathbf{C}_1\,]\\ &= \text{trace}\,[\,\mathbf{B}_1^*\Sigma_1\mathbf{B}_1\,] - \text{trace}\,[\,\mathbf{C}_1\Sigma_1\mathbf{C}_1^*\,] - \text{trace}\,[\,\mathbf{B}_1^*\mathbf{Y}_1\mathbf{B}_1 + \Sigma_1\mathbf{A}_{11}^*\mathbf{Y}_1 + \Sigma_1\mathbf{Y}_1\mathbf{A}_{11}\,]\\ &= \mathbf{0} - \text{trace}\,[\,\left(\mathbf{B}_1\mathbf{B}_1^* + \Sigma_1\mathbf{A}_{11}^* + \mathbf{A}_{11}\Sigma_1\right)\mathbf{Y}_1\,] = \mathbf{0} - \mathbf{0} = \mathbf{0}.\end{aligned}$$

This concludes the proof of (7.9). ■

Proof of the corollary. First we notice that $\mathbf{Y} - \begin{pmatrix}\Sigma_1\\ \mathbf{0}\end{pmatrix}$ satisfies the Sylvester equation:

$$\mathbf{A}^*\begin{pmatrix}\mathbf{Y}_1 - \Sigma_1\\ \mathbf{Y}_2\end{pmatrix} + \begin{pmatrix}\mathbf{Y}_1 - \Sigma_1\\ \mathbf{Y}_2\end{pmatrix}\mathbf{A}_{11} = \begin{pmatrix}\mathbf{0}\\ \Sigma_2\end{pmatrix}\mathbf{A}_{21}.$$

Thus, using the expression (6.14) derived for the solution of the Sylvester equation, and multiplying on the left by $[\mathbf{0}\;\; \mathbf{A}_{12}\Sigma_2]$, we obtain

$$\mathbf{A}_{12}\Sigma_2\mathbf{Y}_2 = \sum_{i=1}^{k}\begin{pmatrix}\mathbf{0} & \mathbf{A}_{12}\Sigma_2\end{pmatrix}(\mu_i\mathbf{I} + \mathbf{A})^{-1}\begin{pmatrix}\mathbf{0}\\ \Sigma_2\mathbf{A}_{21}\end{pmatrix}\mathbf{w}_i\tilde{\mathbf{w}}_i^*,$$

where μ_i, $\mathbf{w}_i$, $\tilde{\mathbf{w}}_i$, respectively, are the eigenvalues, right, left eigenvectors of the $\mathbf{A}$-matrix of the reduced-order system: $\mathbf{A}_{11}$. Thus we can derive an upper bound for the $\mathcal{H}_2$ of the error in terms of the $\mathcal{H}_\infty$-norm of the auxiliary system $\mathbf{\Sigma}_{\text{aux}}$ defined earlier. In particular, each term in the above sum is bounded from above by the $\mathcal{H}_\infty$-norm of this auxiliary system. Hence the sum is bounded by k times the same norm. We thus obtain an upper bound for the $\mathcal{H}_2$-norm. ■

Remark 7.2.3. If we assume that the gramian $\mathcal{Q}$ is not diagonal, instead of formula (7.9) we get the following expression for the $\mathcal{H}_2$-error:

$$\| \mathbf{\Sigma}_{\text{e}} \|_{\mathcal{H}_2}^2 = \text{trace} \left[\begin{pmatrix} \mathbf{B}_1^* & \mathbf{B}_2^* & \mathbf{B}_1^* \end{pmatrix} \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} & -\mathbf{Y}_1 \\ \mathcal{Q}_{12}^* & \mathcal{Q}_{22} & -\mathbf{Y}_2 \\ -\mathbf{Y}_1^* & -\mathbf{Y}_2^* & \hat{\mathcal{Q}} \end{pmatrix} \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{B}_1 \end{pmatrix} \right],$$

where $\mathbf{A}_{11}^* \hat{\mathcal{Q}} + \mathcal{Q}\mathbf{A}_{11} + \mathbf{C}_1^*\mathbf{C}_1 = \mathbf{0}$. Using manipulations similar to those above, it can be shown that the error reduces to the following expression:

$$\| \mathbf{\Sigma}_{\text{e}} \|_{\mathcal{H}_2}^2 = \text{trace}\,[\mathbf{B}_2^* \mathcal{Q}_{22} \mathbf{B}_2] + \text{trace}\,[\mathbf{B}_1^*(\hat{\mathcal{Q}} - \mathcal{Q}_{11})\mathbf{B}_1] + 2\,\text{trace}\,[\mathbf{A}_{12}^*(\mathcal{Q}_{2,:}\,\mathbf{Y})]. \tag{7.11}$$

Interpretation. The first term in the above expression is the $\mathcal{H}_2$-norm of the neglected subsystem of the original system. The second term is the difference between the $\mathcal{H}_2$-norms of the reduced-order system and the dominant subsystem of the original system. The third term is twice the trace of the inner product of the second block row of the observability gramian $\mathcal{Q}$ with $\mathbf{Y}$ weighted by the block off-diagonal entry of $\mathbf{A}$. Finally, $\hat{\mathcal{Q}} - \mathcal{Q}_{11}$ satisfies the Sylvester equation:

$$\mathbf{A}_{11}^*(\hat{\mathcal{Q}} - \mathcal{Q}_{11}) + (\hat{\mathcal{Q}} - \mathcal{Q}_{11})\mathbf{A}_{11} = \mathbf{A}_{12}^*\mathcal{Q}_{21} + \mathcal{Q}_{12}\mathbf{A}_{12}.$$

This implies that if either the gramian $\mathcal{Q}$ or $\mathbf{A}$ has small (zero) off-diagonal elements, then $\hat{\mathcal{Q}} - \mathcal{Q}_{11}$ will be small (zero). Clearly, formula (7.11) reduces to (7.9) when the gramian is block diagonal, i.e., $\mathcal{Q}_{12} = \mathbf{0}$ and $\mathcal{Q}_{21} = \mathbf{0}$.

7.3 Numerical issues: Four algorithms

Model reduction by balanced truncation, discussed above, requires balancing the whole system $\mathbf{\Sigma}$, followed by truncation. This approach may turn out to be numerically inefficient and ill-conditioned, especially for large-scale problems. The reason is that often $\mathcal{P}$ and $\mathcal{Q}$ have *numerically* low rank compared to n. This is due in many cases to the rapid decay of the eigenvalues of $\mathcal{P}$, $\mathcal{Q}$ as well as the singular values $\sigma_i(\mathbf{\Sigma})$. (See section 9.4 for details.) Therefore, it is important to avoid formulas involving matrix inverses, because of ill-conditioning due to the rapid decay of the eigenvalues of the gramians.

In this section we list several algorithms for balancing and balanced truncation, which although in theory are identical, in practice yield algorithms with quite different numerical properties. For a more in-depth account, see [55], [340].

The (infinite) gramians of a reachable, observable, and stable system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ of dimension n are $n \times n$ positive definite matrices which we have denoted by $\mathcal{P}$, $\mathcal{Q}$. Consequently, they have a Cholesky decomposition:

$$\mathcal{P} = \mathbf{U}\mathbf{U}^*, \quad \mathcal{Q} = \mathbf{L}\mathbf{L}^*, \quad \text{where } \mathbf{U}: \text{ upper triangular}, \mathbf{L}: \text{ lower triangular}. \tag{7.12}$$

The eigenvalue decomposition of $\mathbf{U}^*\mathcal{Q}\mathbf{U}$ produces the orthogonal matrix $\mathbf{K}$ and the diagonal matrix Σ which is composed of the Hankel singular values of $\boldsymbol{\Sigma}$ (this follows from the fact that $\mathbf{U}^*\mathcal{Q}\mathbf{U}$ is similar to $\mathcal{P}\mathcal{Q}$):

$$\mathbf{U}^*\mathcal{Q}\mathbf{U} = \mathbf{K}\Sigma^2\mathbf{K}^*. \tag{7.1}$$

In addition, the Hankel singular values are the *singular values* of the product of the triangular matrices $\mathbf{U}^*$, $\mathbf{L}$. The SVD of this product produces the orthogonal matrices $\mathbf{W}$ and $\mathbf{V}$:

$$\mathbf{U}^*\mathbf{L} = \mathbf{W}\Sigma\mathbf{V}^*. \tag{7.13}$$

Finally, we need the QR-factorization of the products $\mathbf{UW}$ and $\mathbf{LV}$:

$$\mathbf{UW} = \mathbf{X}\Phi, \ \ \mathbf{LV} = \mathbf{Y}\Psi, \tag{7.14}$$

where $\mathbf{X}$, $\mathbf{Y}$ are orthogonal and Φ, Ψ are upper triangular.

To summarize, from the gramians $\mathcal{P}$, $\mathcal{Q}$ we generate the *orthogonal matrices* $\mathbf{K}$, $\mathbf{W}$, $\mathbf{V}$, $\mathbf{X}$, $\mathbf{Y}$; the upper triangular matrices $\mathbf{U}$, $\mathbf{L}^*$, Φ, Ψ; and the diagonal matrix Σ.

We are now ready to derive various balancing transformations as well as projections that produce systems obtained from $\boldsymbol{\Sigma}$ by balanced truncation.

1. The first transformation and its inverse are

$$\boxed{\mathbf{T} = \Sigma^{1/2}\mathbf{K}^*\mathbf{U}^{-1}} \text{ and } \boxed{\mathbf{T}^{-1} = \mathbf{UK}\Sigma^{-1/2}}. \tag{7.2}$$

2. *Square root algorithm.* The second transformation follows from (7.13):

$$\boxed{\mathbf{T} = \Sigma^{-1/2}\mathbf{V}^*\mathbf{L}^*} \text{ and } \boxed{\mathbf{T}^{-1} = \mathbf{UW}\Sigma^{-1/2}}. \tag{7.15}$$

It is readily checked that indeed $\mathbf{T}_i = \mathbf{T}^{-1}$ is the inverse of $\mathbf{T}$, and $\mathbf{T}\mathcal{P}\mathbf{T}^* = \Sigma^{-1/2}\mathbf{V}^*\mathbf{L}^*\mathbf{UU}^*\mathbf{LV}\Sigma^{-1/2} = \Sigma$, and $\mathbf{T}_i^*\mathcal{Q}\mathbf{T}_i = \Sigma^{-1/2}\mathbf{W}^*\mathbf{U}^*\mathbf{LL}^*\mathbf{UW}\Sigma^{-1/2} = \Sigma$. Therefore, the system $\left(\begin{array}{c|c} \mathbf{TAT}_i & \mathbf{TB} \\ \hline \mathbf{CT}_i & \mathbf{D} \end{array}\right)$ is balanced. Furthermore, if we partition

$$\mathbf{W} = [\mathbf{W}_1 \ \ \mathbf{W}_2], \ \Sigma = \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix}, \ \mathbf{V} = [\mathbf{V}_1 \ \ \mathbf{V}_2],$$

where $\mathbf{W}_1$, $\mathbf{V}_1$ have k columns and $\Sigma_1 \in \mathbb{R}^{k\times k}$, the transformations

$$\boxed{\mathbf{T}_1 = \Sigma_1^{-1/2}\mathbf{V}_1^*\mathbf{L}^* \in \mathbb{R}^{k\times n}} \text{ and } \boxed{\mathbf{T}_{i1} = \mathbf{UW}_1\Sigma_1^{-1/2} \in \mathbb{R}^{n\times k}} \tag{7.16}$$

satisfy $\mathbf{T}_1\mathbf{T}_{i1} = \mathbf{I}_k$ and applied on $\boldsymbol{\Sigma}$ yield the reduced-order system of dimension k, obtained from $\boldsymbol{\Sigma}$ by *balanced truncation*:

$$\boldsymbol{\Sigma}_1 = \left(\begin{array}{c|c} \mathbf{T}_1\mathbf{AT}_{i1} & \mathbf{T}_1\mathbf{B} \\ \hline \mathbf{CT}_{i1} & \mathbf{D} \end{array}\right). \tag{7.17}$$

3. The third transformation is derived from the previous one by dropping the term $\Sigma^{-1/2}$. This yields a system that is balanced only *up to diagonal scaling*. The transformations are

$$\boxed{\mathbf{T}_3 = \mathbf{V}^*\mathbf{L}^* \in \mathbb{R}^{n\times n}} \text{ and } \boxed{\mathbf{T}_{3i} = \mathbf{UW} \in \mathbb{R}^{n\times n}}, \tag{7.18}$$

which again satisfy $\mathbf{T}_3\mathbf{T}_{3i} = \mathbf{T}_{3i}\mathbf{T}_3 = \Sigma$. Clearly, the system $\left(\begin{array}{c|c} \mathbf{A}_3 & \mathbf{B}_3 \\ \hline \mathbf{C}_3 & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{T}_3\mathbf{A}\mathbf{T}_{3i} & \mathbf{T}_3\mathbf{B} \\ \hline \mathbf{C}\mathbf{T}_{3i} & \mathbf{D} \end{array}\right)$ is still balanced but is no longer the same as Σ. However, if we scale by $\Sigma^{-1/2}$, that is,

$$\left(\begin{array}{c|c} \Sigma^{-1/2}\mathbf{A}_3\Sigma^{-1/2} & \Sigma^{-1/2}\mathbf{B}_3 \\ \hline \mathbf{C}_3\Sigma^{-1/2} & \mathbf{D} \end{array}\right),$$

it becomes balanced and is equal to Σ. Removing the factor $\Sigma^{-1/2}$ may in some cases make the transformation better conditioned.

4. *Balancing-free square root algorithm.* Our goal here is to transform the system by means of an orthogonal transformation in such a way that it is similar to a balanced system *up to a triangular transformation.* This method was introduced by Varga [340].

   Toward this goal we compute the QR factorizations of $\mathbf{UW}$ and $\mathbf{LV}$, as in (7.14), where $\mathbf{X}$ and $\mathbf{Y}$ are unitary matrices and Φ, Ψ are upper triangular. The transformations are

$$\boxed{\mathbf{T}_4 = (\mathbf{Y}^*\mathbf{X})^{-1}\mathbf{Y}^* = \mathbf{X}^*} \quad \text{and} \quad \boxed{\mathbf{T}_{4i} = \mathbf{X}}\,. \tag{7.19}$$

   Then $\left(\begin{array}{c|c} \mathbf{X}^*\mathbf{A}\mathbf{X} & \mathbf{X}^*\mathbf{B} \\ \hline \mathbf{C}\mathbf{X} & \mathbf{D} \end{array}\right)$ is balanced *up to an upper triangular matrix.* Therefore, the truncated system is similar to the system obtained by balanced truncation, up to an upper triangular matrix. This upper triangular transformation is $\mathbf{K} = \Sigma^{1/2}\Phi$. Clearly, $\mathbf{K}$ can be ill-conditioned due to $\Sigma^{1/2}$.

   Next we truncate the QR factorizations of $\mathbf{UW}$ and $\mathbf{LV}$:

$$\boxed{\mathbf{UW}_1 = \mathbf{X}_k\Phi_k,\ \mathbf{X}_k \in \mathbb{R}^{n\times k}} \quad \text{and} \quad \boxed{\mathbf{LV}_1 = \mathbf{Y}_k\Psi_k,\ \mathbf{Y}_k \in \mathbb{R}^{n\times k}}\,, \tag{7.20}$$

   where $\mathbf{X}_k^*\mathbf{X}_k = \mathbf{I}_k$ and $\mathbf{Y}_k^*\mathbf{Y}_k = \mathbf{I}_k$, while Φ_k, Ψ_k are the $k \times k$ leading submatrices of Φ, Ψ, respectively, and consequently are upper triangular matrices. We thus obtain the following transformations:

$$\boxed{\hat{\mathbf{T}}_1 = (\mathbf{Y}_k^*\mathbf{X}_k)^{-1}\mathbf{Y}_k^* \in \mathbb{R}^{k\times n}} \quad \text{and} \quad \boxed{\hat{\mathbf{T}}_{i1} = \mathbf{X}_k \in \mathbb{R}^{n\times k}}\,. \tag{7.21}$$

   Then the system

$$\hat{\Sigma}_1 = \left(\begin{array}{c|c} \hat{\mathbf{T}}_1\mathbf{A}\hat{\mathbf{T}}_{i1} & \hat{\mathbf{T}}_1\mathbf{B} \\ \hline \mathbf{C}\hat{\mathbf{T}}_{i1} & \mathbf{D} \end{array}\right)$$

   is similar to Σ_1. The similarity transformation is $\mathbf{K} = \Sigma^{1/2}\Phi^{-1}$:

$$\left(\begin{array}{c|c} \mathbf{K} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right)\left(\begin{array}{c|c} \hat{\mathbf{A}}_1 & \hat{\mathbf{B}}_1 \\ \hline \hat{\mathbf{C}}_1 & \mathbf{D} \end{array}\right)\left(\begin{array}{c|c} \mathbf{K}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D} \end{array}\right).$$

Remark 7.3.1. **(a)** Summarizing, there are four different ways to compute a balanced realization and consequently a reduced system obtained by balanced truncation. In general, the transformations have different condition numbers. In particular, $\mathbf{T}_1$ has the largest condition number, followed by $\mathbf{T}_2$. The transformations $\mathbf{T}_3$ and $\mathbf{T}_4$ have almost the same condition number, which is in general much lower than those of the first two transformations.

(b) As shown in section 6.3.3, the Cholesky factors **U** and **L** of the gramians can be computed *directly*, that is, *without* first computing the gramians $\mathcal{P}$, $\mathcal{Q}$ and subsequently computing their Cholesky decompositions. This method for solving the Lyapunov equation was first proposed by Hammarling [164]. See also the paper of Safonov and Chiang [283].

(c) The number of operations required to obtain a balanced realization is of the order $\mathcal{O}(n^3)$, where n is the dimension of the system under investigation, while the storage required is $\mathcal{O}(n^2)$ (the storage of $n \times n$ matrices is required).

Summary of balancing transformations

The table below summarizes the four balancing transformations discussed earlier. $\mathbf{T}_1$ is the original one, $\mathbf{T}_2$ the square root balancing transformation, $\mathbf{T}_3$ the square root transformation which balances up to scaling, and $\mathbf{T}_4$ yields balanced systems up to an upper triangular transformation. Their inverses are denoted by $\mathbf{T}_{ji}$, $j = 1, 2, 3, 4$. In the transformed bases, if we partition Σ as in (7.3) and **K**, **U**, **W**, **X**, **Y** conformally, we obtain the right projectors denoted by $\hat{\mathbf{T}}_j$ and the associated left projectors $\hat{\mathbf{T}}_{ji}$, $j = 1, 2, 3, 4$. These are the *balance and truncate* transformations; the second is the *square root balance and truncate*, the third is the up-to-diagonal scaling balance and truncate, and the last is the *square root balancing free* transformation.

Factorization	Properties	Transformation	(Left) Inverse
$\mathcal{P} = \mathbf{U}\mathbf{U}^*$, $\mathcal{Q} = \mathbf{L}\mathbf{L}^*$	**U** upper, **L** lower		
$\mathbf{U}^*\mathcal{Q}\mathbf{U} = \mathbf{K}\Sigma^2\mathbf{K}^*$	**K** unitary	$\mathbf{T}_1 = \Sigma^{1/2}\mathbf{K}^*\mathbf{U}^{-1}$	$\mathbf{T}_{1i} = \mathbf{U}\mathbf{K}\Sigma^{-1/2}$
		$\hat{\mathbf{T}}_1 = \Sigma_1^{1/2}\mathbf{K}_1^*\mathbf{U}^{-1}$	$\hat{\mathbf{T}}_{1i} = \mathbf{U}\mathbf{K}_1\Sigma_1^{-1/2}$
$\mathbf{U}^*\mathbf{L} = \mathbf{W}\Sigma\mathbf{V}^*$	**W**, **V** unitary	$\mathbf{T}_2 = \Sigma^{-1/2}\mathbf{V}^*\mathbf{L}^*$	$\mathbf{T}_{2i} = \mathbf{U}\mathbf{W}\Sigma^{-1/2}$
		$\hat{\mathbf{T}}_2 = \Sigma_1^{-1/2}\mathbf{V}_1^*\mathbf{L}^*$	$\hat{\mathbf{T}}_{2i} = \mathbf{U}\mathbf{W}_1\Sigma_1^{-1/2}$
		$\mathbf{T}_3 = \mathbf{V}^*\mathbf{L}^*$	$\mathbf{T}_{3i} = \mathbf{U}\mathbf{W}$
		$\hat{\mathbf{T}}_3 = \mathbf{V}_1^*\mathbf{L}^*$	$\hat{\mathbf{T}}_{3i} = \mathbf{U}\mathbf{W}_1$
$\mathbf{U}\mathbf{W} = \mathbf{X}\Phi$ $\mathbf{L}\mathbf{V} = \mathbf{Y}\Psi$	**X**, **Y** : unitary Φ, Ψ : upper	$\mathbf{T}_4 = (\mathbf{Y}^*\mathbf{X})^{-1}\mathbf{Y}^*$	$\mathbf{T}_{4i} = \mathbf{X}$
		$\hat{\mathbf{T}}_4 = (\mathbf{Y}_1^*\mathbf{X}_1)^{-1}\mathbf{Y}_1^*$	$\hat{\mathbf{T}}_{4i} = \mathbf{X}_1$

An example

We conclude this section with numerical experiments performed on continuous-time low-pass Butterworth filters; these can be obtained in MATLAB as follows:

$$[\texttt{A}, \texttt{B}, \texttt{C}, \texttt{D}] = \texttt{butter}(\texttt{n}, \texttt{1}, '\texttt{s}').$$

The order n of the filter varies, $n = 1 : 30$. The Hankel singular values for $n = 30$ have condition number close to 10^{19}, and the first six are nearly equal to 1; these singular values

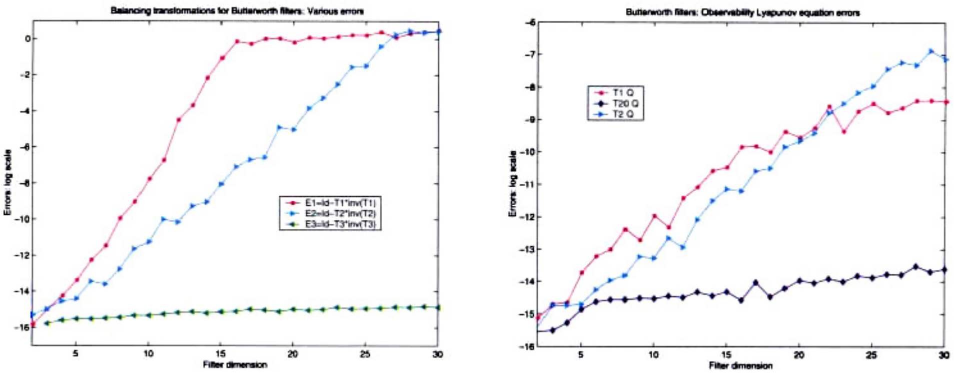


Figure 7.1. *Balancing transformations for Butterworth filters of order* 1–30. *Left: errors of three balancing transformations. Right: errors of three observability Lyapunov equations. The lower curve in both plots belongs to* $\mathbf{T}_2$.

were computed by means of the square root algorithm, that is, (7.13). Our first goal is to compare the first three balancing transformations $\mathbf{T}_j$, $j = 1, 2, 4$, and the corresponding balanced triples $(\mathbf{C}_j, \mathbf{A}_j, \mathbf{B}_j)$. The transformations are $\mathbf{T}_1$, the usual balancing transformation; $\mathbf{T}_2$, the square-root balancing transformation; and $\mathbf{T}_4$, the square root balancing free transformation. The errors that we look at are the 2-norms of the differences between the identity and the product of each transformation and its inverse, and the norm of the left-hand side of the observability Lyapunov equations for three balancing transformations:

$$\mathbf{E}_j^{(1)} = \text{norm}\,(\mathbf{I}_n - \mathbf{T}_j\mathbf{T}_{ji}), \qquad j = 1, 2, 4,$$

$$\mathbf{E}_j^{(2)} = \text{norm}\,(\mathbf{A}_j^*\Sigma + \Sigma\mathbf{A}_j + \mathbf{C}_j^*\mathbf{C}_j), \qquad j = 1, 2, 4.$$

All these quantities are theoretically zero. The actual values of the above errors in floating point arithmetic are plotted as a function of the order n of the filter. The first set is depicted in the left side of Figure 7.1, while the second is depicted in the right side of Figure 7.1. A similar example is explored in Problem 47.

7.4 A canonical form for continuous-time balanced systems

In what follows, we present a canonical form for systems that are continuous-time, stable, reachable, observable, and balanced. We will deal with only the SISO case. For the general case, see the original source [248] and references therein.

Let $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B}, \mathbf{C}^* \in \mathbb{R}^n$, be stable, reachable, observable, and balanced with Hankel singular values as given by (5.22):

$$\Sigma = \begin{pmatrix} \sigma_1\mathbf{I}_{m_1} & & & \\ & \sigma_2\mathbf{I}_{m_2} & & \\ & & \ddots & \\ & & & \sigma_q\mathbf{I}_{m_q} \end{pmatrix},$$

where $\sigma_j > \sigma_{j+1} > 0$ and m_j is the multiplicity of σ_j. Consequently, the following Lyapunov equations are satisfied:

$$\mathbf{A}\Sigma + \Sigma\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}, \tag{7.22}$$

$$\Sigma\mathbf{A} + \mathbf{A}^*\Sigma + \mathbf{C}^*\mathbf{C} = \mathbf{0}. \tag{7.23}$$

Before stating the result valid for the general case, we examine two special cases. The first is when the *multiplicity of all singular values is equal to one*, i.e., the singular values are distinct. Let the entries of $\mathbf{B}$ be denoted by γ_i:

$$\mathbf{B} = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix}.$$

It follows from (7.22) that the entries of $\mathbf{A}$ are:

$$a_{ij} = \frac{-\gamma_i\gamma_j}{s_is_j\sigma_i + \sigma_j}, \quad \text{where } s_i = \pm 1,\ i,\ j = 1, \ldots, n.$$

Moreover, from (7.23) it follows that

$$\mathbf{C} = (s_1\gamma_1 \ \cdots \ s_n\gamma_n).$$

In other words, the s_i are signs associated with the singular values σ_i. Finally, notice that due to the reachability of $(\mathbf{A}, \mathbf{B})$ and to the observability $(\mathbf{C}, \mathbf{A})$, the γ_i must be different from zero. Summarizing, from (7.22) and (7.23) the triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ has the following form:

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccc|c} \frac{-\gamma_1^2}{2\sigma_1} & \frac{-\gamma_1\gamma_2}{s_1s_2\sigma_1+\sigma_2} & \cdots & \frac{-\gamma_1\gamma_n}{s_1s_n\sigma_1+\sigma_n} & \gamma_1 \\ \frac{-\gamma_2\gamma_1}{s_2s_1\sigma_2+\sigma_1} & \frac{-\gamma_2^2}{2\sigma_2} & \cdots & \frac{-\gamma_2\gamma_n}{s_2s_n\sigma_2+\sigma_n} & \gamma_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{-\gamma_n\gamma_1}{s_ns_1\sigma_n+\sigma_1} & \frac{-\gamma_n\gamma_2}{s_ns_2\sigma_n+\sigma_2} & \cdots & \frac{-\gamma_n^2}{2\sigma_n} & \gamma_n \\ \hline s_1\gamma_1 & s_2\gamma_2 & \cdots & s_n\gamma_n & \end{array}\right). \tag{7.24}$$

The second special case is that of one singular value of multiplicity n, i.e., $m_1 = n$, $m_i = 0, i > 1$. For simplicity, we denote this single singular value σ. Following Corollary 7.4, the balanced representation is unique up to *orthogonal* transformation in the state space. It follows that one solution of (7.22) and (7.23) is

$$\mathbf{B}\mathbf{B}^* = \text{diag}\,(\gamma, 0, \ldots, 0) = \mathbf{C}^*\mathbf{C}, \qquad \gamma > 0,$$

$$\mathbf{A} = \frac{-1}{2\sigma}\text{diag}\,(\gamma, 0, \ldots, 0) + \mathbf{L},$$

where $\mathbf{L}$ is an arbitrary *skew symmetric* matrix. To further simplify this form, we are allowed to use orthogonal transformations $\mathbf{U}$ of the type $\mathbf{U} = \operatorname{diag}(1, \mathbf{U}_2)$, where $\mathbf{U}_2$ is an arbitrary orthogonal matrix of size $n-1$. The final form in this case is

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccccc|c} \frac{-\gamma^2}{2\sigma} & \alpha_1 & 0 & \cdots & 0 & 0 & \gamma \\ -\alpha_1 & 0 & \alpha_2 & & 0 & 0 & 0 \\ 0 & -\alpha_2 & 0 & & 0 & 0 & 0 \\ & \vdots & & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \alpha_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & -\alpha_{n-1} & 0 & 0 \\ \hline s\gamma & 0 & 0 & \cdots & 0 & 0 & \end{array}\right), \tag{7.25}$$

where $\sigma > 0$, $\gamma > 0$, $\alpha_i > 0$, $s = \pm 1$. As we will see later, these systems are precisely the *all-pass* systems for an appropriate choice of $\mathbf{D}$.

The *general case* is a combination of the above two special cases. Let there be q distinct singular values each of multiplicity m_i; the canonical form is

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right) = \left(\begin{array}{cccc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1q} & \mathbf{B}_1 \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2n} & \mathbf{B}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{A}_{q1} & \mathbf{A}_{q2} & \cdots & \mathbf{A}_{qq} & \mathbf{B}_q \\ \hline \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_q & \end{array}\right). \tag{7.26}$$

The following hold true:

$$\left(\begin{array}{c|c} \mathbf{A}_{ii} & \mathbf{B}_i \\ \hline \mathbf{C}_i & \end{array}\right) = \left(\begin{array}{cccccc|c} \frac{-\gamma_i^2}{2\sigma_i} & \alpha_{i,1} & 0 & \cdots & 0 & 0 & \gamma_i \\ -\alpha_{i,1} & 0 & \alpha_{i,2} & & 0 & 0 & 0 \\ 0 & -\alpha_{i,2} & 0 & & 0 & 0 & 0 \\ & \vdots & & \ddots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \alpha_{i,m_i-1} & 0 \\ 0 & 0 & 0 & \cdots & -\alpha_{i,m_i-1} & 0 & 0 \\ \hline s_i\gamma_i & 0 & 0 & \cdots & 0 & 0 & \end{array}\right), \tag{7.27}$$

$$\left(\begin{array}{c|c} \mathbf{A}_{ij} & \mathbf{B}_i \\ \hline \mathbf{C}_j & \end{array}\right) = \left(\begin{array}{cccc|c} \frac{-\gamma_i\gamma_j}{s_i s_j \sigma_i + \sigma_j} & 0 & \cdots & 0 & \gamma_i \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ \hline s_j\gamma_j & 0 & \cdots & 0 & 0 \end{array}\right). \tag{7.28}$$

The parameters that enter the forms described by (7.26), (7.27), and (7.28) are

$$\left.\begin{array}{l} \sigma_i > 0,\ m_i \geq 0,\ \gamma_i > 0,\ s_i = \pm 1,\ i = 1, \ldots, q \\ \sum_{i=1}^{q} m_i = n \\ \alpha_{ij} > 0,\ i = 1, \ldots, q,\ j = 1, \ldots, m_i \end{array}\right\}. \tag{7.29}$$

We are now ready to state the main result of this section.

Theorem 7.15. *There is a one-to-one correspondence between the family of stable, reachable, observable, SISO, continuous-time systems with Hankel singular values σ_i of multiplicity m_i and the family of systems parametrized by* (7.26)–(7.29).

Remark 7.4.1. **(a)** It is interesting to notice that a class of stable and *minimal* systems is parametrized in terms of *positive* parameters $\sigma_i, m_i, \gamma_i, \alpha_{ij}$ and the sign parameters s_i. Recall that the parametrization of minimal (let alone stable) systems in terms of other canonical forms is complicated.

(b) Using the parametrization of balanced systems derived above, one can obtain an alternative proof of part 2 of Theorem 7.9. It thus becomes clear under what conditions the truncated system is minimal, even if the condition given in part 2 of the theorem is *not* satisfied.

Example 7.16. A third-order discrete-time FIR (finite impulse response) system described by the transfer function

$$\mathbf{H}(z) = \frac{z^2+1}{z^3}$$

is considered. We investigate approximation by balanced truncation, first in discrete time and then in continuous time, using the bilinear transformation of section 4.3.3, and compare the results. Hankel-norm approximation for the same system is investigated in Example 8.9.

A balanced realization of this system is given by

$$\boldsymbol{\Sigma}_d = \left(\begin{array}{c|c} \mathbf{F} & \mathbf{G} \\ \hline \mathbf{H} & \mathbf{J} \end{array}\right) = \left(\begin{array}{ccc|c} 0 & \alpha & 0 & \beta \\ \alpha & 0 & -\alpha & 0 \\ 0 & \alpha & 0 & -\gamma \\ \hline \beta & 0 & \gamma & 0 \end{array}\right),$$

where $\alpha = 5^{-\frac{1}{4}}$, $\beta = \frac{\sqrt{3\sqrt{5}+5}}{\sqrt{10}}$, $\gamma = \frac{\sqrt{3\sqrt{5}-5}}{\sqrt{10}}$, and the reachability and observability gramians are equal and diagonal:

$$\mathcal{P} = \mathcal{Q} = \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \sigma_3), \ \ \sigma_1 = \frac{\sqrt{5}+1}{2}, \ \sigma_2 = 1, \ \sigma_3 = \frac{\sqrt{5}-1}{2}.$$

The second- and first-order balanced truncated systems are

$$\boldsymbol{\Sigma}_{d,2} = \left(\begin{array}{c|c} \mathbf{F}_2 & \mathbf{G}_2 \\ \hline \mathbf{H}_2 & \mathbf{J}_2 \end{array}\right) = \left(\begin{array}{cc|c} 0 & \alpha & \beta \\ \alpha & 0 & 0 \\ \hline -\beta & 0 & 0 \end{array}\right), \ \ \boldsymbol{\Sigma}_{d,1} = \left(\begin{array}{c|c} \mathbf{F}_1 & \mathbf{G}_1 \\ \hline \mathbf{H}_1 & \mathbf{J}_1 \end{array}\right) = \left(\begin{array}{c|c} 0 & \beta \\ \hline -\beta & 0 \end{array}\right).$$

Notice that $\boldsymbol{\Sigma}_{d,2}$ is *balanced* but has singular values that are *different* from σ_1, σ_2. $\boldsymbol{\Sigma}_{d,1}$ is also balanced since $\mathbf{G}_1 = -\mathbf{H}_2$, but its gramians are not equal to σ_1.

Let $\mathbf{\Sigma}_c$ denote the continuous-time system obtained from $\mathbf{\Sigma}_d$ by means of the bilinear transformation described in section 4.3.3:

$$\mathbf{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left(\begin{array}{ccc|c} -1-2\alpha^2 & 2\alpha & 2\alpha^2 & \delta_+ \\ 2\alpha & -1 & -2\alpha & -\sqrt{2} \\ -2\alpha^2 & -2\alpha & -1+2\alpha^2 & \delta_- \\ \hline -\delta_+ & \sqrt{2} & \delta_- & 2 \end{array}\right),$$

where $\delta_\pm = \sqrt{2}(\alpha \pm \beta)$. Notice that $\mathbf{\Sigma}_c$ is balanced. We now compute first and second reduced order systems $\mathbf{\Sigma}_{c,1}$, $\mathbf{\Sigma}_{c,2}$ by truncating $\mathbf{\Sigma}_c$:

$$\mathbf{\Sigma}_{c,2} = \left(\begin{array}{c|c} \mathbf{A}_2 & \mathbf{B}_2 \\ \hline \mathbf{C}_2 & \mathbf{D}_2 \end{array}\right) = \left(\begin{array}{cc|c} -1-2\alpha^2 & 2\alpha & \delta_+ \\ 2\alpha & -1 & -\sqrt{2} \\ \hline -\delta_+ & \sqrt{2} & 2 \end{array}\right),$$

$$\mathbf{\Sigma}_{c,1} = \left(\begin{array}{c|c} \mathbf{A}_1 & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D}_1 \end{array}\right) = \left(\begin{array}{c|c} -1-2\alpha^2 & \delta_+ \\ \hline -\delta_+ & 2 \end{array}\right).$$

Let $\bar{\mathbf{\Sigma}}_{d,2}$, $\bar{\mathbf{\Sigma}}_{d,1}$ be the discrete-time systems obtained by transforming $\mathbf{\Sigma}_{c,2}$, $\mathbf{\Sigma}_{c,1}$ back to discrete time:

$$\bar{\mathbf{\Sigma}}_{d,2} = \left(\begin{array}{c|c} \bar{\mathbf{F}}_2 & \bar{\mathbf{G}}_2 \\ \hline \bar{\mathbf{H}}_2 & \bar{\mathbf{J}}_2 \end{array}\right) = \left(\begin{array}{cc|c} 0 & \alpha & \beta \\ \alpha & \alpha^2 & -\alpha\gamma \\ \hline -\beta & \alpha\gamma & -\gamma^2 \end{array}\right),$$

$$\bar{\mathbf{\Sigma}}_{d,1} = \left(\begin{array}{c|c} \bar{\mathbf{F}}_1 & \bar{\mathbf{G}}_1 \\ \hline \bar{\mathbf{H}}_1 & \bar{\mathbf{J}}_1 \end{array}\right) = \left(\begin{array}{c|c} -\sigma_3/2 & (\alpha+\beta)\sigma_3/2\alpha^2 \\ \hline -(\alpha+\beta)\sigma_3/2\alpha^2 & 2-(\alpha^2+\beta)^2/(1+\alpha^2) \end{array}\right).$$

The conclusion is that $\bar{\mathbf{\Sigma}}_{d,2}$, $\bar{\mathbf{\Sigma}}_{d,1}$ are balanced and different from $\mathbf{\Sigma}_{d,2}$, $\mathbf{\Sigma}_{d,1}$. It is interesting to notice that the singular value of $\mathbf{\Sigma}_{d,1}$ is $\beta^2 = \frac{1+\sigma_1}{\sqrt{5}}$, while that of $\bar{\mathbf{\Sigma}}_{d,1}$ is σ_1; β^2 satisfies $\sigma_2 < \beta^2 < \sigma_1$. Furthermore, the singular values of $\mathbf{\Sigma}_{d,2}$ are $5\beta^2/4$, $\sqrt{5}\beta^2/4$, which satisfy the following interlacing inequalities:

$$\sigma_3 < \sqrt{5}\beta^2/4 < \sigma_2 < 5\beta^2/4 < \sigma_1.$$

The numerical values of the quantities above are $\sigma_1 = 1.618034$, $\sigma_3 = 0.618034$, $\alpha = 0.66874$, $\beta = 1.08204$, $\gamma = 0.413304$, $\delta_+ = 2.47598$, $\delta_- = .361241$.

7.5 Other types of balancing*

In general terms, balancing consists of the *simultaneous diagonalization* of two appropriately chosen positive definite matrices. This problem has been studied in linear algebra; see, e.g., [79]. Given a reachable and observable system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, these positive definite matrices are solutions either of Lyapunov equations or Riccati equations. Here is a list of four types of balancing:

Type	Equations	Error
Lyapunov	$\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^*=\mathbf{0}$, $\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}+\mathbf{C}^*\mathbf{C}=\mathbf{0}$	$\Vert\mathbf{H}-\mathbf{H}_k\Vert\leq 2\sum_{i=k+1}^{n}\sigma_i$
Stochastic	$\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^*=\mathbf{0}$, $\hat{\mathbf{B}}=\mathcal{P}\mathbf{C}^*+\mathbf{B}\mathbf{D}^*$, $\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}+(\mathbf{C}-\hat{\mathbf{B}}\mathcal{Q})^*(\mathbf{D}\mathbf{D}^*)^{-1}(\mathbf{C}-\hat{\mathbf{B}}\mathcal{Q})=\mathbf{0}$	$\Vert\mathbf{H}^{-1}(\mathbf{H}-\mathbf{H}_r)\Vert$ $\leq 2\prod_{i=k+1}^{n}\frac{1+\sigma_i}{1-\sigma_i}-1$
BR	$\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^*$ $+(\mathcal{P}\mathbf{C}^*+\mathbf{B}\mathbf{D}^*)(\mathbf{I}-\mathbf{D}\mathbf{D}^*)^{-1}(\mathcal{P}\mathbf{C}^*+\mathbf{B}\mathbf{D}^*)^*=\mathbf{0}$, $\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}+\mathbf{C}^*\mathbf{C}$ $+(\mathcal{Q}\mathbf{B}+\mathbf{C}^*\mathbf{D})(\mathbf{I}-\mathbf{D}^*\mathbf{D})^{-1}(\mathcal{Q}\mathbf{B}+\mathbf{C}^*\mathbf{D})^*=\mathbf{0}$	$\Vert\mathbf{H}-\mathbf{H}_k\Vert\leq 2\sum_{i=k+1}^{n}\sigma_i$
PR	$\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*$ $+(\mathcal{P}\mathbf{C}^*-\mathbf{B})(\mathbf{D}+\mathbf{D}^*)^{-1}(\mathcal{P}\mathbf{C}^*-\mathbf{B})^*=\mathbf{0}$, $\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}$ $+(\mathcal{Q}\mathbf{B}-\mathbf{C}^*)(\mathbf{D}^*+\mathbf{D})^{-1}(\mathcal{Q}\mathbf{B}-\mathbf{C}^*)^*=\mathbf{0}$	$\Vert(\mathbf{H}+\mathbf{D}^*)^{-1}-(\mathbf{H}_k+\mathbf{D}^*)^{-1}\Vert$ $\leq 2\,\Vert(\mathbf{D}+\mathbf{D}^*)^{-1}\Vert\sum_{i=k+1}^{n}\sigma_i$

In the sequel these will be briefly explored.

7.5.1 Lyapunov balancing*

This is the method that was discussed in detail in the preceding sections. In this case, the solution of the Lyapunov equations $\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^*=\mathbf{0}$ and $\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}+\mathbf{C}^*\mathbf{C}=\mathbf{0}$ are simultaneously diagonalized. Reduced models, obtained by simple truncation as in (7.4), preserve stability and satisfy an $\mathcal{H}_\infty$-error bound (see Theorems 7.9 and 7.10).

7.5.2 Stochastic balancing*

The starting point is a system $\mathbf{\Sigma}=\left(\begin{array}{c|c}\mathbf{A}&\mathbf{B}\\\hline\mathbf{C}&\mathbf{D}\end{array}\right)$, which satisfies the following properties: (i) it is square, i.e., $m=p$; (ii) it is stable (all eigenvalues of $\mathbf{A}$ are in the left half of the complex plane); and (iii) $\mathbf{D}$ is nonsingular. If $\mathbf{H}(s)=\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$ is the transfer function of $\mathbf{\Sigma}$, we denote by $\mathbf{\Phi}$ its *power spectrum* and by $\mathbf{W}(s)$ a *minimum phase* right spectral factor; furthermore, the *phase system* whose transfer function is $\mathbf{Z}(s)$ is introduced. These quantities satisfy

$$\mathbf{\Phi}(s)=\mathbf{H}(s)\mathbf{H}^*(-s)=\mathbf{W}^*(-s)\mathbf{W}(s)=\mathbf{Z}(s)+\mathbf{Z}^*(-s).$$

Let $\mathcal{P}$ be the usual reachability gramian of $\mathbf{\Sigma}$ satisfying $\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^*=\mathbf{0}$; define $\bar{\mathbf{B}}=\mathcal{P}\mathbf{C}^*+\mathbf{B}\mathbf{D}^*$, and let $\mathcal{Q}$ be a stabilizing solution of the following Riccati equation:

$$\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}+(\mathbf{C}-\bar{\mathbf{B}}^*\mathcal{Q})^*\left[\mathbf{D}\mathbf{D}^*\right]^{-1}(\mathbf{C}-\bar{\mathbf{B}}^*\mathcal{Q})=\mathbf{0}.$$

It can be shown that a realization of $\mathbf{W}$ and $\mathbf{Z}$ is given in terms of the two gramians $\mathcal{P}$ and $\mathcal{Q}$ as follows:

$$\mathbf{\Sigma}_{\mathbf{W}}=\left(\begin{array}{c|c}\bar{\mathbf{A}}&\bar{\mathbf{B}}\\\hline\bar{\mathbf{C}}&\bar{\mathbf{D}}\end{array}\right)=\left(\begin{array}{c|c}\mathbf{A}&\mathcal{P}\mathbf{C}^*+\mathbf{B}\mathbf{D}^*\\\hline\mathbf{D}^{-1}(\mathbf{C}-\bar{\mathbf{B}}\mathcal{Q})&\mathbf{D}^*\end{array}\right),$$

$$\mathbf{\Sigma}_{\mathbf{Z}}=\left(\begin{array}{c|c}\hat{\mathbf{A}}&\hat{\mathbf{B}}\\\hline\hat{\mathbf{C}}&\hat{\mathbf{D}}\end{array}\right)=\left(\begin{array}{c|c}\mathbf{A}&\mathcal{P}\mathbf{C}^*+\mathbf{B}\mathbf{D}^*\\\hline\mathbf{C}&\mathbf{D}\mathbf{D}^*/2\end{array}\right).$$

By construction, the eigenvalues of the product $\mathcal{PQ}$ are the squares of the Hankel singular values σ_i of the following auxiliary system:

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathcal{P}\mathbf{C}^* + \mathbf{BD}^* & \mathbf{0} \end{array}\right).$$

This method proceeds by computing a balancing transformation $\mathbf{T}$ for this auxiliary system, which is then applied to the original system $\boldsymbol{\Sigma}$; reduction by truncation follows as usual. If we let the neglected distinct singular values be σ_i, the following relative error bound holds:

$$\sigma_{k+1} \leq \left\|\mathbf{H}(s)^{-1}\left[\mathbf{H}(s) - \mathbf{H}_k(s)\right]\right\|_{\mathcal{H}_\infty} \leq \prod_{i=k+1}^{n} \frac{1+\sigma_i}{1-\sigma_i} - 1,$$

where only the distinct singular values enter the above formula. Furthermore, the $\mathcal{H}_\infty$-norm of the relative error $\mathbf{H}_k(s)^{-1}\left[\mathbf{H}(s) - \mathbf{H}_k(s)\right]$ satisfies this same error bound. It is worth noting that $\sigma_i \leq 1$ and the number of σ_i which are equal to one is equal to the number of right half plane (i.e., unstable) zeros of $\boldsymbol{\Sigma}$.

Remark 7.5.1. *The multiplicative error bound.* It is pointed out in Chapter 2 of [252] that the multiplicative error measure above is focusing on Bode diagram errors (errors in log-magnitudes and phases). The following considerations are relevant in this regard. In the SISO case, let $\Delta = \frac{\mathbf{H}-\mathbf{H}_k}{\mathbf{H}}$ be the relative error; then $\frac{\mathbf{H}_k}{\mathbf{H}} = 1 - \Delta$, which in turn implies

$$\ln\left(\frac{\mathbf{H}_k}{\mathbf{H}}\right) = \ln(1-\Delta) = \Delta + \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 + \cdots .$$

By neglecting higher-order terms, we have $\ln(\mathbf{H}_k/\mathbf{H}) \approx \Delta$; from $\ln(\mathbf{H}_k/\mathbf{H}) = \ln|\mathbf{H}_k/\mathbf{H}| + i\arg(\mathbf{H}_k/\mathbf{H})$, it follows that

$$\mid \ln \mid \frac{\mathbf{H}_k}{\mathbf{H}} \mid \mid \ \leq \ \mid \Delta \mid \ \text{ and } \ \mid \arg \frac{\mathbf{H}_k}{\mathbf{H}} \mid \ \leq \ \mid \Delta \mid .$$

Thus transforming these relationships to $\log_{10}$, we get

$$20\log_{10} \mid \frac{\mathbf{H}_k}{\mathbf{H}} \mid = 20\log_{10} e \ \ln \mid \frac{\mathbf{H}_k}{\mathbf{H}} \mid \leq 20\log_{10} e \mid \Delta \mid = 8.69 \mid \Delta \mid \text{ dB},$$

and similarly

$$\mid \arg(\mathbf{H}) - \arg(\mathbf{H}_k) \mid \leq \frac{1}{2} \mid \Delta \mid \text{ radians.}$$

Thus assuming that Δ is small, the Bode plot (both amplitude and phase) of the relative approximant $\mathbf{H}_k/\mathbf{H}$ is bounded up to a constant by the product given above.

It is of interest to discuss the special case when $\boldsymbol{\Sigma}$ is minimum phase. The first consequence is that with $\mathcal{Q}$ the usual observability gramian of $\boldsymbol{\Sigma}$, the difference $\mathcal{R} = \mathcal{Q}^{-1} - \mathcal{P}$ is nonsingular, and it satisfies the Lyapunov equation

$$(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})\mathcal{R} + \mathcal{R}(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C})^* + (\mathbf{D}^{-1}\mathbf{C})^*(\mathbf{D}^{-1}\mathbf{C}) = \mathbf{0}.$$

Thus $\mathcal{R}$ is the observability gramian of the inverse system and no Riccati equation needs to be solved. As a consequence of these facts, the stochastic singular values ν_i are related to the Hankel singular values σ_i of $\mathbf{\Sigma}$:

$$\sigma_i = \frac{\nu_i}{\sqrt{1+\nu_i^2}}.$$

Since $\nu_i < 1$, we have $\sigma_i < 2^{-\frac{1}{2}}$. Thus in this case, the minimum phase property is preserved by balanced stochastic truncation. Stochastic balancing was introduced by Desai and Pal [95]; the relative error bound is due to Green [147, 148]. For a recent overview, see the book by Obinata and Anderson [252], and for numerical algorithms, see [54].

Stochastic balanced truncation can be applied to all asymptotically stable dynamical systems that are square and nonsingular. For application to singular systems, see [342] and [140]. In the former reference, it is stated that stochastic balanced truncation yields a uniformly good approximant over the whole frequency range rather than yielding small absolute errors. Also, Zhou [369] showed that for minimal phase systems, stochastic balanced truncation is the same as the self-weighted balanced truncation, where only input weighting exists and is given by $\mathbf{H}^{-1}$. This issue will be discussed in section 7.6.1 in more detail.

7.5.3 Bounded real balancing*

The asymptotically stable system $\mathbf{\Sigma}$ is bounded real if its $\mathcal{H}_\infty$-norm is no greater than one, that is, its transfer function $\mathbf{H}$ satisfies $\mathbf{I} - \mathbf{H}^*(-i\omega)\mathbf{H}(i\omega) \geq 0$, $\omega \in \mathbb{R}$. It is called strictly bounded real if this inequality is strict. For simplicity, we will be concerned with strictly bounded real systems only.

Bounded real systems were discussed in detail in section 5.9.2. Recall in particular the bounded real lemma (5.58), which implies that (strict) bounded realness is equivalent to the existence of a positive definite solution $\mathcal{Y} = \mathcal{Y}^* > 0$ to the Riccati equation,

$$\mathbf{A}^*\mathcal{Y} + \mathcal{Y}\mathbf{A} + \mathbf{C}^*\mathbf{C} + (\mathcal{Y}\mathbf{B} + \mathbf{C}^*\mathbf{D})(\mathbf{I} - \mathbf{D}^*\mathbf{D})^{-1}(\mathcal{Y}\mathbf{B} + \mathbf{C}^*\mathbf{D})^* = \mathbf{0}. \tag{7.30}$$

Any solution $\mathcal{Y}$ lies between two extremal solutions: $\mathbf{0} < \mathcal{Y}_{\min} \leq \mathcal{Y} \leq \mathcal{Y}_{\max}$. $\mathcal{Y}_{\min}$ is the unique solution to (7.30) such that $\mathbf{A} + \mathbf{B}(\mathbf{I} - \mathbf{D}^*\mathbf{D})^{-1}(\mathbf{B}^*\mathcal{Y}_{\min} + \mathbf{D}^*\mathbf{C})$ is asymptotically stable. Furthermore, bounded realness is also equivalent to the existence of a positive definite solution $\mathcal{Z}$ of the dual Riccati equation

$$\mathbf{A}\mathcal{Z} + \mathcal{Z}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* + (\mathcal{Z}\mathbf{C}^* + \mathbf{B}\mathbf{D}^*)(\mathbf{I} - \mathbf{D}\mathbf{D}^*)^{-1}(\mathcal{Z}\mathbf{C}^* + \mathbf{B}\mathbf{D}^*)^* = \mathbf{0}. \tag{7.31}$$

Also in this case, any solution lies between two extremal solutions: $\mathbf{0} < \mathcal{Z}_{\min} \leq \mathcal{Z} \leq \mathcal{Z}_{\max}$. (7.30) and (7.31) are the ***bounded real Riccati equations*** of the system $\mathbf{\Sigma}$. It follows that $\mathcal{Y} = \mathcal{Y}^* > 0$ is a solution to (7.30) if $\mathcal{Z} = \mathcal{Y}^{-1}$ is a solution to (7.31), and conversely. Hence $\mathcal{Z}_{\min} = \mathcal{Y}_{\max}^{-1}$ and $\mathcal{Z}_{\max} = \mathcal{Y}_{\min}^{-1}$.

Bounded real balancing is obtained by simultaneous diagonalization of $\mathcal{Y}_{\min}$ and $\mathcal{Y}_{\max}^{-1}$ or, equivalently, $\mathcal{Y}_{\min}$ and $\mathcal{Z}_{\min}$. The realization of a bounded real system $\mathbf{\Sigma}$ is called ***bounded real balanced*** if

$$\mathcal{Y}_{\min} = \mathcal{Z}_{\min} = \mathcal{Y}_{\max}^{-1} = \mathcal{Z}_{\max}^{-1} = \operatorname{diag}(\xi_1\mathbf{I}_{l_1}, \ldots, \xi_q\mathbf{I}_{l_q}),$$

where $1 \geq \xi_1 > \xi_2 > \cdots > \xi_q > 0$, m_i, $i = 1, \ldots, q$, are the multiplicities of ξ_i, and $m_1 + \cdots + m_q = n$. We will call ξ_i the *bounded real singular values* of Σ.

Truncation now follows as usual after transformation to the balanced basis, by eliminating the states that correspond to small singular values. Let the reduced-order model $\mathbf{\Sigma}_r = \left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D} \end{array}\right)$ be obtained by bounded real balanced truncation. Also let $\mathbf{W}(s)$ and $\mathbf{V}(s)$ be stable and minimum phase spectral factors, that is, $\mathbf{W}^*(s)\mathbf{W}(s) = \mathbf{I} - \mathbf{G}^*(s)\mathbf{G}(s)$ and $\mathbf{V}(s)\mathbf{V}^*(s) = \mathbf{I} - \mathbf{G}(s)\mathbf{G}^*(s)$. Define $\mathbf{W}_r(s)$ and $\mathbf{V}_r(s)$ similarly for $\mathbf{\Sigma}_r$. Then $\mathbf{\Sigma}_r$ is asymptotically stable, minimal, and bounded real balanced and satisfies

$$\max\left\{ \left\| \begin{pmatrix} \mathbf{H} - \mathbf{H}_r \\ \mathbf{W} - \mathbf{W}_r \end{pmatrix} \right\|_{\mathcal{H}_\infty}, \left\| \begin{pmatrix} \mathbf{H} - \mathbf{H}_r \\ \mathbf{V} - \mathbf{V}_r \end{pmatrix} \right\|_{\mathcal{H}_\infty} \right\} \leq 2 \sum_{i=k+1}^{q} \xi_i. \tag{7.32}$$

Thus if $2\sum_{i=k+1}^{q} \xi_i$ is small, not only are $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_r$ close, but also the reduced spectral factors $\mathbf{W}_r$ and $\mathbf{V}_r$ are guaranteed to be close to the full-order spectral factors $\mathbf{W}$ and $\mathbf{V}$, respectively. Bounded real balancing and the above error bounds are due to Opdenacker and Jonckheere [251].

There is also a canonical form associated with bounded real balancing. In the SISO case, and assuming the corresponding singular values are distinct $\xi_i \neq \xi_j$, $i \neq j$, we have

$$\mathbf{A}_{ij} = \frac{-\gamma_i\gamma_j}{1 - d^2}\left[\frac{1 + s_i s_j \xi_i \xi_j}{s_i s_j \xi_i + \xi_j} + \xi_j d\right], \quad \mathbf{B}_i = \gamma_i, \quad \mathbf{C}_j = s_j\gamma_j, \quad \mathbf{D} = d,$$

where $\gamma_i > 0$ and $s_i = \pm 1$. For details on the general case of this canonical form, see [249].

7.5.4 Positive real balancing*

An important class of linear dynamical systems is that whose transfer function is positive real. The definition and properties of such systems are discussed in section 5.9.1. Furthermore, Corollary 5.32 states that a necessary and sufficient condition for a system $\mathbf{\Sigma}$ to have this property is that the Riccati equation (**PRARE**) have a positive (semi) definite solution:

$$\mathbf{A}^*\mathcal{K} + \mathcal{K}\mathbf{A} + (\mathcal{K}\mathbf{B} - \mathbf{C}^*)(\mathbf{D} + \mathbf{D}^*)^{-1}(\mathcal{K}\mathbf{B} - \mathbf{C}^*)^* = \mathbf{0}. \tag{PRARE}$$

The dual Riccati equation in this case is

$$\mathbf{A}\mathcal{L} + \mathcal{L}\mathbf{A}^* + (\mathcal{L}\mathbf{C}^* - \mathbf{B})(\mathbf{D} + \mathbf{D}^*)^{-1}(\mathcal{L}\mathbf{C}^* - \mathbf{B})^* = \mathbf{0}.$$

These are the *positive real Riccati equations* of the passive system $\mathbf{\Sigma}$. The solutions $\mathcal{K}$ and $\mathcal{L}$ of these equations lie between two extremal solutions, i.e., $\mathbf{0} < \mathcal{K}_{\min} \leq \mathcal{K} \leq \mathcal{K}_{\max}$ and $\mathbf{0} < \mathcal{L}_{\min} \leq \mathcal{L} \leq \mathcal{L}_{\max}$. If $\mathcal{K} = \mathcal{K}^*$ is a solution of the former, $\mathcal{L} = \mathcal{K}^{-1}$ is a solution of the latter; hence $\mathcal{K}_{\min} = \mathcal{L}_{\max}^{-1}$ and $\mathcal{K}_{\max} = \mathcal{L}_{\min}^{-1}$. A balancing transformation is obtained by *simultaneous diagonalization* of the minimal solutions $\mathcal{K}_{\min}$, $\mathcal{L}_{\min}$:

$$\mathcal{K}_{\min} = \mathcal{L}_{\min} = \mathcal{K}_{\max}^{-1} = \mathcal{L}_{\max}^{-1} = \operatorname{diag}(\pi_1 \mathbf{I}_{s_1}, \ldots, \pi_q \mathbf{I}_{s_q}) = \Pi,$$

where $1 \geq \pi_1 > \pi_2 > \cdots > \pi_q > 0$, m_i, $i = 1, \ldots, q$, and $m_1 + \cdots + m_q = n$, are the multiplicities of π_i, which are termed the *positive real singular values* of $\mathbf{\Sigma}$. There is a canonical form for the positive real balanced system, which in the SISO case is

$$\mathbf{A}_{ij} = \frac{-\gamma_i\gamma_j(1-s_i\pi_i)(1-s_j\pi_j)}{2d(s_is_j\pi_i+\pi_j)}, \quad \mathbf{B}_i = \gamma_i > 0, \quad \mathbf{C}_j = s_j\gamma_j, \quad \mathbf{D} = d > 0,$$

where $s_i = \pm 1$, $i = 1, \ldots, n$; for details see [249].

The reduced-order model $\mathbf{\Sigma}_r$ obtained by positive real balanced truncation is asymptotically stable, minimal, and positive real balanced. It also satisfies an error bound [159].

Proposition 7.17. *The reduced-order model $\mathbf{\Sigma}_r$, obtained by the positive real balanced truncation, satisfies*

$$\|(\mathbf{D}^* + \mathbf{H}(s))^{-1} - (\mathbf{D}^* + \mathbf{H}_r(s))^{-1}\|_{\mathcal{H}_\infty} \leq 2\|\mathbf{R}\|^2 \sum_{i=k+1}^{q} \pi_i,$$

where $\mathbf{R}^2 = (\mathbf{D}+\mathbf{D}^*)^{-1}$.

Proof. We can assume that $\mathbf{\Sigma}$ is in the positive real balanced basis. Hence the following two Riccati equations result: $\mathbf{A}\Pi + \Pi\mathbf{A}^* + (\Pi\mathbf{C}^* - \mathbf{B})(\mathbf{D}+\mathbf{D}^*)^{-1}(\Pi\mathbf{C}^* - \mathbf{B})^* = \mathbf{0}$ and $\mathbf{A}^*\Pi + \Pi\mathbf{A} + (\Pi\mathbf{B} - \mathbf{C}^*)(\mathbf{D}+\mathbf{D}^*)^{-1}(\Pi\mathbf{B} - \mathbf{C}^*)^* = \mathbf{0}$. These can be written as

$$\underbrace{(\mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C})}_{=\hat{\mathbf{A}}}\Pi + \Pi(\mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C})^* + \Pi\mathbf{C}^*\mathbf{R}\underbrace{\mathbf{R}\mathbf{C}}_{=\hat{\mathbf{C}}}\Pi + \underbrace{\mathbf{B}\mathbf{R}}_{=\hat{\mathbf{B}}}\mathbf{R}\mathbf{B}^* = \mathbf{0},$$

$$(\mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C})^*\Pi + \Pi(\mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C}) + \Pi\mathbf{B}\mathbf{R}^2\mathbf{B}^*\Pi + \mathbf{C}^*\mathbf{R}^2\mathbf{C} = \mathbf{0}.$$

It follows that the system $\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \mathbf{0} \end{array}\right)$ is bounded real balanced with the bounded real gramian Π. Let

$$\hat{\mathbf{A}} = \begin{pmatrix} \hat{\mathbf{A}}_{11} & \hat{\mathbf{A}}_{12} \\ \hat{\mathbf{A}}_{21} & \hat{\mathbf{A}}_{22} \end{pmatrix}, \quad \hat{\mathbf{B}} = \begin{pmatrix} \hat{\mathbf{B}}_1 \\ \hat{\mathbf{B}}_2 \end{pmatrix}, \quad \hat{\mathbf{C}} = \begin{pmatrix} \hat{\mathbf{C}}_1 & \hat{\mathbf{C}}_2 \end{pmatrix}, \quad \Pi = \begin{pmatrix} \Pi_1 & \\ & \Pi_2 \end{pmatrix},$$

where $\Pi_1 = \operatorname{diag}(\pi_1\mathbf{I}_{m_1}, \ldots, \pi_k\mathbf{I}_{m_k})$, $\Pi_2 = \operatorname{diag}(\pi_{k+1}\mathbf{I}_{m_{k+1}}, \ldots, \pi_q\mathbf{I}_{m_q})$, and define the bounded real reduced system $\hat{\mathbf{\Sigma}}_r = \left(\begin{array}{c|c} \hat{\mathbf{A}}_{11} & \hat{\mathbf{B}}_1 \\ \hline \hat{\mathbf{C}}_1 & \mathbf{0} \end{array}\right)$. From (7.32) we conclude that $\|\hat{\mathbf{H}}(s) - \hat{\mathbf{H}}_r(s)\|_{\mathcal{H}_\infty} \leq 2\sum_{i=k+1}^{q}\pi_i$. Since $\|\,\mathbf{R}\,(\hat{\mathbf{H}}(s) - \hat{\mathbf{H}}_r(s)\,(-\mathbf{R})\,\|_{\mathcal{H}_\infty} \leq \|\mathbf{R}\|^2\|\hat{\mathbf{H}}(s) - \hat{\mathbf{H}}_r(s)\|_{\mathcal{H}_\infty}$, this leads to

$$\|\mathbf{R}(\hat{\mathbf{H}}(s) - \hat{\mathbf{H}}_r(s))(-\mathbf{R})\|_{\mathcal{H}_\infty} = \|\underbrace{\mathbf{R}\hat{\mathbf{H}}(s)(-\mathbf{R})}_{=\Theta(s)} - \underbrace{\mathbf{R}\hat{\mathbf{H}}_r(s)(-\mathbf{R})}_{=\Theta_r(s)}\|_{\mathcal{H}_\infty} \leq 2\|\mathbf{R}\|^2 \sum_{i=k+1}^{q} \pi_i.$$

A realization for $\Theta(s)$ and $\Theta_r(s)$ can be obtained as

$$\Theta(s) = \left(\begin{array}{c|c} \mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C} & -\mathbf{B}\mathbf{R}^2 \\ \hline \mathbf{R}^2\mathbf{C} & \mathbf{0} \end{array}\right) \text{ and } \Theta_r(s) = \left(\begin{array}{c|c} \mathbf{A}_{11} - \mathbf{B}_1\mathbf{R}^2\mathbf{C}_1 & -\mathbf{B}_1\mathbf{R}^2 \\ \hline \mathbf{R}^2\mathbf{C}_1 & \mathbf{0} \end{array}\right).$$

Thus $\|\Theta(s) - \Theta_r(s)\|_{\mathcal{H}_\infty} = \|(\Theta(s) + \mathbf{R}^2) - (\Theta_r(s) + \mathbf{R}^2)\|_{\mathcal{H}_\infty} \leq 2\|\mathbf{R}\|^2 \sum_{i=k+1}^{q} \pi_i$. The result follows by noting that

$$\left(\begin{array}{c|c} \mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C} & -\mathbf{B}\mathbf{R}^2 \\ \hline \mathbf{R}^2\mathbf{C} & \mathbf{R}^2 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{R}^{-2} \end{array}\right)^{-1}$$

$$\text{and} \quad \left(\begin{array}{c|c} \mathbf{A}_{11} - \mathbf{B}_1\mathbf{R}^2\mathbf{C}_1 & -\mathbf{B}_1\mathbf{R}^2 \\ \hline \mathbf{R}^2\mathbf{C}_1 & \mathbf{R}^2 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{R}^{-2} \end{array}\right)^{-1},$$

where $\mathbf{R}^{-2} = \mathbf{D} + \mathbf{D}^*$. □

Remark 7.5.2. This error bound is equivalent to

$$\|[\mathbf{D}^* + \mathbf{H}(s)]^{-1}[\mathbf{H}(s) - \mathbf{H}_r(s)][\mathbf{D}^* + \mathbf{H}_r(s)]^{-1}\|_{\mathcal{H}_\infty} \leq 2\|\mathbf{R}\|^2 \sum_{i=k+1}^{q} \pi_i,$$

which is a frequency weighted bound for the error system $\mathbf{\Sigma} - \mathbf{\Sigma}_r$, where the input and the output weightings are $(\mathbf{D}^* + \mathbf{H}(s))^{-1}$ and $(\mathbf{D}^* + \mathbf{H}_r(s))^{-1}$, respectively.

A modified positive real balancing method with an absolute error bound

We will now introduce a modified positive real balancing method for a subclass of positive real systems. Then, based on Proposition 7.17, we will derive an absolute error bound for this reduction method.

Given the system $\mathbf{\Sigma}$ with positive real transfer function $\mathbf{H}$, let $\tilde{\mathbf{\Sigma}}$ be defined by means of its transfer function, as follows:

$$\tilde{\mathbf{H}}(s) = \left(\mathbf{D}^* + \mathbf{H}(s)\right)^{-1} - \frac{1}{2}\mathbf{R}^2.$$

A state-space representation of $\tilde{\mathbf{\Sigma}}$ is easily computed as $\left(\begin{array}{c|c} \mathbf{A} - \mathbf{B}\mathbf{R}^2\mathbf{C} & -\mathbf{B}\mathbf{R}^2 \\ \hline \mathbf{R}^2\mathbf{C} & \mathbf{R}^2/2 \end{array}\right)$. The assumption in this section is that $\tilde{\mathbf{H}}(s)$ is *positive real.*[6] In this case, the positive real balanced truncation discussed above is applied to $\tilde{\mathbf{\Sigma}}$. Let the positive real singular values of $\tilde{\mathbf{\Sigma}}$ be $\bar{\Pi} = \operatorname{diag}(\bar{\pi}_1\mathbf{I}_{s_1}, \ldots, \bar{\pi}_q\mathbf{I}_{s_k}, \bar{\pi}_{k+1}\mathbf{I}_{s_{k+1}}, \ldots, \bar{\pi}_q\mathbf{I}_{s_q})$, and let $\tilde{\mathbf{\Sigma}}_r$ denote the reduced positive real system obtained by keeping the first k positive real singular values $\bar{\pi}_i$. The resulting $\tilde{\mathbf{\Sigma}}_r$ is the intermediate reduced model. The final reduced-order model $\bar{\mathbf{\Sigma}}_r$ is obtained from $\tilde{\mathbf{\Sigma}}_r$ by means of the equation $\bar{\mathbf{D}}_r + \bar{\mathbf{H}}_r(s) = (\mathbf{R}^2/2 + \tilde{\mathbf{H}}_r(s))^{-1}$, where $\bar{\mathbf{D}}_r = \mathbf{D}$. The following bound holds.

Proposition 7.18. *Given the positive real system $\mathbf{\Sigma}$, let $\bar{\mathbf{\Sigma}}_r$ be obtained by the modified positive real balancing method discussed above. Then $\bar{\mathbf{\Sigma}}_r$ is asymptotically stable and positive real and satisfies*

$$\|\mathbf{H}(s) - \bar{\mathbf{H}}_r(s)\|_{\mathcal{H}_\infty} \leq 2\|\mathbf{R}^{-1}\|^2 \sum_{i=k+1}^{q} \bar{\pi}_i.$$

[6] This condition is not always satisfied. For example, if $\mathbf{H}(s) = 1 + \frac{1}{s+p}$, where $p \geq 0$, $\tilde{\mathbf{H}}(s) = \frac{2s+2p-1}{4(2s+2p+1)}$, and the above condition is satisfied for $p > \frac{1}{2}$.

Proof. Asymptotic stability and positive realness follow by construction. Since $\tilde{\mathbf{\Sigma}}_r$ is obtained from $\tilde{\mathbf{\Sigma}}$ by positive real balanced truncation, Proposition 7.17 yields $\|(\tilde{\mathbf{D}}^*+\tilde{\mathbf{H}}(s))^{-1} - (\tilde{\mathbf{D}}^*+\tilde{\mathbf{H}}_r(s))^{-1}\|_{\mathcal{H}_\infty} \leq 2\|\tilde{\mathbf{R}}\|^2 \sum_{i=k+1}^{q} \tilde{\pi}_i$, where $\tilde{\mathbf{R}}^2 = (\tilde{\mathbf{D}}+\tilde{\mathbf{D}}^*)^{-1} = \mathbf{R}^{-2}$. By construction we have $(\tilde{\mathbf{D}}^* + \tilde{\mathbf{H}}(s))^{-1} = \mathbf{D}^* + \mathbf{H}(s)$, and $(\tilde{\mathbf{D}}^* + \tilde{\mathbf{H}}_r(s))^{-1} = \mathbf{D}^* + \mathbf{H}_r(s)$, which implies the desired inequality. □

7.6 Frequency weighted balanced truncation*

The balancing methods discussed above aim at approximating the system $\mathbf{\Sigma}$ over all frequencies. In many cases, however, a good approximation is required only in a specific frequency range. This leads to approximation by *frequency weighted balanced truncation*. Given an *input weighting* $\mathbf{\Sigma}_i$ and an *output weighting* $\mathbf{\Sigma}_o$, the problem is to compute a reduced-order system $\mathbf{\Sigma}_r$, such that the weighted error

$$\| \mathbf{H}_o(s)(\mathbf{H}(s) - \mathbf{H}_r(s))\mathbf{H}_i(s) \|_{\mathcal{H}_\infty}$$

is small. Several methods have been proposed for frequency weighted model reduction. They consist of incorporating suitable output and input frequency weights in the computation of the truncated matrices. Varga and Anderson proposed in [346] the application of square root and balancing free square root techniques for accuracy enhancement. The next section gives a brief overview of existing frequency weighted model reduction methods and discusses how the square root and balancing free square root techniques can be incorporated.

Gramians for weighted systems

Consider the system $\mathbf{\Sigma}$, a system $\mathbf{\Sigma}_i$, which is the *input weight*, and a system $\mathbf{\Sigma}_o$, which is the *output weight*:

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right), \quad \mathbf{\Sigma}_i = \left(\begin{array}{c|c} \mathbf{A}_i & \mathbf{B}_i \\ \hline \mathbf{C}_i & \mathbf{D}_i \end{array}\right), \quad \mathbf{\Sigma}_o = \left(\begin{array}{c|c} \mathbf{A}_o & \mathbf{B}_o \\ \hline \mathbf{C}_o & \mathbf{D}_o \end{array}\right).$$

A realization of the systems $\hat{\mathbf{\Sigma}}_i$, $\hat{\mathbf{\Sigma}}_o$ whose transfer functions are $\hat{\mathbf{H}}_i = \mathbf{H}\mathbf{H}_i$, $\hat{\mathbf{H}}_o = \mathbf{H}_o\mathbf{H}$, respectively, is

$$\hat{\mathbf{\Sigma}}_i = \left[\begin{array}{c|c} \hat{\mathbf{A}}_i & \hat{\mathbf{B}}_i \\ \hline \hat{\mathbf{C}}_i & \hat{\mathbf{D}}_i \end{array}\right] = \left[\begin{array}{cc|c} \mathbf{A} & \mathbf{B}\mathbf{C}_i & \mathbf{B}\mathbf{D}_i \\ \mathbf{0} & \mathbf{A}_i & \mathbf{B}_i \\ \hline \mathbf{C} & \mathbf{D}\mathbf{C}_i & \mathbf{D}\mathbf{D}_i \end{array}\right],$$

$$\hat{\mathbf{\Sigma}}_o = \left[\begin{array}{c|c} \hat{\mathbf{A}}_o & \hat{\mathbf{B}}_o \\ \hline \hat{\mathbf{C}}_o & \hat{\mathbf{D}}_o \end{array}\right] = \left[\begin{array}{cc|c} \mathbf{A} & \mathbf{0} & \mathbf{B} \\ \mathbf{B}_o\mathbf{C} & \mathbf{A}_o & \mathbf{B}_o\mathbf{D} \\ \hline \mathbf{D}_o\mathbf{C} & \mathbf{C}_o & \mathbf{D}_o\mathbf{D} \end{array}\right].$$

Recall the definition of the system gramians in the frequency domain,

$$\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\, d\omega, \tag{4.51}$$

where $(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ is the *input-to-state* map, and

$$\mathcal{Q} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*\mathbf{C}(i\omega\mathbf{I} - \mathbf{A})^{-1}\, d\omega, \tag{4.52}$$

where $\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}$ is the *state-to-output* map. Given that the input to the system has to be weighted, that is, go through the system $\mathbf{\Sigma}_o$, the *weighted input-to-state* map becomes $(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{H}_i(s)$, while the *weighted state-to-output map* becomes $\mathbf{H}_o(s)\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}$. Consequently, the ***reachability and observability weighted gramians*** are

$$\boxed{\mathcal{P}_i = \frac{1}{2\pi}\int_{-\infty}^{\infty}(i\omega\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\,\mathbf{H}_i(i\omega)\mathbf{H}_i^*(-i\omega)\,\mathbf{B}^*(-i\omega\mathbf{I}-\mathbf{A}^*)^{-1}\,d\omega,} \tag{7.33}$$

$$\boxed{\mathcal{Q}_o = \frac{1}{2\pi}\int_{-\infty}^{\infty}(-i\omega\mathbf{I}-\mathbf{A}^*)^{-1}\mathbf{C}^*\,\mathbf{H}_o^*(-i\omega)\mathbf{H}_o(i\omega)\,\mathbf{C}(i\omega\mathbf{I}-\mathbf{A})^{-1}\,d\omega.} \tag{7.34}$$

We can now find a balancing transformation $\mathbf{T}$ which will simultaneously diagonalize these two gramians

$$\mathbf{T}\mathcal{P}_i\mathbf{T}^* = \mathbf{T}^{-*}\mathcal{Q}_o\mathbf{T}^{-1} = \Sigma = \operatorname{diag}(\sigma_1,\ldots,\sigma_n).$$

Subsequently, order reduction takes place exactly as for balanced truncation; to that effect, recall (7.3), (7.4).

The question that arises is whether these weighted gramians can be obtained as solutions of Lyapunov equations. Let $\hat{\mathcal{P}}_i$, $\hat{\mathcal{Q}}_o$ be the reachability, observability, gramians of $\hat{\Sigma}_i$, $\hat{\Sigma}_o$, respectively, that is,

$$\hat{\mathbf{A}}_i\hat{\mathcal{P}}_i + \hat{\mathcal{P}}_i\hat{\mathbf{A}}_i^* + \hat{\mathbf{B}}_i\hat{\mathbf{B}}_i^* = \mathbf{0}, \quad \hat{\mathbf{A}}_o^*\hat{\mathcal{Q}}_o + \hat{\mathcal{Q}}_o\hat{\mathbf{A}}_o + \hat{\mathbf{C}}_o^*\hat{\mathbf{B}}_o = \mathbf{0}. \tag{7.35}$$

The gramians are partitioned in two-by-two blocks, so that the dimension of the $(1,1)$ block is the same as that of $\mathbf{A}$, i.e., the order of the to-be-reduced system,

$$\hat{\mathcal{P}}_i = \begin{bmatrix} \hat{\mathcal{P}}_{11} & \hat{\mathcal{P}}_{12} \\ \hat{\mathcal{P}}_{21} & \hat{\mathcal{P}}_{22} \end{bmatrix}, \quad \hat{\mathcal{Q}}_o = \begin{bmatrix} \hat{\mathcal{Q}}_{11} & \hat{\mathcal{Q}}_{12} \\ \hat{\mathcal{Q}}_{21} & \hat{\mathcal{Q}}_{22} \end{bmatrix}.$$

Proposition 7.19. *The frequency weighted reachability, observability, gramians defined by* (7.33), (7.34), *are equal to the* $(1,1)$ *block of the gramians* $\hat{\mathcal{P}}_i$, $\hat{\mathcal{Q}}_o$, *respectively:*

$$\mathcal{P}_i = \hat{\mathcal{P}}_{11}, \quad \mathcal{Q}_o = \hat{\mathcal{Q}}_{11}. \tag{7.36}$$

Proof. The proof follows by noticing that

$$\begin{aligned}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\,\mathbf{H}_i(s) &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \end{pmatrix}\begin{pmatrix} s\mathbf{I}-\mathbf{A} & -\mathbf{B}\mathbf{C}_i \\ \mathbf{0} & s\mathbf{I}-\mathbf{A}_i \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{B}\mathbf{D}_i \\ \mathbf{B}_i \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I} & \mathbf{0} \end{pmatrix}(s\mathbf{I}-\hat{\mathbf{A}}_i)^{-1}\hat{\mathbf{B}}_i,\end{aligned}$$

$$\begin{aligned}\mathbf{H}_o(s)\,\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1} &= \begin{pmatrix} \mathbf{D}_o\mathbf{C} & \mathbf{C}_o \end{pmatrix}\begin{pmatrix} s\mathbf{I}-\mathbf{A} & \mathbf{0} \\ -\mathbf{B}_o\mathbf{C} & s\mathbf{I}-\mathbf{A}_o \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} \\ &= \hat{\mathbf{C}}_o(s\mathbf{I}-\hat{\mathbf{A}}_o)^{-1}\begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix}.\end{aligned}$$

Therefore, the required gramians are the $(1, 1)$ blocks of the gramians obtained by solving the Lyapunov equations (7.35). □

7.6.1 Frequency weighted balanced truncation*

Once the weighted gramians $\mathcal{P}_i$ and $\mathcal{Q}_o$ have been defined, a balanced realization is obtained by determining the transformation which simultaneously diagonalizes these two gramians,

$$\mathcal{P}_i = \mathcal{Q}_o = \operatorname{diag}(\sigma_1 \mathbf{I}_{m_1}, \ldots, \sigma_k \mathbf{I}_{m_k}, \sigma_{k+1} \mathbf{I}_{m_{k+1}}, \ldots, \sigma_q \mathbf{I}_{m_q}),$$

where m_i are the multiplicities of σ_i with $m_1 + \cdots + m_q = n$. Truncation in the balanced basis of the states, which correspond to small singular values, yields a system $\boldsymbol{\Sigma}_r$ of reduced order obtained by *weighted balanced truncation.* The following result holds.

Lemma 7.20. *Given the stable and minimal system $\boldsymbol{\Sigma}$, together with input, output, weights $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Sigma}_o$, respectively, let $\boldsymbol{\Sigma}_r$ be obtained by frequency weighted balanced truncation as above. If either $\mathbf{H}_i = \mathbf{I}$ or $\mathbf{H}_o = \mathbf{I}$, $\boldsymbol{\Sigma}_r$ is stable. In general, whenever $\boldsymbol{\Sigma}_r$ is stable, the error is bounded by the expression*

$$\|\mathbf{H}_o(s)(\mathbf{H}(s) - \mathbf{H}_r(s))\mathbf{H}_i(s)\|_{\mathcal{H}_\infty} \leq 2 \sum_{i=k+1}^{q} \sqrt{\sigma_k^2 + (\alpha_k + \beta_k)\sigma_k^{3/2} + \alpha_k \beta_k \sigma_k},$$

where α_k and β_k are the $\mathcal{H}_\infty$-norms of transfer functions, depending on the weights and the reduced system.

For details on the computation of α_k and β_k, see [201].

The first attempt to deal with weightings in model reduction was by Enns [107]. However, this method does not guarantee stability of the reduced model for the case of two-sided weighting. The original work of Enns also did not provide an error bound.

The method by Lin and Chiu

To add stability for the case of two-sided weighting, Lin and Chiu [229] compute $\mathcal{P}$, $\mathcal{Q}$ as above but replace the gramians (7.36) by their Schur complements:

$$\mathcal{P} = \mathcal{P}_{11} - \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathcal{P}_{21}, \quad \mathcal{Q} = \mathcal{Q}_{11} - \mathcal{Q}_{12}\mathcal{Q}_{22}^{-1}\mathcal{Q}_{21}.$$

If the realizations are minimal, these gramians are nonsingular. Furthermore, stability of the reduced system is guaranteed. The main drawback of this method is that the realizations of $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_o$ as given above must be minimal. Finally, the weighted error system satisfies an error bound of the form

$$\| \mathbf{H}_o(s)(\mathbf{H}(s) - \mathbf{H}_r(s))\mathbf{H}_i(s) \|_{\mathcal{H}_\infty} \leq 2 \sum_{i=k+1}^{q} \sqrt{(\tilde{\sigma}_k^2 + \alpha_k + \lambda_k)(\tilde{\sigma}_k + \beta_k + \omega_k)},$$

where α_k, β_k, λ_k, and ω_k denote the $\mathcal{H}_\infty$-norms of transfer functions depending on the data. For details, see [229].

Zhou's self-weighted method

Zhou's method [369] is applicable to any stable $\mathbf{\Sigma}$ that has a stable right inverse and $\mathbf{D}$ is full rank. For simplicity, we will discuss only the case where $\mathbf{H}(s)$ is square, nonsingular, $\det(\mathbf{D}) \neq 0$, and $\mathbf{H}^{-1}$ is asymptotically stable, i.e., the system is minimum phase. This method is a special case of Enns' method, where

$$\mathbf{H}_i(s) = \mathbf{I}, \quad \mathbf{H}_o(s) = \mathbf{H}^{-1}(s), \text{ where } \mathbf{\Sigma}_o = \left(\begin{array}{c|c} \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} & -\mathbf{B}\mathbf{D}^{-1} \\ \hline \mathbf{D}^{-1}\mathbf{C} & \mathbf{D}^{-1} \end{array}\right).$$

The gramians to be diagonalized are therefore

$$\mathbf{A}\mathcal{P}_i + \mathcal{P}_i\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}, \quad \mathcal{Q}_o(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) + (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^*\mathcal{Q}_o + (\mathbf{D}^{-1}\mathbf{C})^*(\mathbf{D}^{-1}\mathbf{C}) = \mathbf{0}.$$

The self-weighted balanced realization is then obtained by simultaneously diagonalizing $\mathcal{P}_{11}$ and $\mathcal{Q}_{11}$, i.e.,

$$\mathcal{P}_{11} = \mathcal{Q}_{11} = \operatorname{diag}(\sigma_1 I_{n_1}, \ldots, \sigma_k I_{n_k}, \sigma_{k+1} I_{n_{k+1}}, \ldots, \sigma_q I_{n_q}).$$

Lemma 7.21. *Let* $\mathbf{H}(s)$ *be an asymptotically stable, square, nonsingular, and minimum phase system. Also, let* $\mathbf{H}_r(s)$ *be obtained by the self-weightedfrequency truncation method. Then* $\mathbf{H}_r(s)$ *is asymptotically stable and minimum phase and satisfies*

$$\|\mathbf{H}_r^{-1}(\mathbf{H} - \mathbf{H}_r)\|_{\mathcal{H}_\infty}, \ \|\mathbf{H}^{-1}(\mathbf{H} - \mathbf{H}_r)\|_{\mathcal{H}_\infty} \leq \prod_{i=k+1}^{q} \left(1 + 2\sigma_i\sqrt{1 + \sigma_i^2} + 2\sigma_i^2\right) - 1.$$

It can be shown that if $\mathbf{H}(s)$ is square, asymptotically stable, nonsingular, and minimum phase as in the above theorem, this method is equivalent to stochastic balancing. In this case, the σ_i and the stochastic singular values μ_i are related by $\mu_i = \frac{\sigma_i}{\sqrt{1+\sigma_i^2}}$.

The method by Wang, Sreeram, and Liu

A method due to Wang, Sreeram, and Liu [351] provides computation of an a priori computable error bound for the weighted error. This method also guarantees stability for the case of two-sided weighting. It does so by forcing the gramians to satisfy a definite Lyapunov equation (i.e., $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{Q} = \mathbf{0}$, where $\mathbf{Q} \geq \mathbf{0}$). Then stability is a result of the inertia property for Lyapunov equations discussed in section 6.2. Let

$$\mathbf{X} = -\mathbf{A}\mathcal{P}_{11} - \mathcal{P}_{11}\mathbf{A}^*, \quad \mathbf{Y} = -\mathbf{A}^*\mathcal{Q}_{11} - \mathcal{Q}_{11}\mathbf{A}.$$

If these two matrices $\mathbf{X}$ and $\mathbf{Y}$ are positive (semi) definite, the reduced system is guaranteed to be stable. Otherwise, let the eigenvalue decomposition of the former be $\mathbf{X} = \mathbf{U}\Lambda\mathbf{U}^*$, where $\mathbf{U}$ is orthogonal and Λ is diagonal with real entries λ_i on the diagonal. Let

$$\mid \Lambda \mid = \operatorname{diag}(|\lambda_1|, \ldots, |\lambda_n|).$$

Let $\mid \mathbf{X} \mid = \mathbf{U} \mid \Lambda \mid \mathbf{U}^*$, and similarly for $\mid \mathbf{Y} \mid$. The gramians are now computed by solving the Lyapunov equations,

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mid \mathbf{X} \mid = \mathbf{0}, \quad \mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mid \mathbf{Y} \mid = \mathbf{0}.$$

Subsequently, balancing and truncation are applied as earlier.

Although the computation in this method is more involved, stability of the reduced system is guaranteed. Furthermore, an error bound can be computed. (See [351] for details.)

The combination method by Varga and Anderson

The combination method proposed in [346] is formulated by combining Enns' method with Lin and Chiu's method. Using this approach, the reduced model retains the features of both methods. This method introduces two parameters, α, β, and computes the frequency weighted gramians as follows:

$$\mathcal{P} = \mathcal{P}_{11} - \alpha^2 \mathcal{P}_{12}\mathcal{P}_{22}^{-1}\mathcal{P}_{21}, \quad \mathcal{Q} = \mathcal{Q}_{11} - \beta^2 \mathcal{Q}_{12}\mathcal{Q}_{22}^{-1}\mathcal{Q}_{21}.$$

Since for $\alpha = \beta = 1$, the reduced-order system is stable, while for $\alpha = \beta = 0$ it may not be, by choosing the parameters in the neighborhood of one, a family of stable weighted reduced-order models is obtained.

In the same paper, a modification of Wang's method is presented, consisting in generating a semidefinite $\mid \mathbf{X} \mid$ from $\mathbf{X}$ by setting all its negative eigenvalues to zero (similarly for $\mid \mathbf{Y} \mid$.

Remark 7.6.1. To improve accuracy in all methods mentioned above, square root and balancing free square root techniques can be used. One issue that needs to be addressed is the sensitivity of the reduced model to the choice of the input and output weightings.

7.6.2 Frequency weighted balanced reduction without weights*

The goal of the frequency weighted reduction problems is to find a system $\boldsymbol{\Sigma}_r$ of order less than $\boldsymbol{\Sigma}$ such that the weighted error $\|\mathbf{H}_o(s)(\mathbf{H}(s) - \mathbf{H}_r(s))\mathbf{H}_i(s)\|_{\mathcal{H}_\infty}$ is *small*. It should be mentioned that in many cases the input and output weightings $\mathbf{H}_i(s)$ and $\mathbf{H}_o(s)$ are not given. Instead the problem is to approximate $\boldsymbol{\Sigma}$ over a given frequency range $[\omega_1, \omega_2]$. As shown by Gawronski and Juang [135], this problem can be attacked directly, without constructing input and output weights. This is achieved by using the frequency domain representation of the gramians, with the limits of integration appropriately restricted:

$$\boxed{\mathcal{P}(\omega) = \frac{1}{2\pi}\int_{-\omega}^{\omega} (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\, d\omega,} \tag{7.37}$$

$$\boxed{\mathcal{Q}(\omega) = \frac{1}{2\pi}\int_{-\omega}^{\omega} (-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\mathbf{C}^*\mathbf{C}(i\omega\mathbf{I} - \mathbf{A})^{-1}\, d\omega.} \tag{7.38}$$

Thus if we are interested in the frequency interval $[0, \omega]$, we simultaneously diagonalize the two gramians $\mathcal{P}(\omega)$ and $\mathcal{Q}(\omega)$ defined above and proceed as before. If the frequency interval $[\omega_1, \omega_2]$ is of interest, the gramians are defined as follows:

$$\mathcal{P}(\omega_1, \omega_2) = \mathcal{P}(\omega_2) - \mathcal{P}(\omega_1) \text{ and } \mathcal{Q}(\omega_1, \omega_2) = \mathcal{Q}(\omega_2) - \mathcal{Q}(\omega_1). \tag{7.39}$$

It is readily seen that these expressions are positive (semi) definite. Thus simultaneous diagonalization of these gramians and subsequent truncation provide a reduced-order system which is expected to yield small errors in the interval $[\omega_1, \omega_2]$.

The question now is whether these gramians satisfy Lyapunov equations. To answer this question, we introduce the following notation:

$$\mathbf{S}(\omega) = \frac{1}{2\pi}\int_{-\omega}^{\omega}(i\omega\mathbf{I}-\mathbf{A})^{-1}\,d\omega = -\frac{i}{2\pi}\ln\left[(i\omega\mathbf{I}-\mathbf{A})(-i\omega\mathbf{I}-\mathbf{A})^{-1}\right], \tag{7.40}$$

$$\mathbf{W}_r(\omega) = \mathbf{S}(\omega)\mathbf{B}\mathbf{B}^* + \mathbf{B}\mathbf{B}^*\mathbf{S}^*(-\omega), \quad \mathbf{W}_o(\omega) = \mathbf{S}^*(-\omega)\mathbf{C}^*\mathbf{C} + \mathbf{C}^*\mathbf{C}\mathbf{S}(\omega),$$

$$\mathbf{W}_r(\omega_1,\omega_2) = \mathbf{W}_r(\omega_2) - \mathbf{W}_r(\omega_1), \quad \mathbf{W}_o(\omega_1,\omega_2) = \mathbf{W}_o(\omega_2) - \mathbf{W}_o(\omega_1).$$

Lemma 7.22. *With the notation introduced above, the gramians satisfy the following Lyapunov equations:*

$$\mathbf{A}\mathcal{P}(\omega) + \mathcal{P}(\omega)\mathbf{A}^* + \mathbf{W}_r(\omega) = \mathbf{0}, \quad \mathbf{A}^*\mathcal{Q}(\omega) + \mathcal{Q}(\omega)\mathbf{A} + \mathbf{W}_o(\omega) = \mathbf{0}.$$

Furthermore,

$$\mathbf{A}\mathcal{P}(\omega_1,\omega_2) + \mathcal{P}(\omega_1,\omega_2)\mathbf{A}^* + \mathbf{W}_r(\omega_1,\omega_2) = \mathbf{0},$$
$$\mathbf{A}^*\mathcal{Q}(\omega_1,\omega_2) + \mathcal{Q}(\omega_1,\omega_2)\mathbf{A} + \mathbf{W}_o(\omega_1,\omega_2) = \mathbf{0}.$$

The computations of the various gramians require the evaluation of a matrix logarithm, in addition to the solution of two Lyapunov equations. For small- to medium-scale problem for which an exact balancing transformation can be computed, $\mathbf{S}(\omega)$ can be efficiently determined as well. However, for large-scale problems, this issue is still under investigation. But we note that computing an exact solution to a Lyapunov equation in large-scale settings is an ill-conditioned problem itself.

Balancing involves the simultaneous diagonalization of the gramians $\mathcal{P}(\omega_1,\omega_2)$ and $\mathcal{Q}(\omega_1,\omega_2)$,

$$\mathcal{P}(\omega_1,\omega_2) = \mathcal{Q}(\omega_1,\omega_2) = \mathrm{diag}(\sigma_1\mathbf{I}_{m_1},\ldots,\sigma_q\mathbf{I}_{m_q}),$$

where as before m_i are the multiplicities of each distinct singular value σ_i. The reduced model is obtained as usual, by truncation. However, since $\mathbf{W}_r$ and $\mathbf{W}_o$ are not guaranteed to be positive definite, stability of the reduced model cannot be guaranteed. In [159], a modified reduction method is presented, where the Lyapunov equations are forced to be definite (much the same way as the method in [351]). This also results in error bounds. The details are omitted here.

Remark 7.6.2. **(a)** In section 6.1.2, a Lyapunov equation was derived for an expression of the form (6.11), where the eigenvalues of $\mathbf{A}$ may lie both in the left- and the right-hand side of the complex plane. We recognize this as the gramian $\mathcal{P}(\infty)$ discussed above. In that section, a Lyapunov equation satisfied by this gramian was given. This equation is in terms of the spectral projector Π onto the stable invariant subspace of $\mathbf{A}$. This projector is related to $\mathbf{S}(\infty)$ defined above. It can be shown that $\Pi = \frac{1}{2}\mathbf{I} + \mathbf{S}(\infty)$. For an illustration of this fact, see the example at the end of this section.

(b) The above discussion reveals the connection between Enns' and Gawronski and Juang's frequency weighted balancing methods. The latter is obtained from the former by choosing $\mathbf{H}_i(s)$ and $\mathbf{H}_o(s)$ as the perfect bandpass filters over the frequency range of interest. However, the realizations of the weightings are never computed. We note that an infinite-dimensional realization will be needed to obtain perfect band-pass filters. Hence, in Enns' method, these band-pass filters are approximated by lower-order band-pass filters. The resulting Lyapunov equations have dimension $n + n_i$, where n_i is the order of $\mathbf{H}_i(s)$. It can be shown by means of examples that as the order of the weightings increases, i.e., as they get closer to a perfect band-pass filter, the two methods produce similar results.

7.6.3 Balanced reduction using time-limited gramians*

Yet another set of gramians is obtained by restricting in time the gramians $\mathcal{P}$, $\mathcal{Q}$ given by (4.43), (4.44). For $T = [\, t_1,\ t_2\,]$, the time-limited gramians are defined as

$$\mathcal{P}(T) = \int_{t_1}^{t_2} e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^*\tau} d\tau \quad \text{and} \quad \mathcal{Q}(T) = \int_{t_1}^{t_2} e^{\mathbf{A}^*\tau}\mathbf{C}^*\mathbf{C} e^{\mathbf{A}\tau} d\tau. \tag{7.41}$$

These quantities are positive (semi) definite and therefore qualify as gramians. Similarly the frequency-limited gramians, $\mathcal{P}(T)$, $\mathcal{Q}(T)$ satisfy Lyapunov equations. Let

$$\mathbf{V}_r(T) = e^{\mathbf{A}t_1}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^* t_1} - e^{\mathbf{A}t_2}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^* t_2}, \quad \mathbf{V}_o(T) = e^{\mathbf{A}^* t_1}\mathbf{C}^*\mathbf{C} e^{\mathbf{A}t_1} - e^{\mathbf{A}^* t_2}\mathbf{C}^*\mathbf{C} e^{\mathbf{A}t_2}.$$

Lemma 7.23. *The gramians $\mathcal{P}(T)$ and $\mathcal{Q}(T)$ are solutions of the following Lyapunov equations:*

$$\mathbf{A}\mathcal{P}(T) + \mathcal{P}(T)\mathbf{A}^* + \mathbf{V}_r(T) = \mathbf{0}, \quad \mathbf{A}^*\mathcal{Q}(T) + \mathcal{Q}(T)\mathbf{A} + \mathbf{V}_o(T) = \mathbf{0}.$$

Balancing in this case is obtained by simultaneous diagonalization of the time-limited gramians $\mathcal{P}(T)$ and $\mathcal{Q}(T)$:

$$\mathcal{P}(T) = \mathcal{Q}(T) = \operatorname{diag}(\sigma_{n_1}\mathbf{I}_{n_1}, \ldots, \sigma_{n_q}\mathbf{I}_{n_q}).$$

The reduced model follows by truncation in this basis; the impulse response of the reduced model is expected to match that of the full-order model in the time interval $T = [\, t_1,\ t_2\,]$; see [135]. However, as in the frequency weighted case, the reduced model is not guaranteed to be stable; this can be fixed as shown in [159].

7.6.4 Closed-loop gramians and model reduction*

Here, the given system $\mathbf{\Sigma}$ is part of a closed loop with controller $\mathbf{\Sigma}_c$. In the simplest case, the overall system is characterized by the transfer function $\phi(\mathbf{H}, \mathbf{H}_c) = \mathbf{H}(s)(\mathbf{I} + \mathbf{H}_c(s)\mathbf{H}(s))^{-1}$. A reduced-order system $\hat{\mathbf{\Sigma}}$ is sought such that the error

$$\|\phi(\mathbf{H}, \mathbf{H}_c) - \phi(\hat{\mathbf{H}}, \mathbf{H}_c)\|$$

is kept small or minimized. Alternatively, one may also wish to find a reduced-order $\hat{\mathbf{\Sigma}}_c$ such that

$$\|\phi(\mathbf{H}, \mathbf{H}_c) - \phi(\mathbf{H}, \hat{\mathbf{H}}_c)\| \quad \text{or} \quad \|\phi(\mathbf{H}, \mathbf{H}_c) - \phi(\hat{\mathbf{H}}, \hat{\mathbf{H}}_c)\|$$

is kept small or minimized.

We now turn our attention to the definition of gramians for closed-loop systems. The approach follows [362]. A related approach applicable to open-loop prefiltered systems is described in [133].

Consider the system $\mathbf{\Sigma}$ in the usual feedback loop together with the controller $\mathbf{\Sigma}_c$; let the transfer function be $\phi(\mathbf{H}, \mathbf{H}_c)$, as defined above. This system with controller $\mathbf{\Sigma}_c$, can be considered as a weighted composite system with input weighting,

$$\mathbf{H}_i(s) = (\mathbf{I} + \mathbf{H}_c(s)\mathbf{H}(s))^{-1}.$$

The gramians of the system $\mathbf{\Sigma}$ *with respect to the given closed loop* with the controller $\mathbf{\Sigma}_c$ are now defined as shown earlier, by means of (7.33), for the input weighting case. Similarly, the gramians of the controller $\mathbf{H}_c(s)$ in the same closed loop can be defined by noting that the transfer function of interest is $\mathbf{H}_c(s)\mathbf{H}(s)\ (\mathbf{I} + \mathbf{H}_c(s)\mathbf{H}(s))^{-1}$. Consequently, this can be considered as a weighted system with input weighting,

$$\mathbf{H}_i(s) = \mathbf{H}(s)\ (\mathbf{I} + \mathbf{H}_c(s)\mathbf{H}(s))^{-1}.$$

The resulting gramians of $\mathbf{H}_c(s)$ *with respect to the given closed loop* can be used to reduce the order of the controller in the closed loop.

Finally, as expected, the closed-loop gramians defined above in the frequency domain can be computed by solving Lyapunov equations. Let

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{0} \end{array}\right), \quad \mathbf{\Sigma}_c = \left(\begin{array}{c|c} \mathbf{A}_c & \mathbf{B}_c \\ \hline \mathbf{C}_c & \mathbf{0} \end{array}\right)$$

be the system quadruples; then a realization $(\mathbf{A}_t, \mathbf{B}_t, \mathbf{C}_t)$ for the compound $\phi(\mathbf{H}, \mathbf{H}_c)$ is given by

$$\left(\begin{array}{c|c} \mathbf{A}_t & \mathbf{B}_t \\ \hline \mathbf{C}_t & \mathbf{0} \end{array}\right) = \left(\begin{array}{cc|c} \mathbf{A} & -\mathbf{B}\mathbf{C}_c & \mathbf{B} \\ \mathbf{B}_c\mathbf{C} & \mathbf{A}_c & \mathbf{0} \\ \hline \mathbf{C} & \mathbf{0} & \mathbf{0} \end{array}\right).$$

Thus to get an approximation by balanced truncation of the system $\mathbf{\Sigma}$ in the sense that the closed-loop norm is kept small or minimized, one can again use the $(1, 1)$ blocks of the two gramians of the partitioned system above.

7.6.5 Balancing for unstable systems*

The problem with balancing of unstable systems is that the infinite gramians $\mathcal{P}$, $\mathcal{Q}$ given by (4.43), (4.44) are not defined, and therefore one cannot talk of simultaneous diagonalization. There are two remedies to this situation.

The first is to use *time-limited gramians*, which are defined by (7.41) over any finite interval $[t_1, t_2]$. Balancing $\mathcal{P}(T)$ and $\mathcal{Q}(T)$ followed by truncation would then provide a reduced-order system which is expected to approximate well the impulse response over the chosen interval.

The second approach proposed in [372] consists of using the expression for the gramians in the frequency domain, e.g., (4.51), (4.52). These frequency domain expressions are indeed defined even if the system is unstable (they are not defined when the system has poles on the imaginary axis) and are positive definite. Therefore, they can be simultaneously diagonalized. Actually, as noticed earlier, these gramians satisfy Lyapunov equations given in Lemma 7.22. Furthermore, as noted in [372], balancing based on simultaneous diagonalization of the infinite gramians consists of separate balancing of the stable and of the antistable parts of the system.

Example 7.24. We will now illustrate some of the properties of the reduction of unstable systems by balanced truncation with the following example:

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cccc|c} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 4 & 1/5 & 0 & -1/10 & 1 \\ \hline 1 & 0 & 0 & 0 & 0 \end{array}\right].$$

The eigenvalues of $\mathbf{A}$ are $\lambda_{1,2} = \pm\sqrt{2}$, $\lambda_{3,4} = -1/20 \pm i\sqrt{799}/20$, and the transfer function is

$$\mathbf{H}(s) = \frac{5}{799}\left[\frac{s - \frac{399}{10}}{s^2 + \frac{1}{10}s + 2} - \frac{a}{s + \sqrt{2}} + \frac{b}{s - \sqrt{2}}\right],$$

$$\text{where } a = 10\sqrt{2} + \frac{1}{2},\ b = 10\sqrt{2} - \frac{1}{2}.$$

We will balance this system for $\omega = 1$, $\omega = 100$, and $\omega = \infty$. For $\omega = 1$, the singular values are .0181, .0129, .0010, .0009. Hence the balanced system is

$$\left(\begin{array}{c|c} \mathbf{A}_{\text{bal}} & \mathbf{B}_{\text{bal}} \\ \hline \mathbf{C}_{\text{bal}} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cccc|c} -2.7440 & 3.1204 & -2.9117 & 3.0559 & .8471 \\ -3.1204 & 2.0320 & -2.2698 & 2.5908 & .7696 \\ -2.9117 & 2.2698 & -3.6929 & 3.5878 & .8881 \\ -3.0559 & 2.5908 & -3.5878 & 4.3049 & .9560 \\ \hline -.8471 & .7696 & -.8881 & .9560 & 0 \end{array}\right].$$

For $\omega = 100$, the singular values are .9203, .8535, .0299, .0277, and the balanced system is

$$\left(\begin{array}{c|c} \mathbf{A}_{\text{bal}} & \mathbf{B}_{\text{bal}} \\ \hline \mathbf{C}_{\text{bal}} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cccc|c} -.0534 & 1.4097 & -.0013 & -.0924 & .3124 \\ -1.4097 & -.0529 & .0005 & .1017 & .3017 \\ .0013 & .0005 & 1.4143 & .0128 & -.2937 \\ -.0924 & -.1017 & -.0128 & -1.4080 & .2822 \\ \hline -.3124 & .3017 & -.2934 & -.2822 & 0 \end{array}\right].$$

Finally, for $\omega \to \infty$, the balanced realization of the system consists of the direct sum of the balanced realization of the stable part and that of the instable part. The singular values are .9206, .8538, .0302, .0280. Consequently the balancing transformation is

$$\mathbf{T} = \begin{bmatrix} -.9439 & -.5426 & .4139 & .3123 \\ .7987 & -.6061 & -.3974 & .3016 \\ -.8264 & -.6257 & -.4424 & -.2922 \\ -.7938 & .6037 & -.4270 & .2807 \end{bmatrix}.$$

Thus $\mathbf{A}_{\text{bal}} = \mathbf{TAT}^{-1}$, $\mathbf{B}_{\text{bal}} = \mathbf{TB}$, $\mathbf{C}_{\text{bal}} = \mathbf{CT}^{-1}$:

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A}_{\text{bal}} & \mathbf{B}_{\text{bal}} \\ \hline \mathbf{C}_{\text{bal}} & \end{array}\right) = \left[\begin{array}{cccc|c} -.0529 & 1.4096 & 0 & -.0924 & .3123 \\ -1.4096 & -.0533 & 0 & .1025 & .3016 \\ 0 & 0 & 1.4142 & 0 & -.2922 \\ -.0924 & -.1025 & 0 & -1.4080 & .2807 \\ \hline -.3123 & .3016 & -.2922 & -.2807 & 0 \end{array}\right].$$

This is the direct sum of the balanced realizations of the stable subsystem $\mathbf{\Sigma}^-$ and of the antistable subsystem $\mathbf{\Sigma}^+$:

$$\mathbf{\Sigma} = \left(\begin{array}{cc|c} \mathbf{A}_{\text{bal}}^- & \mathbf{0} & \mathbf{B}_{\text{bal}}^- \\ \mathbf{0} & \mathbf{A}_{\text{bal}}^+ & \mathbf{B}_{\text{bal}}^+ \\ \hline \mathbf{C}_{\text{bal}}^- & \mathbf{C}_{\text{bal}}^+ & \end{array}\right),$$

$$\mathbf{\Sigma}^- = \left(\begin{array}{c|c} \mathbf{A}_{\text{bal}}^- & \mathbf{B}_{\text{bal}}^- \\ \hline \mathbf{C}_{\text{bal}}^- & \end{array}\right) = \left[\begin{array}{ccc|c} -.0529 & 1.4096 & -.0924 & .3123 \\ -1.4096 & -.0533 & .1025 & .3016 \\ -.0924 & -.1025 & -1.4080 & .2807 \\ \hline -.3123 & .3016 & -.2807 & 0 \end{array}\right],$$

$$\mathbf{\Sigma}^+ = \left(\begin{array}{c|c} \mathbf{A}_{\text{bal}}^+ & \mathbf{B}_{\text{bal}}^+ \\ \hline \mathbf{C}_{\text{bal}}^+ & \end{array}\right) = \left[\begin{array}{c|c} 1.4142 & -.2922 \\ \hline -.2922 & 0 \end{array}\right].$$

Finally, we recall that $\mathbf{S}(\infty)$ defines the spectral projectors related to $\mathbf{A}$, namely, $\Pi_+ = \frac{1}{2}\mathbf{I} + \mathbf{S}(\infty)$, is an (oblique) projection onto the invariant subspace of $\mathbf{A}$ corresponding to the eigenvalues in the right half plane (unstable eigenvalues), and $\Pi_- = \frac{1}{2}\mathbf{I} - \mathbf{S}(\infty)$ is a projection onto the invariant subspace corresponding to the eigenvalues in the left half plane (stable eigenvalues). The eigenvalues of $\mathbf{S}(1)$ are $-.1956$, $.0159 \pm .2797i$, $.1956$; those of $\mathbf{S}(100)$ are $-.4955$, $.4998 \pm .0045i$ $.4955$; and the eigenvalues of $\mathbf{S}(\infty)$ are $-\frac{1}{2}$ and $\frac{1}{2}$ with multiplicity 3. It follows that $\Pi_+ = \mathbf{vw}^*$, where $\mathbf{v}^* = [.2582 \ \ .3651 \ \ .5164 \ \ .7303]$, $\mathbf{w}^* = [.9352 \ \ .7080 \ \ .5006 \ \ .3306]$; this implies $\mathbf{w}^*\mathbf{Av} = 1.4142$. With

$$\mathbf{V} = \begin{bmatrix} .2582 & .0273 + .2567i & .0273 - .2567i \\ -.3651 & -.3642 + .0258i & -.3642 - .0258i \\ .5164 & -.0182 - .5161i & -.0182 + .5161i \\ -.7303 & .7303 & .7303 \end{bmatrix},$$

$$\mathbf{W} = \begin{bmatrix} 1.0037 & .0343 - .9688i & .0343 + .9688i \\ -.6596 & -.6838 - .0484i & -.6838 + .0484i \\ .4664 & -.0171 + .4844i & -.0171 - .4844i \\ -.3549 & .3419 + .0242i & .3419 - .0242i \end{bmatrix},$$

$\Pi_- = \mathbf{VW}^*$, which implies that the projected $\mathbf{A}$ is $\mathbf{W}^*\mathbf{AV} = \operatorname{diag}[-1.4142, \ -.05 + 1.4133i, \ -.05 - 1.4133i]$.

Finally, balanced truncation, leads to reduced-order systems of the unstable system $\mathbf{\Sigma}$. We will compare the reduced-order systems of orders two and three computed for $\omega = 1$, $\omega = 100$. The top pane of Figure 7.2, the frequency response of the original system is plotted together with the frequency responses of the second- and third-order approximants, which are labeled 2a, 3a for $\omega = 1$ and 2b, 3b for $\omega = 100$. The frequency responses are plotted from .3 Hz to 7 Hz. The plots show that the approximants for $\omega = 1$ approximate

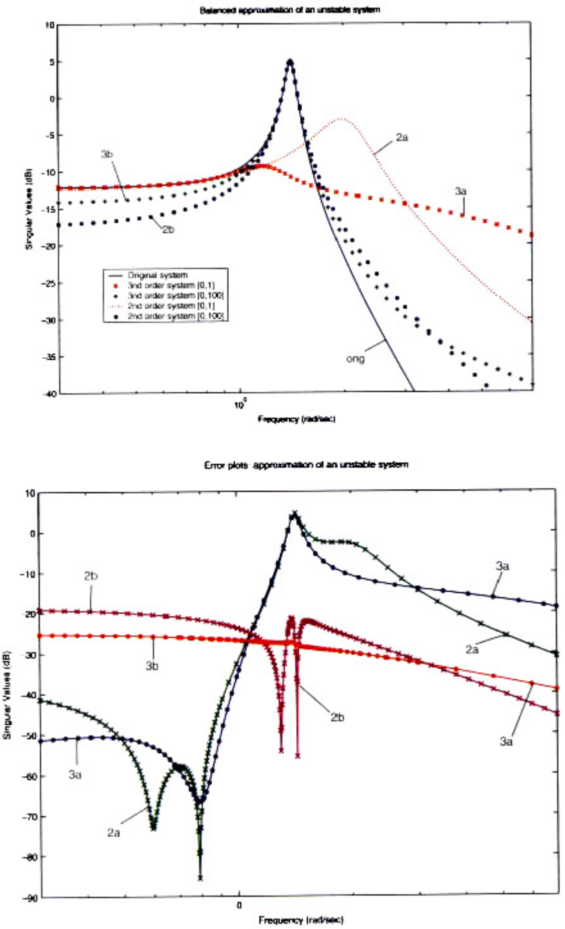


Figure 7.2. *Approximation of the unstable system in Example 7.24 by balanced truncation.*

the frequency response well for low frequencies up to 1 Hz, while the approximants for $\omega = 100$ do a better job in approximating the peak. These results can also be observed in the frequency response plots of the error systems, which are depicted in the bottom pane of Figure 7.2.

Example 7.25. We consider a second unstable system defined as follows:

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cccccc|c} -1 & 0.0200 & -1.9800 & 2.9399 & -1.0201 & 2.0402 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right].$$

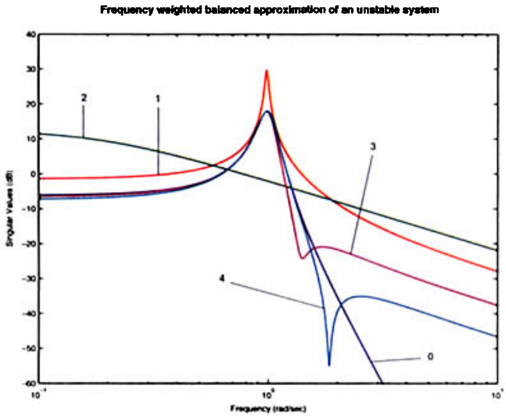


Figure 7.3. *Approximation of the unstable system in Example 7.25 by balanced truncation.*

The poles of the system are 1, -2, $-1 \pm i$, $1 \pm i$. We compute second- and fourth-order approximants obtained by frequency weighted balanced truncation with horizon $[-1, 1]$ and with infinite horizon. The original system is denoted with the subscript 0, the second-order approximant obtained with infinite horizon with 1, and the one with horizon $[-1, 1]$ by 2. Subsequently, 3 is the fourth-order system with horizon $[-1, 1]$ and 4 the infinite horizon fourth-order model. The resulting frequency responses are shown in Figure 7.3. Notice that the second-order infinite horizon approximant misses the resonance while the finite horizon one does not. Both fourth-order approximants reproduce the peak, but the infinite horizon fourth-order one gives the overall better approximation.

7.7 Chapter summary

A balanced realization of a system is one in which states that are difficult (easy) to reach are also difficult (easy) to observe. In this case, the minimal energy required to reach a given state is the inverse of the observation energy generated by the same state. Mathematically, balancing amounts to the simultaneous diagonalization of two positive (semi) definite gramians. These facts form the basis for one of the most effective approximation methods, namely, *approximation by balanced truncation*. It has the following properties: (i) stability is preserved and (ii) an a priori computable error bound exists. The latter provides a trade-off between *accuracy of the approximation* and *complexity of the approximant*, just as the singular values of a (constant) matrix provide a trade-off between accuracy and complexity of low rank approximants.

The main results and their proofs can be found in section 7.2. The third section is devoted to numerical issues; the difficulty in applying balanced truncation in large-scale settings is the fact that $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ storage are required. The main method used for computing the gramians is the square root (Hammarling's) method, which, as discussed in the previous chapter, yields directly the square roots (Cholesky factors) of the gramians, without computing them first. A canonical form for continuous-time balanced systems follows.

Subsequently, two specialized sections are presented, namely, additional types of balancing and frequency weighted balancing. They can be omitted at first reading. Since balancing consists of the simultaneous diagonalization of two positive (semi) definite matrices, different kinds of balancing, can be obtained by considering various Riccati equations. Besides the usual kind of balancing, called *Lyapunov balancing*, in section 7.5, we discuss *stochastic balancing*, *bounded real balancing*, and *positive real balancing*. Section 7.6 explores the issue of balancing which is *frequency selective*, both with and without the choice of frequency weights. The section concludes with time-limited balancing and balancing for unstable systems.

This page intentionally left blank

Chapter 8

Hankel-norm Approximation

As mentioned in section 7.1, a balanced state-space representation of a linear system $\boldsymbol{\Sigma}$ can be derived by minimizing an appropriate criterion. Nevertheless, approximation by balanced truncation does not seem to minimize any norm. A refinement leads to an approximation method which is optimal with respect to the 2-induced norm of the Hankel operator, known as, *optimal approximation in the Hankel-norm.* The resulting theory is the generalization to dynamical systems of the optimal approximation in the 2-induced norm of finite-dimensional matrices and operators discussed in section 3.2.4. It provides explicit formulas for optimal and suboptimal approximants as well as an error bound in the related $\mathcal{H}_\infty$-norm (the 2-induced norm of the convolution operator). The developments in this chapter are due to Adamjan, Arov, and Krein [1], [2] and especially to Glover [139]. See also [237], [137] and the books [149] and [370].

The theory of optimal/suboptimal approximation in the Hankel norm is presented in the first five sections of this chapter. Section 8.6 gives an exposition of a polynomial approach to Hankel-norm approximations. Although this is limited to SISO systems, it provides additional insights into the theory presented in the first part of the chapter. It shows, for instance, that the all-pass dilation system has additional structure, since it can be constructed using rational interpolation at the poles of $\boldsymbol{\Sigma}$.

8.1 Introduction

To successfully apply the SVD approach to the approximation of dynamical systems, we need to come up with the SVD of some input-output operator associated with such systems. For surveys, see [14], [17].

Given a linear system $\boldsymbol{\Sigma}$, the natural choice for an *input-output* operator is the *convolution operator* $\mathcal{S}$, defined by (4.5):

$$\mathcal{S} : \mathbf{u} \mapsto \mathbf{y} = \mathcal{S}(\mathbf{u})(t) = \int_{-\infty}^{\infty} \mathbf{h}(t-\tau)\mathbf{u}(\tau)d\tau, \qquad t \in \mathbb{R},$$

where $\mathbf{h}$ is the impulse response of the system. If the system is stable (the eigenvalues of $\mathbf{A}$ are in the left half of the complex plane), the 2-induced norm of $\mathcal{S}$ turns out to be equal to the $\mathcal{H}_\infty$-norm of the Laplace transform $\mathbf{H}(s)$ of $\mathbf{h}(t)$, that is, of the transfer function of $\mathbf{\Sigma}$ (see (5.16)). This is referred to as the $\mathcal{H}_\infty$-norm of the system. The operator $\mathcal{S}$ is not compact, however, and it does not have a *discrete* SVD (see section 5.2). The consequence of this lack of a discrete SVD is that the problem of approximating $\mathcal{S}$ by an operator of lower complexity that minimizes the 2-induced norm of the error *cannot* be solved at present, except in very special cases.

Searching for an operator associated with $\mathbf{\Sigma}$ that has a *discrete* SVD, we define one that makes use of the same convolution sum or integral but that has domain and range *different* from those of $\mathcal{S}$:

$$\mathcal{H} : \mathbf{u}_- \mapsto \mathbf{y}_+ = \mathcal{H}(\mathbf{u}_-)(t) = \int_{-\infty}^{0} \mathbf{h}(t-\tau)\mathbf{u}(\tau)d\tau, \qquad t \in \mathbb{R}_+.$$

This is the *Hankel operator* of $\mathbf{\Sigma}$ defined by means of (5.20); it maps *past inputs* $\mathbf{u}_-$ into *future outputs* $\mathbf{y}_+$. If the system is finite-dimensional, $\mathcal{H}$ has finite rank (that is, its range is a finite-dimensional space). Therefore (since it is bounded and compact), it possesses a *discrete* SVD, and in principle one could try to solve the optimal approximation problem, as defined above, in the 2-induced norm of the Hankel operator. As discussed in section 5.4.2, this induced norm is called the *Hankel-norm*, and the corresponding singular values are the *Hankel singular values* of the system.

It is a nontrivial fact that the optimal approximation problem in the 2-induced norm of $\mathcal{H}$ *can be solved*; the formulas obtained are explicit in the system parameters. This solution originated in the work of Adamjan, Arov, and Krein [1], [2], with subsequent contributions by Glover [139].

The main attributes of optimal approximation in the Hankel-norm are (i) preservation of stability and (ii) the existence of an error bound. The $\mathcal{H}_\infty$-norm of the error system is bounded by twice the sum of the neglected Hankel singular values $2(\sigma_{k+1} + \cdots + \sigma_n)$; this bound is the same as the one valid in the balanced truncation case and is reminiscent of the bound that holds for the optimal of approximation constant matrices in the 2-norm. It should also be mentioned that in some cases, balanced truncation is superior to Hankel-norm approximation, in the $\mathcal{H}_\infty$-norm. The reason is that Hankel-norm approximants are *optimal* in the Hankel-norm but not in the $\mathcal{H}_\infty$-norm.

8.1.1 Main ingredients

Hankel-norm approximation theory is based on the following four facts. Although we are seeking to approximate stable systems, the theory involves unstable systems. Therefore, formulas (5.38) and (5.39) of the 2-norm (which is equal to the $\mathcal{L}_\infty$-norm of the associated transfer function) and of the Hankel-norm of such systems should be kept in mind.

Fact I. Given stable systems $\mathbf{\Sigma}$, $\mathbf{\Sigma}'$ of McMillan degree n, k, respectively, where $n > k$, there holds

$$\| \mathbf{\Sigma} - \mathbf{\Sigma}' \|_H \geq \sigma_{k+1}(\mathbf{\Sigma}). \tag{8.1}$$

See section 8.4.2.

Fact II. The 2-norm of any $\mathcal{L}_2$-system $\boldsymbol{\Sigma}$ is no less than the Hankel-norm of its stable part $\boldsymbol{\Sigma}_+$:

$$\| \boldsymbol{\Sigma} \|_2 \geq \| \boldsymbol{\Sigma}_+ \|_H . \tag{8.2}$$

Fact III. Given a stable system $\boldsymbol{\Sigma}$, there exists a system $\hat{\boldsymbol{\Sigma}}$, which is in general not stable, having exactly k stable poles, such that

$$\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \|_2 = \sigma_{k+1}(\boldsymbol{\Sigma}). \tag{8.3}$$

Furthermore, $\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}$ is *all-pass*. This construction is discussed in section 8.4.5.

Fact IV. Given $k \in \{0, 1, 2, \ldots, n-1\}$, a stable system $\boldsymbol{\Sigma}$, and a positive real number ϵ such that

$$\sigma_k(\boldsymbol{\Sigma}) > \epsilon \geq \sigma_{k+1}(\boldsymbol{\Sigma}) \tag{8.4}$$

with $\sigma_0(\boldsymbol{\Sigma}) = \infty$, there exists a system $\hat{\boldsymbol{\Sigma}}$ having k stable poles, such that

$$\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}} \|_2 = \epsilon. \tag{8.5}$$

Furthermore, $\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}$ is *all-pass*. This construction, depicted in Figure 8.1, is discussed in section 8.4.4.

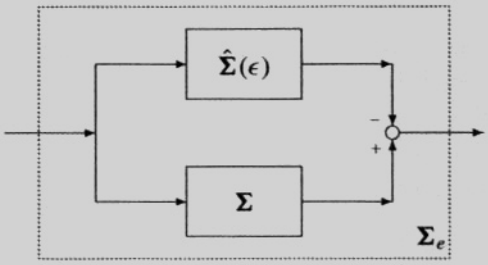


Figure 8.1. *Construction of approximants that are optimal and suboptimal in the Hankel-norm.*

Conclusion. The solution of the Hankel-norm approximation problem for both the optimal and the suboptimal cases follows as a consequence of the facts listed above: given $\boldsymbol{\Sigma}$ which is stable, construct $\hat{\boldsymbol{\Sigma}}$ so that the parallel connection $\boldsymbol{\Sigma}_e$ is all-pass (having both stable and unstable poles) and 2-norm equal to ϵ; then the stable part of $\hat{\boldsymbol{\Sigma}}$ is the optimal/suboptimal approximant of $\boldsymbol{\Sigma}$, and the associated error system has a Hankel-norm between σ_{k+1} and ϵ.

In the following sections we report the results that are valid for continuous-time systems. These results can also be used for approximating discrete-time systems by means of the bilinear transformation of subsection 4.3.3.

8.2 The Adamjan–Arov–Krein theorem

Consider the stable systems $\mathbf{\Sigma}$, $\mathbf{\Sigma}'$ of dimensions n, k, respectively. As shown earlier, the associated Hankel operator $\mathcal{H}_{\Sigma}$ has rank n and $\mathcal{H}_{\Sigma'}$ has rank k. Therefore, the Schmidt–Eckart–Young–Mirsky theorem implies that

$$\|\mathcal{H}_{\Sigma} - \mathcal{H}_{\Sigma'}\|_{2-\mathrm{ind}} \geq \sigma_{k+1}(\mathcal{H}_{\Sigma}). \tag{8.6}$$

The question that arises is to find the infimum of the above norm, given the fact that the approximant is structured (block Hankel matrix): $\inf_{\mathrm{rank}\,\Sigma'=k} \|\mathcal{H}_{\Sigma} - \mathcal{H}_{\Sigma'}\|_{2-\mathrm{ind}}$. A result due to Adamjan, Arov, and Krein (AAK) asserts that this lower bound is indeed attained for some $\mathbf{\Sigma}'$ of dimension k. The original sources for this result are [1] and [2].

Theorem 8.1 (AAK theorem). *Given the sequence of $p \times m$ matrices $\mathbf{h} = (\mathbf{h}(k))_{k>0}$, such that the associated Hankel matrix $\mathcal{H}$ has finite rank n, there exists a sequence of $p \times m$ matrices $\mathbf{h}_* = (\mathbf{h}_*(k))_{k>0}$ such that the associated Hankel matrix $\mathcal{H}_*$ has rank k and, in addition*

$$\|\mathcal{H} - \mathcal{H}_*\|_{2-\mathrm{ind}} = \sigma_{k+1}(\mathcal{H}). \tag{8.7}$$

If $p = m = 1$, the optimal approximant is unique.

This result says that every stable and causal system $\mathbf{\Sigma}$ can be optimally approximated by a stable and causal system $\mathbf{\Sigma}_*$ of lower dimension. *Optimal* here means

$$\inf \|\mathcal{H} - \mathcal{H}_*\| = \inf \|\mathcal{H} - \mathcal{K}\|,$$

where the first infimum is taken over all *Hankel* matrices $\mathcal{H}_*$ and the second over all *arbitrary* matrices $\mathcal{K}$. The optimality is with respect to the 2-induced norm of the associated Hankel operator. Notice that the above theorem holds for continuous-time systems as well.

8.3 The main result

In this section we present the main result. As it turns out, one can consider both suboptimal and optimal approximants within the same framework. Actually, the formulas for suboptimal approximants are simpler than their optimal counterparts. Both continuous- and discrete-time systems can also be handled within the same framework.

Problem 8.3.1. *Given a stable system $\mathbf{\Sigma}$, we seek stable approximants $\mathbf{\Sigma}_*$ of dimension k, satisfying*

$$\sigma_{k+1}(\mathbf{\Sigma}) \leq \|\mathbf{\Sigma} - \mathbf{\Sigma}_*\|_H \leq \epsilon < \sigma_k(\mathbf{\Sigma}).$$

This is a generalization of the optimal approximation in the Hankel-norm, solved by the AAK theorem. The concept introduced in the next definition is the key to its solution.

Definition 8.2. *Let $\mathbf{\Sigma}_e$ be the parallel connection of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$: $\mathbf{\Sigma}_e = \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}$. If $\mathbf{\Sigma}_e$ is an all-pass system with norm ϵ, $\hat{\mathbf{\Sigma}}$ is called an ϵ-all-pass dilation of $\mathbf{\Sigma}$.*

As a consequence of the inertia result of section 6.2, all-pass dilation systems have the following crucial property.

Lemma 8.3.1 (main lemma). *Let* $\hat{\Sigma}$, *with* $\dim \hat{\Sigma} \leq \dim \Sigma$, *be an* ϵ*-all-pass dilation of* Σ, *where* ϵ *satisfies* (8.4). *Then* $\hat{\Sigma}$ *has exactly* k stable *poles, i.e.,* $\dim \hat{\Sigma}_+ = k$.

Recall that stability in the continuous-time case means that the eigenvalues of $\mathbf{A}$ are in the left half of the complex plane, while in the discrete-time case, they are inside the unit disc. The analogue of the Schmidt–Eckart–Young–Mirsky result (8.6) for dynamical systems is stated next; it is proved in section 8.4.2.

Proposition 8.3. *Given the stable system* Σ, *let* Σ' *have at most k stable poles. There holds*

$$\|\Sigma - \Sigma'\|_H \geq \sigma_{k+1}(\Sigma).$$

This means that the 2-induced norm of the Hankel operator of the difference between Σ and Σ' is no less than the $(k+1)$st singular value of the Hankel operator of Σ. Finally, recall that if a system has both stable and unstable poles, its Hankel-norm is that of its stable part. We are now ready for the main result.

Theorem 8.4. *Let* $\hat{\Sigma}$ *be an* ϵ*-all-pass dilation of the linear, stable, discrete-, or continuous-time system* Σ, *where*

$$\sigma_{k+1}(\Sigma) \leq \epsilon < \sigma_k(\Sigma). \tag{8.8}$$

It follows that $\hat{\Sigma}_+$ *has exactly k stable poles and consequently*

$$\sigma_{k+1}(\Sigma) \leq \|\Sigma - \hat{\Sigma}\|_H < \epsilon. \tag{8.9}$$

In case $\sigma_{k+1}(\Sigma) = \epsilon$,

$$\sigma_{k+1}(\Sigma) = \|\Sigma - \hat{\Sigma}\|_H.$$

Proof. The result is a consequence of the following sequence of equalities and inequalities:

$$\sigma_{k+1}(\Sigma) \leq \|\Sigma - \hat{\Sigma}_+\|_H = \|\Sigma - \hat{\Sigma}\|_H \leq \|\Sigma - \hat{\Sigma}\|_\infty = \epsilon.$$

The first inequality on the left side is a consequence of Lemma 8.3.1, the equality follows by definition, the second inequality follows from (5.21), and the last equality holds by construction, since $\Sigma - \hat{\Sigma}$ is ϵ-all-pass. □

Remark 8.3.1. **(a)** An important special case of the above problem is obtained for $k = 0$ in (8.4), (8.5) and for $\epsilon = \sigma_1(\Sigma)$ in (8.8). This is the *Nehari problem.* It seeks to find the distance of a stable system from the set of antistable systems (i.e., systems whose poles are all unstable). The construction mentioned above implies that this distance is equal to the Hankel-norm of Σ:

$$\inf \| \Sigma - \hat{\Sigma} \|_2 \;=\; \| \Sigma \|_H = \sigma_1(\Sigma), \tag{8.10}$$

where the infimum is taken over all antistable systems $\hat{\Sigma}$. The Nehari problem and associated construction of $\hat{\Sigma}$ form one of the cornerstones of Hankel-norm approximation theory.

(b) We are given a stable system Σ and seek to compute an approximant in the same class (i.e., stable). To achieve this, the construction given above takes us *outside* this class of systems, since the all-pass dilation system $\hat{\Sigma}$ has poles that are both stable and unstable. In terms of matrices, we start with a system whose convolution operator $\mathcal{S}_\Sigma$ is a (block) lower triangular Toeplitz matrix. We then compute a (block) Toeplitz matrix $\mathcal{S}_{\hat{\Sigma}}$, which is no longer lower triangular, such that the difference is unitary. It then follows that the lower left portion of $\mathcal{S}_{\hat{\Sigma}}$, which is the Hankel matrix $\mathcal{H}_{\hat{\Sigma}}$, has rank r and approximates the Hankel matrix $\mathcal{H}_\Sigma$, so that the 2-norm of the error satisfies (8.8).

(c) The suboptimal and optimal approximants can be constructed using explicit formulas. For continuous-time systems, see sections 8.4.4 and 8.4.5.

(d) It is interesting to notice that while the AAK result settles the problem of optimal approximation of *infinite-dimensional* Hankel matrices of finite rank, with Hankel matrices of lower rank, the problem of optimal approximation concerning *finite-dimensional* Hankel matrices is still open. A special case of this problem is treated in [13].

8.4 Construction of approximants

We are ready to give some of the formulas for the construction of suboptimal and optimal Hankel-norm approximants. As mentioned earlier, all formulas describe the construction of all-pass dilation systems (see 8.4). We will discuss the following cases:

- An input-output construction applicable to scalar systems. Both optimal and suboptimal approximants are treated. The advantage of this approach is that the equations can be set up in a straightforward manner using the numerator and denominator polynomials of the transfer function of the given system (section 8.4.1). The drawback is that the proof of the main lemma, Lemma 8.3.1, is *not* easy in this framework.

- A state space based construction method for suboptimal approximants (section 8.4.4) and for optimal approximants (section 8.4.5) of square systems.

- A state space based parametrization of *all* suboptimal approximants for general (i.e., not necessarily square) systems (section 8.4.6).

- The optimality of the approximants is with respect to the Hankel-norm (2-induced norm of the Hankel operator). Section 8.5 gives an account of error bounds for the infinity norm of the approximants (2-induced norm of the convolution operator).

8.4.1 A simple input-output construction method

Given the polynomials $\mathbf{a} = \sum_{i=0}^{\alpha} a_i s^i$, $\mathbf{b} = \sum_{i=0}^{\beta} b_i s^i$, $\mathbf{c} = \sum_{i=0}^{\gamma} c_i s^i$ satisfying $\mathbf{c} = \mathbf{ab}$, the coefficients of the product $\mathbf{c}$ are a linear combination of those of $\mathbf{b}$:

$$\underline{\mathbf{c}} = \mathbb{T}(\mathbf{a})\underline{\mathbf{b}},$$

where $\underline{\mathbf{c}} = (c_\gamma \; c_{\gamma-1} \; \cdots \; c_1 \; c_0)^* \in \mathbb{R}^{\gamma+1}$, $\underline{\mathbf{b}} = (b_\beta \; \cdots \; b_0)^* \in \mathbb{R}^{\beta+1}$, and $\mathbb{T}(\mathbf{a})$ is a Toeplitz matrix with first column $(a_\alpha \; \cdots \; a_0 \; 0 \; \cdots \; 0)^* \in \mathbb{R}^{\gamma+1}$ and first row $(a_\alpha \; 0 \; \cdots \; 0) \in \mathbb{R}^{1\times(\beta+1)}$.

We also define the sign matrix whose last diagonal entry is 1,

$$\mathbb{J} = \mathrm{diag}\,(\ldots, 1, -1, 1),$$

of appropriate size. Given a polynomial $\mathbf{c}$ with real coefficients, the polynomial $\mathbf{c}^*$ is defined as $\mathbf{c}(s)^* = \mathbf{c}(-s)$. This means that

$$\underline{\mathbf{c}}^* = \mathbb{J}\,\underline{\mathbf{c}}.$$

The basic construction given in Theorem 8.4 hinges on the construction of an ϵ-all-pass dilation $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$. Let

$$\mathbf{H}_{\Sigma}(s) = \frac{\mathbf{p}(s)}{\mathbf{q}(s)}, \quad \mathbf{H}_{\hat{\Sigma}}(s) = \frac{\hat{\mathbf{p}}(s)}{\hat{\mathbf{q}}(s)}.$$

We require that the difference $\mathbf{H}_{\Sigma} - \mathbf{H}_{\hat{\Sigma}} = \mathbf{H}_{\Sigma_e}$ be ϵ-all-pass. Therefore, the problem is given ϵ, and the polynomials $\mathbf{p}$, $\mathbf{q}$, such that $\deg(\mathbf{p}) \le \deg(\mathbf{q}) = n$, find polynomials $\hat{\mathbf{p}}$, $\hat{\mathbf{q}}$, of degree at most n such that

$$\frac{\mathbf{p}}{\mathbf{q}} - \frac{\hat{\mathbf{p}}}{\hat{\mathbf{q}}} = \epsilon\frac{\mathbf{q}^*\hat{\mathbf{q}}^*}{\mathbf{q}\hat{\mathbf{q}}} \quad \Leftrightarrow \quad \mathbf{p}\hat{\mathbf{q}} - \mathbf{q}\hat{\mathbf{p}} = \epsilon\mathbf{q}^*\hat{\mathbf{q}}^*.$$

This polynomial equation can be rewritten as a matrix equation involving the quantities defined above:

$$\mathbb{T}(\mathbf{p})\underline{\hat{\mathbf{q}}} - \mathbb{T}(\mathbf{q})\underline{\hat{\mathbf{p}}} = \epsilon\mathbb{T}(\mathbf{q}^*)\underline{\hat{\mathbf{q}}^*} = \epsilon\mathbb{T}(\mathbf{q}^*)\mathbb{J}\underline{\hat{\mathbf{q}}}.$$

Collecting terms, we have

$$\left(\mathbb{T}(\mathbf{p}) - \epsilon\mathbb{T}(\mathbf{q}^*)\mathbb{J}, \quad -\mathbb{T}(\mathbf{q})\right)\begin{pmatrix} \underline{\hat{\mathbf{q}}} \\ \underline{\hat{\mathbf{p}}} \end{pmatrix} = \mathbf{0}.$$

The solution of this set of linear equations provides the coefficients of the ϵ-all-pass dilation system $\hat{\mathbf{\Sigma}}$. Furthermore, this system can be solved for both the suboptimal $\epsilon \neq \sigma_i$ and the optimal $\epsilon = \sigma_i$ cases. We illustrate the features of this approach by means of a simple example. For an alternative approach along similar lines, see [121].

Example 8.5. Let $\mathbf{\Sigma}$ be a second-order system, i.e., $n = 2$. If we normalize the coefficient of the highest power of $\hat{q}$, i.e., $\hat{q}_2 = 1$, we obtain the following system of equations:

$$\underbrace{\begin{pmatrix} 0 & 0 & q_2 & 0 & 0 \\ p_2 - \epsilon q_2 & 0 & q_1 & q_2 & 0 \\ p_1 + \epsilon q_1 & p_2 + \epsilon q_2 & q_0 & q_1 & q_2 \\ p_0 - \epsilon q_0 & p_1 - \epsilon q_1 & 0 & q_0 & q_1 \\ 0 & p_0 + \epsilon q_0 & 0 & 0 & q_0 \end{pmatrix}}_{\mathbf{W}(\epsilon)}\begin{pmatrix} \hat{q}_1 \\ \hat{q}_0 \\ \hat{p}_2 \\ \hat{p}_1 \\ \hat{p}_0 \end{pmatrix} = -\begin{pmatrix} p_2 + \epsilon q_2 \\ p_1 - \epsilon q_1 \\ p_0 + \epsilon q_0 \\ 0 \\ 0 \end{pmatrix}.$$

This can be solved for all ϵ that are not roots of the equation $\det \mathbf{W}(\epsilon) = 0$. The latter is a polynomial equation of second degree; there are thus two values of ϵ, ϵ_1 and ϵ_2, for which the determinant of $\mathbf{W}$ is zero. It can be shown that the roots of this determinant are the ***eigenvalues*** of the Hankel operator $\mathcal{H}_{\Sigma}$; since in the SISO case $\mathcal{H}_{\Sigma}$ is self-adjoint (symmetric), the absolute values of ϵ_1, ϵ_2 are the singular values of $\mathcal{H}_{\Sigma}$. Thus both suboptimal and optimal approximants can be computed this way.

8.4.2 An analogue of Schmidt–Eckart–Young–Mirsky

In this section, we prove Proposition 8.3. Let $(\mathbf{H}, \mathbf{F}, \mathbf{G})$ and $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ be realizations of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$, respectively, where $\mathbf{\Sigma}$ has McMillan degree n and $\mathbf{\Sigma}'$ has McMillan degree $k < n$. A realization of the difference $\mathbf{\Sigma}_e = \mathbf{\Sigma} - \mathbf{\Sigma}'$ is given by

$$\mathbf{H}_e = (\mathbf{H} \quad -\mathbf{C}), \quad \mathbf{F}_e = \begin{pmatrix} \mathbf{F} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{pmatrix}, \quad \mathbf{G}_e = \begin{pmatrix} \mathbf{G} \\ \mathbf{B} \end{pmatrix}.$$

By definition,

$$\| \mathbf{\Sigma} - \mathbf{\Sigma}' \|_H^2 = \lambda_{max}(\mathcal{P}_e \mathcal{Q}_e),$$

where $\mathcal{P}_e$, $\mathcal{Q}_e$ are the reachability, observability, gramians of $\mathbf{\Sigma}_e$, satisfying

$$\mathbf{F}_e\mathcal{P}_e + \mathcal{P}_e\mathbf{F}_e^* + \mathbf{G}_e\mathbf{G}_e^* = \mathbf{0}, \ \mathbf{F}_e^*\mathcal{Q}_e + \mathcal{Q}_e\mathbf{F}_e + \mathbf{H}_e^*\mathbf{H}_e = \mathbf{0}, \ \mathcal{P}_e, \mathcal{Q}_e \in \mathbb{R}^{(n+k)\times(n+k)}.$$

Consider the Cholesky factorization of $\mathcal{P}_e$:

$$\mathcal{P}_e = \mathbf{R}^*\mathbf{R}, \quad \text{where } \mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{R}_{22} \end{pmatrix} \text{ and } \mathcal{P}_e = \begin{pmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} \\ \mathcal{P}_{12}^* & \mathcal{P}_{22} \end{pmatrix}.$$

The following relationships hold between the P_{ij} and R_{ij}:

$$\mathcal{P}_{11} = \mathbf{R}_{11}^*\mathbf{R}_{11}, \quad \mathcal{P}_{12} = \mathbf{R}_{11}^*\mathbf{R}_{12}, \quad \mathcal{P}_{22} = \mathbf{R}_{12}^*\mathbf{R}_{12} + \mathbf{R}_{22}^*\mathbf{R}_{22}.$$

Notice that

$$\lambda_{max}(\mathcal{P}_e\mathcal{Q}_e) = \lambda_{max}(\mathbf{R}\mathbf{R}^*\mathcal{Q}_e) = \lambda_{max}(\mathbf{R}^*\mathcal{Q}_e\mathbf{R}) \geq \lambda_{max}(\mathbf{R}^*\mathcal{Q}_e\mathbf{R})_{11},$$

where the subscript $(\ \cdot\ )_{11}$ denotes the block $(1, 1)$ element of $(\ \cdot\ )$. Furthermore,

$$(\mathbf{R}^*\mathcal{Q}_e\mathbf{R})_{11} = (\mathbf{R}_{11} \quad \mathbf{R}_{12})\mathcal{Q}_e(\mathbf{R}_{11} \quad \mathbf{R}_{12})^* = \mathbf{R}_{11}(\mathcal{Q}_e)_{11}\mathbf{R}_{11}^* + \mathbf{X},$$

where

$$\mathbf{X} = (\mathbf{R}_{11} \quad \mathbf{R}_{12}) \begin{pmatrix} \mathbf{0} & \mathcal{Q}_{12} \\ \mathcal{Q}_{12}^* & \mathcal{Q}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{11}^* \\ \mathbf{R}_{12}^* \end{pmatrix} \geq \mathbf{0}, \quad \operatorname{rank} \mathbf{X} \leq k.$$

Since $\operatorname{rank} \mathbf{X} \leq k$, we conclude that

$$\lambda_{max}(\mathbf{R}^*\mathcal{Q}_e\mathbf{R})_{11} \geq \lambda_{max}(\mathbf{R}_{11}(\mathcal{Q}_e)_{11}\mathbf{R}_{11}^* + \mathbf{X}) \geq \lambda_{k+1}(\mathbf{R}_{11}(\mathcal{Q}_e)_{11}\mathbf{R}_{11}^*) = \sigma_{k+1}(\mathbf{\Sigma}).$$

This completes the proof of Proposition 8.3.

8.4.3 Unitary dilation of constant matrices

In this section, we investigate the problem of augmenting a matrix so that it becomes unitary or augmenting two matrices so that their product becomes the identity. We proceed from the simple to the more involved cases.

1. Given $\sigma \in \mathbb{R}$, $|\sigma| < 1$, find $\mathbf{M}$, where

$$\mathbf{M} = \begin{pmatrix} \sigma & ? \\ ? & ? \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

such that $\mathbf{M}\mathbf{M}^* = \mathbf{I}$. *Solution*:

$$\mathbf{M} = \begin{pmatrix} \sigma & -\sqrt{1-\sigma^2} \\ \sqrt{1-\sigma^2} & \sigma \end{pmatrix}.$$

2. Given $\sigma \in \mathbb{R}$, $|\sigma| > 1$, find $\mathbf{M}, \mathbf{N}$, where

$$\mathbf{M} = \begin{pmatrix} \sigma & ? \\ ? & ? \end{pmatrix} \in \mathbb{R}^{2\times 2},\ \mathbf{N} = \begin{pmatrix} \sigma & ? \\ ? & ? \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

such that $\mathbf{MN} = \mathbf{I}$. *Solution*:

$$\mathbf{M} = \begin{pmatrix} \sigma & 1-\sigma^2 \\ 1-\sigma^2 & -\sigma(1-\sigma^2) \end{pmatrix},\ \mathbf{N} = \begin{pmatrix} \sigma & 1 \\ 1 & -\frac{\sigma}{1-\sigma^2} \end{pmatrix}.$$

3. Given $p, q \in \mathbb{R}$, $pq \neq 1$, find $\mathbf{M}, \mathbf{N}$, where

$$\mathbf{M} = \begin{pmatrix} p & ? \\ ? & ? \end{pmatrix},\ \mathbf{N} = \begin{pmatrix} q & ? \\ ? & ? \end{pmatrix}$$

such that $\mathbf{MN} = \mathbf{I}$. *Solution*:

$$\mathbf{M} = \begin{pmatrix} p & 1-pq \\ 1-pq & -q(1-pq) \end{pmatrix},\ \mathbf{N} = \begin{pmatrix} q & 1 \\ 1 & -\frac{p}{1-pq} \end{pmatrix}.$$

4. Given $\Sigma = \mathrm{diag}\,(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_i \neq 1$, find

$$\mathbf{M} = \begin{pmatrix} \Sigma & ? \\ ? & ? \end{pmatrix} \in \mathbb{R}^{2n\times 2n},\ \mathbf{N} = \begin{pmatrix} \Sigma & ? \\ ? & ? \end{pmatrix} \in \mathbb{R}^{2n\times 2n}$$

such that $\mathbf{MN} = \mathbf{I}$. *Solution*: Let $\Gamma = \mathbf{I} - \Sigma^2$,

$$\mathbf{M} = \begin{pmatrix} \Sigma & \Gamma \\ \Gamma & -\Sigma\Gamma \end{pmatrix},\ \mathbf{N} = \begin{pmatrix} \Sigma & \mathbf{I} \\ \mathbf{I} & -\Gamma^{-1}\Sigma \end{pmatrix}.$$

5. Given $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$, $\lambda_i(\mathbf{PQ}) \neq 1$, find

$$\mathbf{P}_e = \begin{pmatrix} \mathbf{P} & ? \\ ? & ? \end{pmatrix},\ \mathbf{Q}_e = \begin{pmatrix} \mathbf{Q} & ? \\ ? & ? \end{pmatrix}$$

such that $\mathbf{P}_e\mathbf{Q}_e = \mathbf{I}$. *Solution*: Let $\Gamma = \mathbf{I} - \mathbf{PQ}$,

$$\mathbf{P}_e = \begin{pmatrix} \mathbf{P} & \Gamma \\ \Gamma^* & -\mathbf{Q}\Gamma \end{pmatrix},\ \mathbf{Q}_e = \begin{pmatrix} \mathbf{Q} & \mathbf{I} \\ \mathbf{I} & -\Gamma^{-1}\mathbf{P} \end{pmatrix}. \tag{8.11}$$

The last dilation, (8.11), is now used to construct suboptimal approximants.

8.4.4 Unitary system dilation: Suboptimal case

Recall the definition of all-pass or unitary systems given in section 5.8.3.

Problem 8.4.1. *Given* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathcal{P}$, $\mathcal{Q}$, $p = m$, *find* $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{D}}$ *such that*

$$\mathbf{A}_e = \begin{pmatrix} \mathbf{A} & \\ & \hat{\mathbf{A}} \end{pmatrix}, \ \mathbf{B}_e = \begin{pmatrix} \mathbf{B} \\ \hat{\mathbf{B}} \end{pmatrix}, \ \mathbf{C}_e = (\mathbf{C} \quad -\hat{\mathbf{C}}), \ \mathbf{D}_e = -\hat{\mathbf{D}}, \tag{8.12}$$

$\boldsymbol{\Sigma}_e = \left(\begin{array}{c|c} \mathbf{A}_e & \mathbf{B}_e \\ \hline \mathbf{C}_e & \mathbf{D}_e \end{array}\right)$ *is unitary (all-pass).*

Solution. Assume $\sigma_k > 1 > \sigma_{k+1}$. Following the dilation (8.11), we define

$$\mathcal{Q}_e = \begin{pmatrix} \mathcal{Q} & \mathbf{I} \\ \mathbf{I} & -\Gamma^{-1}\mathcal{P} \end{pmatrix}, \ \Gamma = \mathbf{I} - \mathcal{P}\mathcal{Q}. \tag{8.13}$$

The conditions for the dilation to be all-pass given in part 3 of Theorem 5.23 are

$$\begin{aligned} &\text{(i)} && \mathbf{D}_e^*\mathbf{D}_e = \mathbf{I}, \\ &\text{(ii)} && \mathbf{A}_e^*\mathcal{Q}_e + \mathcal{Q}_e\mathbf{A}_e + \mathbf{C}_e^*\mathbf{C}_e = \mathbf{0}, \\ &\text{(iii)} && \mathcal{Q}_e\mathbf{B}_e + \mathbf{C}_e^*\mathbf{D}_e = \mathbf{0}. \end{aligned}$$

Solving these relationships for the quantities sought, we obtain

$$\boxed{\begin{aligned} \hat{\mathbf{B}} &= -\mathcal{Q}\mathbf{B} + \mathbf{C}^*\hat{\mathbf{D}}, \\ \hat{\mathbf{C}} &= (\mathbf{C}\mathcal{P} - \hat{\mathbf{D}}\mathbf{B}^*)(\Gamma^*)^{-1}, \\ \hat{\mathbf{A}} &= -\mathbf{A}^* + \mathbf{C}^*\hat{\mathbf{C}}. \end{aligned}} \tag{8.14}$$

Furthermore, $\hat{\mathbf{A}}^*\hat{\mathcal{Q}} + \hat{\mathcal{Q}}\hat{\mathbf{A}} + \hat{\mathbf{C}}^*\hat{\mathbf{C}} = \mathbf{0}$, where $\hat{\mathcal{Q}} = -(\Gamma^*)^{-1}\mathcal{P}$. Hence, if $\sigma_{k+1} < 1 < \sigma_k$, it follows that $-\Gamma$ has k positive eigenvalues, and consequently $-\Gamma^{-1}\mathcal{P}$ has k positive eigenvalues. We conclude that $\hat{\mathbf{A}}$ has k eigenvalues in the left half of the complex place $\mathbb{C}_-$.

Suboptimal Hankel-norm approximation

Let $\hat{\mathbf{H}}(s) = \hat{\mathbf{D}} + \hat{\mathbf{C}}(s\mathbf{I} - \hat{\mathbf{A}})^{-1}\hat{\mathbf{B}}$. Decompose $\hat{\mathbf{H}}$ in a stable and an antistable part $\hat{\mathbf{H}} = \hat{\mathbf{H}}_+ + \hat{\mathbf{H}}_-$. Then due to Theorem 8.4, the desired

$$\sigma_{k+1} < \| \mathbf{H} - \hat{\mathbf{H}}_+ \|_H < 1$$

holds true.

8.4.5 Unitary system dilation: Optimal case

We now discuss the construction of solutions in the optimal case, i.e., the case where 1 is a singular value of multiplicity r of the system to be approximated. Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathcal{P}$, $\mathcal{Q}$,

$p = m$, be partitioned as follows:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} \\ \mathbf{B}_{21} \end{pmatrix}, \ \mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \end{pmatrix},$$

$$\mathcal{P} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathcal{P}_{22} \end{pmatrix}, \ \mathcal{Q} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathcal{Q}_{22} \end{pmatrix}, \tag{8.15}$$

where $\mathbf{A}_{11} \in \mathbb{R}^{r \times r}$, $\mathbf{B}_{11}, \mathbf{C}_{11}^* \in \mathbb{R}^r$, and $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. First, notice that $\mathbf{A}_{11}$, $\mathbf{B}_{11}$, and $\mathbf{C}_{11}$ satisfy

$$\mathbf{A}_{11} + \mathbf{A}_{11}^* + \mathbf{B}_{11}\mathbf{B}_{11}^* = \mathbf{0}, \ \mathbf{A}_{11} + \mathbf{A}_{11}^* + \mathbf{C}_{11}^*\mathbf{C}_{11} = \mathbf{0}.$$

Consequently, $\mathbf{B}_{11}\mathbf{B}_{11}^* = \mathbf{C}_{11}^*\mathbf{C}_{11}$, which implies the existence of a unitary matrix $\hat{\mathbf{D}}$ such that

$$\mathbf{B}_{11}\hat{\mathbf{D}} = \mathbf{C}_{11}^*.$$

Using formulas (8.14), we construct an all-pass dilation of the subsystem $(\mathbf{A}_{22}, \mathbf{B}_{21}, \mathbf{C}_{12})$, namely,

$$\left.\begin{aligned} \hat{\mathbf{B}} &= -\mathcal{Q}_{22}\mathbf{B}_{21} + \mathbf{C}_{12}^*\hat{\mathbf{D}} \\ \hat{\mathbf{C}} &= (\mathbf{C}_{12}\mathcal{P}_{22} - \hat{\mathbf{D}}\mathbf{B}_{21}^*)(\Gamma_2^*)^{-1} \\ \hat{\mathbf{A}} &= -\mathbf{A}_{22}^* + \mathbf{C}_{12}^*\hat{\mathbf{C}} \end{aligned}\right\}, \tag{8.16}$$

where $\Gamma_2 = \mathbf{I} - \mathcal{P}_{22}\mathcal{Q}_{22} \in \mathbb{R}^{(n-r)\times(n-r)}$, and $\hat{\mathbf{D}}$ is the unitary matrix defined above. Hence, unlike the suboptimal case, $\hat{\mathbf{D}}$ is not arbitrary. In the SISO case, $\hat{\mathbf{D}}$ is completely determined by the above relationship (it is either $+1$ or -1), and hence there is a unique optimal approximant. This is not true for systems with more than one input and/or more than one output.

Lemma 8.6. *The system* $\mathbf{A}_e$, $\mathbf{B}_e$, $\mathbf{C}_e$, $\mathbf{D}_e$ *defined by* (8.12) *where* $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$, $\hat{\mathbf{D}}$ *is defined by* (8.16) *is all-pass. Moreover,* $\hat{\mathbf{A}}$ *has k eigenvalues in the left half plane and $n - r - k$ eigenvalues in the right half plane.*

Example 8.7. Consider a system of McMillan degree 4 with Hankel singular values $\sigma_1 > \sigma_2 > \sigma_3 > \sigma_4 > 0$. Let δ denote the McMillan degree of $\hat{\Sigma}$ and δ_+ the McMillan degree of $\hat{\Sigma}_+$; i.e., δ denotes the number of poles of the dilation system, while δ_+ denotes the number of *stable* poles of the same system. Figure 8.2 shows δ and δ_+ as a function of γ.

8.4.6 All suboptimal solutions

Above, we provided a way to construct one suboptimal approximant of a square system. It is, however, rather straightforward to provide a parametrization of all suboptimal approximants of an arbitrary system Σ. The exposition below follows section 24.3 in the book by Ball, Gohberg, and Rodman [38].

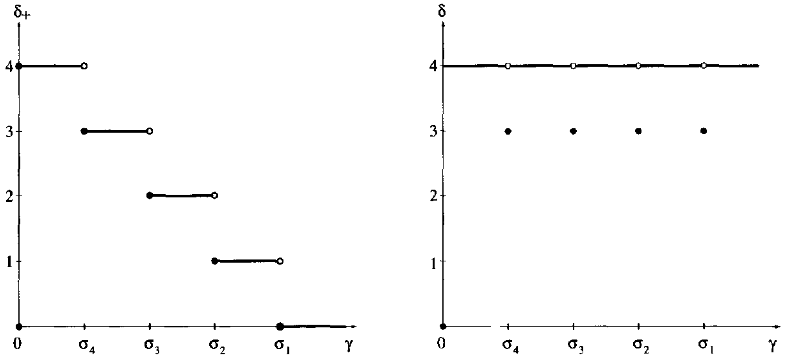


Figure 8.2. *Hankel-norm approximation: dimension δ, δ_+ of all-pass dilation systems, stable part of all-pass dilation systems; the system to be approximated has dimension $n = 4$. The abscissa is γ.*

Given the system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ with m inputs and p outputs, let $\sigma_k(\mathbf{\Sigma}) > \epsilon > \sigma_{k+1}(\mathbf{\Sigma})$. The following rational matrix $\Theta(s)$ has dimension $(p+m) \times (p+m)$:

$$\Theta(s) = \mathbf{I}_{p+m} - \mathbf{H}_\Theta(s\mathbf{I} - \mathbf{F}_\Theta)^{-1}\mathcal{Q}_\Theta^{-1}\mathbf{H}_\Theta^*\mathbf{J},$$

where

$$\mathbf{H}_\Theta = \begin{pmatrix} \frac{1}{\epsilon}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^* \end{pmatrix} \in \mathbb{R}^{(p+m)\times 2n}, \quad \mathbf{F}_\Theta = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}^* \end{pmatrix} \in \mathbb{R}^{2n\times 2n},$$

$$\Gamma = \left(\mathbf{I} - \frac{1}{\epsilon^2}\mathcal{P}\mathcal{Q}\right), \quad \mathbf{J} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_m \end{pmatrix} \in \mathbb{R}^{(p+m)\times(p+m)}, \tag{8.17}$$

$$\mathcal{Q}_\Theta = \begin{pmatrix} \frac{1}{\epsilon^2}\mathcal{Q} & \mathbf{I}_n \\ \mathbf{I}_n & \mathcal{P} \end{pmatrix} \Rightarrow \mathcal{Q}_\Theta^{-1} = \begin{pmatrix} -\Gamma^{-1}\mathcal{P} & \Gamma^{-1} \\ \Gamma^{-*} & -\frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1} \end{pmatrix} \in \mathbb{R}^{2n\times 2n};$$

thus

$$\Theta(s) = \begin{pmatrix} \mathbf{I}_p + \frac{1}{\epsilon^2}\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\Gamma^{-1}\mathcal{P}\mathbf{C}^* & \frac{1}{\epsilon}\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\Gamma^{-1}\mathbf{B} \\ -\frac{1}{\epsilon}\mathbf{B}^*(s\mathbf{I}+\mathbf{A}^*)^{-1}\Gamma^{-*}\mathbf{C}^* & \mathbf{I}_m - \frac{1}{\epsilon^2}\mathbf{B}^*(s\mathbf{I}+\mathbf{A}^*)^{-1}\mathcal{Q}\Gamma^{-1}\mathbf{B} \end{pmatrix}.$$

Notice that by construction Θ is $\mathbf{J}$-unitary on the $j\omega$-axis. Define

$$\begin{pmatrix} \Phi_1(\Delta) \\ \Phi_2(\Delta) \end{pmatrix} = \Theta\begin{pmatrix} \Delta \\ \mathbf{I} \end{pmatrix} = \begin{pmatrix} \Theta_{11}(s)\Delta(s) + \Theta_{12}(s) \\ \Theta_{21}(s)\Delta(s) + \Theta_{22}(s) \end{pmatrix}. \tag{8.18}$$

The following result holds.

Theorem 8.8. *$\hat{\mathbf{\Sigma}}$ is an approximant of $\mathbf{\Sigma}$ with k stable poles satisfying $\| \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_+ \|_H < \epsilon$ if and only if the associated transfer functions satisfy*

$$\mathbf{H}(s) - \hat{\mathbf{H}}(s) = \Phi_1(\Delta)\Phi_2(\Delta)^{-1} = \mathbf{Z}(s),$$

where $[\,\cdot\,]_+$ denotes the stable part of $[\,\cdot\,]$ and $\Delta(s)$ is a $p \times m$ antistable contraction, i.e.,

$$\| \Delta(s) \|_{\mathcal{H}_\infty^-} < 1.$$

Outline of proof. The proof can be divided into three parts.

(a) Let $\mathbf{Z}(\Delta) = \Phi_1(\Delta)\Phi_2(\Delta)^{-1}$, where $\Phi_i(\Delta)$, $i = 1, 2$, are defined by (8.18) and the $\mathcal{L}_\infty$-norm of Δ is less than 1. We show that the number of LHP poles of $\mathbf{Z}$ is $n + k + \nu$, where ν is the number of LHP poles of Δ. To prove this fact, we write $\mathbf{Z}$ as follows:

$$\mathbf{Z} = \underbrace{(\Theta_{11}(s)\Delta(s) + \Theta_{12}(s))}_{\psi_1}\underbrace{(\Theta_{22}(s)^{-1}\Theta_{21}(s)\Delta(s) + \mathbf{I})^{-1}}_{\psi_2^{-1}}\underbrace{\Theta_{22}(s)^{-1}}_{\psi_3}.$$

We examine the poles of each of the above terms. ψ_1 has $n + \nu$ poles in the LHP. By a homotopy argument, ψ_2^{-1} has no zeros in the LHP and hence ψ_2 has no poles there. The last term can be treated as follows. A realization for ψ_3^{-1} is

$$\psi_3^{-1} = \left(\begin{array}{c|c} -\mathbf{A}^* & \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B} \\ \hline -\mathbf{B}^* & \mathbf{I}_m \end{array}\right) \Rightarrow \psi_3 = \left(\begin{array}{c|c} -\mathbf{A}^* + \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B}\mathbf{B}^* & \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B} \\ \hline \mathbf{B}^* & \mathbf{I}_m \end{array}\right).$$

Using an inertia argument, $-\mathbf{A}^* + \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B}\mathbf{B}^*$ has exactly k eigenvalues in the LHP. This concludes the proof of (a).

(b) With $\mathbf{Z}$ as above and Δ antistable, we can write $\mathbf{Z} = \mathbf{H} - \hat{\mathbf{H}}$ for some $\hat{\mathbf{H}}$ having k stable poles. We show this only for $\Delta = \mathbf{0}$. In this case,

$$\begin{aligned}\mathbf{Z}(s) &= \frac{1}{\epsilon}\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\Gamma^{-1}\mathbf{B}\left[\mathbf{I}_m - \frac{1}{\epsilon^2}\mathbf{B}^*(s\mathbf{I} + \mathbf{A}^*)^{-1}\mathcal{Q}\Gamma^{-1}\mathbf{B}\right]^{-1} \\ &= \frac{1}{\epsilon}\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\Gamma^{-1}\mathbf{B}\left[\mathbf{I}_m + \frac{1}{\epsilon^2}\mathbf{B}^*\left(s\mathbf{I} + \mathbf{A}^* - \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B}\mathbf{B}^*\right)^{-1}\mathcal{Q}\Gamma^{-1}\mathbf{B}\right].\end{aligned}$$

Thus a realization of $\mathbf{Z}$ is given by

$$\left(\begin{array}{cc|c} -\mathbf{A}^* + \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B}\mathbf{B}^* & \mathbf{0} & \frac{1}{\epsilon}\mathcal{Q}\Gamma^{-1}\mathbf{B} \\ \frac{1}{\epsilon}\Gamma^{-1}\mathbf{B}\mathbf{B}^* & \mathbf{A} & \Gamma^{-1}\mathbf{B} \\ \hline \mathbf{0} & \mathbf{C} & \mathbf{0} \end{array}\right).$$

Applying the equivalence transformation $\mathbf{T} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \epsilon^{-1}\mathcal{P} & \mathbf{I} \end{pmatrix}$ (the new state is $\mathbf{T}$ times the old state), we obtain the realization

$$\left(\begin{array}{cc|c} -\mathbf{A}^* + \frac{1}{\epsilon^2}\mathcal{Q}\Gamma^{-1}\mathbf{B}\mathbf{B}^* & \mathbf{0} & \frac{1}{\epsilon}\mathcal{Q}\Gamma^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{A} & \mathbf{B} \\ \hline \frac{1}{\epsilon^2}\mathbf{C}\mathcal{P} & \frac{1}{\epsilon}\mathbf{C} & \mathbf{0} \end{array}\right).$$

This proves the desired decomposition for $\Delta = \mathbf{0}$.

(c) Given $\hat{\mathbf{H}}$ with k stable poles such that $\mathbf{H} - \hat{\mathbf{H}}$ is all-pass, there exists an antistable contractive Δ such that $\mathbf{H} - \hat{\mathbf{H}} = \mathbf{Z}(\Delta)$. Let $\Phi_1 = \mathbf{H} - \hat{\mathbf{H}}$, $\Phi_2 = \mathbf{I}$. Solving (8.18), we obtain

$$(\mathbf{I}\ \ \Delta) = (\mathbf{I}\ \ \Phi_1)\underbrace{\begin{pmatrix} -\Theta_{11} & \Theta_{12} \\ \Theta_{21} & -\Theta_{22} \end{pmatrix}}_{\bar{\Theta}}.$$

Since Θ is $\mathbf{J}$-unitary, so is $\bar{\Theta}$. Hence Δ has the correct $\mathcal{L}_\infty$-norm (i.e., size on the $j\omega$-axis). It remains to show that Δ is antistable. This follows from part (a) of the proof. □

8.5 Error bounds for optimal approximants

We now concentrate on error bounds that exist for optimal Hankel-norm approximants. For details, see section 9.2 of [139].

Given is a square system $\mathbf{\Sigma}$ $(m = p)$ which is stable. Let its distinct Hankel singular values be as defined by (5.22):

$$\sigma_1(\mathbf{\Sigma}) > \cdots > \sigma_q(\mathbf{\Sigma}) \text{ each with multiplicity } m_i,\ i = 1, \ldots, q,\ \sum_{i=1}^{q} m_i = n. \quad (5.22)$$

Furthermore, let $\hat{\mathbf{\Sigma}}$ be an optimal all-pass dilation of $\mathbf{\Sigma}$ defined by (8.16). This system is decomposed into its stable and antistable parts,

$$\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_+ + \hat{\mathbf{\Sigma}}_-,$$

where the poles of $\mathbf{\Sigma}_+$ are all in the LHP and those of $\mathbf{\Sigma}_-$ are in the RHP. It follows that $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Sigma}}_+$, i.e., the approximant is stable, and, furthermore, $\mathbf{H}(s) - \hat{\mathbf{H}}(s)$ is *all-pass* with magnitude σ_q, and

$$\sigma_i(\hat{\mathbf{\Sigma}}) = \sigma_i(\mathbf{\Sigma}), \qquad i = 1, 2, \ldots, n - r_q.$$

This fact implies the following decomposition of the transfer function $\mathbf{H}$:

$$\mathbf{H}(s) = \mathbf{D}_0 + \sigma_1\mathbf{H}_1(s) + \cdots + \sigma_q\mathbf{H}_q(s), \qquad (8.19)$$

where $\mathbf{H}_k(s)$, $k = 1, 2, \ldots, q$, are stable, all-pass, with McMillan degree

$$\delta\left\{\sum_{i=1}^{k} \sigma_i\mathbf{H}_i(s)\right\} = \sum_{i=1}^{k} m_i, \qquad k = 1, 2, \ldots, q.$$

From the above decomposition, we can derive the following upper bound of the $\mathcal{H}_\infty$-norm of $\mathbf{\Sigma}$. Assume that $\mathbf{H}(\infty) = \mathbf{0}$; the following inequality holds:

$$\boxed{\parallel \mathbf{\Sigma} \parallel_{\mathcal{H}_\infty} \leq 2(\sigma_1 + \cdots + \sigma_q).} \qquad (8.20)$$

Furthermore, there exists $D_0 \in \mathbb{R}^{m \times m}$ such that

$$\parallel \mathbf{H}(s) - \mathbf{D}_0 \parallel_{\mathcal{H}_\infty} \leq \sigma_1 + \cdots + \sigma_q.$$

In addition to $\hat{\mathbf{H}}$ constructed above, which will be denoted by $\mathbf{H}_h$ in what follows, consider the balanced truncated system $(\mathbf{A}_{22}, \mathbf{B}_{21}, \mathbf{C}_{12})$, whose transfer function will be denoted by $\mathbf{H}_b$. It follows that $\mathbf{H}_b(s) - \mathbf{H}_h(s)$ is all-pass with norm σ_q, and thus both the $\mathcal{H}_\infty$- and the Hankel-norms have the same upper bound:

$$\parallel \mathbf{H}(s) - \mathbf{H}_b(s) \parallel_{\mathcal{H}_\infty, H} \leq 2\sigma_q.$$

A consequence of the above inequalities is that the error for reduction by balanced truncation can be upper bounded by means of the singular values of $\mathbf{\Sigma}$. Let $\mathbf{H}_{b,k}$ denote a balanced approximant of McMillan degree $m_1 + \cdots + m_k$; then both the Hankel- and the $\mathcal{H}_\infty$-norms of the error system have the same upper bound, namely,

$$\boxed{\| \mathbf{H}(s) - \mathbf{H}_{b,k}(s) \|_{\mathcal{H}_\infty, H} \leq 2(\sigma_{k+1} + \cdots + \sigma_q).} \tag{8.21}$$

Below, we assume that each Hankel singular value has multiplicity equal to one: $m_i = 1$. Let $\mathbf{H}_{h,k}$ denote an optimal Hankel approximant of McMillan degree k. The Hankel singular values of the error can be explicitly determined:

$$\begin{aligned} \sigma_i(\mathbf{H} - \mathbf{H}_{h,k}) &= \sigma_{k+1}(\mathbf{H}), \qquad i = 1, \ldots, 2k+1, \\ \sigma_{2k+1+j}(\mathbf{H} - \mathbf{H}_{h,k}) &= \sigma_j(\mathbf{H}_e^-) \leq \sigma_{k+1+j}(\mathbf{H}), \qquad j = 1, \ldots, n-k-1. \end{aligned}$$

The final set of inequalities concerning the error of the approximation in the Hankel-norm follows from (8.20). Since we know explicitly the singular values of the stable part of the error system $\mathbf{\Sigma}_e^-$, we conclude

$$\| \mathbf{H} - \mathbf{H}_{h,k} \|_{\mathcal{H}_\infty} \leq 2(\sigma_{k+1} + \cdots + \sigma_n).$$

Notice that the above formulas remain valid if we scale the singular values by γ, i.e., replace σ_i by $\frac{\sigma_i}{\gamma}$.

Example 8.9 (*continuation of Example* 7.16). Consider the system described by the transfer function:

$$\mathbf{Z}(z) = \frac{z^2+1}{z^3}.$$

As already mentioned, this is a finite impulse response (FIR) system. We investigate the optimal approximation of this system in the Hankel-norm. It is interesting to notice that the optimal approximant of order 2 is no longer FIR; it is inifinite impulse response (IIR) instead. The corresponding Hankel matrix is

$$\mathcal{H} = \begin{pmatrix} 1 & 0 & 1 & 0 & \\ 0 & 1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 0 & \\ & \vdots & & & \ddots \end{pmatrix}.$$

The SVD of the 3×3 submatrix of $\mathcal{H}$ is

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ \sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & -\sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ -\sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{bmatrix}, \tag{8.22}$$

where

$$\sigma_1 = \frac{\sqrt{5}+1}{2}, \quad \sigma_2 = 1, \quad \sigma_3 = \frac{\sqrt{5}-1}{2}.$$

It is tempting to conjecture that the optimal second-order approximant is obtained by setting $\sigma_3 = 0$ in (8.22). The problem with this procedure is that the resulting approximant does *not* have Hankel structure.

To compute the optimal approximant, we proceed as follows. First, transform the system to a continuous-time system using the transformation of subsection 4.3.3; we obtain the transfer function

$$\mathbf{G}(s) = \mathbf{Z}\left(\frac{1+s}{1-s}\right) = \frac{2(s^3 - s^2 + s - 1)}{(s+1)^3}.$$

Applying the theory discussed in the preceding subsections, we obtain the following second-order continuous-time optimal approximant:

$$\mathbf{G}_2(s) = \frac{-(s^2-1)}{(1-\sigma_3)s^2 + 2\sigma_1 s + (1-\sigma_3)}.$$

Again using the transformation of subsection 4.3.3, we obtain the following discrete-time optimal approximant:

$$\mathbf{Z}_2(z) = \mathbf{G}_2\left(\frac{z-1}{z+1}\right) = \frac{z}{z^2 - \sigma_3}.$$

Notice that the optimal approximant is *not* a FIR system. It has poles at $\pm\sqrt{\sigma_3}$. Furthermore, the error $\mathbf{Z}(z) - \mathbf{Z}_2(z)$ is all-pass with magnitude equal to σ_3 on the unit circle:

$$\mathbf{Z}(z) - \mathbf{Z}_2(z) = \sigma_3\left[\frac{1-\sigma_3 z^2}{z^3(z^2-\sigma_3)}\right].$$

The corresponding optimal Hankel matrix of rank 2 is

$$\hat{\mathcal{H}} = \begin{pmatrix} 1 & 0 & \sigma_3 & 0 & \sigma_3^2 & \\ 0 & \sigma_3 & 0 & \sigma_3^2 & 0 & \cdots \\ \sigma_3 & 0 & \sigma_3^2 & 0 & \sigma_3^3 & \\ 0 & \sigma_3^2 & 0 & \sigma_3^3 & 0 & \\ \sigma_3^2 & 0 & \sigma_3^3 & 0 & \sigma_3^4 & \\ & \vdots & & & & \ddots \end{pmatrix}.$$

In this particular case, the 3×3 submatrix of $\hat{\mathcal{H}}$ is also an optimal approximant of the corresponding submatrix of $\mathcal{H}$,

$$\begin{bmatrix} 1 & 0 & \sigma_3 \\ 0 & \sigma_3 & 0 \\ \sigma_3 & 0 & \sigma_3^2 \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ \sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & -\sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{bmatrix} \begin{bmatrix} 1+\sigma_3^2 & 0 & 0 \\ 0 & \sigma_3 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{\sigma_1}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_3}{\sqrt{5}}} \\ 0 & 1 & 0 \\ -\sqrt{\frac{\sigma_3}{\sqrt{5}}} & 0 & \sqrt{\frac{\sigma_1}{\sqrt{5}}} \end{bmatrix}. \tag{8.23}$$

This decomposition is an application of formula (3.16), which provides a class of minimizers; it is straightforward to verify that $\eta_1 = \sigma_1 - \sigma_3^2 - 1 = \sqrt{5} - 2$, which implies $\eta_1 < \sigma_3$, while $\eta_2 = \sigma_2 - \sigma_3 = (3 - \sqrt{5})/2$, which implies $\eta_2 \leq \sigma_3$.

Finally, it is readily checked that the Hankel matrix consisting of 1 as the $(1, 1)$ entry and 0 everywhere else is the optimal approximant of $\mathcal{H}$ of rank one. The decomposition (8.19) of $\mathbf{Z}$ is

$$\mathbf{Z}(z) = \sigma_1 \underbrace{\begin{bmatrix} 1 \\ \overline{z} \end{bmatrix}}_{\mathbf{Z}_1(z)} + \sigma_2 \underbrace{\begin{bmatrix} \frac{1-\sigma_3 z^2}{z(z^2-\sigma_3)} \end{bmatrix}}_{\mathbf{Z}_2(z)} + \sigma_3 \underbrace{\begin{bmatrix} \frac{-1+\sigma_3 z^2}{z^3(z^2-\sigma_3)} \end{bmatrix}}_{\mathbf{Z}_3(z)}.$$

Notice that each $\mathbf{Z}_i$ is all-pass, and the McMillan degree of $\mathbf{Z}_1$ is one, that of $\sigma_1\mathbf{Z}_1 + \mathbf{Z}_2$ is two, and, finally, that of all three summands is three.

Example 8.10. *A continuous-time suboptimal approximation.* Consider the system $\mathbf{\Sigma}$ given by (4.13), where $n = 2$, $m = p = 1$, and

$$\mathbf{A} = -\begin{pmatrix} \frac{1}{2\sigma_1} & \frac{1}{\sigma_1+\sigma_2} \\ \frac{1}{\sigma_1+\sigma_2} & \frac{1}{2\sigma_2} \end{pmatrix}, \; \mathbf{B} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \; \mathbf{C} = \begin{pmatrix} 1 & 1 \end{pmatrix}, \; \mathbf{D} = \mathbf{0},$$

where $\sigma_1 > \sigma_2$. This system is in *balanced canonical form*; this means that the gramians are $\mathcal{P} = \mathcal{Q} = \operatorname{diag}(\sigma_1, \sigma_2) = \Sigma$. This canonical form is a special case of the forms discussed in section 8.6.3.

To determine the suboptimal Hankel-norm approximants for $\sigma_1 > \epsilon > \sigma_2$, we compute the limit of this family for $\epsilon \to \sigma_2$ and $\epsilon \to \sigma_1$ and show that the system obtained is indeed the optimal approximant. From (8.13), (8.17), $\Gamma = \epsilon^2 - \Sigma^2 = \operatorname{diag}(\epsilon^2 - \sigma_1^2, \epsilon^2 - \sigma_2^2)$; the inertia of Γ is $\{1, 0, 1\}$. Furthermore, from (8.14),

$$\hat{\mathbf{A}} = \begin{pmatrix} \frac{\epsilon-\sigma_1}{2\sigma_1(\epsilon+\sigma_1)} & \frac{\epsilon-\sigma_1}{(\sigma_1+\sigma_2)(\epsilon+\sigma_2)} \\ \frac{\epsilon-\sigma_2}{(\sigma_1+\sigma_2)(\epsilon+\sigma_1)} & \frac{\epsilon-\sigma_2}{2\sigma_2(\epsilon+\sigma_2)} \end{pmatrix}, \; \hat{\mathbf{B}} = \begin{pmatrix} \epsilon - \sigma_1 \\ \epsilon - \sigma_2 \end{pmatrix}, \; \hat{\mathbf{C}} = \begin{pmatrix} \frac{-1}{\epsilon+\sigma_1} & \frac{-1}{\epsilon+\sigma_2} \end{pmatrix},$$

and $\hat{\mathbf{D}} = \epsilon$. Since the inertia of $\hat{\mathbf{A}}$ is equal to the inertia of $-\Gamma$, $\hat{\mathbf{A}}$ has one stable and one unstable pole. (This can be checked directly by noticing that the determinant of $\hat{\mathbf{A}}$ is negative.) As $\epsilon \to \sigma_2$, we obtain

$$\hat{\mathbf{A}} = \begin{pmatrix} \frac{\sigma_2-\sigma_1}{2\sigma_1(\sigma_1+\sigma_2)} & \frac{\sigma_2-\sigma_1}{2\sigma_2(\sigma_1+\sigma_2)} \\ 0 & 0 \end{pmatrix}, \; \hat{\mathbf{B}} = \begin{pmatrix} \sigma_2 - \sigma_1 \\ 0 \end{pmatrix}, \; \hat{\mathbf{C}} = \begin{pmatrix} \frac{-1}{\sigma_1+\sigma_2} & \frac{-1}{2\sigma_2} \end{pmatrix},$$

and $\hat{\mathbf{D}} = \sigma_2$. This system is not reachable but observable (i.e., there is a pole-zero cancellation in the transfer function). A state-space representation of the reachable and observable subsystem is

$$\bar{\mathbf{A}} = \frac{\sigma_2 - \sigma_1}{2\sigma_1(\sigma_1 + \sigma_2)}, \quad \bar{\mathbf{B}} = \sigma_2 - \sigma_1, \quad \bar{\mathbf{C}} = \frac{-1}{\sigma_1 + \sigma_2}, \quad \bar{\mathbf{D}} = \sigma_2.$$

The formulas (8.14) depend on the choice of $\hat{\mathbf{D}}$. If we choose it to be $-\epsilon$, the limit still exists and gives a realization of the optimal system, which is equivalent to $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, $\bar{\mathbf{C}}$, $\bar{\mathbf{D}}$, given above.

Finally, if $\epsilon \to \sigma_1$, after a pole-zero cancellation, we obtain the following reachable and observable approximant:

$$\bar{\mathbf{A}} = \frac{\sigma_1 - \sigma_2}{2\sigma_1(\sigma_1 + \sigma_2)}, \quad \bar{\mathbf{B}} = \sigma_1 - \sigma_2, \quad \bar{\mathbf{C}} = \frac{-1}{\sigma_1 + \sigma_2}, \quad \bar{\mathbf{D}} = \sigma_1.$$

This is the best antistable approximant of $\mathbf{\Sigma}$, i.e., the Nehari solution (see (8.10)).

Example 8.11. *Balancing and Hankel-norm approximation applied to low-pass filters.* Four types of analog filters will be approximated next, by means of balanced truncation and Hankel-norm approximation. The filters are

Σ_{B}	-	BUTTERWORTH	
Σ_{C1}	-	CHEBYSHEV-1:	1% ripple in the pass band
Σ_{C2}	-	CHEBYSHEV-2:	1% ripple in the stop band
Σ_{E}	-	ELLIPTIC:	0.1% ripple both in the pass and stop bands

In each case we consider 20th-order low-pass filters, with pass band gain equal to 1 and cutoff frequency normalized to 1. Figure 8.3 shows the Hankel singular values of the full-order models. It follows from these plots that to obtain roughly comparable approximation errors, Σ_{B} and Σ_{C2} have to be approximated by systems of lower order than Σ_{C1} and Σ_{E}; we thus choose to approximate Σ_{B}, Σ_{C2} by *8th-order* models and Σ_{C1}, Σ_{E} by *10th-order* models. Observe that for the Chebyshev-2 filter, the difference of the singular values $\sigma_{13} - \sigma_{20}$ is of the order 10^{-7}. Thus Σ_{C2} has an 8th-order (approximately) all-pass subsystem of magnitude .05. Similarly, the Chebyshev-1 filter has an all-pass subsystem of order 3.

In this example, the subscript bal stands for approximation by balanced truncation; the subscript hank stands for optimal Hankel-norm approximation; FOM stands for full order model; ROM stands for reduced order model.

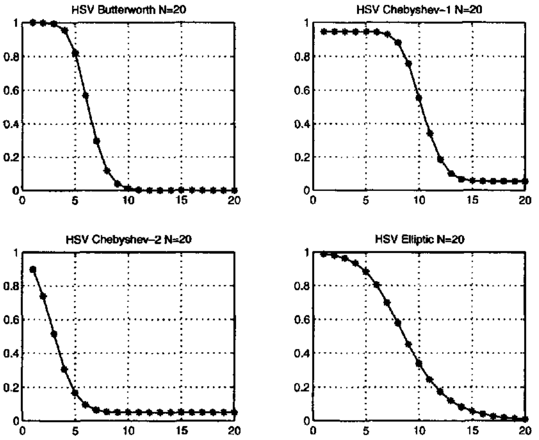


Figure 8.3. *Analogue filter approximation: Hankel singular values.*

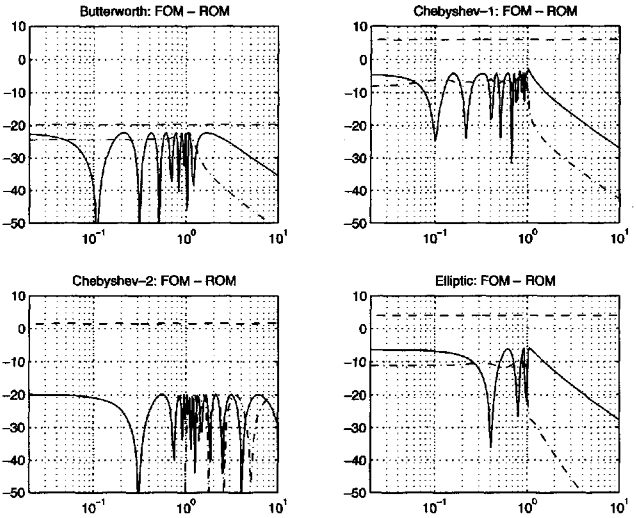


Figure 8.4. *Analog filter approximation: Bode plots of the error systems for model reduction by optimal Hankel-norm approximation (continuous curves), balanced truncation (dash-dot curves), and the upper bound* (8.21) *(dash-dash curves).*

Figure 8.4 gives the amplitude Bode plots of the error systems and tabulates their $\mathcal{H}_\infty$-norms and upper bounds. We observe that the 10th-order Hankel-norm approximants of Σ_{C1} and Σ_E are not very good in the stop band. One way to improve them is to increase the approximation order; another is to compute weighted approximants.

The following table compares the peak values of these Bode plots with the lower and upper bounds predicted by the theory:

	$\mathcal{H}_\infty$ Norm of the Error and Bounds			
	σ_9	$\|\Sigma - \Sigma_{\text{hank}}\|_\infty$	$\|\Sigma - \Sigma_{\text{bal}}\|_\infty$	$2(\sigma_9 + \cdots + \sigma_{20})$
BUTTERWORTH	0.0383	0.0388	0.0779	0.1035
CHEBYSHEV-2	0.0506	0.1008	0.0999	1.2015
	σ_{11}	$\|\Sigma - \Sigma_{\text{hank}}\|_\infty$	$\|\Sigma - \Sigma_{\text{bal}}\|_\infty$	$2(\sigma_{11} + \cdots + \sigma_{20})$
CHEBYSHEV-1	0.3374	0.4113	0.6508	1.9678
ELLIPTIC	0.2457	0.2700	0.3595	1.5818

8.6 Balanced and Hankel-norm approximation: A polynomial approach*

In Chapter 7 and in the preceding sections of the current chapter, we presented the theory of balanced and Hankel-norm approximations by employing state space methods. In the following, we present an approach based on polynomial methods. This approach provides new insights and connections between the concepts discussed above. The exposition is based on the work of Fuhrmann [119].

8.6.1 The Hankel operator and its SVD*

In this section, we restrict our attention to the SISO case. Recall that

$$\mathcal{L}_2(i\mathbb{R}) = \mathcal{H}_2(\mathbb{C}_-) \oplus \mathcal{H}_2(\mathbb{C}_+),$$

where $\mathcal{H}_2(\mathbb{C}_-)$, $\mathcal{H}_2(\mathbb{C}_+)$ contain functions that are analytic in $\mathbb{C}_-$, $\mathbb{C}_+$, respectively. The projections onto $\mathcal{H}_2(\mathbb{C}_-)$, $\mathcal{H}_2(\mathbb{C}_+)$ are denoted by $\mathbf{P}_-$, $\mathbf{P}_+$. Given $\phi \in \mathcal{L}_2(i\mathbb{R})$, we use ϕ^* to denote

$$\phi(s)^* = \phi^*(-s).$$

The *Hankel operator with symbol* $\phi \in \mathcal{L}_2(j\mathbb{R})$ is defined as follows:

$$\mathcal{H}_\phi : \mathcal{H}_2(\mathbb{C}_-) \rightarrow \mathcal{H}_2(\mathbb{C}_+), \quad \mathbf{f} \mapsto \mathcal{H}_\phi \mathbf{f} = \mathbf{P}_+(\phi \mathbf{f}).$$

The *dual Hankel operator* is defined as

$$\mathcal{H}_\phi^* : \mathcal{H}_2(\mathbb{C}_+) \rightarrow \mathcal{H}_2(\mathbb{C}_-), \quad \mathbf{f} \mapsto \mathcal{H}_\phi^* \mathbf{f} = \mathbf{P}_-(\phi^* \mathbf{f}).$$

An immediate consequence of the above definitions is ϕ, $\psi \in \mathcal{H}_2(\mathbb{C}_-)$, implies $\mathbf{P}_+\psi\mathcal{H}_\phi\mathbf{f} = \mathcal{H}_\phi\psi\mathbf{f}$. For the rest of this section we consider Hankel operators with stable, rational symbol:

$$\mathcal{H}_\phi, \; \phi = \frac{\mathbf{n}}{\mathbf{d}} \in \mathcal{H}_2(\mathbb{C}_+), \; \gcd(\mathbf{n}, \mathbf{d}) = 1. \tag{8.24}$$

To state the next result, we need to introduce the space $\mathbf{X}^{\mathbf{d}}$ of all strictly proper rational functions with denominator $\mathbf{d}$:

$$\mathbf{X}^{\mathbf{d}} = \left\{\frac{\mathbf{a}}{\mathbf{d}} : \deg \mathbf{a} < \deg \mathbf{d}\right\}.$$

$\mathbf{X}^{\mathbf{d}}$ is a finite-dimensional linear space with dimension $\deg \mathbf{d}$. The *shift operator* in this space is defined as

$$\mathbf{F}^{\mathbf{d}}(\mathbf{h}(s)) = \pi_-\left[\frac{1}{\mathbf{d}(s)}\pi_+\left[\frac{\mathbf{d}(s)}{s}\mathbf{h}(s)\right]\right],$$

where π_+, π_- are projections onto the polynomials, strictly proper rational functions, respectively. It readily follows that the characteristic polynomial of this shift operator is $\mathbf{d}$.

Proposition 8.12. *Consider a Hankel operator $\mathcal{H}_\phi$ satisfying* (8.24) *and its adjoint $\mathcal{H}_\phi^*$.*

1. *The kernel of this Hankel operator is* $\ker \mathcal{H}_\phi = \frac{\mathbf{d}}{\mathbf{d}^*}\mathcal{H}_2(\mathbb{C}_+)$.

2. *The kernel of the adjoint Hankel operator is* $\ker \mathcal{H}_\phi^* = \frac{\mathbf{d}^*}{\mathbf{d}}\mathcal{H}_2(\mathbb{C}_-)$.

3. *The image of $\mathcal{H}_\phi$ is* $\operatorname{im} \mathcal{H}_\phi = \mathbf{X}^{\mathbf{d}}$.

4. *The image of the adjoint operator is* $\operatorname{im} \mathcal{H}_\phi^* = \mathbf{X}^{\mathbf{d}^*}$.

Next, we compute the singular values and corresponding singular vectors of the Hankel operator, also known as Schmidt pairs. A pair of functions $\mathbf{f} \in \mathcal{H}_2(\mathbb{C}_-)$, $\mathbf{g} \in \mathcal{H}_2(\mathbb{C}_+)$ is a Schmidt pair of $\mathcal{H}_\phi$ corresponding to the singular value σ, provided that the following relationships are satisfied:

$$\mathcal{H}_\phi \mathbf{f} = \sigma \mathbf{g}, \ \ \mathcal{H}_\phi^* \mathbf{g} = \sigma \mathbf{f}.$$

Since $\mathbf{g} \in \operatorname{im} \mathcal{H}_\phi$ and $\mathbf{f} \in \operatorname{im} \mathcal{H}_\phi^*$, we obtain the following equations:

$$P_+ \frac{\mathbf{n}}{\mathbf{d}}\frac{\mathbf{p}}{\mathbf{d}^*} = \sigma \frac{\hat{\mathbf{p}}}{\mathbf{d}} \Rightarrow \frac{\mathbf{n}}{\mathbf{d}}\frac{\mathbf{p}}{\mathbf{d}^*} = \sigma \frac{\hat{\mathbf{p}}}{\mathbf{d}} + \frac{\pi}{\mathbf{d}^*} \ \Rightarrow \ \boxed{\mathbf{n}\mathbf{p} = \sigma \mathbf{d}^* \hat{\mathbf{p}} + \mathbf{d}\pi} \qquad (8.25)$$

$$P_- \frac{\mathbf{n}^*}{\mathbf{d}^*}\frac{\hat{\mathbf{p}}}{\mathbf{d}} = \sigma \frac{\mathbf{p}}{\mathbf{d}^*} \Rightarrow \frac{\mathbf{n}^*}{\mathbf{d}^*}\frac{\hat{\mathbf{p}}}{\mathbf{d}} = \sigma \frac{\mathbf{p}}{\mathbf{d}^*} + \frac{\xi}{\mathbf{d}} \ \Rightarrow \ \boxed{\mathbf{n}^* \hat{\mathbf{p}} = \sigma \mathbf{d}\mathbf{p} + \mathbf{d}^* \xi} \qquad (8.26)$$

The quantities $\mathbf{p}$, $\hat{\mathbf{p}}$, π, ξ are polynomials having degree less than the degree of $\mathbf{d}$,

$$\deg \mathbf{p}, \ \deg \hat{\mathbf{p}}, \ \deg \pi, \ \deg \xi < \deg \mathbf{d} = \nu.$$

By conjugating the second equation and eliminating $\mathbf{n}$, we obtain

$$\mathbf{0} = \sigma \mathbf{d}^*(\hat{\mathbf{p}}\hat{\mathbf{p}}^* - \mathbf{p}\mathbf{p}^*) + \mathbf{d}(\pi \hat{\mathbf{p}}^* - \xi^* \mathbf{p}),$$

while by conjugating the first and eliminating $\mathbf{n}^*$, we obtain

$$\mathbf{0} = \sigma \mathbf{d}(\hat{\mathbf{p}}\hat{\mathbf{p}}^* - \mathbf{p}\mathbf{p}^*) + \mathbf{d}^*(\pi^* \hat{\mathbf{p}} - \xi \mathbf{p}^*).$$

Since $\mathbf{d}$ is stable (roots in the left half plane), $\mathbf{d}^*$ is antistable, and hence $\mathbf{d}$, $\mathbf{d}^*$ are coprime. Hence the first equation above implies that $\mathbf{d}$ divides $\hat{\mathbf{p}}\hat{\mathbf{p}}^* - \mathbf{p}\mathbf{p}^*$, while the second implies that $\mathbf{d}^*$ divides the same quantity; therefore,

$$\mathbf{d}\mathbf{d}^* \mid \left(\hat{\mathbf{p}}\hat{\mathbf{p}}^* - \mathbf{p}\mathbf{p}^*\right) \ \Rightarrow \ \hat{\mathbf{p}}\hat{\mathbf{p}}^* - \mathbf{p}\mathbf{p}^* = 0.$$

The above implication follows because the degree of $\mathbf{d}\mathbf{d}^*$ is 2ν, while that of $\hat{\mathbf{p}}\hat{\mathbf{p}}^* - \mathbf{p}\mathbf{p}^*$ is less than 2ν. This in turn implies that

$$\frac{\mathbf{p}}{\hat{\mathbf{p}}}\left[\frac{\mathbf{p}}{\hat{\mathbf{p}}}\right]^* = 1 \Rightarrow \hat{\mathbf{p}} = \epsilon \mathbf{p}^*, \qquad \epsilon = \pm 1.$$

Proposition 8.13. *The Schmidt pairs of* $\mathcal{H}_\phi$ *satisfying* (8.24) *have the form*

$$\boxed{\left\{\frac{\mathbf{p}}{\mathbf{d}^*}, \epsilon \frac{\mathbf{p}^*}{\mathbf{d}}\right\}, \qquad \epsilon = \pm 1,}$$

where $\deg \mathbf{p} < \deg \mathbf{d}$.

Next, we discuss the issue of the multiplicity of singular values and characterize the corresponding Schmidt pairs. Using an argument similar to the above, we can show that if

$$\left\{\frac{\mathbf{p}}{\mathbf{d}^*}, \frac{\hat{\mathbf{p}}}{\mathbf{d}}\right\}, \ \left\{\frac{\mathbf{q}}{\mathbf{d}^*}, \epsilon \frac{\hat{\mathbf{q}}}{\mathbf{d}}\right\}$$

are Schmidt pairs corresponding to the same singular value σ, the quotient of $\mathbf{p}$ by $\hat{\mathbf{p}}$ remains independent of the particular Schmidt pair:

$$\frac{\mathbf{p}}{\hat{\mathbf{p}}} = \frac{\mathbf{q}}{\hat{\mathbf{q}}}. \tag{8.27}$$

Let $(\mathbf{p}, \hat{\mathbf{p}})$ be a *minimal* degree solution of (8.25) for some fixed σ. Then all other solutions of this equation for the same σ are given by

$$(\mathbf{q}, \hat{\mathbf{q}}) = (\mathbf{a}\mathbf{p}, \mathbf{a}\hat{\mathbf{p}}), \ \ \deg \mathbf{a} < \deg \mathbf{d} - \deg \mathbf{p}.$$

Thus

$$\ker \left(\mathcal{H}_\phi^* \mathcal{H}_\phi - \sigma^2 \mathbf{I}\right) = \left\{ \frac{\mathbf{a}\mathbf{p}}{\mathbf{d}^*} : \ \deg \mathbf{a} < \deg \mathbf{d} - \deg \mathbf{p} \right\}.$$

This implies that the multiplicity of a singular value is equal to the degree of $\mathbf{d}$ minus the degree of the minimal degree $\mathbf{p}$ which satisfies (8.25):

$$\dim \ker \left(\mathcal{H}_\phi^* \mathcal{H}_\phi - \sigma^2 \mathbf{I}\right) = \deg \mathbf{d} - \deg \mathbf{p}.$$

Combining the above results, we can state the main result with regard to the SVD of $\mathcal{H}_\phi$.

Theorem 8.14. *With the notation above, σ is a singular value of the Hankel operator $\mathcal{H}_\phi$, $\phi = \frac{\mathbf{n}}{\mathbf{d}} \in \mathcal{H}_2(\mathbb{C}_+)$, with $\mathbf{n}$, $\mathbf{d}$ coprime, and $\left\{ \frac{\mathbf{p}}{\mathbf{d}^*}, \epsilon \frac{\mathbf{p}^*}{\mathbf{d}} \right\}$ a corresponding Schmidt pair if and only if the following equation is satisfied:*

$$\boxed{\mathbf{n}\mathbf{p} = \lambda \mathbf{d}^* \mathbf{p}^* + \mathbf{d}\pi, \qquad \lambda = \epsilon\sigma,} \tag{8.28}$$

where $\deg \pi = \deg \mathbf{p}$.

Equation (8.28) is an *eigenvalue equation modulo the polynomial* $\mathbf{d}$. It can be converted into a matrix equation, and subsequently solved, as follows. Let $\mathbf{e}_i$ be a basis of the space $\mathbf{X}_\mathbf{d}$ of all polynomials of degree less than $\deg \mathbf{d}$, and let $\mathbf{F}_\mathbf{d}$ be the shift in $\mathbf{X}_\mathbf{d}$, modulo $\mathbf{d}$:

$$\mathbf{F}_\mathbf{d}(\mathbf{r}) = s \cdot \mathbf{r}(s) \bmod \mathbf{d}(s).$$

Notice that the characteristic polynomial of $\mathbf{F}_\mathbf{d}$ is $\mathbf{d}$. $\underline{\mathbf{p}}$, $\underline{\mathbf{p}^*}$ denote the vector representations of $\mathbf{p}$, $\mathbf{p}^*$, respectively in the basis $\mathbf{e}_i$. Then from (8.28) we obtain the matrix equation

$$\mathbf{n}(\mathbf{F}_\mathbf{d})\underline{\mathbf{p}} = \lambda \mathbf{d}^*(\mathbf{F}_\mathbf{d})\underline{\mathbf{p}^*}.$$

Since $\mathbf{d}$, $\mathbf{d}^*$ are coprime, there exist polynomials $\mathbf{a}$, $\mathbf{b}$ such that

$$\mathbf{a}\mathbf{d} + \mathbf{b}\mathbf{d}^* = 1.$$

Thus the inverse of $\mathbf{d}^*(\mathbf{F}_\mathbf{d})$ is $\mathbf{b}(\mathbf{F}_\mathbf{d})$. This implies

$$\mathbf{b}(\mathbf{F}_\mathbf{d})\mathbf{n}(\mathbf{F}_\mathbf{d})\underline{\mathbf{p}} = \lambda \underline{\mathbf{p}^*}.$$

Finally let $\mathbf{K}$ be the map such that $\mathbf{K}\mathbf{p}^* = \mathbf{p}$; the matrix representation of $\mathbf{K}$ is a sign matrix. Hence, with $\nu = \deg \mathbf{d}$, there holds

$$\mathbf{M}\underline{\mathbf{p}} = \lambda\underline{\mathbf{p}}, \quad \text{where} \quad \mathbf{M} = \mathbf{K}\mathbf{b}(\mathbf{F_d})\mathbf{n}(\mathbf{F_d}) \in \mathbb{R}^{\nu\times\nu} \quad \text{and} \quad \underline{\mathbf{p}} \in \mathbb{R}^{\nu}. \tag{8.29}$$

This is the eigenvalue equation that has to be solved to compute the singular values and singular vectors of the Hankel operator $\mathcal{H}_\phi$.

Corollary 8.15. *Let λ_i, with multiplicities μ_i, $i = 1, \ldots, r$, $\sum_{i=1}^{r} \mu_i = \nu$, be the solutions of the eigenvalue equation defined by* (8.28)*. Furthermore, let $\mathbf{p}_i$, $i = 1, \ldots, r$, be the minimal degree solutions of this equation. The singular values of $\mathcal{H}_\phi$ and their multiplicities are*

$$\sigma_i = |\lambda_i|, \; \epsilon_i = \frac{\sigma_i}{\lambda_i}, \; \mu_i, \qquad i = 1, \ldots, r,$$

while the Schmidt pairs of $\mathcal{H}_\phi$ corresponding to σ_i have the form

$$\left\{ \frac{\mathbf{p}_i \mathbf{a}_i}{\mathbf{d}^*}, \epsilon_i \frac{\mathbf{p}_i^* \mathbf{a}_i^*}{\mathbf{d}} \right\}, \quad \deg \mathbf{a}_i < \deg \mathbf{d} - \deg \mathbf{p}_i, \qquad i = 1, \ldots, r.$$

Example 8.16. We consider again the system defined by

$$\phi(s) = \frac{1}{s^2 + s + 1}.$$

We chose the basis $1, s$ in $\mathbf{X_d}$, $\mathbf{d} = s^2 + s + 1$. Furthermore, $\mathbf{b}(s) = \frac{1}{2}(s + 1)$. Thus

$$\mathbf{F_d} = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \; \mathbf{n}(\mathbf{F_d}) = \mathbf{I}_2, \; \mathbf{b}(\mathbf{F_d}) = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \; \mathbf{K} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Hence $\mathbf{M}$ in (8.29) is

$$\mathbf{M} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 0 \end{pmatrix}.$$

The eigenvalues of $\mathbf{M}$ are the roots of $\lambda^2 - \frac{1}{2}\lambda - \frac{1}{4}$:

$$\lambda_1 = \frac{1+\sqrt{5}}{4}, \; \lambda_2 = \frac{1-\sqrt{5}}{4} \Rightarrow \sigma_1 = \lambda_1, \; \sigma_2 = -\lambda_2, \; \epsilon_1 = 1, \; \epsilon_2 = -1.$$

The corresponding (unnormalized) eigenvectors are

$$\begin{pmatrix} -2\lambda_1 \\ 1 \end{pmatrix}, \begin{pmatrix} -2\lambda_2 \\ 1 \end{pmatrix}.$$

Therefore, $\mathbf{p}_1(s) = s - 2\lambda_1$, which implies $\pi_1 = \lambda_1 s + \lambda_2$, and $\mathbf{p}_2(s) = s - 2\lambda_2$, which implies $\pi_2 = \lambda_2 s + \lambda_1$.

8.6.2 The Nehari problem and Hankel-norm approximants*

For simplicity we assume that the Hankel operator is generic in the sense that it has ν distinct singular values of multiplicity one: $\sigma_1 > \sigma_2 > \cdots > \sigma_\nu > 0$. Thus there are polynomials $\mathbf{p}_i, \pi_i, i = 1, \ldots, \nu$, of degree equal to $\nu - 1$ and signs $\epsilon_i = \pm 1$ such that (8.28) is satisfied,

$$\mathbf{n}\mathbf{p}_i = \lambda_i \mathbf{d}^* \mathbf{p}_i^* + \mathbf{d}\pi_i, \qquad \lambda_i = \epsilon_i \sigma_i. \tag{8.30}$$

On dividing by $\mathbf{d} \cdot \mathbf{p}_i$, this equation becomes

$$\frac{\mathbf{n}}{\mathbf{d}} - \frac{\pi_i}{\mathbf{p}_i} = \lambda_i \frac{\mathbf{d}^*}{\mathbf{d}} \frac{\mathbf{p}_i^*}{\mathbf{p}_i}, \qquad i = 1, \ldots, \nu.$$

We now have the following main result.

Theorem 8.17. *The kth-order stable and optimal Hankel-norm approximant ϕ_k of $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ is given by*

$$\phi_k = \left[\frac{\pi_{k+1}}{\mathbf{p}_{k+1}}\right]_+, \qquad k = 1, \ldots, \nu - 1,$$

where $[\ ]_+$ denotes the stable part of $[\ ]$.

Corollary 8.18. Nehari's theorem. *The best antistable approximant of $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ is*

$$\phi_0 = \frac{\pi_1}{\mathbf{p}_1}.$$

The $\mathcal{L}_\infty$-norm of the difference $\phi - \phi_0$ is equal to the Hankel-norm of ϕ, namely, σ_1.

Corollary 8.19. One-step reduction. *The best stable approximant of $\phi = \frac{\mathbf{n}}{\mathbf{d}}$ of order $\nu - 1$ is*

$$\phi_{\nu-1} = \frac{\pi_\nu}{\mathbf{p}_\nu}.$$

The difference $\phi - \phi_{\nu-1}$ is stable and all-pass; its $\mathcal{H}_\infty$- and Hankel-norms are equal to σ_ν.

Proof. To prove Corollary 8.18, it suffices to show that the zeros of $\mathbf{p}_1$ in (8.30) are all unstable. Let $\mathbf{p}_1$ have stable zeros; we can factorize $\mathbf{p}_1$ in its stable/antistable parts: $\mathbf{p}_1 = [\mathbf{p}_1]_+ [\mathbf{p}_1]_-$. Then

$$\underbrace{\frac{\pi_1}{\mathbf{p}_1}}_{\mathbf{f}_1} = \underbrace{\frac{[\mathbf{p}_1]_+^*}{[\mathbf{p}_1]_+}}_{\mathbf{f}_i} \underbrace{\frac{\pi_1}{[\mathbf{p}_1]_+^* [\mathbf{p}_1]_-}}_{\mathbf{f}_o},$$

where $\mathbf{f}_i$ is the inner factor (i.e., stable and all-pass), while $\mathbf{f}_o$ is the outer factor (i.e., antistable). Let $(\mathbf{f}_1, \mathbf{f}_1^*)$ be a Schmidt pair corresponding to the largest singular value of $\mathcal{H}_\phi$:

$$\mathcal{H}_\phi \mathbf{f}_1 = \lambda_1 \mathbf{f}_1^* \text{ and } \| \mathcal{H}_\phi \|_{2-ind} = \sigma_1.$$

Hence $\| \mathcal{H}_\phi \mathbf{f}_1 \| = \sigma_1 \| \mathbf{f}_1 \|$. The following string of equalities and inequalities holds:

$$\begin{aligned} \| \mathcal{H}_\phi \| \cdot \| \mathbf{f}_o \| &= \| \mathcal{H}_\phi \| \cdot \| \mathbf{f}_1 \| = \| \mathcal{H}_\phi \mathbf{f}_1 \| = \| \mathcal{H}_\phi \mathbf{f}_i \mathbf{f}_o \| \\ &= \| \mathbf{P}_+ \mathbf{f}_i \mathbf{P}_+ \phi \mathbf{f}_o \| = \| \mathbf{P}_+ \mathbf{f}_i \mathcal{H}_\phi \mathbf{f}_o \| \\ &\leq \underbrace{\| \mathbf{P}_+ \mathbf{f}_i \|}_{=1} \cdot \| \mathcal{H}_\phi \mathbf{f}_o \| \leq \| \mathcal{H}_\phi \| \cdot \| \mathbf{f}_o \| . \end{aligned}$$

The above relationships imply that $\mathbf{f}_o$ is a singular vector of the Hankel operator corresponding to the singular value σ_1, whose poles are antistable. □

Proof. The proof of Corollary 8.19 is based on the fact that the Hankel operator $\mathcal{H}_\mathbf{r}$ with rational symbol $\mathbf{r}$ defined by

$$\mathcal{H}_\mathbf{r}, \ \mathbf{r} = \frac{1}{\lambda_\nu} \frac{\mathbf{d}^*}{\mathbf{d}} \frac{\mathbf{p}_\nu^*}{\mathbf{p}_\nu},$$

has singular values, $\sigma_1^{-1} < \sigma_2^{-1} < \cdots < \sigma_\nu^{-1}$, and corresponding Schmidt pairs

$$\left\{ \frac{\mathbf{p}_i^*}{\mathbf{d}^*}, \epsilon_i \frac{\mathbf{p}_i}{\mathbf{d}} \right\}, \qquad i = 1, \ldots, \nu.$$

By Corollary 8.18, however, we know that $\mathbf{p}_\nu^*$ is antistable. This shows that $\mathbf{p}_\nu$ is stable with degree $\nu - 1$. Hence the difference between $\frac{\mathbf{n}}{\mathbf{d}}$ and $\frac{\pi_\nu}{\mathbf{p}_\nu}$ is all-pass with norm σ_ν:

$$\frac{\mathbf{n}}{\mathbf{d}} - \frac{\pi_\nu}{\mathbf{p}_\nu} = \lambda_\nu \frac{\mathbf{d}^*}{\mathbf{d}} \frac{\mathbf{p}_\nu^*}{\mathbf{p}_\nu} \Rightarrow \left\| \frac{\mathbf{n}}{\mathbf{d}} - \frac{\pi_\nu}{\mathbf{p}_\nu} \right\|_H = \left\| \frac{\mathbf{n}}{\mathbf{d}} - \frac{\pi_\nu}{\mathbf{p}_\nu} \right\|_{\mathcal{H}_\infty} = \sigma_\nu .$$

The proof is thus complete. □

Example 8.20 (*continuation*). Consider again the Hankel operator $\mathcal{H}_\phi$ with symbol

$$\phi(s) = \frac{1}{s^2 + s + 1}.$$

The best antistable approximant and the best stable first-order approximant are, respectively,

$$\frac{\pi_1}{\mathbf{p}_1} = \frac{\lambda_1 s + \lambda_2}{s - 2\lambda_1} = \lambda_1 + \frac{1}{s - 2\lambda_1}, \quad \frac{\pi_2}{\mathbf{p}_2} = \frac{\lambda_2 s + \lambda_1}{s - 2\lambda_2} = \lambda_2 + \frac{1}{s - 2\lambda_2}.$$

8.6.3 Balanced canonical form*

Recall (8.30). Let the $\mathbf{p}_i$ be normalized so that

$$\left\| \frac{\mathbf{p}_i^*}{\mathbf{d}} \right\|^2 = \sigma_i .$$

The function ϕ can always be decomposed as a linear combination of the singular vectors of the Hankel operator $\mathcal{H}_\phi$; namely, there exist γ_i, $i = 1, \ldots, \nu$, such that

$$\boxed{\frac{\mathbf{n}}{\mathbf{d}} = \gamma_1 \frac{\mathbf{p}_1^*}{\mathbf{d}} + \gamma_2 \frac{\mathbf{p}_2^*}{\mathbf{d}} + \cdots + \gamma_\nu \frac{\mathbf{p}_\nu^*}{\mathbf{d}}.} \tag{8.31}$$

Furthermore, let

$$\sigma_i = \epsilon_i \lambda_i .$$

Then ϕ has a minimal realization $\left(\begin{array}{c|c} \mathbf{A}_{\text{bal}} & \mathbf{B}_{\text{bal}} \\ \hline \mathbf{C}_{\text{bal}} & \mathbf{D}_{\text{bal}} \end{array}\right) \in \mathbb{R}^{(\nu+1)\times(\nu+1)}$, which has the form given by (7.24), namely,

$$\left(\begin{array}{c|c} \mathbf{A}_{\text{bal}} & \mathbf{B}_{\text{bal}} \\ \hline \mathbf{C}_{\text{bal}} & \mathbf{D}_{\text{bal}} \end{array}\right) = \left(\begin{array}{cccc|c} \frac{-\gamma_1^2}{2\sigma_1} & \frac{-\gamma_1\gamma_2}{\epsilon_1\epsilon_2\sigma_1+\sigma_2} & \cdots & \frac{-\gamma_1\gamma_\nu}{\epsilon_1\epsilon_\nu\sigma_1+\sigma_\nu} & \gamma_1 \\ \frac{-\gamma_2\gamma_1}{\epsilon_2\epsilon_1\sigma_2+\sigma_1} & \frac{-\gamma_2^2}{2\sigma_2} & \cdots & \frac{-\gamma_2\gamma_\nu}{\epsilon_2\epsilon_\nu\sigma_2+\sigma_\nu} & \gamma_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{-\gamma_\nu\gamma_1}{\epsilon_\nu\epsilon_1\sigma_\nu+\sigma_1} & \frac{-\gamma_\nu\gamma_2}{\epsilon_\nu\epsilon_2\sigma_\nu+\sigma_2} & \cdots & \frac{-\gamma_\nu^2}{2\sigma_\nu} & \gamma_\nu \\ \hline \epsilon_1\gamma_1 & \epsilon_2\gamma_2 & \cdots & \epsilon_\nu\gamma_\nu & \phi(\infty) \end{array}\right). \tag{8.32}$$

Remark 8.6.1. The above result shows an additional close connection between the Hankel operator $\mathcal{H}_\phi$ and balancing. (8.31) shows, namely, that the balanced realization is the realization obtained by using as a basis in $\mathbf{X}^{\mathbf{d}}$ (this space can be chosen as a state space for a realization of ϕ) the Schmidt vectors of $\mathcal{H}_\phi$.

Making use of (8.32), we see that the solution to the Nehari problem, namely, $\frac{\pi_1}{\mathbf{p}_1}$, has a realization which can be written explicitly in terms of the quantities introduced above. We assume without loss of generality that $\lambda_1 > 0$; define

$$\omega_i = \sqrt{\frac{\lambda_1 - \lambda_i}{\lambda_1 + \lambda_i}}, \qquad i = 2, \ldots, \nu.$$

Then $\frac{\pi_1}{p_1}$ has the following realization:

$$\left(\begin{array}{c|c} \mathbf{A}_{\text{neh}} & \mathbf{B}_{\text{neh}} \\ \hline \mathbf{C}_{\text{neh}} & \mathbf{D}_{\text{neh}} \end{array}\right) = \left(\begin{array}{ccc|c} \frac{\omega_2^2\gamma_2^2}{2\sigma_2} & \cdots & \frac{\omega_2\omega_\nu\gamma_2\gamma_\nu}{\epsilon_2\epsilon_\nu\sigma_2+\sigma_\nu} & \omega_2\gamma_2 \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\omega_\nu\omega_2\gamma_\nu\gamma_2}{\epsilon_\nu\epsilon_2\sigma_\nu+\sigma_2} & \cdots & \frac{\omega_\nu^2\gamma_\nu^2}{2\sigma_\nu} & \omega_\nu\gamma_\nu \\ \hline \epsilon_2\omega_2\gamma_2 & \cdots & \epsilon_\nu\omega_\nu\gamma_\nu & \epsilon_1\sigma_1 \end{array}\right) \in \mathbb{R}^{\nu\times\nu}. \tag{8.33}$$

Remark 8.6.2. Another important connection between balancing and the Schmidt pairs of $\mathcal{H}_\phi$ is that the Nehari extension, that is, the solution of the Hankel-norm approximation problem for antistable approximants, can be written down explicitly in terms of the balanced realization of ϕ.

Example 8.21. In the example discussed above,

$$\omega_2 = \sqrt{\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}} = 5^{\frac{1}{4}} = \frac{1}{\gamma}.$$

Recalling that $\lambda_1 \lambda_2 = -\frac{1}{4}$, the realization of $\frac{\pi_1}{\mathbf{p}_1}$ according to (8.33) is

$$\left(\begin{array}{c|c} 2\lambda_1 & 1 \\ \hline 1 & \lambda_1 \end{array} \right),$$

which agrees with the expression for $\frac{\pi_1}{\mathbf{p}_1}$ obtained previously.

8.6.4 Error formulas for balanced and Hankel-norm approximations*

Let

$$\mathbf{g}_b = \frac{\mathbf{n}_b}{\mathbf{d}_b}, \quad \mathbf{g}_h = \frac{\mathbf{n}_h}{\mathbf{d}_h} = \frac{\pi_n}{\mathbf{p}_n}$$

be the reduced-order models of order $\nu - 1$ obtained from ϕ by balanced truncation, Hankel-norm approximation, respectively. The following relations hold:

$$\sigma_\nu(\mathbf{d}\mathbf{d}_b^* + \mathbf{d}^*\mathbf{d}_b) = \mathbf{p}_\nu \mathbf{p}_\nu^*.$$

Furthermore,

$$\mathbf{e}_b = \phi - \mathbf{g}_b = \lambda_\nu \frac{\mathbf{p}_\nu^*}{\mathbf{p}_\nu} \left(\frac{\mathbf{d}^*}{\mathbf{d}} + \frac{\mathbf{d}_b^*}{\mathbf{d}_b} \right),$$

$$\mathbf{e}_h = \phi - \mathbf{g}_h = \frac{\pi_\nu}{\mathbf{p}_\nu} = \lambda_\nu \frac{\mathbf{d}^*\mathbf{p}_\nu^*}{\mathbf{d}\,\mathbf{p}_\nu} \Rightarrow \mathbf{e}_{hb} = \mathbf{g}_h - \mathbf{g}_b = \lambda_\nu \frac{\mathbf{d}_b^*\mathbf{p}_\nu^*}{\mathbf{d}_b\,\mathbf{p}_\nu}.$$

These relationships imply that error $\mathbf{e}_b$ is the sum of two all-pass functions and its norm is $2\sigma_\nu$. Furthermore, the errors $\mathbf{e}_h$ and $\mathbf{e}_{hb}$ are all-pass with norm σ_ν.

8.7 Chapter summary

Although approximation by balanced truncation enjoys two important properties (namely, preservation of stability and existence of an error bound), it is not optimal in any norm. The most natural norm in which to seek optimality is the 2-induced norm of the convolution operator $\mathcal{S}$ associated with the linear system $\boldsymbol{\Sigma}$ in question (in other words, the $\mathcal{H}_\infty$-norm of $\boldsymbol{\Sigma}$). This problem, however (except in very special cases), remains unsolved. If this norm is replaced by the 2-induced norm of the Hankel operator $\mathcal{H}$ associated with $\boldsymbol{\Sigma}$, the optimal approximation problem can be solved. Recall that the Hankel-norm is the 2-induced norm of the map $\mathcal{H}$ which assigns past inputs into future outputs and is different from the $\mathcal{H}_\infty$-norm (and less than or equal in value). The preceding chapter was dedicated to a detailed discussion of optimal and suboptimal system approximation in the Hankel-norm.

The key construction for solving this problem is the *all-pass or unitary dilation* of the system Σ which is to be approximated. This means that Σ is embedded in a larger system (i.e., a system with more states) which is all-pass (unitary) and denoted by Σ_e. The beauty of the theory lies in the fact that the *norm* of the all-pass system determines the order of the optimal (or suboptimal) approximants. It should be stressed that as far as the $\mathcal{H}_\infty$-norm of the error system is concerned, the optimal approximation in the Hankel-norm need not provide better approximants than balanced approximation.

In the latter part of the chapter, a polynomial approach to this theory is presented that, although it uses a framework not discussed in this book, is presented because it yields new insights; it can be omitted without loss of continuity. For instance, it is interesting to note that the decomposition of the transfer function in terms of singular (Schmidt) vectors yields the coefficients of a balanced realization of the underlying system.

Chapter 9

Special Topics in SVD-based Approximation Methods

A model reduction method in wide use is proper orthogonal decomposition (POD). We describe this method and show its connection with both the SVD and balanced truncation. Subsequently, another widely used approximation method, *modal approximation*, is discussed. This is an EVD-based approximation method and may lead to approximants that are not close to the original system, if $\mathbf{A}$ has Jordan blocks. In section 9.3 we provide a general view of approximation by *truncation* and the associated approximation by *residualization*.

The chapter concludes with a study of the decay rate of the Hankel singular values of a system $\boldsymbol{\Sigma}$. This is important because the sum of the neglected Hankel singular values provides error bounds for balanced and Hankel-norm approximants; thus the faster the decay of these singular values, the tighter the error bound.

9.1 Proper orthogonal decomposition

This section discusses a model reduction method known as POD (proper orthogonal decomposition). This can be considered as an application of the SVD to the approximation of general *dynamical systems*. A starting point for this method is an input function or an initial condition. The resulting state trajectory that lives in $\mathbb{R}^n$, where n is the number of state variables needed to describe the system, is measured. The issue then becomes whether it is possible to approximate this state trajectory with one living in a lower-dimensional space $\mathbb{R}^k$, where k is smaller than n. Toward this goal, a projector known as *Galerkin* or *Petrov–Galerkin* is constructed, and the original dynamical system of dimension n is reduced to one that has dimension k. The problem that arises is to determine how well this reduced system approximates trajectories other than the measured one.

In the next two sections, we provide a brief exposition of the POD method and the associated Galerkin projection. Section 9.1.3 examines the similarities between POD and balancing and proposes a method for obtaining a global error bound; this approach (inspired by balanced truncation) concludes that projection on the dominant eigenspace of the product of *two* gramians leads in the linear case to a global error bound, that is, an error bound that is valid for trajectories other than the ones used to obtain the projector.

9.1.1 POD

Given a function $\mathbf{x}(t, \mathbf{w})$ of time t and the vector valued variable $\mathbf{w}$, let $\mathbf{x}(t_i, \mathbf{w}_j)$ be a finite number of samples of $\mathbf{x}$ for $t = t_1, \ldots, t_N$ and $\mathbf{w} = \mathbf{w}_1, \ldots, \mathbf{w}_n$. The data are collected in *time-snapshots* denoted by

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}(t_i, \mathbf{w}_1) \\ \vdots \\ \mathbf{x}(t_i, \mathbf{w}_n) \end{bmatrix} \in \mathbb{R}^n, \qquad i = 1, \ldots, N.$$

We are looking for a set of orthonormal basis vectors $\mathbf{u}_j \in \mathbb{R}^n$, $j = 1, 2, \ldots, N$, such that $\mathbf{x}_i = \sum_{j=1}^{n} \gamma_{ji}\mathbf{u}_j$, $i = 1, 2, \ldots, N$, that is,

$$\underbrace{[\mathbf{x}_1 \cdots \mathbf{x}_N]}_{\mathbf{X}} = \underbrace{[\mathbf{u}_1 \cdots \mathbf{u}_N]}_{\mathbf{U}} \overbrace{\begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{N1} & \cdots & \gamma_{NN} \end{bmatrix}}^{\Gamma}, \quad \mathbf{U}^*\mathbf{U} = \mathbf{I}_N. \tag{9.1}$$

The $\mathbf{u}_i$ are sometimes referred to as *empirical eigenfunctions* or *principal directions* of the "cloud" of data $\{\mathbf{x}_i\}$. In addition, it is required that the truncated elements $\hat{\mathbf{x}}_i = \sum_{j=1}^{k} \gamma_{ji}\mathbf{u}_j$, $i = 1, 2, \ldots, N$, that is, the snapshots reconstructed from only k empirical eigenfunctions,

$$\underbrace{[\hat{\mathbf{x}}_1 \cdots \hat{\mathbf{x}}_N]}_{\hat{\mathbf{X}}} = \underbrace{[\mathbf{u}_1 \cdots \mathbf{u}_k]}_{\mathbf{U}_k} \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{k1} & \cdots & \gamma_{kN} \end{bmatrix} \in \mathbb{R}^{n \times N}, \qquad k < N,$$

approximate the elements in the family $\{\mathbf{x}_i\}$ optimally, in some *average* sense. Often it is required that the 2-induced norm of the difference $\|\mathbf{X} - \hat{\mathbf{X}}\|_2$ be minimized. Equation (9.1) is known as *proper orthogonal decomposition* of the family $\{\mathbf{x}_i\}$. The above optimality condition can also be defined by means of the *gramian* (or *autocorrelation* matrix) of the data, namely,

$$\mathbf{P} = \sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^* = \mathbf{P}^* \in \mathbb{R}^{n \times n}.$$

The optimization problem can now be formulated as a matrix approximation problem, namely, find

$$\hat{\mathbf{P}} = \sum_{i=1}^{N} \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^* \;\text{ with }\; k = \operatorname{rank}\hat{\mathbf{P}} < \operatorname{rank}\mathbf{P}$$

such that the 2-induced norm of the error $\|\mathbf{P} - \hat{\mathbf{P}}\|_2$ is minimized.

The problem just stated is precisely the one solved by the Schmidt–Mirsky–Eckard–Young result, Theorem 3.6. Given $\mathbf{X}$, let $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^*$ be its SVD. The columns of $\mathbf{U}$ are the desired *empirical eigenfunctions* and the coefficients are given by $\Gamma = \Sigma \mathbf{V}^*$. If the number of spatial samples n is bigger than that of time samples N, assuming that $\Sigma \in \mathbb{R}^{N \times N}$, the computation of the SVD of $\mathbf{X}$ can be achieved by solving the (small) eigenvalue problem $\mathbf{X}^*\mathbf{X} = \mathbf{V}\Sigma^2\mathbf{V}^*$; then the columns of $\mathbf{XV}$ are, up to scaling, the requited eigenvectors, i.e., $\mathbf{U} = \mathbf{XV}\Sigma^{-1}$.

9.1.2 Galerkin and Petrov–Galerkin projections

The next concept is that of a *projection*. Recall the dynamical system defined by (1.1):

$$\Sigma : \frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t),\ \mathbf{u}(t)),\ \ \mathbf{y}(t) = \mathbf{g}(\mathbf{x}(t),\ \mathbf{u}(t)). \tag{1.1}$$

Using the projection $\Pi = \mathbf{VW}^*$, where $\mathbf{W}^*\mathbf{V} = \mathbf{I}_k$, $\mathbf{V},\ \mathbf{W} \in \mathbb{R}^{n \times k}$, we obtain the reduced-order dynamical system

$$\hat{\Sigma} : \frac{d}{dt}\hat{\mathbf{x}} = \mathbf{W}^*\mathbf{f}(\mathbf{V}\hat{\mathbf{x}},\ \mathbf{u}),\ \ \mathbf{y} = \mathbf{g}(\mathbf{V}\hat{\mathbf{x}}, \mathbf{u}), \tag{1.7}$$

whose trajectories $\hat{\mathbf{x}} = \mathbf{W}^*\mathbf{x}$ evolve in a k-dimensional subspace. If the $\mathbf{V} = \mathbf{W}$, that is, the columns of $\mathbf{V}$ form as orthonormal set, Π is orthogonal and is called a *Galerkin* projection. Otherwise, if $\mathbf{V} \neq \mathbf{W}$, we speak of a *Petrov–Galerkin* projection.

Using the considerations developed in the previous section, Π is chosen using a POD method applied on a number of snapshots (i.e., samples of the state computed at given time instants) of the original system. Let these snapshots $\mathbf{x}(t_j)$, $j = 1, \ldots, N$, be provided. Given the matrix of snapshots $\mathbf{X} = [\mathbf{x}(t_1)\ \mathbf{x}(t_2)\ \cdots\ \mathbf{x}(t_N)]$, we compute the SVD $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^*$. Assume that the singular values of $\mathbf{X}$ decay rapidly and only k of them are significant for the application in question. The Galerkin projection consists of the leading k left singular vectors of the snapshot matrix $\mathbf{X}$, that is,

$$\mathbf{V} = \mathbf{W} = \mathbf{U}_k \in \mathbb{R}^{n \times k}.$$

Thus $\mathbf{V}\hat{\mathbf{x}} = \mathbf{VV}^*\mathbf{x}$ is the projection of $\mathbf{x}$ onto span col $\mathbf{U}_k$.

9.1.3 POD and balancing

The method just described is POD combined with a Galerkin projection. It is in wide use for the simulation and control of nonlinear dynamical systems and systems described by PDEs.

We will now show that the method of approximation of linear systems by *balanced truncation* discussed in Chapter 7 is a POD method combined with a Petrov–Galerkin projection applied to the *impulse response* of the system.

Consider a linear system of the usual form, namely,

$$\dot{\mathbf{x}}(t) = \mathbf{Ax}(t) + \mathbf{Bu}(t),\ \ \mathbf{y}(t) = \mathbf{Cx}(t) + \mathbf{Du}(t).$$

One choice for the input function is the *impulse* $\delta(t)$. If $m = 1$, the corresponding state is $\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{B}$. If $m > 1$, each input channel is excited separately, $\mathbf{u}(t) = \delta(t)\mathbf{e}_i$, where

$\mathbf{e}_i \in \mathbb{R}^m$ is the canonical ith unit vector. The resulting state is $\mathbf{x}_i(t) = e^{\mathbf{A}t}\mathbf{B}_i$, where $\mathbf{B}_i$ is the ith column of $\mathbf{B}$. In this case, we collect all such state trajectories $\mathbf{X}(t) = [\mathbf{x}_1(t) \cdots \mathbf{x}_m(t)]$ and compute the *gramian*:

$$\int_0^T \mathbf{X}(t)\mathbf{X}^*(t)\,dt = \int_0^T e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^*t}\,dt = \mathcal{P}(T).$$

We recognize this quantity as the reachability gramian of the system discussed in section 4.2.1 and in particular (4.28). If $\mathbf{A}$ happens to have eigenvalues in the left half of the complex plane (i.e., the system is stable), we can let T approach infinity, in which case we recover the infinite reachability gramian $\mathcal{P}$ introduced in (4.43).

POD now proceeds by computing an eigenvalue decomposition of this gramian:

$$\mathcal{P} = \mathbf{U}\Lambda\mathbf{U}^* = [\mathbf{U}_1 \ \ \mathbf{U}_2]\begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix}[\mathbf{U}_1 \ \ \mathbf{U}_2]^* = \mathbf{U}_1\Lambda_1\mathbf{U}_1^* + \mathbf{U}_2\Lambda_2\mathbf{U}_2^*,$$

where it is assumed that Λ_2 contains the *small* or *negligible* eigenvalues of $\mathcal{P}$. The Galerkin projection is thus defined by the leading eigenvectors of this gramian, namely, $\mathbf{U}_1$, and the reduced system is described by

$$\hat{\mathbf{A}} = \mathbf{U}_1^*\mathbf{A}\mathbf{U}_1, \quad \hat{\mathbf{B}} = \mathbf{U}_1^*\mathbf{B}, \quad \hat{\mathbf{C}} = \mathbf{C}\mathbf{U}_1, \quad \hat{\mathbf{D}} = \mathbf{D}.$$

According to the analysis presented earlier, this is the subsystem of the original system which retains those states which are easier to reach while discarding the states which are the most difficult to reach. Recall that by (4.55), the energy required to steer the system to a given state (starting from the zero state) is related to the inverse of $\mathcal{P}$; therefore, discarding the small eigenvalues corresponds to discarding the states which are most difficult to reach. In conclusion, for linear systems, POD approximation using the impulse response is related to balanced truncation, where in POD approximation only the degree of reachability of the states is considered.

Balanced truncation, however, suggests that if, in addition to the degree of reachability of the states, we also consider their degree of observability, a *global approximation error bound* can be derived (see (7.5)). Therefore, let $\mathcal{Q}$ be the (infinite) observability gramian which satisfies as usual the Lyapunov equation $\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$. To obtain error bounds, according to Theorem 7.9, we need to project onto the dominant eigenspace of the product of the two gramians $\mathcal{P}\mathcal{Q}$. It readily follows from (7.16) that if we partition $\Sigma = \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix}$ so that Σ_2 contains the (square root of the) negligible eigenvalues of $\mathcal{P}\mathcal{Q}$, the projection $\Pi = \mathbf{T}_{i1}\mathbf{T}_1$, where

$$\mathbf{T}_1 = \Sigma_1^{-1/2}\mathbf{V}_1^*\mathbf{L}^* \in \mathbb{R}^{k\times n}, \quad \mathbf{T}_{i1} = \mathbf{U}\mathbf{W}_1\Sigma_1^{-1/2} \in \mathbb{R}^{n\times k},$$

leads to the following reduced system:

$$\hat{\mathbf{A}} = \mathbf{T}_1\mathbf{A}\mathbf{T}_{i1}, \quad \hat{\mathbf{B}} = \mathbf{T}_1\mathbf{B}, \quad \hat{\mathbf{C}} = \mathbf{C}\mathbf{T}_{i1}, \quad \hat{\mathbf{D}} = \mathbf{D}.$$

The projected system is balanced, since $\mathcal{P}\mathbf{T}_1^* = \mathbf{T}_{i1}\Sigma_1$ and $\mathcal{Q}\mathbf{T}_{i1} = \mathbf{T}_1^*\Sigma_1$ imply

$$\mathbf{0} = \mathbf{T}_1(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)\mathbf{T}_1^* = (\mathbf{T}_1\mathbf{A}\mathbf{T}_{i1})\Sigma_1 + \Sigma_1(\mathbf{T}_{i1}^*\mathbf{A}^*\mathbf{T}_1^*) + (\mathbf{T}_1\mathbf{B})(\mathbf{B}^*\mathbf{T}_1^*) = \mathbf{0},$$

and also

$$\mathbf{0} = \mathbf{T}_{i1}^*(\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C})\mathbf{T}_{i1} = (\mathbf{T}_{i1}^*\mathbf{A}^*\mathbf{T}_1^*)\Sigma_1 + \Sigma_1(\mathbf{T}_1\mathbf{A}\mathbf{T}_{i1}) + (\mathbf{T}_{i1}^*\mathbf{C}^*)(\mathbf{C}\mathbf{T}_{i1}) = \mathbf{0}.$$

If we denote by $\hat{\mathbf{y}}$ the output of the reduced system, the error bound obtained tells us that the $\mathcal{L}_2$-norm of the difference $\mathbf{y} - \hat{\mathbf{y}}$ is upper bounded by twice the sum of the neglected singular values for all inputs $\mathbf{u}$ of unit norm:

$$\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \leq 2(\sigma_{k+1} + \cdots + \sigma_n).$$

It should be stressed at this point that even if the output of the system is the state $\mathbf{y} = \mathbf{x}$, in other words, $\mathbf{C} = \mathbf{I}_n$, to obtain an error bound on the behavior of the approximate system, the observability gramian has to be considered. In this case, the corresponding Lyapunov equation becomes $\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{I} = \mathbf{0}$. The projected state is $\hat{\mathbf{x}} = \mathbf{T}_1\mathbf{x}$, and the output or reconstructed state is $\tilde{\mathbf{x}} = \mathbf{T}_{i1}\hat{\mathbf{x}} = \mathbf{T}_{i1}\mathbf{T}_1\mathbf{x}$. The error bound then asserts that

$$\|\mathbf{x} - \tilde{\mathbf{x}}\|_2 = \|(\mathbf{I}_n - \mathbf{T}_{i1}\mathbf{T}_1)\mathbf{x}\|_2 \leq 2(\sigma_{k+1} + \cdots + \sigma_n)$$

for all input functions of unit energy.

Remark 9.1.1. (a) While the projection onto the dominant eigenspace of one gramian (either $\mathcal{P}$ or $\mathcal{Q}$) is orthogonal (Galerkin projection), the projection onto the dominant eigenspace of their product $\mathcal{P}\mathcal{Q}$ is an oblique (Petrov–Galerkin) projection.

(b) If the dynamical system in question is *autonomous*, that is, in the linear case $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$, the snapshots result from a preassigned initial condition $\mathbf{x}(0)$. In this case, the snapshots of the state of the autonomous system are the same as the snapshots of the impulse response of the dynamical system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, where $\mathbf{B} = \mathbf{x}(0)$.

(c) If the snapshots result from an input other than the impulse, weighted gramians become relevant. (See section 7.6 for details.)

(d) The snapshots needed to construct the observability gramian can be interpreted using the *adjoint system*. According to Theorem 4.23, the following holds: *the reachability gramian of the adjoint (dual) system is the same as the observability gramian of the original.* If the system in question is linear, the adjoint system running backward in time (see (5.15)) is

$$\Sigma^* : \ \frac{d}{dt}\mathbf{p}(t) = -\mathbf{A}^*\mathbf{p}(t) - \mathbf{C}^*\mathbf{y}(t), \ \ \mathbf{u}(t) = \mathbf{B}^*\mathbf{p}(t) + \mathbf{D}^*\mathbf{y}(t),$$

where $\mathbf{p}$ is the state of the adjoint system and $\mathbf{y}$, $\mathbf{u}$ are the input and output of Σ^*, respectively.

In numerous optimal fluid control problems, the adjoint equation enters the formulation of the problem in a natural way through the optimization. Consequently, collecting snapshots of the adjoint system is straightforward. Furthermore, this procedure can be implemented even if the dynamics are not linear. Therefore, we would like to propose instead of an orthogonal Galerkin projection, the following:

Use an oblique (Petrov–Galerkin) projection onto the dominant eigenspace of the *product* of two gramians. These gramians can be constructed from snapshots of the original system and snapshots of the adjoint system.

(e) Some historical remarks on the POD. Given its conceptual simplicity, POD is in wide use for model reduction of systems described in general by nonlinear PDEs or by nonlinear ODEs. This method, also known as the method of *empirical eigenfunctions* for dynamical systems, was introduced by Lorenz in 1956 for the study of weather prediction [231]. Subsequently, Lumley, Sirovich, and Holmes made important contributions; see [298], [63]. POD is also related to the Karhunen–Loève expansion introduced in the theory of stochastic processes in the late 1940s.

There is an extensive literature on POD methods. The papers mentioned below are a few rather randomly chosen contributions. The original work was geared toward capturing the open-loop dynamics with differential equations having few degrees of freedom (low dimension); in this regard, we mention the work of Kevrekidis and coworkers as exemplified in [98]; for more recent work in this direction, see [277] and [247]. The connection between POD and balancing has also been noted in [356]. POD methods have also been applied to control problems. We mention here the work of Burns, King and coworkers [81], [29], [30], [82], [203], [204]; Banks, Tran, and coworkers [49], [199], [200]; Kevrekidis and coworkers [4], [296]; Kunisch and Volkwein [349], [212], [213]; and Hinze [3], [174]. We would also like to mention [160], [273]. Model reduction of nonlinear systems using empirical gramians has been studied in [215], [216], [162], [163]. Also worth mentioning is the work of Fujimoto and Scherpen on balancing for nonlinear systems [124] and the work of Bamieh [40], [41] on a system theoretic approach to the study of fluids.

9.2 Modal approximation

One widespread application of the truncation approach to model reduction, besides balanced truncation, is *modal truncation*. Modal approximation is not derived from balancing. It is derived by looking at the EVD not of the product of gramians $\mathcal{PQ}$ but simply of $\mathbf{A}$. In particular, the transfer function is decomposed in terms of partial fractions and only those partial fractions are retained that have their poles closest to the imaginary axis. For details, see [341].

Assuming that $\mathbf{A}$ is diagonalizable, we can transform the state-space representation of $\mathbf{\Sigma}$ into the basis composed of the eigenvectors of $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} -\lambda_1 & & \\ & \ddots & \\ & & -\lambda_n \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{1,:} \\ \vdots \\ \mathbf{B}_{n,:} \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_{:,1} & \cdots & \mathbf{C}_{:,n} \end{bmatrix},$$

where $\mathbf{B}_{i,:} \in \mathbb{R}^{1\times m}$ denotes the ith row of $\mathbf{B}$ and $\mathbf{C}_{:,j} \in \mathbb{R}^{p\times 1}$ denotes the jth column of $\mathbf{C}$. If the system is stable, the eigenvalues of $\mathbf{A}$ can be ordered so that $\mathcal{R}e(\lambda_{i+1}) \leq \mathcal{R}e(\lambda_i) < 0$; the reduced system $\mathbf{\Sigma}_r$ obtained by truncation now preserves the k *dominant poles* (i.e., the k eigenvalues with largest real part). It can then be shown that

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}_r\|_{\mathcal{H}_\infty} \leq \sum_{i=k+1}^{n} \frac{\|\mathbf{C}_{:,i}\mathbf{B}_{i,:}\|}{|\mathcal{R}e(\lambda_i)|}.$$

In some cases, e.g., when the poles have physical significance, this can be a meaningful model reduction method. However, if the eigenvalue decomposition contains Jordan blocks, this method may fail.

The principle of choosing a subsystem, namely, the one consisting of the *dominant poles*, may not be appropriate for model reduction. We present a simple example that shows that the *nearness* of the poles to the *imaginary axis* need bear no relationship to the dominant behavior of the frequency response (i.e., to the location of the resonances of the system).

It is well known that polynomial-exponential functions of the form $t^k e^{-\lambda t}$, where $k = 0, 1, 2, \ldots$, form a basis for $\mathcal{L}_2[0, \infty]$. If we orthonormalize this family of functions, we obtain the *Laguerre basis*. Recall the following Laplace transform pairs:

$$\mathcal{L}\left\{\frac{t^n}{n!}\right\} = \frac{1}{s^{n+1}}, \quad \mathcal{L}\left\{\frac{t^n}{n!}e^{-\lambda t}\right\} = \frac{1}{(s+\lambda)^{n+1}}.$$

The (unnormalized) Laguerre functions, expressed in the Laplace domain, are as follows:

$$\mathbf{E}_0(s) = \frac{1}{s+\lambda}, \quad \mathbf{E}_1(s) = \frac{1}{s+\lambda}\frac{s-\lambda}{s+\lambda}, \quad \ldots, \quad \mathbf{E}_n(s) = \frac{1}{s+\lambda}\frac{(s-\lambda)^n}{(s+\lambda)^n}, \quad \ldots.$$

It can be shown that the above functions have the same norm, namely, $\frac{1}{\sqrt{2\lambda}}$. We now wish to expand $e^{-\mu t}$ in terms of the above functions; in the frequency domain we have

$$\frac{1}{s+\mu} = \frac{\gamma_0}{s+\lambda} + \frac{\gamma_1}{s+\lambda}\frac{s-\lambda}{s+\lambda} + \frac{\gamma_2}{s+\lambda}\frac{(s-\lambda)^2}{(s+\lambda)^2} + \cdots + \frac{\gamma_n}{s+\lambda}\frac{(s-\lambda)^n}{(s+\lambda)^n} + \cdots.$$

Taking inner products in the time domain, we obtain the following values:

$$\gamma_k(\mu) = \frac{2\lambda}{\mu+\lambda}\frac{(\mu-\lambda)^k}{(\mu+\lambda)^k}.$$

Notice that this is the transfer function of the kth basis function evaluated as $s = \mu$. We thus obtain the following expansion:

$$\begin{aligned}\frac{1}{s+\mu} &= \frac{2\lambda}{\mu+\lambda}\frac{1}{s+\lambda} + \frac{2\lambda}{\mu+\lambda}\frac{\mu-\lambda}{\mu+\lambda}\frac{1}{s+\lambda}\frac{s-\lambda}{s+\lambda} + \frac{2\lambda}{\mu+\lambda}\frac{(\mu-\lambda)^2}{(\mu+\lambda)^2}\frac{1}{s+\lambda}\frac{(s-\lambda)^2}{(s+\lambda)^2} + \cdots \\ &= \frac{2\lambda}{\mu+\lambda}\frac{1}{s+\lambda}\left[1 + \frac{\mu-\lambda}{\mu+\lambda}\frac{s-\lambda}{s+\lambda} + \frac{(\mu-\lambda)^2}{(\mu+\lambda)^2}\frac{(s-\lambda)^2}{(s+\lambda)^2} + \cdots\right] \\ &= \frac{2\lambda}{\mu+\lambda}\frac{1}{s+\lambda}\left[\frac{1}{1 - \frac{\mu-\lambda}{\mu+\lambda}\frac{s-\lambda}{s+\lambda}}\right] = \frac{2\lambda}{(\mu+\lambda)(s+\lambda) - (\mu-\lambda)(s-\lambda)} = \frac{1}{s+\mu}.\end{aligned}$$

As an example, we consider the expansion of the second-order transfer function having a pair of complex conjugate poles μ and $\bar{\mu}$:

$$\mathbf{H}(s) = \frac{1}{2i}\left[\frac{1}{s+\bar{\mu}} - \frac{1}{s+\mu}\right] = \frac{\mathcal{I}m\,\mu}{s^2 + 2\mathcal{R}e(\mu)\,s + |\mu|^2}.$$

The coefficients of the expansion of $\mathbf{H} = \sum_{i\geq 0}\alpha_i E_i$ in terms of the above orthonormal basis are

$$\alpha_k = \mathcal{I}m\,\gamma_k(\bar{\mu}).$$

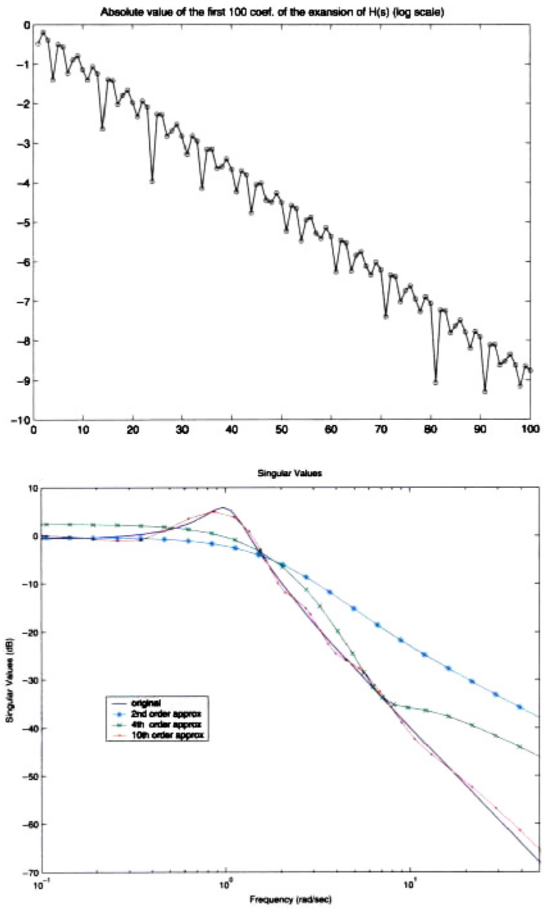


Figure 9.1. *Top: absolute value of the first* 100 *coefficients of the expansion of* **H** *(log scale). Bottom: Bode plots of the original system and three approximants.*

In particular, we take $\mu = \sigma + i\omega$, where $\sigma = 1/4$ and $\omega = 1$. The frequency response has a maximum at frequency $\frac{\sqrt{3}}{2}$, which is 2. The expansion of this transfer function in terms of $\frac{1}{s+2}$ is shown in Figure 9.1. Notice that the magnitude of the coefficients of this expansion decays exponentially, and the $\mathcal{H}_\infty$-norm of the error system between the original and a 4th-order approximation is 1.498, while that with a 10th-order approximation drops to 1.126.

Conclusion. Let the to-be-approximated system be $\mathbf{G}_N(s) = \sum_{k=1}^{N} \alpha_k \mathbf{E}_k(s)$; for large N, its frequency response comes close to that of the second-order system $\mathbf{H}(s)$. Following the above considerations, approximation of $\mathbf{G}_N$ by modal truncation gives slow convergence, although the magnitude of the coefficients α_k is exponentially decreasing. Therefore, modal truncation has difficulty capturing the behavior due to multiple poles, and as argued above, any behavior can be approximated arbitrarily by means of a single pole given high-enough multiplicity.

9.3 Truncation and residualization

Given $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, let us consider a partitioning of the state $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$, together with a conformal partitioning of the state space representation,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \quad \mathbf{C} = (\mathbf{C}_1 \quad \mathbf{C}_2).$$

As discussed in section 7.2, a reduced-order model can be obtained by eliminating $\mathbf{x}_2$, i.e., *truncating* the state. The resulting reduced system is

$$\mathbf{\Sigma}_r = \left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{B}_1 \\ \hline \mathbf{C}_1 & \mathbf{D} \end{array}\right).$$

We showed earlier (see Theorem 7.9) that if $\mathbf{\Sigma}_r$ is constructed by balanced truncation, it enjoys stability and an error bound. In general, however, the only property of the reduced system $\mathbf{\Sigma}_r$ is that its transfer function at infinity is equal to the transfer function of the original system $\mathbf{\Sigma}$ at infinity:

$$\mathbf{H}(\infty) = \mathbf{H}_r(\infty) = \mathbf{D}.$$

An alternative to state truncation is *state residualization*. If $\mathbf{A}_{22}$ is nonsingular, we can define the following reduced-order system:

$$\mathbf{\Sigma}_{\text{resid}} = \left(\begin{array}{c|c} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{B}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{B}_2 \\ \hline \mathbf{C}_1 - \mathbf{C}_2\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{D} - \mathbf{C}_2\mathbf{A}_{22}^{-1}\mathbf{B}_2 \end{array}\right).$$

This reduced-order model is obtained by *residualizing* $\mathbf{x}_2$, that is, by setting $\dot{\mathbf{x}}_2 = 0$. Residualization is equivalent to *singular perturbation approximation*; for an overview, see [208], and for a more recent account, see [51]. An important attribute of this model reduction method is that it preserves the steady state gain of the system:

$$\mathbf{H}(0) = \mathbf{H}_{\text{resid}}(0) = \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}.$$

This fact is not surprising. In [230] it is shown that reduction by truncation and reduction by residualization are related through the bilinear transformation $\frac{1}{s}$. Thus while the former provides a reduced-order system which approximates the original well at high frequencies, the latter provides a good approximation at low frequencies.

It should be mentioned that residualization can be applied to any realization of the original system $\mathbf{\Sigma}$. In particular, it can be applied to the one obtained by balancing. In the above-mentioned reference, it was shown that in this case, stability is preserved and the same error bound exists as for the reduced-order system obtained by truncating the balanced state.

9.4 A study of the decay rates of the Hankel singular values*

The issue of decay of the Hankel singular values is of interest in model reduction by means, for instance, of balanced truncation, since the sum of the neglected singular values provides an upper bound for an appropriate norm of the approximation error. This section follows [21].

The decay rate involves a new set of invariants associated with a linear system, which are obtained by evaluating a modified transfer function at the poles of the system. These considerations are equivalent to studying the decay rate of the eigenvalues of the product of the solutions of two Lyapunov equations. The related problem of determining the decay rate of the eigenvalues of the solution to one Lyapunov equation is also addressed. Very often these eigenvalues, like the Hankel singular values, are rapidly decaying. This fact has motivated the development of several algorithms for computing low rank approximate solutions to Lyapunov equations. However, until now, conditions ensuring rapid decay have not been well understood. Such conditions are derived here by relating the solution to a numerically low rank Cauchy matrix determined by the poles of the system. Bounds explaining rapid decay rates are obtained under some mild conditions.

9.4.1 Preliminary remarks on decay rates*

In the theory of function approximation by means of truncated Fourier series expansions or truncated wavelet expansions, there are explicit results relating the approximation error to the decay rate of the Fourier or wavelet coefficients. For example, the following result can be found in [100]. Consider functions f defined on the interval $[0, 1]$, which are possibly discontinuous but have *bounded variation.* If we approximate f by means of a truncated n-term Fourier series expansion or by means of a truncated n-term wavelet expansion, the *approximation error decays* asymptotically as $n^{-\frac{1}{2}}$, n^{-1}, respectively. Furthermore, if additional smoothness assumptions on f are made, the decay is faster.

Our purpose here is to explore the issue of decay of the approximation error as applied to *linear dynamical systems.* Recall that the Hankel singular values are the square roots of the eigenvalues of the product of two gramians $\mathcal{P}$, $\mathcal{Q}$ which are positive definite. As mentioned earlier, the error bound is $\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_\infty \leq 2\sum_{i=k+1}^{n}\sigma_i$. Thus, the smaller the sum of the tail of the Hankel singular values, the better the approximation.

More precisely, our purpose is first to investigate the *decay rate of the eigenvalues of one gramian*, and second to investigate the *decay rate of the eigenvalues of the product of the two gramians*, that is, the *decay rate of the Hankel singular values.*

We begin by considering only one of the two Lyapunov equations, namely,

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}. \tag{4.45}$$

In general and especially when n is large, it is unwise or even impossible to solve for $\mathcal{P}$ directly since this requires $O(n^3)$ flops and $O(n^2)$ storage. Many have observed that the eigenvalues of $\mathcal{P}$ generally decay very fast [155], [267]. Because of this, $\mathcal{P}$ may be approximated by a low rank matrix (see section 12.4). Several iterative methods for computing a low rank approximation to $\mathcal{P}$ have been proposed [155], [176], [187], [267], [284]. See [20] for a recent survey of such methods. There are some results on the eigenvalue bounds for the solution of Lyapunov equations [214], [108], but these do not explain why Lyapunov equations permit the very low rank approximate solutions observed in practice. The eigenvalue bounds surveyed in [214] focus mainly on $\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{S} = \mathbf{0}$ with $\mathbf{S} \in \mathbb{R}^{n\times n}$ positive semidefinite, and the special low rank structure of $\mathbf{S} = \mathbf{B}\mathbf{B}^*$ is not fully exploited. Moreover, the lower bounds for small eigenvalues of $\mathcal{P}$ are trivially zeros. Penzl [266] took into account the low rank structure of $\mathbf{B}\mathbf{B}^*$. He established upper bounds on the ratios

$\frac{\lambda_k(\mathcal{P})}{\lambda_1(\mathcal{P})}$ for symmetric $\mathbf{A}$, but this approach depends heavily on symmetry and is not easily generalized to the nonsymmetric case.

First, decay rates are derived that are direct estimates of the error of the best rank k approximation to one gramian $\mathcal{P}$. In contrast to the Penzl estimates, our results do not establish explicit bounds for the eigenvalues of $\mathcal{P}$. Instead, we obtain an outer product representation of the solution of the form

$$\mathcal{P} = \sum_{j=1}^{n} \delta_j \mathbf{z}_j \mathbf{z}_j^* \text{ with } \delta_1 \geq \delta_2 \geq \cdots \geq \delta_n > 0.$$

When $\mathbf{A}$ has an eigenvector basis that is not too ill-conditioned, the norms of vectors $\mathbf{z}_j$ are uniformly bounded by a modest constant, and hence the ratio δ_{k+1}/δ_1 gives an order of magnitude relative error estimate for a rank k approximation to $\mathcal{P}$.

These results lead directly to an explanation of why it is often possible to approximate $\mathcal{P}$ with a very low rank matrix. They are closely related to the eigenvalue decay rates of Penzl [266]; these results are stated in terms of the condition number of $\mathbf{A}$ which is assumed symmetric. Our bounds are functions of the eigenvalues of $\mathbf{A}$ and make no symmetry assumptions. We give some numerical results and compare the two. These results show that the bounds given here seem to give a significantly better indication of the actual decay rate of the eigenvalues of $\mathcal{P}$.

Next we turn our attention to the *Hankel singular values* of $\mathbf{\Sigma}$. Due to their importance in model reduction, in particular, balanced model reduction of large-scale systems, there has been some activity recently on the issue of the decay rate of these singular values. It has been observed that in many cases these quantities decay very fast, and therefore the corresponding systems are easy to approximate. Two recent approaches are [45] and [48].

Second, the problem of determining the decay rate of the *Hankel singular values* is discussed. It is based on a new set of system invariants. If the transfer function of the system in question is $\mathbf{H} = \frac{\mathbf{p}}{\mathbf{q}}$, the invariants are the magnitudes of $\frac{\mathbf{p}}{\mathbf{q}^*}$ evaluated at the poles of $\mathbf{H}$. The main result states that these invariants and the Hankel singular values are related by means of multiplicative majorization relations.

9.4.2 Cauchy matrices*

Our results depend heavily on properties of Cauchy matrices, which appear fundamentally in direct formulas for solutions of Lyapunov equations. Moreover, there are closed form expressions for elements of Cholesky factors and inverses of Cauchy matrices that lend themselves to derivation of the decay estimates we seek.

Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, let ξ_i, η_i denote their ith entry, respectively. The *Cauchy matrix* $\mathcal{C}(\mathbf{x}, \mathbf{y}) \in \mathbb{C}^{n\times n}$ is defined as follows:

$$\mathcal{C}(\mathbf{x}, \mathbf{y})_{ij} = \frac{1}{\xi_i + \eta_j^*}, \qquad i, j = 1, \dots, n. \tag{9.2}$$

It readily follows that $\mathcal{C}(\mathbf{x}, \mathbf{y})^* = \mathcal{C}(\mathbf{y}, \mathbf{x})$, where the superscript $*$ denotes *complex conjugation* followed by *transposition*. Define the vector $\mathbf{d}(\mathbf{x}, \mathbf{y}) \in \mathbb{C}^n$ as follows:

$$\mathbf{d}(x, y)_i = \frac{\prod_{k\neq i}(\xi_i - \xi_k)}{\prod_k (\xi_i + \eta_k^*)} \text{ and } \mathcal{D}(\mathbf{x}, \mathbf{y}) = \operatorname{diag}(\mathbf{d}(\mathbf{x}, \mathbf{y})). \tag{9.3}$$

If the components of $\mathbf{x}$ and $\mathbf{y}$ are such that no entry of $\mathbf{d}(\mathbf{x}, \mathbf{y})$ takes a value of 0 or ∞, then $\mathcal{C}(\mathbf{x}, \mathbf{y})$ is nonsingular and the following result holds.

Lemma 9.1. *With the notation established above,*

$$\mathcal{C}(\mathbf{x}, \mathbf{y}) \left[\mathcal{D}(\mathbf{y}^*, \mathbf{x}^*)\right]^{-1} \mathcal{C}(\mathbf{y}^*, \mathbf{x}^*) \left[\mathcal{D}(\mathbf{x}, \mathbf{y})\right]^{-1} = \mathbf{I}_n. \tag{9.4}$$

The proof of this result follows from the closed form expression of the inverse of the Cauchy matrix; see, e.g., section 26.1 of [171]. Actually, this result is quoted for real $\mathbf{x}$ and $\mathbf{y}$, and a slight extension is necessary to obtain the correct formula for complex x and y.

The special case $\mathbf{x} = \mathbf{y} = \ell = (\lambda_1, \lambda_2, \dots, \lambda_n)^*$ is of interest here. It turns out that when $\mathcal{C}$ is positive definite, there are explicit formulas, due to Gohberg and Koltracht [143], for the elements of the Cholesky factors.

Lemma 9.2. *If $\mathcal{R}e\,(\ell) < 0$, then $-\mathcal{C}(\ell, \ell)$ is positive definite. Moreover, if $-\mathcal{C}(\ell, \ell) = \mathbf{L}\Delta\mathbf{L}^*$ is the Cholesky factorization, then*

$$\delta_k = \frac{-1}{2\mathcal{R}e\,(\lambda_k)} \prod_{j=1}^{k-1} \left|\frac{\lambda_k - \lambda_j}{\lambda_k^* + \lambda_j}\right|^2, \tag{9.5}$$

where $\Delta = \operatorname{diag}(\delta_1, \delta_2, \dots, \delta_n)$ and $\mathbf{L}$ is lower triangular with ones on the diagonal.

Remark 9.4.1. Formula (9.4) implies that if $\mathbf{x}$ and $\mathbf{y}$ are real, a diagonally scaled version of $\mathcal{C}(\mathbf{x}, \mathbf{y})$ is $\mathbf{J}$-unitary. Let the operations $|\cdot|$ and $\sqrt{\cdot}$ on vectors or matrices be meant to apply element-wise. Let also $d(\mathbf{x}, \mathbf{y})_i = \rho_i e^{i\theta_i}$ and $\mathbf{d}(\mathbf{y}, \mathbf{x})_i = \hat{\rho}_i e^{i\hat{\theta}_i}$. Also let $\mathbf{J}_1, \mathbf{J}_2$ be diagonal matrices with $(\mathbf{J}_1)_{ii} = e^{i\theta_i}$ and $(\mathbf{J}_2)_{ii} = e^{i\hat{\theta}_i}$. There holds

$$\underbrace{\left(|\mathcal{D}(\mathbf{y}, \mathbf{x})|^{-\frac{1}{2}} \mathcal{C}(\mathbf{x}, \mathbf{y})\, |\mathcal{D}(\mathbf{x}, \mathbf{y})|^{-\frac{1}{2}}\right)}_{\Gamma} \mathbf{J}_1 \underbrace{\left(|\mathcal{D}(\mathbf{x}, \mathbf{y})|^{-\frac{1}{2}} \mathcal{C}(\mathbf{y}, \mathbf{x})\, |\mathcal{D}(\mathbf{y}, \mathbf{x})|^{-\frac{1}{2}}\right)}_{\Gamma^*} = \mathbf{J}_2.$$

Thus, if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{J}_1, \mathbf{J}_2$ are diagonal matrices of signs (i.e., ±1), and Γ is $\mathbf{J}_1, \mathbf{J}_2$-unitary.

9.4.3 Eigenvalue decay rates for system gramians*

A canonical gramian and the Cauchy kernel*

In this section, we establish the main result on the decay rates of the eigenvalues of a single Lyapunov equation. We consider the SISO case, so $\mathbf{B}$ is a column vector which will be denoted by $\mathbf{b}$, and the Lyapunov equation is

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{b}\mathbf{b}^* = \mathbf{0}.$$

We assume that $\mathbf{A}$ is a stable matrix (all eigenvalues in the open left half plane). Let $\mathbf{A}_c$ be the companion matrix for $\mathbf{A}$,

$$\mathbf{A}_c = \mathbf{J} - \mathbf{g}\mathbf{e}_n^*,$$

where $\mathbf{J}$ is a left shift matrix with ones on the first subdiagonal and zeros elsewhere. The vector $\mathbf{g}^* = (\alpha_0, \alpha_1, \ldots, \alpha_{n-1})$ defines the characteristic polynomial of $\mathbf{A}$ with $\mathbf{q}(s) = \det(s\mathbf{I} - \mathbf{A}) = \sum_{i=0}^{n} \alpha_i s^i$, $\alpha_n = 1$. Define $\mathbf{G}$ to be the solution of the *canonical Lyapunov equation,*

$$\mathbf{A}_c\mathbf{G} + \mathbf{G}\mathbf{A}_c^* + \mathbf{e}_1\mathbf{e}_1^* = \mathbf{0}.$$

Let $\mathbf{b}$ be any vector such that $(\mathbf{A}, \mathbf{b})$ is reachable and let $\mathcal{R} = [\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{n-1}\mathbf{b}]$ be the *Krylov* or *reachability* matrix. Since $(\mathbf{A}, \mathbf{b})$ is reachable, $\mathcal{R}$ is nonsingular and $\mathcal{R}\mathbf{e}_1 = \mathbf{b}$. It follows easily from the Cayley–Hamilton theorem that $\mathbf{A}\mathcal{R} = \mathcal{R}\mathbf{A}_c$. We immediately have the following lemma.

Lemma 9.3. *$\mathcal{P}$ solves $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{b}\mathbf{b}^* = \mathbf{0}$ if and only if $\mathcal{P} = \mathcal{R}\mathbf{G}\mathcal{R}^*$.*

Proof. This is easily seen by multiplying on the left and right by $\mathcal{R}$ and $\mathcal{R}^*$ to get

$$\mathbf{0} = \mathcal{R}(\mathbf{A}_c\mathbf{G} + \mathbf{G}\mathbf{A}_c^* + \mathbf{e}_1\mathbf{e}_1^*)\mathcal{R}^* = \mathbf{A}\mathcal{R}\mathbf{G}\mathcal{R}^* + \mathcal{R}\mathbf{G}\mathcal{R}^*\mathbf{A}^* + \mathbf{b}\mathbf{b}^*.$$

Since $\mathbf{A}$ is stable, this solution is unique and the lemma is proved. □

This result provides a direct relationship with the *Krylov* or *reachability* matrix $\mathcal{R}$, but further analysis is needed to derive decay rates. These rates are a function of the eigenvalues λ_j of the matrix $\mathbf{A}$. When $\mathbf{A}$ is diagonalizable, one has $\mathbf{Y}\mathbf{A}_c = \Lambda\mathbf{Y}$, where $\mathbf{Y}$ is the Vandermonde matrix of powers of the eigenvalues. The jth row of $\mathbf{Y}$ is

$$\mathbf{e}_j^*\mathbf{Y} = [1, \lambda_j, \lambda_j^2, \ldots, \lambda_j^{n-1}],$$

and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. We shall define the *Cauchy kernel* to be the matrix

$$\mathcal{C} = \mathbf{Y}\mathbf{G}\mathbf{Y}^*.$$

This kernel provides the decay rates we seek due to the following result.

Lemma 9.4. *Let $\mathbf{X}$ be a matrix of right eigenvectors for $\mathbf{A}$ so that $\mathbf{A}\mathbf{X} = \mathbf{X}\Lambda$ and assume the columns of $\mathbf{X}$ each have unit norm. Then*

$$\mathcal{C}_{ij} = \frac{-1}{\lambda_i + \lambda_j^*}$$

is Hermitian positive definite and $\mathcal{P} = \mathbf{X}_b\mathcal{C}\mathbf{X}_b^*$, where $\mathbf{X}_b = \mathbf{X}\operatorname{diag}(\hat{\mathbf{b}})$ with $\hat{\mathbf{b}} = \mathbf{X}^{-1}\mathbf{b}$.

Proof. First observe that

$$\mathcal{R} = [\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{n-1}\mathbf{b}] = [\mathbf{X}\hat{\mathbf{b}}, \mathbf{X}\Lambda\hat{\mathbf{b}}, \mathbf{X}\Lambda^2\hat{\mathbf{b}}, \ldots, \mathbf{X}\Lambda^{n-1}\hat{\mathbf{b}}] = \mathbf{X}_b\mathbf{Y}.$$

From Lemma 9.3 we have $\mathcal{P} = \mathcal{R}\mathbf{G}\mathcal{R}^* = \mathbf{X}_b\mathbf{Y}\mathbf{G}(\mathbf{X}_b\mathbf{Y})^* = \mathbf{X}_b\mathcal{C}\mathbf{X}_b^*$. Since the pair $(\mathbf{A}, \mathbf{b})$ is reachable, $\mathbf{b}$ cannot be orthogonal to any left eigenvector of $\mathbf{A}$. Hence, no component of

$\hat{\mathbf{b}} = \mathbf{X}^{-1}\mathbf{b}$ is zero and the matrix diag $(\hat{\mathbf{b}})$ is nonsingular. Moreover, $\mathbf{A}\mathbf{X}_b = \mathbf{X}_b\Lambda$ and $\mathbf{X}_b^{-1}\mathbf{b} = \mathbf{e}$, where $\mathbf{e}^* = (1, 1, \ldots, 1)$. Therefore,

$$\mathbf{0} = \mathbf{X}_b^{-1}(\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{b}\mathbf{b}^*)\mathbf{X}_b^{-*} = \Lambda\mathcal{C} + \mathcal{C}\Lambda^* + \mathbf{e}\mathbf{e}^*.$$

By inspection, one finds that $\mathcal{C}$ is a Cauchy matrix with $\mathcal{C}_{ij} = \frac{-1}{\lambda_i + \lambda_j^*}$. Since $\mathbf{A}$ is stable, Λ is stable and $\mathcal{C}$ must be positive definite. This concludes the proof. □

We are now prepared to derive a decay rate based on the eigenvalues of $\mathbf{A}$. Since $\mathcal{C}$ is positive definite, it has a Cholesky factorization. Moreover, if diagonal pivoting is included to bring the maximum diagonal element to the pivot position at each stage of the factorization, we may assume an ordering of the eigenvalues of $\mathbf{A}$ and hence a symmetric permutation of the rows and columns of $\mathcal{C}$ such that

$$\mathcal{C} = \mathbf{L}\Delta\mathbf{L}^*$$

with $\Delta = \text{diag}(\delta_1, \delta_2, \ldots, \delta_n)$ and with $\mathbf{L}$ unit lower triangular and such that each column $\mathbf{L}\mathbf{e}_j$ satisfies $\|\mathbf{L}\mathbf{e}_j\|_\infty = 1$. Formula (9.5) of Lemma 9.2 gives the δ_k in terms of the eigenvalues of $\mathbf{A}$, and one may think of the diagonal pivoting in the Cholesky factorization as a means to order (i.e., index) these eigenvalues. Let $\lambda(\mathbf{A})$ denote the spectrum of $\mathbf{A}$. If the first $k-1$ eigenvalues $\mathcal{S}_{k-1} \equiv \{\lambda_j : 1 \le j \le k-1\} \subset \lambda(\mathbf{A})$ have been selected and indexed, then the kth eigenvalue λ_k is selected according to

$$\lambda_k = \text{argmax}\left\{ \frac{-1}{2\mathcal{R}e\,(\lambda)} \prod_{j=1}^{k-1} \left| \frac{\lambda - \lambda_j}{\lambda^* + \lambda_j} \right|^2 : \lambda \in \lambda(A) \setminus \mathcal{S}_{k-1} \right\}. \tag{9.6}$$

We shall call this selection the *Cholesky ordering*. Now, given a fixed λ in the open left half plane, the function

$$\phi(\zeta) = \frac{\lambda - \zeta}{\lambda^* + \zeta}$$

is a linear fractional transformation that maps the open left half plane onto the open unit disk. Thus, $|\phi(\lambda_j)| < 1$ for every $\lambda_j \in \lambda(\mathbf{A})$. From this, we may conclude that the sequence $\{\delta_j\}$ is decreasing with $\delta_1 = \frac{-1}{2\max\{\mathcal{R}e\,(\lambda_i)\}} = \max |\mathcal{C}_{ij}|$. These results may be summarized in the following theorem.

Theorem 9.5. *Let $\mathcal{P}$ solve the Lyapunov equation $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{b}\mathbf{b}^* = \mathbf{0}$. Let $\mathcal{P}_k = \sum_{j=1}^{k} \delta_j \mathbf{z}_j \mathbf{z}_j^*$, with $\delta_1 \ge \delta_2 \ge \cdots \ge \delta_n > 0$, as given above and where $\mathbf{z}_j = \mathbf{X}_b\mathbf{L}\mathbf{e}_j$. Then*

$$\|\mathcal{P} - \mathcal{P}_k\|_2 \le (n-k)^2 \delta_{k+1} (\kappa_2(\mathbf{X})\|\mathbf{b}\|_2)^2.$$

Proof. The previous discussion has established

$$\|\mathcal{P} - \mathcal{P}_k\|_2 = \left\| \sum_{j=k+1}^{n} \delta_j \mathbf{z}_j \mathbf{z}_j^* \right\|_2 \le \delta_{k+1}(n-k) \max_{j>k} \|\mathbf{z}_j\|_2^2,$$

since the δ_j are decreasing and $\|\mathbf{z}_j\mathbf{z}_j^*\|_2 = \|\mathbf{z}_j\|_2^2$. Due to $\|\mathbf{L}\mathbf{e}_j\|_\infty = 1$, we have $\|\mathbf{L}\mathbf{e}_j\|_2 \leq (n-j+1)^{1/2}$, and thus

$$\|\mathbf{z}_j\|_2 \leq \|\mathbf{X}\|_2\|\hat{\mathbf{b}}\|_2\|\mathbf{L}\mathbf{e}_j\|_2 = \|\mathbf{X}\|_2\|\mathbf{X}^{-1}\mathbf{b}\|_2(n-j+1)^{1/2} \leq \kappa_2(\mathbf{X})\|\mathbf{b}\|_2(n-k)^{1/2}, \quad j > k.$$

This completes the proof. □

Corollary 9.6. Decay rate estimate. *When* **A** *has an eigenvector basis that is well-conditioned, the norms of the vectors* $\mathbf{z}_j$ *are uniformly bounded by a modest constant and hence the following estimate is obtained:*

$$\boxed{\frac{\lambda_k(\mathcal{P})}{\lambda_1(\mathcal{P})} \approx \frac{\delta_{k+1}}{\delta_1} = \prod_{j=1}^{k-1}\left|\frac{\lambda_k - \lambda_j}{\lambda_k^* + \lambda_j}\right|^2.} \tag{9.7}$$

This gives an order of magnitude relative error estimate for a rank k approximation to $\mathcal{P}$. *The eigenvalues* $\lambda_k(\mathcal{P})$ *are ordered in the Cholesky ordering, defined by* (9.6).

Departure from normality increases the condition number of **X** and renders this bound useless. One might see the low rank phenomenon accentuated through tiny components of $\hat{\mathbf{b}}$. On the other hand, the components of $\hat{\mathbf{b}}$ may be magnified in a way that cancels the effects of the rapidly decaying δ_j.

Generalization to MIMO*

The bounds of the previous section apply only to the case where $\mathbf{B} = \mathbf{b}$ is a single vector. However, this result still has implications for the general case. Let $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_m]$, where $\mathbf{b}_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, m$. Note that the Lyapunov equation may be written as

$$\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \sum_{i=1}^{m} \mathbf{b}_i\mathbf{b}_i^* = \mathbf{0}.$$

Let $\mathcal{P}_i$ be the solution to

$$\mathbf{A}\mathcal{P}_i + \mathcal{P}_i\mathbf{A}^* + \mathbf{b}_i\mathbf{b}_i^* = \mathbf{0}, \qquad i = 1, 2, \ldots, m.$$

From (9.5) above, we have

$$\mathcal{P}_i = \mathbf{X}_i\mathcal{C}\mathbf{X}_i^* = \mathbf{X}_i\mathbf{L}\Delta\mathbf{L}^*\mathbf{X}_i^*, \quad \text{where } \mathbf{X}_i = \mathbf{X}\mathrm{diag}\,(\mathbf{X}^{-1}\mathbf{b}_i).$$

Let

$$\mathbf{Z}_j = [\mathbf{X}_1\mathbf{L}\mathbf{e}_j,\ \mathbf{X}_2\mathbf{L}\mathbf{e}_j, \ldots, \mathbf{X}_m\mathbf{L}\mathbf{e}_j] = [\mathbf{z}_{1j},\ \mathbf{z}_{2j}, \ldots, \mathbf{z}_{mj}],$$

where $\mathbf{z}_{ij} = \mathbf{X}\mathrm{diag}\,(\mathbf{X}^{-1}\mathbf{b}_i)\mathbf{L}\mathbf{e}_j$. Linearity of the Lyapunov equation yields

$$\mathcal{P} = \sum_{i=1}^{m}\mathcal{P}_i = \sum_{j=1}^{n}\delta_j\mathbf{Z}_j\mathbf{Z}_j^*.$$

With this notation we may establish the following result.

Theorem 9.7. *Let $\hat{\mathcal{P}}_{km} = \sum_{j=1}^{k} \delta_j \mathbf{Z}_j \mathbf{Z}_j^*$. If $\frac{\delta_{k+1}}{\delta_1} < \epsilon$, then $\hat{\mathcal{P}}_{km}$ is an approximation to $\mathcal{P}$ of rank at most km which satisfies*

$$\|\mathcal{P} - \hat{\mathcal{P}}_{km}\|_2 \leq \epsilon\delta_1 m(n-k)^2(\kappa_2(\mathbf{X})\|\mathbf{B}\|_2)^2$$

with $\delta_1 \approx \|\mathcal{P}\|_2$.

Proof. Since $\mathbf{Z}_j\mathbf{Z}_j^* = \sum_{i=1}^{m} \mathbf{z}_{ij}\mathbf{z}_{ij}^*$, it follows that

$$\begin{aligned}\|\mathbf{Z}_j\mathbf{Z}_j^*\|_2 &\leq \sum_{i=1}^{m} \|\mathbf{z}_{ij}\mathbf{z}_{ij}^*\|_2 \leq m \max_i \|\mathbf{z}_{ij}\mathbf{z}_{ij}^*\|_2 \\ &\leq m \max_i (n-k)(\kappa_2(\mathbf{X})\|\mathbf{b}_i\|_2)^2 \leq m\,(n-k)(\kappa_2(\mathbf{X})\|\mathbf{B}\|_2)^2\end{aligned}$$

for $j > k$, where the estimates of Theorem 9.5 are applied to each $\|\mathbf{z}_{ij}\mathbf{z}_{ij}^*\|_2$ and the final result follows from an argument analogous to that of Theorem 9.5. □

Remark 9.4.2. There is a close connection between these bounds and the Krylov (reachability) sequence $\mathbf{A}^{k-1}\mathbf{B}$, $k = 1, \ldots, n$. In particular, with $\mathbf{A} \in \mathbb{R}^{n\times n}$ stable, $\mathbf{B} \in \mathbb{R}^{n\times m}$, let $\mathcal{P}$ be the unique solution to $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$. Then $\text{rank}\,(\mathcal{P}) = \text{rank}\,(\mathcal{R}_n)$, where $\mathcal{R}_n = [\mathbf{B}, \mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}]$, is the reachability matrix. This fact is numerically reflected in the decay rates.

A bound for the decay rate

Formula (9.7) provides an *estimate* for the decay rate of the eigenvalues of the Lyapunov equation. In Chapter 4 of [373], an *upper bound* for the decay rate was derived, which we state next. The approach used is based on the SVD.

Given the eigenvalues $\lambda_i(\mathbf{A})$, we first define the complex constants $\hat{\nu}_i \in \mathbb{C}$, $i = 1, \ldots, n$, as the solution of the following min-max problem:

$$\{\hat{\nu}_1, \ldots, \hat{\nu}_k\} = \arg \min_{\nu_1,\ldots,\nu_k} \left\{ \max_{1\leq\ell\leq n} \Pi_{i=1}^{k} \left| \frac{\nu_i - \lambda_\ell}{\nu_i^* + \lambda_\ell} \right| \right\}.$$

The following result holds.

Lemma 9.8. *Let* $\mathbf{A}$ *be stable and diagonalizable with* $\mathbf{X}$ *the right eigenvector matrix. Furthermore, let* $\mathbf{B}$ *have m columns. There holds*

$$\frac{\lambda_{mk+1}(\mathcal{P})}{\lambda_1(\mathcal{P})} \leq \kappa^2(\mathbf{X}) \left[\max_{1\leq\ell\leq n} \Pi_{i=1}^{k} \left| \frac{\hat{\nu}_i - \lambda_\ell}{\hat{\nu}_i^* + \lambda_\ell} \right|^2 \right],$$

where $\kappa(\mathbf{X})$ denotes the condition number of $\mathbf{X}$ *and the $\hat{\nu}_i$ solve the min-max problem defined above.*

The min-max problem mentioned above remains in general unsolved. A straightforward choice of these parameters is $\nu_i = \lambda_i(\mathbf{A})$, where the eigenvalues are ordered in the Cholesky ordering. In this case, we obtain the following computable upper bound.

Corollary 9.9. Computable decay rate bound. *With the choice* $\nu_i = \lambda_i(\mathbf{A})$, *the following bound holds:*

$$\boxed{\frac{\lambda_{mk+1}(\mathcal{P})}{\lambda_1(\mathcal{P})} \le \kappa^2(\mathbf{X}) \left[\max_{1\le\ell\le n} \prod_{i=1}^{k} \left|\frac{\lambda_i - \lambda_\ell}{\lambda_i^* + \lambda_\ell}\right|^2\right],} \tag{9.8}$$

where the eigenvalues $\lambda_i(\mathbf{A})$ *are ordered in the Cholesky ordering.*

This expression is similar to (9.7), although the method used to obtain it is different. The proof of the above lemma is given in [373].

9.4.4 Decay rate of the Hankel singular values*

In this section, we present a *bound* for the decay rate based on a new set of system invariants. The theory requires us to consider the transfer function of the system $\mathbf{H} = \frac{\mathbf{p}}{\mathbf{q}}$, $n = \deg \mathbf{q}$. The invariants we introduce are the magnitudes of $\frac{\mathbf{p}}{\mathbf{q}^*}$ evaluated at the poles of $\mathbf{H}$. The main result states that these invariants and the Hankel singular values are related by means of multiplicative majorization relations.

Once again, we consider the SISO case, and we assume for convenience that the eigenvalues λ_i of $\mathbf{A}$ are *distinct* and that the system is stable. The usual assumption of minimality (reachability and observability) is also made.

Recall that the *Hankel singular values* of the system $\mathbf{\Sigma}$ are defined as the square roots of the eigenvalues of the product of the reachability gramian $\mathcal{P}$ and the observability gramian $\mathcal{Q}$: $\sigma_i(\mathbf{\Sigma}) = \sqrt{\lambda_i(\mathcal{PQ})}$, $i = 1, \ldots, n$. These quantities are assumed ordered in decreasing order $\sigma_i \ge \sigma_{i+1}$. These singular values are invariant under state-space transformations.

We now introduce a new set of system invariants as follows. Let the transfer function be

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \frac{\mathbf{p}(s)}{\mathbf{q}(s)}.$$

Due to the minimality of the above realization, the eigenvalues of $\mathbf{A}$ are the same as the poles of $\mathbf{H}(s)$, that is, the roots of the denominator polynomial $\mathbf{q}(s)$, that is, $\mathbf{q}(s) = \det(s\mathbf{I} - \mathbf{A})$ $= \sum_{i=0}^{n} \alpha_i s^i$, $\alpha_n = 1$. We make use of the standard definition,

$$\mathbf{q}(s)^* = \mathbf{q}^*(-s) = \sum_{k=0}^{n} \alpha_k^*(-s)^k.$$

We now define the following quantities:

$$\gamma_i = \frac{\mathbf{p}(\lambda_i)}{\mathbf{q}^*(\lambda_i)} = \left.\frac{\mathbf{p}(s)}{\mathbf{q}(s)^*}\right|_{s=\lambda_i}, \qquad |\gamma_i| \ge |\gamma_{i+1}|,\ i = 1, \ldots, n-1. \tag{9.9}$$

Recall that the Hankel singular values of all-pass systems satisfy the property $\sigma_1 = \cdots = \sigma_n$. A consequence of the above definition is that the same holds for the γ's.

Lemma 9.10. *A system is all-pass (unitary) if and only if the* γ_i *are all equal:* $\gamma_1 = \cdots = \gamma_n$. *In this case, the Hankel singular values of the system and the* γ_i, $i = 1, \ldots, n$, *are all equal.*

The proof of this result is given in section 9.4.4.

Multiplicative majorization

To state the main result of this section, we need the notion of *multiplicative majorization*. This is a partial ordering relation between vectors. Additive and multiplicative majorization is discussed in section 3.5. For convenience, we repeat next the definition which will be needed in the sequel. Let $\gamma, \sigma \in \mathbb{R}^n$ be the vectors whose ith entry is equal to γ_i, σ_i, respectively. According to Definition 3.20, we say that σ majorizes γ multiplicatively, and write $|\gamma| \prec_\mu \sigma$, if the following relations hold:

$$|\gamma| \prec_\pi \sigma \quad \Leftrightarrow \quad \Pi_{i=1}^k |\gamma_i| \le \Pi_{i=1}^k \sigma_i, \ k = 1, \ldots, n-1, \ \text{ and } \ \Pi_{i=1}^n |\gamma_i| = \Pi_{i=1}^n \sigma_i. \quad (9.10)$$

The result quoted next is due to Sherman and Thompson [295].

Lemma 9.11. *Let the matrices* $\Gamma, \mathbf{U}, \mathbf{P}$ *satisfy the relationship* $\Gamma = \mathbf{UP}$, *where* $\mathbf{P} > 0$ *is positive definite and* $\mathbf{U}$ *is unitary. If* $\gamma_i = \lambda_i(\Gamma)$, $|\gamma_i| \ge |\gamma_{i+1}|$, *and* $\sigma_i = \lambda_i(\mathbf{P}) \ge \sigma_{i+1}$, $i = 1, \ldots, n-1$, *then* multiplicative majorization *holds, namely,*

$$|\gamma| \prec_\pi \sigma.$$

The converse is also true. Given $\mathbf{P}$ and γ_i satisfying the above multiplicative majorization relations, there exist a matrix Γ, where γ_i are its eigenvalues, and a unitary matrix $\mathbf{U}$ such that $\Gamma = \mathbf{UP}$. The main result of this section is the next theorem.

Theorem 9.12. Decay rate of the Hankel singular values. *The vector of Hankel singular values* σ majorizes *the absolute value of the vector of new invariants* γ multiplicatively:

$$\boxed{|\gamma| \prec_\pi \sigma.} \quad (9.11)$$

Remark 9.4.3. **(a)** The last relationship of the main result, namely, that the product the $|\gamma_i|$ is equal to the product of the Hankel singular values, was first reported in the discrete-time case by Mullis and Roberts [245].

(b) It follows from the majorization inequalities that $\Pi_{i=1}^k |\gamma_{n-i+1}| \ge \Pi_{i=1}^k \sigma_{n-i+1}$ for $i = 1, 2, \ldots, n-1$ and with equality holding for $i = n$. This implies

$$\sum_{i=1}^k \log |\gamma_{n-i+1}| \ge \sum_{i=1}^k \log \sigma_{n-i+1},$$

that is, the logarithmic sum of the tail of the Hankel singular values can be bounded above by the logarithmic sum of the tail of the γ_i.

Interpretation of Theorem 9.12

The issue in balanced model reduction is how fast the Hankel singular values decay to zero. In many cases, the poles (natural frequencies) of the system are known together with a state space realization $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$. Thus in principle, one can compute the γ_i, with relatively small computational effort. The main theorem then says that the (discrete) curve

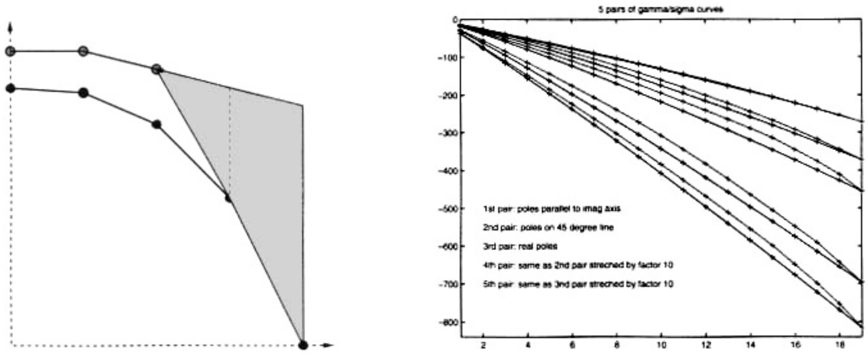


Figure 9.2. *Left:* $\sum_{i=1}^{k} \log |\gamma_i|$ *(upper curve) and* $\sum_{i=1}^{k} \log \sigma_i$ *(lower curve) versus* $k = 1, \ldots, 5$. *Right: five pairs of* γ/σ *curves for all-pole systems of order* 19.

whose kth value is the product $\Pi_{i=1}^{k} |\gamma_i|$ is of importance. It is best to plot the logarithm of this curve, namely, $\sum_{i=1}^{k} \log |\gamma_i|$, because the γ_i tend to decrease rapidly and their product even more so. The main result asserts that given this curve, the corresponding curve for the Hankel singular values $\sum_{i=1}^{k} \log \sigma_i$ remains above, and in addition the two curves have to converge at the last point $\sum_{i=1}^{n} \log \sigma_i = \sum_{i=1}^{n} \log |\gamma_i|$. Furthermore, the curves are monotone decreasing. This is depicted for a fifth-order system on the left-hand side of Figure 9.2. Stated differently, let σ_k be known. Then the σ_ℓ, $\ell > k$, lie in the region bounded from above by $\log \sigma_k$ and $\sum \log |\gamma_k|$. This means that $\sum_{i=1}^{4} \log \sigma_i$ has to lie on the dotted line, while the last two points coincide.

On the right-hand side of Figure 9.2, five pairs of γ/σ curves for all-pole systems of order 19 are depicted. The γ curve is always the lower one. The units on the y-axis are orders of magnitude. The curves are from upper to lower, as follows. The first system has poles with real part equal to -1. The next has poles which are on a 45-degree angle with respect to the negative real axis. The third has real poles only. Finally, the last two pairs have their poles spread apart by a factor of 10 with respect to the previous two pairs.

Remark 9.4.4. A consequence of the main result is that *nonminimum phase systems are harder to approximate than minimum phase ones.* Recall that minimum phase systems are systems with *stable* zeros, i.e., zeros that lie in the left half of the complex plane. This follows from the definition (9.9) since, assuming that the poles are the same, each γ_i for a minimum phase system is smaller than the γ_i for a system with all zeros in the right half of the complex plane.

The proof of Theorem 9.12

Consider again the system $\mathbf{\Sigma}$ where the previous assumptions hold. Because the poles are *distinct*, the transfer function can be decomposed in a partial fraction expansion of the type

$$\mathbf{H}(s) = \frac{\mathbf{p}(s)}{\mathbf{q}(s)} = \frac{\mathbf{p}(s)}{\Pi_{i=1}^{n}(s - \lambda_i)\,\cdot} = \mathbf{D} + \sum_{i=1}^{n} \frac{b_i}{s - \lambda_i}. \tag{9.12}$$

It follows that $\mathbf{p}(s) = \mathbf{D}\mathbf{q}(s) + \sum_{i=1}^{n} b_i \Pi_{j\neq i}(s - \lambda_j)$. Therefore, $\mathbf{p}(\lambda_i) = b_i \Pi_{j\neq i}(\lambda_i - \lambda_j)$. Notice also that

$$\mathbf{q}(s)^*|_{s=\lambda_i} = \Pi_{j=1}^{n} (\lambda_i - \lambda_j^*).$$

From the partial fraction decomposition follows the state-space realization

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right), \quad \mathbf{A} = \operatorname{diag}([\lambda_1, \ldots, \lambda_n]), \ \mathbf{B} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \mathbf{C} = [b_1 \ \cdots \ b_n].$$

Then the reachability gramian is equal to the Cauchy matrix with $\mathbf{x} = \mathbf{y} = \ell$ (the vector of eigenvalues λ_j), while the observability gramian is a diagonally scaled version of the Cauchy matrix with $x = y = \bar{\ell}$ (the vector of complex conjugate eigenvalues). Let $\mathcal{B}$ be the diagonal matrix whose entries on the diagonal are b_i. Then

$$\mathcal{P} = -\mathcal{C}(\ell, \ell) = -\begin{pmatrix} \frac{1}{\lambda_1+\lambda_1^*} & \frac{1}{\lambda_1+\lambda_2^*} & \cdots & \frac{1}{\lambda_1+\lambda_n^*} \\ \frac{1}{\lambda_2+\lambda_1^*} & \frac{1}{\lambda_2+\lambda_2^*} & \cdots & \frac{1}{\lambda_2+\lambda_n^*} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\lambda_n+\lambda_1^*} & \frac{1}{\lambda_n+\lambda_2^*} & \cdots & \frac{1}{\lambda_n+\lambda_n^*} \end{pmatrix} \in \mathbb{C}^{n\times n} \ \Rightarrow \ \mathcal{Q} = -\mathcal{B}\mathcal{C}(\ell^*, \ell^*)\mathcal{B}^*.$$

From the above remarks, it follows that $\mathbf{d}(\ell, \ell)_i = \frac{\mathbf{p}(\lambda_i)}{\mathbf{q}^*(\lambda_i)} = b_i \frac{\Pi_{j\neq i}(\lambda_i - \lambda_j)}{\Pi_j(\lambda_i + \lambda_j^*)}$. Thus from the previous discussion and in particular Lemma 9.1, we conclude that

$$\mathcal{C}(\ell, \ell)\mathcal{D}(\ell^*, \ell^*)^{-1}\mathcal{C}(\ell^*, \ell^*)\mathcal{D}(\ell, \ell)^{-1} = I_n.$$

This implies

$$\mathcal{P}\left[\mathcal{B}\mathcal{D}(\ell^*, \ell^*)\right]^{-1} \mathcal{Q}\left[\mathcal{D}(\ell, \ell)\mathcal{B}^*\right]^{-1} = I_n.$$

From here we proceed as follows. Let $\mathbf{T}$ be a balancing transformation, i.e., $\mathbf{T}\mathcal{P}\mathbf{T}^* = \mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1} = \Sigma$, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, are the Hankel singular values of $\boldsymbol{\Sigma}$. Then

$$\underbrace{\mathbf{T}\mathcal{P}\mathbf{T}^*}_{\Sigma} \underbrace{\mathbf{T}^{-*}(\mathcal{B}\mathcal{D}(\ell^*, \ell^*))^{-1}\mathbf{T}^*}_{\hat{\mathcal{D}}^{-*}} \underbrace{\mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1}}_{\Sigma} \underbrace{\mathbf{T}(\mathcal{D}(\ell, \ell)\mathcal{B}^*)^{-1}\mathbf{T}^{-1}}_{\hat{\mathcal{D}}^{-1}} = \Sigma\hat{\mathcal{D}}^{-1}\Sigma\hat{\mathcal{D}}^{-1} = \mathbf{I}_n$$

$$\Rightarrow \underbrace{\sqrt{\Sigma}\hat{\mathcal{D}}^{-*}\sqrt{\Sigma}}_{\mathbf{U}^*} \underbrace{\sqrt{\Sigma}\hat{\mathcal{D}}^{-1}\sqrt{\Sigma}}_{\mathbf{U}} = \mathbf{I}_n,$$

where $\sqrt{\Sigma}$ denotes the square root of Σ, i.e., $\sqrt{\Sigma}\sqrt{\Sigma} = \Sigma$; then the above equality implies $\sqrt{\Sigma}\,\hat{\mathcal{D}}^{-1}\sqrt{\Sigma} = \mathbf{U}$, where $\mathbf{U}$ is unitary $\mathbf{U}\mathbf{U}^* = \mathbf{U}^*\mathbf{U} = \mathbf{I}_n$. Therefore, $\sqrt{\Sigma}\,\mathbf{U}^*\sqrt{\Sigma} = \hat{\mathcal{D}}$. Then, since, $\lambda_i(\hat{\mathcal{D}}) = \mathbf{d}(\ell, \ell)_i$, and since $\lambda_i(\sqrt{\Sigma}\,\mathbf{U}^*\sqrt{\Sigma}) = \lambda_i(\mathbf{U}\Sigma)$, Lemma 9.11 together with (9.9) yields the desired result (9.11).

A sharpening of the main result

Recall (9.2) and (9.3). We associate with the ξ_i and η_i the following vectors $\delta^{\uparrow} \in \mathbb{R}^n$ and $\delta^{\downarrow} \in \mathbb{R}^n$:

$$\delta_k^{\uparrow} = \frac{|b_k|}{\xi_k + \eta_k^*}\Pi_{i=1}^{k-1}\left|\frac{\xi_k - \xi_i}{\xi_k + \eta_i^*}\right|^2, \quad \delta_k^{\downarrow} = \frac{|b_k|}{\xi_k + \eta_k^*}\Pi_{i=k+1}^{n}\left|\frac{\xi_k - \xi_i}{\xi_k + \eta_i^*}\right|^2, \qquad k = 1, 2, \ldots, n.$$

It is assumed in the following that the ξ_i are arranged so that the entries of $\delta^{\uparrow}$ turn out to be in *decreasing Cauchy ordering*: $\delta_k^{\uparrow} \geq \delta_{k+1}^{\uparrow}$. From our earlier discussion, it follows that if the system is given by (9.12), the γ_i are

$$\gamma_k = \frac{b_k}{\xi_k + \eta_k^*} \Pi_{i \neq k} \frac{\xi_k - \xi_i}{\xi_k + \eta_i^*}.$$

Notice that $\delta^{\uparrow} \odot \delta^{\downarrow} = \gamma \odot \bar{\gamma} = |\gamma|^2$, where $\odot$ denotes pointwise multiplication. It should also be noticed that, as mentioned earlier [143], the lower-diagonal-upper (L-D-U) factorization of the Cauchy matrix can be written explicitly as $\mathcal{C} = \mathbf{L}\Delta\mathbf{U}$, where $\mathbf{L}$ is lower triangular with ones on the diagonal, $\mathbf{U}$ is upper triangular with ones on the diagonal, and $\Delta = \operatorname{diag} \delta^{\uparrow}$.

Because of the above-mentioned fact on the L-D-U factorization, we have the following result due to Horn [180]; it states that the vector of eigenvalues of a positive definite matrix majorizes multiplicatively the vector whose ith entry is the quotient of the determinants of the ith over the $(i-1)$st principal minors of the matrix.

Lemma 9.13. *Let $\mathcal{P}$ be the solution to the Lyapunov equation with eigenvalues $\lambda_i(\mathcal{P})$. Then the vector $\delta^{\uparrow}$ is majorized multiplicatively by the vector of eigenvalues $\lambda(\mathcal{P})$. Furthermore, the former multiplicatively majorizes the vector γ:*

$$|\gamma| \prec_{\pi} \delta^{\uparrow} \prec_{\pi} \lambda(P). \tag{9.13}$$

Remark 9.4.5. **(a)** In section 9.5, it was shown that the optimal conditioning of Cauchy matrices by means of diagonal scaling is given by

$$\beta_k = |\xi_k + \eta_k^*| \prod_{i \neq k} \left| \frac{\xi_k + \eta_i^*}{\xi_k - \xi_i} \right|.$$

This implies that for optimally conditioned Lyapunov equations we must have $\delta^{\uparrow} \odot \delta^{\downarrow} = e^* = [1 \ 1 \ \cdots \ 1]$.

(b) The inequalities (9.13) are a refined version of the result concerning the decay rate of the eigenvalues of a single gramian; see Theorem 9.5.

Connection with Fuhrmann's results

In section 8.6, the signed Hankel singular values μ (i.e., the eigenvalues) of the Hankel operator, and the corresponding eigenvectors, are characterized by means of the polynomial equation (8.28); using the notation of this section, it becomes $\mathbf{p}\mathbf{r} = \mu \mathbf{q}^* \mathbf{r}^* + \mathbf{q}\pi$, where $\mathbf{H} = \frac{\mathbf{p}}{\mathbf{q}}$ is the transfer function of the system, σ is a Hankel singular value, $\mu = \epsilon\sigma$, $\epsilon = \pm 1$ is the corresponding signed singular value, $(\cdot)^*$ is as defined earlier, and $\mathbf{r}$, π are unknown polynomials of degree less than $n = \deg \mathbf{q}$. Let $\mathbf{a}$, $\mathbf{b}$ be polynomials satisfying the Bezout equation $\mathbf{a}\mathbf{q} + \mathbf{b}\mathbf{q}^* = 1$. Then the coefficients of the polynomial $\mathbf{r}$ and the eigenvalues μ of the Hankel operator are determined from the eigenvalue decomposition of the matrix

$$\mathbf{M} = \mathbf{K}\mathbf{b}(\mathbf{A})\mathbf{p}(\mathbf{A}), \quad \text{where } \mathbf{K} = \operatorname{diag}(1, -1, 1, -1, \ldots),$$

and $\mathbf{A}$ has characteristic polynomial equal to $\mathbf{q}$.

We notice that this equation can be rewritten as follows:

$$\frac{\mathbf{p}}{\mathbf{q}^*} = \mu\frac{\mathbf{r}^*}{\mathbf{r}} + \frac{\mathbf{q}}{\mathbf{q}^*}\frac{\pi}{\mathbf{r}} \Rightarrow \gamma_i = \left.\frac{\mathbf{p}}{\mathbf{q}^*}\right|_{s=\lambda_i} = \left.\mu\frac{\mathbf{r}^*}{\mathbf{r}}\right|_{s=\lambda_i},$$

where $-\lambda_i$ are the roots of $\mathbf{q}$ (poles of Σ). It follows, therefore, that γ_i is equal to some singular value times the magnitude of the corresponding all-pass $\frac{\mathbf{r}^*}{\mathbf{r}}$, evaluated at $s = \lambda_i$; to stress the dependency on the jth singular value, we write r_j instead of $\mathbf{r}$,

$$\gamma_i = \frac{\mathbf{p}(\lambda_i)}{\mathbf{q}^*(\lambda_i)} = \sigma_j\frac{\mathbf{r}_j^*(\lambda_i)}{\mathbf{r}_j(\lambda_i)}, \qquad j = 1, \ldots, n. \tag{9.14}$$

Thus the ratio of γ_i and σ_j is a quantity depending on the all-pass function defined by the polynomial $\mathbf{r}$, which in turn defines the eigenvectors of the Hankel operator.

We are now ready to prove Lemma 9.10. Necessity: if the system is all-pass, i.e., $\mathbf{H} = \kappa\frac{\mathbf{q}^*}{\mathbf{q}}$ for some $\kappa \in \mathbb{R}$, by definition (9.9) of the gammas we have $\gamma_i = \kappa$ for all i. Sufficiency: if all the γ_i, $i = 1, \ldots, n$, are equal, since the degree of $\mathbf{r}$ is at most $n - 1$, (9.14) implies $\mathbf{r} = \mathbf{r}^*$. Therefore, $\mathbf{p} = \kappa\mathbf{q}^*$, where $\kappa = \gamma_i = \sigma_j$ for all i, j.

9.4.5 Numerical examples and discussion*

In this section, we illustrate the effectiveness of the *Cholesky estimates* for decay rates of the eigenvalues of the system gramians derived in section 9.4.3 (see Theorem 9.5). We have observed essentially the same quality of results for the decay rate estimates for the Hankel singular values. As the two are intimately related, we report on the results for the eigenvalues of only one gramian here.

Although the results of section 9.4.3 do not establish direct bounds on the eigenvalues of $\mathcal{P}$, they seem to predict the behavior of these eigenvalues quite well. We illustrate this with some computational results and compare our estimates to those derived by Penzl in [266]. To our knowledge, these were the only eigen-decay estimates available prior to the results given here. Penzl's results are only valid for symmetric $\mathbf{A}$. When $m = 1$ (the SISO case), these bounds are

$$\frac{\lambda_{k+1}(\mathcal{P})}{\lambda_1(\mathcal{P})} \leq \left(\prod_{j=0}^{k-1}\frac{\kappa(\mathbf{A})^{(2j+1)/(2k)} - 1}{\kappa(\mathbf{A})^{(2j+1)/(2k)} + 1}\right)^2. \tag{9.15}$$

Our estimates are derived from the diagonal elements of the Cholesky factor of the Cauchy matrix $\mathcal{C}$ defined by (9.2) and do not require symmetry of $\mathbf{A}$. They are of the form

$$\delta_k = \frac{-1}{2\mathcal{R}e\,(\lambda_k)}\prod_{j=1}^{k-1}\left|\frac{\lambda_k - \lambda_j}{\bar{\lambda}_k + \lambda_j}\right|^2, \tag{9.16}$$

where the λ_j have been indexed according to the Cholesky ordering imposed by diagonal pivoting.

Although these two results are derived in very different ways, they are closely related. In fact, Penzl derives his bounds from expressions involving the same linear fractional transformations that led to our results. However, in that case, they arose from alternating direction implicit (ADI) approximations to the solution of the Lyapunov equation (see section 12.4).

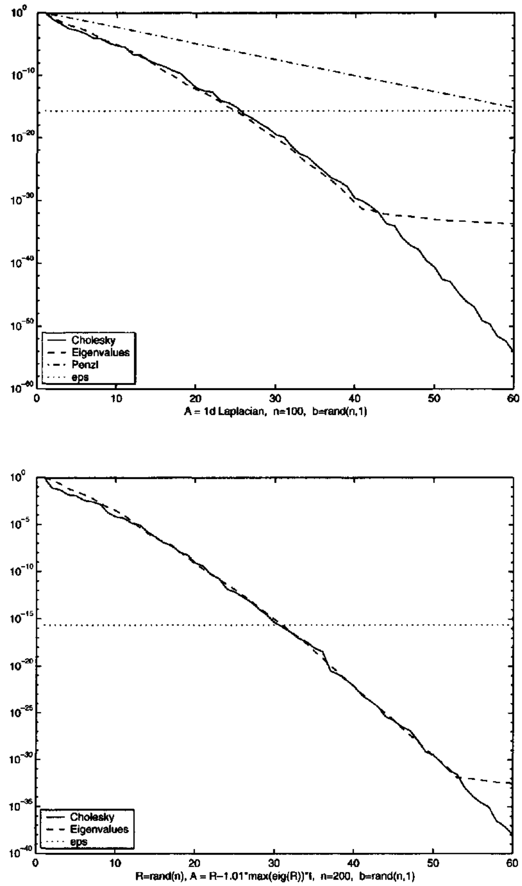


Figure 9.3. *Top: comparison of estimates for the discrete Laplacian. Bottom: comparison of estimates for a nonsymmetric* **A**.

Our computations indicate that a great deal may be lost in replacing estimates involving all of the eigenvalues λ_j of **A** with a condition number. In Figure 9.3, we show the results of our estimates versus Penzl's in comparison to the eigenvalues of $\mathcal{P}$, where **A** is the standard finite difference approximation to the one-dimensional Laplacian of order 100.

The upper pane of Figure 9.3 gives a semilog graph of the Cholesky estimates δ_k/δ_1, the Penzl estimates, and the actual eigenvalue ratios $\lambda_k(\mathcal{P})/\lambda_1(\mathcal{P})$ for $1 \leq k \leq 60$. The horizontal dotted line indicates where these ratios fall below machine precision $eps \approx 10^{-16}$. In the lower pane of Figure 9.3, we show the same comparison for a random nonsymmetric **A** of order 200 with a few eigenvalues near the imaginary axis (distance about .01).

In Figure 9.4, we compare Cholesky estimates to actual eigen-decay on LTI examples that are more closely related to engineering applications. Two of the examples are simple model problems, a finite element model of heat conduction in a plate with boundary controls

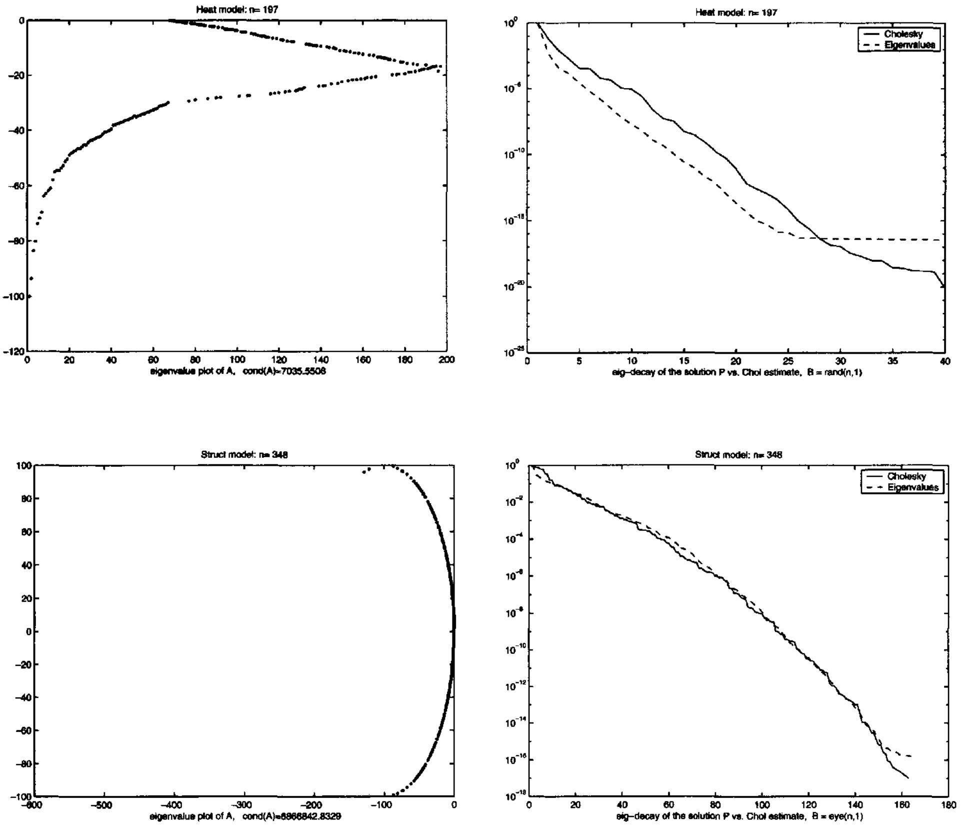


Figure 9.4. *Eigen-decay rate versus Cholesky estimates for some real LTI systems. Upper plots: Heat model $n = 197$. Lower plots: structural model $n = 348$. Left panes: system eigenvalues. Right panes: decay rates.*

and a finite element model of a clamped beam with a control force applied at the free end. These are labeled the heat model and the struct model, respectively. A third example, labeled CD player model, is a simplified simulation of a CD player tracking mechanism that has been described in detail in [153]. The fourth example, labeled ISS 1r-c04 model, is an actual finite element discretization of the flex modes of the Zvezda service module of the ISS. There are three inputs and three outputs, namely, the roll, pitch, yaw jets and the roll, pitch, yaw rate gyros readings, respectively. This example was provided to us by Draper Labs. Since $\mathbf{A}$ is nonsymmetric in all of these systems, the Penzl estimates do not apply. However, the Cholesky ratios give very good approximations to the actual eigen-decay rates for all of these examples.

These results indicate that the condition number $\kappa(\mathbf{A})$ alone may not be enough to determine the decay rate effectively. It seems that the decay rate really depends on the full spectrum of $\mathbf{A}$ as indicated by the Cholesky estimates. Moreover, one can easily see that clustered eigenvalues of $\mathbf{A}$ can make the Cauchy kernel have very low numerical rank (hence

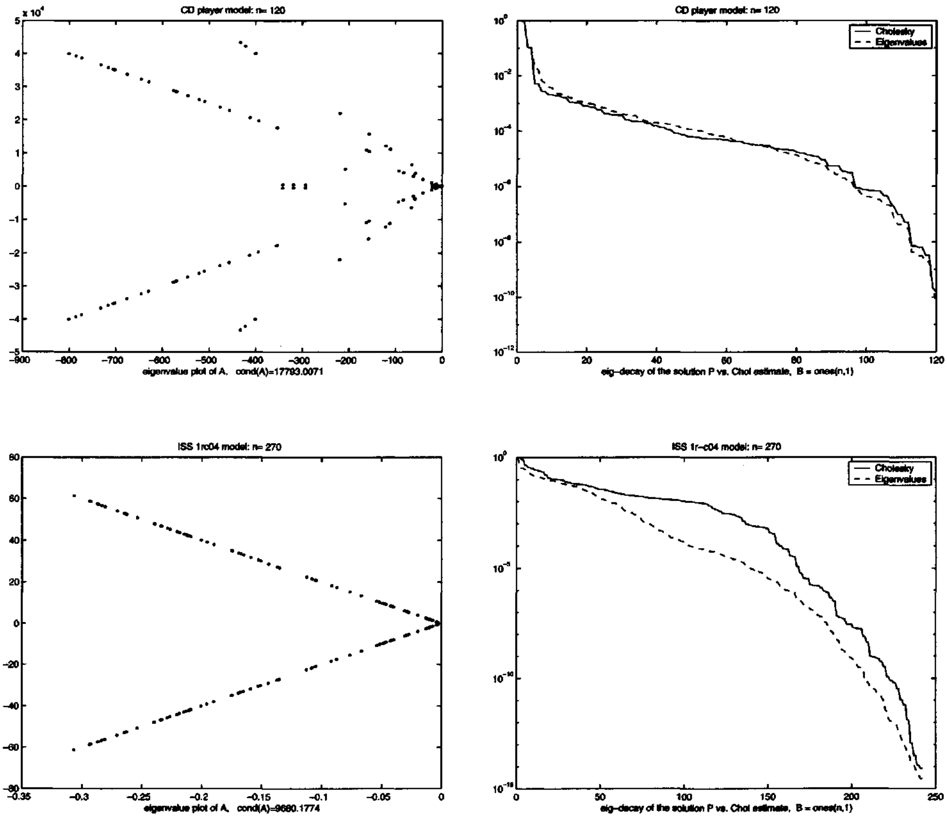


Figure 9.4. *(Continued). Upper plots: CD model. Lower plots: ISS module. Left panes: system eigenvalues. Right panes: decay rates.*

$\mathcal{P}$ has very low numerical rank), but $\kappa(\mathbf{A})$ can be arbitrarily large at the same time. Hence the right-hand side in (9.15) fails to predict the rapid eigen-decay rate. In fact, the eigen-decay rate can be extremely fast even when $\kappa(\mathbf{A})$ is huge. This is illustrated by the example shown in Figure 9.5, where $\mathbf{A}$ has been constructed to have two clustered eigenvalues as follows:

```
for iter = 1 : 4
    A = - 10^(6+iter)*eye(n) + diag(rand(n,1))*100;
    A(n,n)= -1e-5;
end
```

For the nonsymmetric case, Penzl [266] constructed a nice example to show that even in the SISO case, the eigen-decay rate can be arbitrarily slow. In the extreme case, the ratios remain almost constant at 1. His example is as follows: put $n = 2d + 1$, $m = 1$, $\mathbf{A} = \operatorname{diag}(-1, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d) \in \mathbb{R}^{n \times n}$, $\mathbf{B} = [1, \ldots, 1]^* \in \mathbb{R}^n$, where

$$\mathbf{A}_j = \mathbf{A}_j(t) = \begin{bmatrix} -1, & jt/d \\ -jt/d, & -1 \end{bmatrix}, \qquad j = 1, 2, \ldots, d.$$

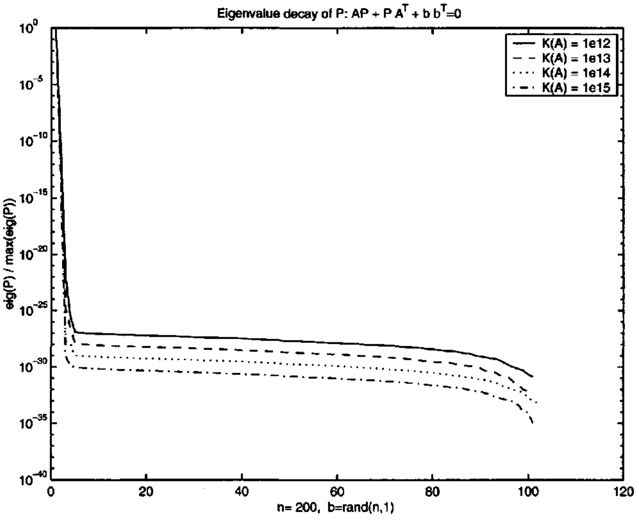


Figure 9.5. *Fast eigen-decay rate for large cond(*$\mathbf{A}$*).*

Penzl considered $d = 50$, $t = 10, 100, 1000$. From the construction of $\mathbf{A}$, $\kappa(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 \simeq t$. He observed that the eigenvalue ratios are almost constant 1 for large t.

We tried this example with $d = 300$, $t = 10, 100, 10^3, 10^4$. As can be seen from the top of Figure 9.6, the eigen-decay rate of $\mathcal{P}$ slows down when $\mathbf{A}$ has eigenvalues with increasingly dominant imaginary parts. In [266], Penzl suggested that the dominant imaginary parts of the eigenvalues cause the decay rate to slow down. This observation is relevant, but further analysis seems to indicate that relative dominance of the imaginary parts to the real parts, together with the absence of clustering in the spectrum of $\mathbf{A}$, are very important factors.

We illustrate this in the bottom of Figure 9.6 by constructing $\mathbf{A}$ with eigenvalues having huge imaginary parts but with all eigenvalues clustered about three points: $\{-1,\ -1+t\,i,\ -1-t\,i\}$. Again we construct $n = 2d+1$, $\mathbf{A} = \operatorname{diag}(-1, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d) \in \mathbb{R}^{n\times n}$, $\mathbf{B} = [1, \ldots, 1]^* \in \mathbb{R}^n$, while we modified $\mathbf{A}_j$ as follows:

$$\mathbf{A}_j = \mathbf{A}_j(t) = \begin{bmatrix} -1, & t+j/d \\ -t-j/d, & -1 \end{bmatrix}, \qquad j = 1, 2, \ldots, d.$$

Again, we take $t = 10, 100, 10^3, 10^4$ for comparison. In this example, despite the presence of eigenvalues with increasingly dominant imaginary parts, the eigen-decay rate of $\mathcal{P}$ does not deteriorate because of the clustered eigenvalues. For each t in this example, $\mathbf{A}$ has only three clustered eigenvalues, and thus the Cauchy kernel (9.2) has low numerical rank for each t. Hence, the eigen-decay rate of $\mathcal{P}$ continues to be fast regardless of the magnitude of t as demonstrated in Figure 9.6 (bottom). We also see that $\kappa(\mathbf{A})$ is irrelevant to the decay rate in this example since here $\kappa(\mathbf{A}) \simeq t$ for each t.

Actually, (9.15) gives an explanation of the nondecay in Penzl's example. In this example, $\mathcal{R}e\,(\lambda)$ remains constant and the increasingly dominant Imag(λ) leads to $\left|\frac{\lambda_k-\lambda_j}{\lambda_k+\lambda_j}\right| \rightarrow 1$.

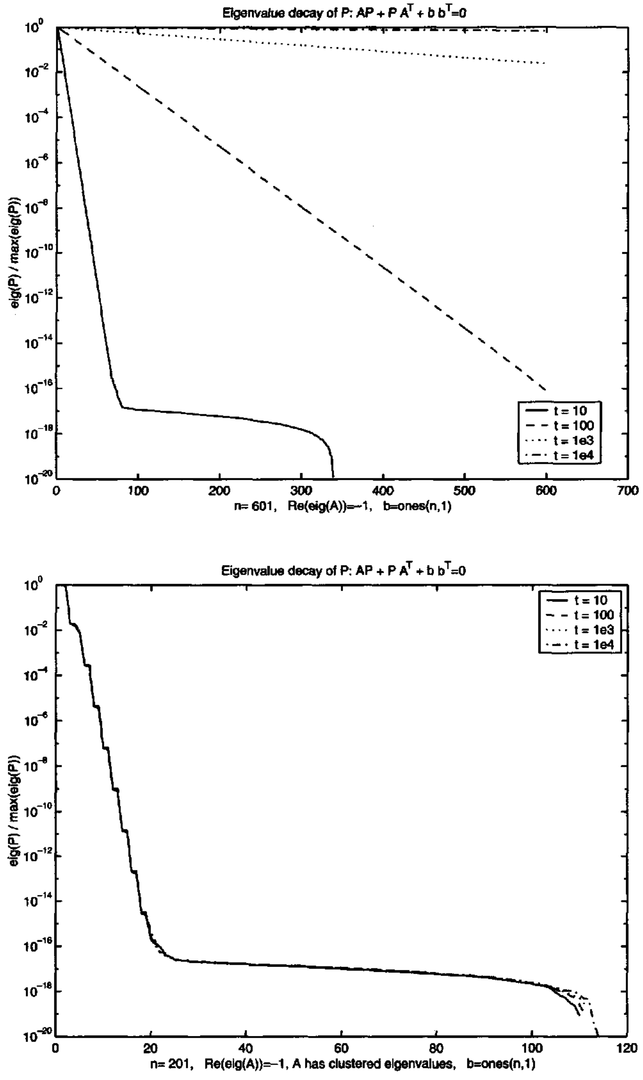


Figure 9.6. *Top: decay rates for Penzl's example. Bottom: fast eigen-decay rate for* **A** *with clustered eigenvalues having dominant imaginary parts.*

Hence, $\{\delta_k\}$ becomes nearly a constant sequence and there is very little decay of the eigenvalues.

The Cholesky estimates are not shown in Figures 9.5 and 9.6 to avoid the clutter of many overlaid plots. However, in all these cases, the Cholesky ratios were computed and again they approximate the actual eigen-decay rates well. All computations were done in MATLAB 5.3.0 on a Sun™ Sparc® Ultra-60 under SunOS 5.6. Machine epsilon is approximately 10^{-16}.

Cases of nondecay or slow decay

Penzl's example is one slow decay case. In fact, the worst case is possible. The Lyapunov equation may have the solution $\mathcal{P} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. As an example, if $\mathbf{T} = -\mathbf{T}^*$, choose $\mathbf{A} = \mathbf{T} - \mu\, \mathbf{e}_1 \mathbf{e}_1^*$, $\mathbf{b} = \sqrt{2\mu}\, \mathbf{e}_1$ for some $\mu > 0$. Then

$$\mathbf{AI} + \mathbf{IA}^* + \mathbf{bb}^* = \mathbf{0}.$$

Thus, even in the SISO case, it is possible for the solution to have no eigen-decay; these cases are related to *all-pass* systems. We also did extensive numerical studies on these nondecay cases; the most relevant reasons we found are (i) most eigenvalues of $\mathbf{A}$ have tiny real parts with not-so-tiny imaginary parts; (ii) most eigenvalues of $\mathbf{A}$ have dominant imaginary parts and the spectrum of $\mathbf{A}$ is not clustered relative to $\max_j |\mathcal{R}e\,(\lambda_j)|$. However, Figure 9.7 shows that even in the slow decay or nondecay case, (9.16) still approximates

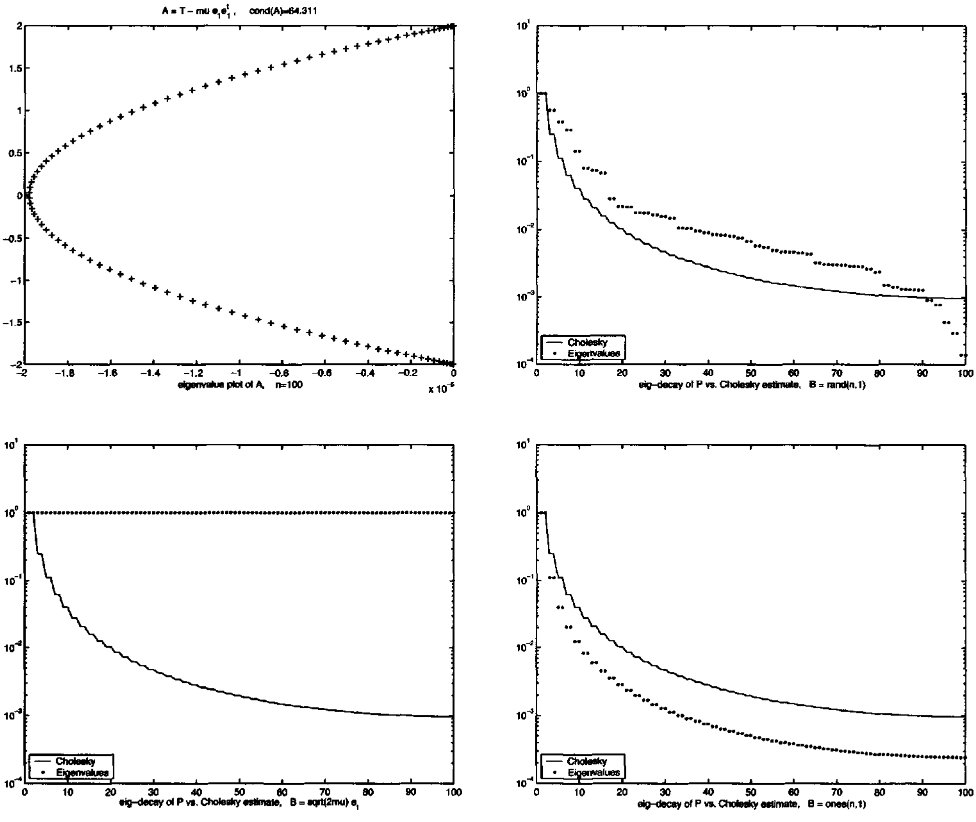


Figure 9.7. *Eigen-decay rate and Cholesky estimates for the slow decay cases. Upper plots: Random* $\mathbf{A}$ *$n = 100$, with* $\mathbf{B} = \text{rand}(n, 1)$. *Lower plots: same* $\mathbf{A}$, $\mathbf{B} = \text{ones}(n, 1)$. *Left panes: eigenvalues of* $\mathbf{A}$. *Right panes: decay rates.*

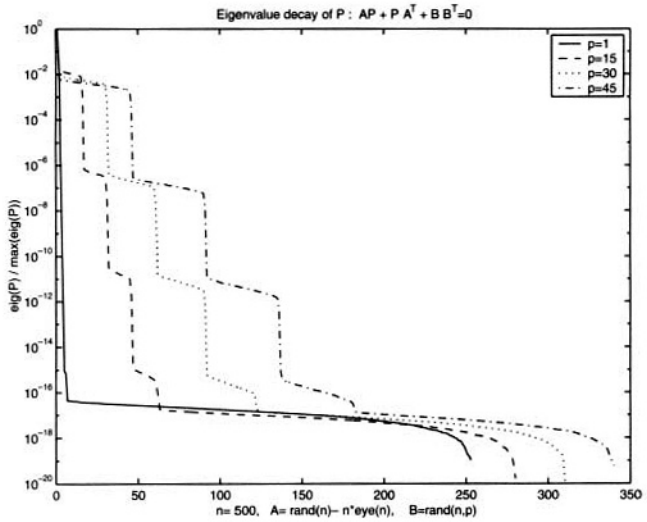


Figure 9.8. *MIMO cases: the effect of different p.*

the actual decay rate well. In engineering applications, it seems to be uncommon for the conditions for slow decay to arise.

Remark 9.4.6. **(a)** The Cholesky ratio (9.16) involves only the eigenvalues of **A**, while $\mathcal{P}$ depends on both **A** and **b**, so there may be considerable discrepancy between the Cholesky ratio and the actual eigen-decay rate for some **b**. However, when **A** has a moderately well-conditioned eigenvector basis, (9.16) usually gives an accurate prediction of the numerical rank of $\mathcal{P}$.

(b) On the other hand, **b** is taken into account in Theorem 9.12 and in the refined bounds (9.13).

The result of Theorem 9.7 suggests that the rank of $\mathcal{P}$ increases with the rank of **B**. Figure 9.8 gives one example of the effect of increasing m (the number of columns of **B**). Note in this example that the spectrum of **A** contains relatively clustered points with dominant real parts, which implies the fast eigen-decay rate.

Further numerical experiments

The plots in this section are due to Embree [105]. The goal in each case is to pick some region in the complex plane, generate optimal interpolation points, and then compare the δ_i obtained from those points to the δ_i obtained from the eigenvalues, keeping in mind that the δ_i are meant to provide an order of magnitude relative error estimate for the numerical rank of $\mathcal{P}$. Three examples are investigated.

Discrete Laplacian in one dimension, $n = 400$. In Figure 9.9, we illustrate the values based on the true eigenvalues of **A** (middle curve) and compare these values to the values

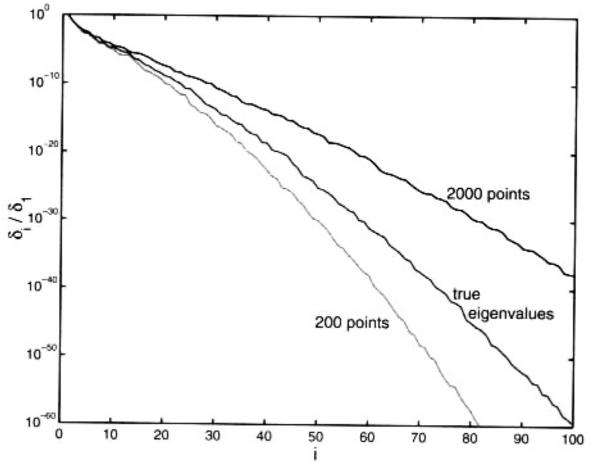


Figure 9.9. *Eigenvalue decay for the discrete Laplacian.*

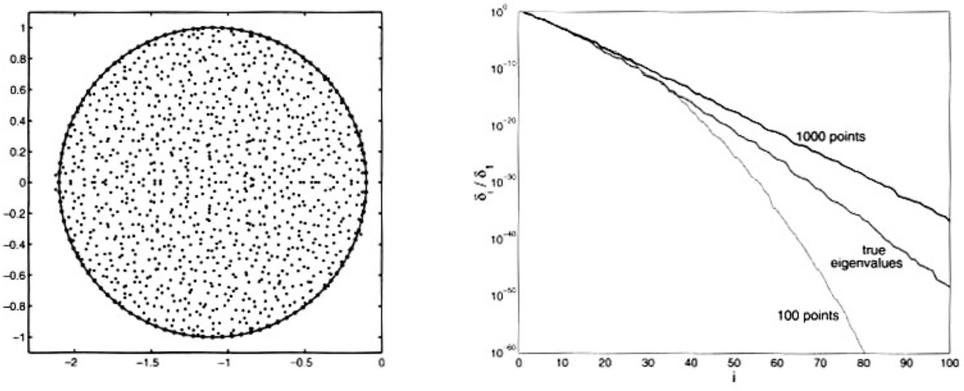


Figure 9.10. *Eigenvalue decay for shifted random matrices.*

generated from 200 points (lower curve) and 2000 points (upper curve) distributed like the Chebyshev points in $[\lambda_{min}, \lambda_{max}]$, that is, filling a sharp eigenvalue inclusion region.

Shifted random matrix, $n = 1000$ (Figure 9.10). Let $\mathbf{A}$ be generated as follows:

$$\mathtt{A} = \mathtt{randn(n)/sqrt(n)} - \mathtt{1.1} * \mathtt{eye(n)}.$$

The entries of $\mathbf{A}$ are thus normally distributed with mean $\mu = -1.1$ and variance $\sigma^2 = n^{-1}$. With this scaling, the spectrum of $\mathbf{A}$ will fill the disk of radius 1 centered at 1.1 in the large n limit. We approximate the spectrum by this disk, as shown in Figure 9.10. Notice that some of the eigenvalues of $\mathbf{A}$ are slightly outside this region. We use 100 uniformly spaced

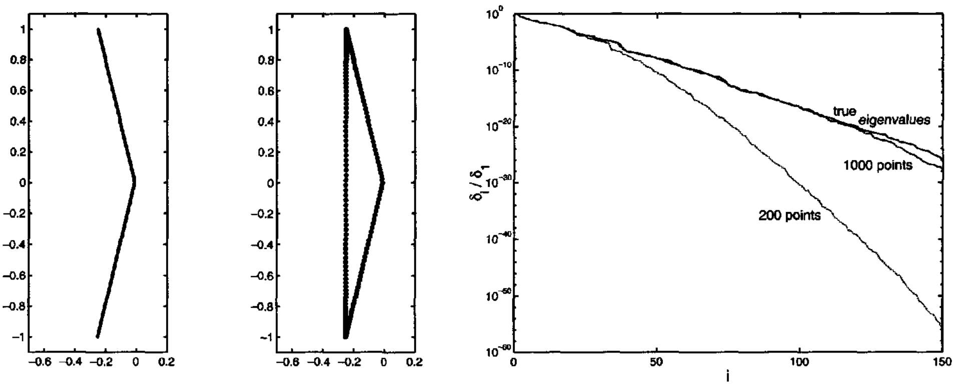


Figure 9.11. *Eigenvalue decay for V-shaped spectrum.*

points on the boundary of the disk to get approximate values. This classic example from random matrix theory exhibits only mild nonnormality.

The right pane of Figure 9.10 compares the decay of the true δ_i values (middle curve) with the δ_i values obtained from the 100 points (lower curve) and 1000 points (upper curve) on the disk boundary. Notice the good agreement for the first 25 or so values of δ_i, at which point discrete effects begin to take over.

Matrix with a V-shaped spectrum, n = 1000 (Figure 9.11). Here, we let A have eigenvalues uniformly distributed on two sides of a triangle, with the rightmost eigenvalue $\lambda = 0.01$ at the vertex, having multiplicity 2. For our comparison, we use 200 points derived from a conformal map of the convex hull of the eigenvalues of A, generated using Driscoll's Schwarz–Christoffel toolbox for MATLAB. The true eigenvalues are shown in the left plot of Figure 9.11; the approximation points are shown in the middle pane.

Lastly (Figure 9.12), we modify the previous example by supposing that we don't have such an accurate inclusion region for the eigenvalues. We bound the spectrum with a pentagon having a vertex at 0.005. The left pane shows the eigenvalues and the set of 200 approximation points. We now see that this new region generally leads to an overestimate of the decay behavior. If in practice one only requires an order of magnitude estimate (and, indeed, is more willing to accept an overestimate of the numerical rank rather than an underestimate), this may still be acceptable.

9.5 Assignment of eigenvalues and Hankel singular values*

We will now briefly address the issue of conditioning of A and of the Hankel singular values of a linear system. *Two* important invariants are associated with a system Σ:

- the natural frequencies or poles $\lambda_i(\Sigma)$, of Σ,
- the Hankel singular values $\sigma_i(\Sigma)$, of Σ.

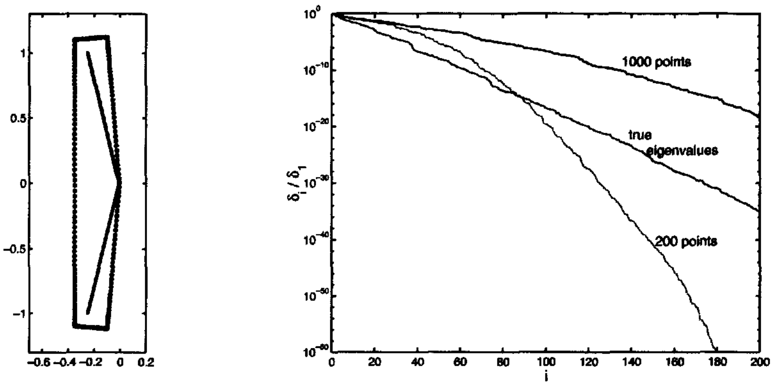


Figure 9.12. *Eigenvalue decay for a spectrum included in a pentagon-shaped region.*

The former quantities are defined as the eigenvalues of $\mathbf{A}$: $\lambda_i(\Sigma) = \lambda_i(\mathbf{A})$, $i = 1, \ldots, n$. The latter are defined as the singular values of the *Hankel operator* $\mathcal{H}_\Sigma$, associated with Σ. Next, we investigate the relationship between these two quantities. This section follows [18].

As pointed out in earlier chapters, the nonzero singular values of Σ can be computed as $\sigma_i(\Sigma) = \sqrt{\lambda_i(\mathcal{P}\mathcal{Q})}$, $i = 1, \ldots, n$, where $\mathcal{P}$, $\mathcal{Q}$ are the reachability, observability gramians, respectively. As before $\Sigma = \text{diag}\ (\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n \times n}$ denotes the Hankel singular values.

The eigenvalues $\lambda_i(\Sigma)$ describe the dynamics of the system, while the Hankel singular values $\sigma_i(\Sigma)$, just as the singular values in the case of constant matrices, describe how well Σ can be approximated by a system of lower dimension using balanced truncation or Hankel-norm approximation. Approximation works in this case by truncating the small Hankel singular values, and these quantities provide a priori computable information on how well a dynamical system can be approximated by a system of prespecified order.

A system representation is balanced if the solution of both Lyapunov equations is equal and diagonal, i.e., $\mathcal{P} = \mathcal{Q} = \Sigma$. As discussed earlier, every system Σ which is stable, reachable, and observable has a balanced representation. We can thus assume without loss of generality that such a system Σ is in balanced form. In this case, the matrices have a special form, which is referred to as *balanced canonical form*; see section 7.4 for details. In the case that the singular values are distinct, this form is

$$\mathbf{B}_i = \gamma_i > 0,\ \mathbf{C}_i = s_i\gamma_i,\ \mathbf{A}_{ij} = \frac{-\gamma_i\gamma_j}{s_i s_j \sigma_i + \sigma_j},\ \text{ where } s_i = \pm 1,\ i = 1, \ldots, n.$$

In other words, the s_i are signs associated with the singular values σ_i; the quantities $s_i\sigma_i$, $i = 1, \ldots, n$, turn out to be the eigenvalues of the Hankel operator $\mathcal{H}$. Finally, recall that due to the reachability of $(\mathbf{A}, \mathbf{B})$ and to the observability $(\mathbf{C}, \mathbf{A})$, the γ_i must be different from zero, and thus without loss of generality they can be chosen positive $\gamma_i > 0$. From the above relationships, it follows that $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{B}_d\mathbf{A}_0\mathbf{B}_d,\quad (\mathbf{A}_0)_{ij} = \frac{-1}{s_i s_j \sigma_i + \sigma_j},\quad \mathbf{B}_d = \text{diag}\,(\gamma_1, \ldots, \gamma_n), \tag{9.17}$$

where $\gamma_i, \sigma_i > 0$, and $s_i = \pm 1$. The problem that arises is the *relationship* between $\lambda_i(\Sigma)$ and $\sigma_i(\Sigma)$.

Numerical experiments. (I) If the eigenvalues $\lambda_i(\Sigma)$ of Σ are fixed, but $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are otherwise randomly chosen, the condition number of the singular values turns out to be on the average of the order of 10^n, where n is the system dimension. (II) If the singular values $\sigma_i(\Sigma)$ of Σ are fixed, but the system is otherwise randomly chosen, the condition number of the resulting $\mathbf{A}$ is on the average of the order of $10^{\frac{n}{2}}$.

Ensuing problems. The above numerical experiments suggest two specific problems: to what extent is it possible to reduce (I) the condition number of the singular values in the first case above, and (II) the condition number of $\mathbf{A}$ in the second case, by *appropriate choice of the quantities which have been chosen randomly*?

From the numerical experiments outlined earlier, we conclude that the Hankel singular values of generic linear systems fall off rapidly, and hence such systems can be approximated well by systems of low dimension (complexity). The issue raised here is thus concerned with determining the extent to which nongeneric systems can be approximated well by low-dimensional systems.

Problem (I) is addressed as follows. Given a pair of matrices $(\mathbf{C}, \mathbf{A})$, with $\mathbf{A}$ stable, following Corollary 5.24, one can always find $\mathbf{B} = -\mathcal{Q}^{-1}\mathbf{C}^*$ such that $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{I} \end{array}\right)$ is all-pass; in this case, the Hankel singular values are all equal to 1 (i.e., there is no decay). Thus the condition number of the Hankel singular values can be minimized optimally by appropriate choice of $\mathbf{B}$.

Problem (II), i.e., the distribution of the eigenvalues of $\mathbf{A}$ for *preassigned* Hankel singular values, can be addressed using the balanced canonical form. Using results of Golub and Varah [145], the following result can be proved.

Theorem 9.14. *Given n distinct positive real numbers σ_i, together with n signs s_i, the condition number of $\mathbf{A}$ defined by* (9.17) *is minimized, for the following choice of the entries $\gamma_i > 0$, of the diagonal scaling matrix $\mathbf{B}_d$:*

$$\gamma_i^2 = |\lambda_i + \lambda_i| \prod_{j \neq i} \left| \frac{\lambda_i + \lambda_j}{\lambda_i - \lambda_j} \right|, \quad \textit{where } \lambda_k = s_k \sigma_k, \ k = 1, \ldots, n.$$

Remark 9.5.1. (a) The improvement of the condition number of $\mathbf{A}$ in (9.17) for optimal conditioning, compared with random choice of the γ_i, is between 10^3 and 10^4. This improvement is not very pronounced, especially for higher n.

(b) Matrices of the type defined by $\mathbf{A}_0$ are *Cauchy* matrices, while for the special case that all signs s_i are positive we have *Hilbert matrices*. Thus the theorem above provides the optimal diagonal scaling of Cauchy and Hilbert matrices.

9.6 Chapter summary

POD is a popular model reduction method in the PDE community. After a brief overview, it was argued that if one can afford to consider a second set of snapshots coming from the adjoint system, then at least in the linear case a global bound for the approximation error results.

The next two sections discussed model reduction by *modal approximation* and by *residualization*. The former works in some cases but may run into trouble if the system has poles of high multiplicity, while the purpose of approximation by residualization (obtained by means of the Schur complement) is to preserve the steady state gain.

The latter part of this chapter presents an extensive discussion of the decay rates of the eigenvalues of a single gramian, which satisfies a Lyapunov equation, as well as the decay rate of the Hankel singular values (which are the square roots of the eigenvalues of the product of two gramians). Three results are presented. The first is an order of magnitude decay rate of the eigenvalues of a single gramian given by (9.7); the key is the Cholesky ordering of the eigenvalues of **A**, which in turn is closely related to the Cauchy matrix. The second is an upper bound on the decay rate given by (9.8). The main ingredient is a min-max problem which is unsolved; but it turns out that this can be bypassed, leading to a computable upper bound. Finally, it is shown that a new set of invariants given by (9.9) and the Hankel singular values satisfy the multiplicative majorization relation (9.11). This leads to an average decay rate for the Hankel singular values. This approach also provides an explanation for the fact that nonminimum phase systems (i.e., systems having unstable zeros) are more difficult to approximate than minimum phase ones. The chapter concludes with a brief account on the assignment of eigenvalues and Hankel singular values.

Part IV

Krylov-based Approximation Methods

This page intentionally left blank

Chapter 10

Eigenvalue Computations

The 2-norm approach to system approximation outlined above requires dense computations of order n^3 and storage of order n^2, where n is the dimension of the original system. Consequently, it can be used for systems of moderate complexity only. A set of *iterative algorithms* known as *Krylov* methods provides an alternative.

The basic Krylov iteration

Given a real $n \times n$ matrix $\mathbf{A}$ and an n-vector $\mathbf{b}$, let $\mathbf{v}_1 = \frac{\mathbf{b}}{\|\mathbf{b}\|}$. At the kth step we have

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{f}_k\mathbf{e}_k^*, \quad \mathbf{V}_k \in \mathbb{R}^{n\times k}, \quad \mathbf{H}_k \in \mathbb{R}^{k\times k}, \quad \mathbf{f}_k \in \mathbb{R}^n, \quad \mathbf{e}_k \in \mathbb{R}^k, \tag{10.1}$$

where $\mathbf{e}_k$ is the kth canonical unit vector in $\mathbb{R}^k$, $\mathbf{V}_k = [\mathbf{v}_1 \cdots \mathbf{v}_k]$ consists of k column vectors which are orthonormal, $\mathbf{V}_k^*\mathbf{V}_k = \mathbf{I}_k$, and $\mathbf{A}$ projected onto the subspace spanned by the columns of $\mathbf{V}_k$ is $\mathbf{H}_k = \mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$; these conditions imply that $\mathbf{v}_{j+1} = \frac{\mathbf{f}_j}{\|\mathbf{f}_j\|}$, $j = 1, \ldots, n-1$. The computational complexity required for k steps of this iteration is $\mathcal{O}(n^2k)$ and the storage is $\mathcal{O}(nk)$; often, the iteration is applied to a *sparse* matrix $\mathbf{A}$, in which case the complexity reduces further. Two algorithms fall under this umbrella, namely, the Lanczos algorithm [221], [222] and the Arnoldi algorithm [28]. For an overview of the Krylov iteration, see [285], [150], [332], and [72].

This iteration has rich structure involving the subspaces $\mathcal{K}_i$ spanned by the sequence of vectors $\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}$, which are known as *Krylov subspaces* in the numerical linear algebra community and as *reachability* or *controllability subspaces* in the control systems community. For arbitrary $\mathbf{A}$, (10.1) is known as the *Arnoldi iteration*; in this case, $\mathbf{H}_k$ is *upper Hessenberg*. For symmetric $\mathbf{A} = \mathbf{A}^*$, (10.1) is known as the symmetric or one-sided *Lanczos iteration*, in which case $\mathbf{H}_k$ is tridiagonal and symmetric. A variant of (10.1) involving two staring vectors can be applied to nonsymmetric matrices $\mathbf{A}$ and is known as the two-sided or nonsymmetric Lanczos iteration. In this case, the projected matrix $\mathbf{H}_k$ is tridiagonal (but not symmetric).

Three uses of the Krylov iteration

The iteration described above has three main uses.

(a) *Iterative solution of* $\mathbf{Ax} = \mathbf{b}$. In this case, we seek to approximate the solution $\mathbf{x}$ in an iterative fashion. The Krylov methods are based on the fact that successive approximants belong to the subspaces $\mathcal{K}_i$ mentioned above. Both the Arnoldi and the one-sided Lanczos algorithms construct iteratively orthonormal bases for these subspaces.

(b) *Iterative approximation of the eigenvalues of* $\mathbf{A}$. In this case, $\mathbf{b}$ is not a priori fixed. The goal is to use the eigenvalues of the projected matrix $\mathbf{H}_k$ as approximants of the dominant eigenvalues of $\mathbf{A}$. The most simple-minded approach to the approximation of eigenvalues is the *power method* where, given $\mathbf{b}$, successive terms $\mathbf{A}^{k-1}b$ are computed. To overcome the slow convergence of this method, *Krylov methods* are used, where at the kth step one makes use of the information contained in the whole sequence $\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{k-1}\mathbf{b}$. An important advance is the *implicitly restartedArnoldi method* (IRAM) developed by Sorensen. This approach was introduced to overcome the often intractable storage and computational requirements of the original Lanczos/Arnoldi method; this turns out to be a truncated form of the implicitly shifted QR algorithm. This approach has been implemented; the software package is called ARPACK. For details, see the ARPACK manual [227]. For a summary of results on eigenvalue computations, see Sorensen [307] and [306].

It should be noted that in the case of symmetric matrices, there exists a theory for the convergence of the eigenvalues of $\mathbf{H}_k$ to the eigenvalues of $\mathbf{A}$, which has been worked out by Saad [285]. In the general case, it is known that convergence of the eigenvalues on the boundary precedes convergence of the rest of the eigenvalues; a theory, however, is emerging in [46].

(c) *Approximation of linear systems by moment matching*. This problem of interest in the present context will be discussed in Chapter 11.

Krylov methods have their origins in eigenvalue computations and in particular eigenvalue estimations. The present chapter is dedicated to an exposition of Krylov methods as they apply to eigenvalue computations.

10.1 Introduction

Given a matrix or operator $\mathbf{A} : \mathbb{X} \rightarrow \mathbb{X}$, the *eigenvalue problem* consists of finding all complex numbers λ and nonzero vectors $\mathbf{x} \in \mathbb{X}$ such that

$$\mathbf{Ax} = \lambda\mathbf{x}. \qquad (10.2)$$

If $(\lambda, \mathbf{x})$ is a solution of (10.2), λ is called an *eigenvalue* and $\mathbf{x}$ is called the corresponding *eigenvector* of $\mathbf{A}$; furthermore, $(\lambda, \mathbf{x})$ is called an *eigenpair* of $\mathbf{A}$. The set of all eigenvalues of $\mathbf{A}$ is called the spectrum of $\mathbf{A}$ and is denoted by $\Lambda(\mathbf{A})$.

The problem defined by (10.2) is nonlinear in λ and the entries of $\mathbf{x}$. Nevertheless, it can be solved in two steps, as follows. First, notice that for a fixed λ, (10.2) has a solution $\mathbf{x}$ if and only if $\mathbf{A} - \lambda\mathbf{I}$ has a nontrivial null space. If we are working in a finite-dimensional space $\mathbb{X} \cong \mathbb{R}^n$, this is equivalent to the determinant of $\mathbf{A} - \lambda\mathbf{I}$ being zero; let

$$\chi_{\mathbf{A}}(\xi) = \det(\xi\mathbf{I} - \mathbf{A}) = \Pi_{i=1}^{k}(\xi - \lambda_i)^{m_i}, \qquad \lambda_i \neq \lambda_j,\ i \neq j,\ \sum_{i=1}^{k} m_i = n$$

be the characteristic polynomial of $\mathbf{A}$. Then λ is an eigenvalue if and only if λ is a root of the characteristic polynomial $\chi_{\mathbf{A}}(\lambda) = 0$. Consequently, $\mathbf{x}$ is an eigenvector provided that $\mathbf{x} \in \ker(\mathbf{A} - \lambda\mathbf{I})$. If $\lambda_i \in \Lambda(\mathbf{A})$, m_i is its *algebraic* multiplicity and the dimension of $\ker(\mathbf{A} - \lambda_i\mathbf{I})$ is its *geometric* multiplicity. The matrix $\mathbf{A}$ is *diagonalizable* if the algebraic multiplicity of each eigenvalue is equal to its geometric multiplicity. If this condition is not satisfied, the *Jordan canonical form*—with nontrivial Jordan blocks—comes into play. (For details, see, e.g., [181].)

Two square matrices $\mathbf{A}$ and $\mathbf{B}$ are called *similar* if there exists a nonsingular matrix $\mathbf{T}$ such that $\mathbf{AT} = \mathbf{TB}$ or, equivalently, $\mathbf{A} = \mathbf{TBT}^{-1}$. It readily follows that two matrices are similar if and only if they have the same spectrum, i.e., $\Lambda(\mathbf{A}) = \Lambda(\mathbf{B})$, and the same Jordan structure. A subspace $\mathcal{V} \subset \mathbb{X}$ is $\mathbf{A}$-invariant if $\mathbf{A}\mathcal{V} \subset \mathcal{V}$. Let the columns of $\mathbf{V} \in \mathbb{R}^{n\times k}$ form a basis for $\mathcal{V}$; then there exists a matrix $\mathbf{H} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{AV} = \mathbf{VH} \;\Rightarrow\; \Lambda(\mathbf{H}) \subset \Lambda(\mathbf{A}),$$

that is, the eigenvalues of $\mathbf{H}$ are a subset of the eigenvalues of $\mathbf{A}$.

Our goal in this section is to discuss methods for obtaining *approximate* invariant subspaces, say, $\mathcal{V}$, where the residual is orthogonal to some k-dimensional subspace $\mathcal{W} = \text{span col}\,\mathbf{W}$:

$$\mathbf{AV} = \mathbf{VH} + \mathbf{R}, \quad \text{where } \mathbf{R} \perp \mathbf{W} \Leftrightarrow \mathbf{W}^*\mathbf{R} = 0. \tag{10.3}$$

This orthogonality property is also known as a *Petrov–Galerkin condition*. The resulting *approximate eigenvalues* are then obtained by appropriately projecting $\mathbf{A}$:

$$\mathbf{H} = \mathbf{W}^*\mathbf{AV}, \quad \text{where } \mathbf{W}^*\mathbf{V} = \mathbf{I}_k. \tag{10.4}$$

In this case, the projection is called *biorthogonal*, and the approximate eigenvalues are the eigenvalues of $\mathbf{H}$. Often we take $\mathcal{W} = \mathcal{V}$, in which case $\mathbf{H} = \mathbf{V}^*\mathbf{AV}$.

10.1.1 Condition numbers and perturbation results

Definition 10.1. *Given a matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\det\mathbf{A} \neq 0$, *and a matrix norm* $\|\cdot\|$, *the quantity*

$$\kappa(\mathbf{A}) = \lim_{\epsilon\to 0^+} \sup_{\|\mathbf{E}\|\leq\epsilon\|\mathbf{A}\|} \frac{\|(\mathbf{A}+\mathbf{E})^{-1} - \mathbf{A}^{-1}\|}{\epsilon\|\mathbf{A}^{-1}\|}$$

is called the condition number of $\mathbf{A}$.

It is well known that the following relationship holds:

$$\kappa(\mathbf{A}) = \|\mathbf{A}\| \, \|\mathbf{A}^{-1}\|.$$

Furthermore, for induced matrix norms, the reciprocal condition number is the normwise distance to singularity (see, e.g., [314]):

$$\min_{\|\mathbf{E}\|} \{\mathbf{A} + \mathbf{E} \text{ is singular}\} = \frac{\|\mathbf{A}\|}{\kappa(\mathbf{A})}.$$

Notice that the condition number is bounded from below by one: $\kappa(\mathbf{A}) \geq 1$. Orthogonal (unitary) matrices $\mathbf{U}$ are perfectly conditioned, since $\mathbf{U}^{-1} = \mathbf{U}^*$, which implies $\kappa(\mathbf{U}) = 1$, in the 2-induced norm. The following result holds for p-norms.

Proposition 10.2 (Bauer–Fike). *Let* $\mathbf{A}$ *be diagonalizable and let* $\mathbf{E}$ *be a perturbation. For every eigenvalue* $\hat{\lambda}$ *of* $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, *there exists an eigenvalue* λ *of* $\mathbf{A}$ *such that*

$$|\hat{\lambda} - \lambda| \leq \kappa(\mathbf{V})\|\mathbf{E}\|,$$

where $\mathbf{V}$ *is the matrix whose columns are the eigenvectors of* $\mathbf{A}$.

Proof (*see* [144]). Let $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$. By assumption, the matrix $\mathbf{V}^{-1}(\hat{\lambda}\mathbf{I} - \mathbf{A} - \mathbf{E})\mathbf{V}$ is singular; the same holds true for the matrices $\hat{\lambda}\mathbf{I} - \Lambda - \mathbf{V}^{-1}\mathbf{E}\mathbf{V}$ and $\mathbf{I} - (\hat{\lambda}\mathbf{I} - \Lambda)^{-1}\mathbf{V}^{-1}\mathbf{E}\mathbf{V}$, assuming $\hat{\lambda} \notin \Lambda(\mathbf{A})$; singularity of this latter expression implies that $(\hat{\lambda}\mathbf{I} - \Lambda)^{-1}\mathbf{V}^{-1}\mathbf{E}\mathbf{V}$ has an eigenvalue equal to one, so its norm must be at least that large. Thus

$$1 \leq \|(\hat{\lambda}\mathbf{I} - \Lambda)^{-1}\mathbf{V}\mathbf{E}\mathbf{V}^{-1}\| \leq \|(\hat{\lambda}\mathbf{I} - \Lambda)^{-1}\| \|\mathbf{V}\| \|\mathbf{E}\| \|\mathbf{V}^{-1}\| = \frac{1}{\min_\lambda(\hat{\lambda} - \lambda)} \kappa(\mathbf{V})\|\mathbf{E}\|.$$

The result thus follows. □

The distance of the eigenvalues of $\mathbf{A}$ and $\mathbf{A} + \mathbf{E}$ is bounded from above by the norm of the perturbation and by the condition number of the eigenvector matrix of $\mathbf{A}$; if $\mathbf{A}$ has almost parallel eigenvectors, the condition number becomes very big. In the limit, the existence of parallel eigenvectors is equivalent to the nondiagonalizability of $\mathbf{A}$, i.e., to the existence of nontrivial Jordan blocks; this is also equivalent to the condition number of the eigenvector matrix being infinite $\kappa(\mathbf{V}) = \infty$. A consequence of this fact is that the Jordan form is not stably computable. Instead, what is stably computable is the distance of the given matrix to various Jordan forms; for details, see [190].

Next, we consider in more detail what happens to $\mathbf{A}$ under the influence of a perturbation. We discuss here only a simple case. Consider

$$\mathbf{A}_\epsilon = \mathbf{A} + \epsilon\mathbf{E}.$$

It is assumed that $\mathbf{A}$ has simple eigenvalues ($m_i = 1$), $\mathbf{E}$ is fixed, and ϵ is a varying parameter, which is small enough so that $\mathbf{A}_\epsilon$ has simple eigenvalues as well. Let $\mathbf{A}\mathbf{V} = \mathbf{V}\Lambda$; then $\mathbf{V}^{-1}\mathbf{A} = \Lambda\mathbf{V}^{-1}$. Thus if the columns of $\mathbf{V}$ are the *right eigenvectors* of $\mathbf{A}$, the rows of

$\mathbf{W}^* = \mathbf{V}^{-1}$ are its *left eigenvectors*; let $\mathbf{v}_i$, $\mathbf{w}_i$ denote the ith column of $\mathbf{V}$, $\mathbf{W}$, respectively:

$$\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i, \quad \mathbf{w}_i^* \mathbf{A} = \lambda_i \mathbf{w}_i^*, \quad \mathbf{w}_j^* \mathbf{v}_i = \delta_{j,i},$$

where $\delta_{j,i}$ is the Kronecker symbol (equal to one if the subscripts are equal and zero otherwise).

Moreover, let λ_ϵ, $\mathbf{v}_\epsilon$, and $\mathbf{w}_\epsilon$ be an eigenvalue, right, and left eigenvector of $\mathbf{A}_\epsilon$, respectively. Since $\mathbf{A}_0 = \mathbf{A}$ it follows that $\lambda_0 = \lambda_i$, i.e., the perturbed eigenvalue chosen, evaluated, for $\epsilon = 0$, is equal to an eigenvalue of $\mathbf{A}$; the same holds true for the right and left eigenvectors: $\mathbf{v}_0 = \mathbf{v}_i$, $\mathbf{w}_0 = \mathbf{w}_i$; we also assume that $\mathbf{w}_j^* \mathbf{v}_\epsilon = 0$ and $\mathbf{w}_\epsilon^* \mathbf{v}_i = 0$ for all $i \neq j$ and for all permissible values of the parameter ϵ. For details, see the book by Wilkinson [355].

Proposition 10.3. *Under the above assumptions, the Taylor series expansions of* λ_ϵ *and* $\mathbf{v}_\epsilon$ *are*

$$\lambda_\epsilon = \lambda_i + \left[\left. \frac{d\lambda_\epsilon}{d\epsilon} \right|_{\epsilon=0} \right] \epsilon + \left[\left. \frac{d^2\lambda_\epsilon}{d\epsilon^2} \right|_{\epsilon=0} \right] \frac{\epsilon^2}{2} + O(\epsilon^3),$$

$$\mathbf{v}_\epsilon = \mathbf{v}_i + \left[\left. \frac{d\mathbf{v}_\epsilon}{d\epsilon} \right|_{\epsilon=0} \right] \epsilon + \left[\left. \frac{d^2\mathbf{v}_\epsilon}{d\epsilon^2} \right|_{\epsilon=0} \right] \frac{\epsilon^2}{2} + O(\epsilon^3).$$

The first and the second derivatives of the quantities above are

$$\left. \begin{aligned} \left. \frac{d\lambda_\epsilon}{d\epsilon} \right|_{\epsilon=0} &= \frac{\mathbf{w}_i^* \mathbf{E} \mathbf{v}_i}{\mathbf{w}_i^* \mathbf{v}_i}, \quad \left. \frac{d\mathbf{v}_\epsilon}{d\epsilon} \right|_{\epsilon=0} = \sum\nolimits_{k \neq i} \frac{\mathbf{w}_k^* \mathbf{E} \mathbf{v}_i}{\lambda_k - \lambda_i} \mathbf{v}_k, \\ \left. \frac{1}{2} \frac{d^2\lambda_\epsilon}{d\epsilon^2} \right|_{\epsilon=0} &= \frac{1}{\mathbf{w}_i^* \mathbf{v}_i} \sum\nolimits_{k \neq i} \frac{(\mathbf{w}_i^* \mathbf{E} \mathbf{v}_k)\,(\mathbf{w}_k^* \mathbf{E} \mathbf{v}_i)}{\lambda_i - \lambda_k}, \\ \left. \frac{1}{2} \frac{d^2\mathbf{v}_\epsilon}{d\epsilon^2} \right|_{\epsilon=0} &= \sum\nolimits_{k,j \neq i} \frac{(\mathbf{w}_k^* \mathbf{E} \mathbf{v}_j)\,(\mathbf{w}_j^* \mathbf{E} \mathbf{v}_i)}{(\lambda_i - \lambda_k)(\lambda_j - \lambda_i)} \mathbf{v}_j. \end{aligned} \right\} \tag{10.5}$$

Proof. We prove only the formulas for the first derivatives; the corresponding formulas for the second derivatives follow similarly. For details, see [220], [144]. For a sufficiently small $\epsilon > 0$, there holds

$$(\mathbf{A}_\epsilon - \lambda_\epsilon \mathbf{I})\mathbf{v}_\epsilon = \mathbf{0}, \quad \mathbf{w}_\epsilon^* (\mathbf{A}_\epsilon - \lambda_\epsilon \mathbf{I}) = \mathbf{0},$$

where $\mathbf{A}_\epsilon = \mathbf{A} + \epsilon \mathbf{E}$, $\|\mathbf{v}_k(\epsilon)\| = 1$, $\|\mathbf{w}_k(\epsilon)\| = 1$, $\langle \mathbf{w}_\epsilon, \mathbf{v}_\epsilon \rangle = 1$. The derivative of the above expression with respect to ϵ is equal to the following expression; for simplicity, the derivative of $\mathbf{f}$ with respect to ϵ is denoted by $\dot{\mathbf{f}}$:

$$(\dot{\mathbf{A}}_\epsilon - \dot{\lambda}_\epsilon \mathbf{I})\mathbf{v}_\epsilon + (\mathbf{A}_\epsilon - \lambda_\epsilon \mathbf{I})\dot{\mathbf{v}}_\epsilon = \mathbf{0}.$$

The first formula in (10.5) is obtained by taking the inner product of the above expression by $\mathbf{w}_\epsilon$:

$$\mathbf{w}_\epsilon^* (\dot{\mathbf{A}}_\epsilon - \dot{\lambda}_\epsilon \mathbf{I})\mathbf{v}_\epsilon + \mathbf{w}_\epsilon^* (\mathbf{A}_\epsilon - \lambda_\epsilon \mathbf{I})\dot{\mathbf{v}}_\epsilon = \mathbf{0}.$$

The second summand above vanishes and hence the first one yields the expression for $\dot{\lambda}_\epsilon|_{\epsilon=0}$ given in (10.5).

To prove the formula for $\dot{\mathbf{v}}_\epsilon$, we calculate the inner product of the expression above with a left eigenvector $\mathbf{w}_k$ of $\mathbf{A}$ that corresponds to the eigenvalue $\lambda_k \neq \lambda_i$:

$$\mathbf{w}_k^*(\dot{\mathbf{A}}_\epsilon - \dot{\lambda}_\epsilon \mathbf{I})\mathbf{v}_\epsilon + \mathbf{w}_k^*(\mathbf{A}_\epsilon - \lambda_\epsilon \mathbf{I})\dot{\mathbf{v}}_\epsilon = \mathbf{0}.$$

Taking the limit $\epsilon \to 0$, the expression becomes $(\lambda_k - \lambda_i)\mathbf{w}_k^*\dot{\mathbf{v}}_i = \mathbf{w}_k^*\mathbf{E}\mathbf{v}_i, k \neq i$; combining these equalities with $\mathbf{w}_i^*\dot{\mathbf{v}}_\epsilon = 0$ yields the desired second part of (10.5). □

Remark 10.1.1. **(a)** The condition number κ_{λ_k} of the eigenvalue λ_k is now defined as the maximum of $\frac{\|\mathbf{w}_k^*\mathbf{E}\mathbf{v}_k\|}{\mathbf{w}_k^*\mathbf{v}_k}$, over all $\mathbf{E}$ satisfying $\|\mathbf{E}\| = 1$. First notice that $\|\mathbf{w}_k^*\mathbf{E}\mathbf{v}_k\| \leq \|\mathbf{w}_k^*\|\|\mathbf{v}_k\|$; furthermore, this upper bound is attained for the choice of the perturbation matrix $\mathbf{E} = \frac{1}{\|\mathbf{w}_k\|\|\mathbf{v}_k\|}\mathbf{w}_k\mathbf{v}_k^*$. Thus

$$\kappa_{\lambda_k} = \frac{\|\mathbf{w}_k\|\|\mathbf{v}_k\|}{\mathbf{w}_k^*\mathbf{v}_k} = \frac{1}{\cos\angle(\mathbf{w}_k, \mathbf{v}_k)}.$$

The condition number is large if the left and right eigenvectors are close to being orthogonal $\langle \mathbf{w}_i, \mathbf{v}_i \rangle \approx 0$. Notice that if $\mathbf{A}$ is in Jordan form and λ_i corresponds to a nontrivial Jordan block with ones above the diagonal, the right eigenvector is $\mathbf{v}_i = \mathbf{e}_1$, while the left one is $\mathbf{w}_i = \mathbf{e}_n$; therefore $\langle \mathbf{w}_i, \mathbf{v}_i \rangle = 0$.

(b) To study the sensitivity of the eigenvectors, we consider the perturbed matrix $\mathbf{A} + \epsilon\mathbf{E}$, where ϵ is a small positive parameter and $\mathbf{E}$ is arbitrary subject to the norm constraint $\|\mathbf{E}\| = 1$; the second formula (10.5) gives an expression for the rate of change of the right eigenvector of $\mathbf{A}$ as a function of the perturbation $\mathbf{E}$. Thus, the eigenvector $\mathbf{v}_i$ is most sensitive in the direction of eigenvectors corresponding to eigenvalues that are close to λ_i. Therefore, eigenvectors corresponding to *clustered eigenvalues* are difficult to compute.

(c) If $\mathbf{A}$ is symmetric (Hermitian), there is an orthonormal set of eigenvectors, and hence the condition number of every eigenvalue is one. Perturbations of the unitary eigenproblem are studied in [67].

10.2 Pseudospectra*

An important tool for analyzing matrices whose basis of eigenvectors is not well conditioned is the concept of *pseudospectra*. For details on this topic, see the work of Trefethen [326], [327]. See also the paper by Embree and Trefethen [106] and references therein. Below we just give the definition and present a few illustrative examples. These examples have been computed using the *pseudospectra* graphical user interface (GUI), which can be downloaded from

http://web.comlab.ox.ac.uk/projects/pseudospectra/psagui.

Definition 10.4. *The ϵ-pseudospectrum of* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is*

$$\Lambda_\epsilon(\mathbf{A}) = \{\Lambda\,(\mathbf{A} + \mathbf{E}) : \|\mathbf{E}\| \leq \epsilon\} = \left\{z \in \mathbb{C} : \|(z\mathbf{I} - \mathbf{A})^{-1}\| \geq \epsilon^{-1}\right\}.$$

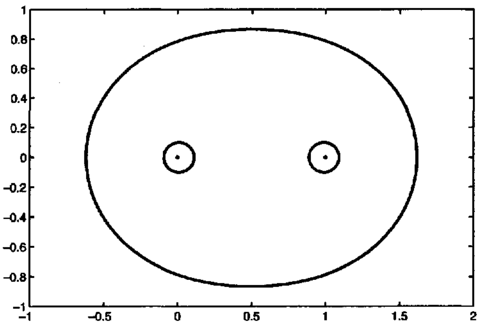


Figure 10.1. 10^{-3} *and* 10^{-2} *pseudospectra of the* 2×2 *matrix of Example* 3.5.

With the above definition, the Bauer–Fike result (Proposition 10.2) can be expressed as

$$\Lambda_\epsilon(\mathbf{A}) \subseteq \Lambda\,(\mathbf{A}) + \Delta_{\kappa(\mathbf{V})},$$

where $\Delta_\kappa(\mathbf{V})$ is the disk of radius $\kappa(\mathbf{V})$. In our first example, Figure 10.1, we revisit the matrix discussed in Example 3.5. Here we take $\epsilon = 100$. Next we consider the following matrices:

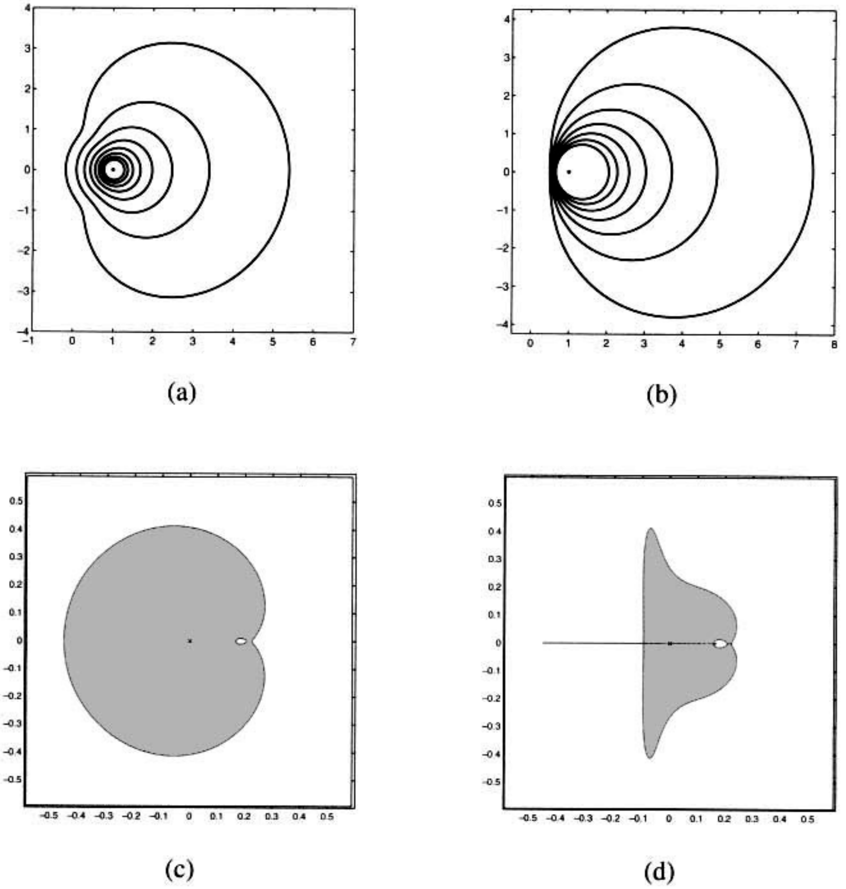
A = compan(poly(ones(10, 1))), A = triu(ones(50, 1));

and A = [01 − 6; 001; 000].

Figure 10.2 shows the ϵ-pseudospectra of these matrices. The two top plots of this figure are due to Embree, while the two bottom plots are due to Karow [196].

Example 10.5. *Pseudospectra.* Consider a Jordan block $\mathbf{J}$ of size $n \times n$ with ones above the diagonal and zeros everywhere else. Consider a perturbed matrix $\mathbf{J}_\epsilon$ which differs from $\mathbf{J}$ in that $\mathbf{J}_\epsilon(n, 1) = \epsilon$; a simple calculation shows that the eigenvalues of $\mathbf{J}_\epsilon$ lie on a circle of radius $\epsilon^{\frac{1}{n}}$. For example, if $n = 10^3$ and $\epsilon = 10^{-18}$, the radius of this circle is 0.9594 (see Figure 10.3). It is interesting to note that while the eigenvalues are highly perturbed with small perturbations of the Jordan block, the singular values are not. The singular values of the original Jordan block are $\sigma_k = 1, k = 1, \ldots, n-1$, and $\sigma_n = 0$. It readily follows that whether just one entry or all entries of $\mathbf{J}$ are perturbed, the singular values are the same as those of the unperturbed $\mathbf{J}$ to within the norm of the perturbation. The resulting shape in Figure 10.3 is the *pseudospectrum* of $\mathbf{J}$ for $n = 12$.

In more detail, Figure 10.3 illustrates the perturbation of a 12×12 Jordan block $\mathbf{J}$ with ones above the diagonal and zeros everywhere else. In the left side of the figure, the (12, 1) element of $\mathbf{J}$ is perturbed from zero to $\epsilon = 10^{-3}, 10^{-2}, 10^{-1}, 1$. The corresponding eigenvalues lie on a circle of radius equal to the 12th root of the perturbation, namely, 0.56, 0.68, 0.82, 1.00, respectively. In the right side of the figure, $\mathbf{J}$ is perturbed by adding a matrix with random entries which are normally distributed, with variance 10^{-6} and 10^{-9}, respectively. Five thousand experiments are performed and the resulting eigenvalues plotted.



(a) (b)

(c) (d)

Figure 10.2. (a) *ϵ-pseudospectra of a* 10×10 *companion matrix for* $\epsilon = 10^{-9}, 10^{-8}, \ldots, 10^{-2}$. (b) *Pseudospectra of a* 50×50 *upper triangular matrix for* $\epsilon = 10^{-14}, 10^{-12}, \ldots, 10^{-2}$. (c) *Complex pseudospectrum of the* 3×3 *matrix above for* $\epsilon = 0.025$. (d) *Real pseudospectrum of the same matrix for the same* ϵ.

The superimposed dots are the eigenvalues obtained as in the left figure with the (12, 1) element being 10^{-6} and 10^{-9}, respectively. If the perturbations of $\mathbf{J}$ are allowed to be complex, disks covering the star-like figures result.

10.3 Iterative methods for eigenvalue estimation

The following are well-known facts.

Proposition 10.6.

Schur triangular form. *Every square matrix is unitarily equivalent to a triangular matrix:* $\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^*$, *where* $\mathbf{Q}$ *is orthogonal (unitary) and* $\mathbf{T}$ *is upper triangular.*

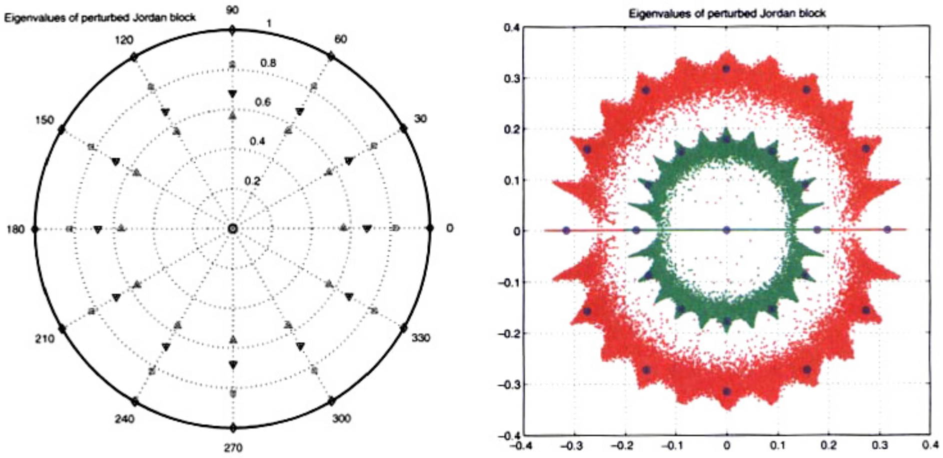


Figure 10.3. *Perturbation of the eigenvalues of a* 12×12 *Jordan block (left) and its pseudospectra for real perturbations with* $\epsilon = 10^{-9}, 10^{-6}$ *(right).*

Spectral theorem for Hermitian matrices. *Given the* $n \times n$ *matrix* $\mathbf{A} = \mathbf{A}^*$, *there exists an orthogonal (unitary) matrix* $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_n]$, *and real numbers* λ_i *such that*

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^* = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^*.$$

The columns/rows of $\mathbf{U}$ *are the right/left eigenvectors of* $\mathbf{A}$ *and form an orthonormal basis for* $\mathbb{R}^n$ $(\mathbb{C}^n)$.

The QR factorization. *Given a matrix* $\mathbf{A} \in \mathbb{R}^{n\times k}$, $k \leq n$, *there exists a matrix* $\mathbf{Q} \in \mathbb{R}^{n\times k}$ *whose columns are orthonormal, i.e.,* $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_k$, *and an upper triangular matrix* $\mathbf{R} \in \mathbb{R}^{k\times k}$ *such that* $\mathbf{A} = \mathbf{QR}$.

Definition 10.7. *The* Rayleigh quotient *of a square symmetric matrix* $\mathbf{A} = \mathbf{A}^* \in \mathbb{R}^{n\times n}$ *at* $\mathbf{x} \in \mathbb{R}^n$ *is*

$$\rho(\mathbf{x}) = \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}, \qquad \mathbf{x} \neq 0.$$

If we order the eigenvalues of $\mathbf{A}$ as $\lambda_1 \geq \cdots \geq \lambda_n$, it readily follows that for all nonzero $\mathbf{x}$,

$$\lambda_1 \geq \rho(\mathbf{x}) \geq \lambda_n.$$

Furthermore, the extremal values are attained by choosing $\mathbf{x}$ to be the (right) eigenvector corresponding to the largest, smallest, eigenvalue, respectively.

10.3.1 The Rayleigh–Ritz procedure

Given $\mathbf{A} = \mathbf{A}^* \in \mathbb{R}^{n\times n}$, let the columns of $\mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_r] \in \mathbb{R}^{n\times r}$, $r \leq n$, be orthonormal, i.e., $\mathbf{V}^*\mathbf{V} = \mathbf{I}_r$. Consider the eigenvalue problem projected onto the subspace spanned by the columns of $\mathbf{V}$; let $\mathbf{x} = \mathbf{V}\mathbf{y}$:

$$\mathbf{A}\mathbf{x} = \mu\mathbf{x} + \mathbf{w} \text{ with } \mathbf{w} \perp \mathbf{V} \Rightarrow \mathbf{A}\mathbf{V}\mathbf{y} = \mu\mathbf{V}\mathbf{y} + \mathbf{w} \;\Rightarrow \mathbf{V}^*\mathbf{A}\mathbf{V}\mathbf{y} = \mu\mathbf{y}.$$

The matrix $\hat{\mathbf{A}} = \mathbf{V}^*\mathbf{A}\mathbf{V}$ is a symmetric matrix; let its eigenvalues be $\mu_1 \geq \cdots \geq \mu_r$. By the Cauchy interlacing theorem (see Proposition 3.26), we have

$$\lambda_i \geq \mu_i \geq \lambda_{n-r+i}, \qquad i = 1, 2, \ldots, r.$$

Let $(\hat{\lambda}, \hat{\mathbf{x}})$ be an eigenpair of $\hat{\mathbf{A}}$. Then $\hat{\lambda}$ is called a *Ritz value* and $\mathbf{V}\hat{\mathbf{x}}$ a *Ritz vector* of $\mathbf{A}$. There follows the next proposition.

Proposition 10.8. *Given $\alpha \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\| = 1$, let $\beta = \|\mathbf{A}\mathbf{v} - \alpha\mathbf{v}\|$. Then $\mathbf{A}$ has an eigenvalue in the closed interval $[\alpha - \beta,\ \alpha + \beta]$. The interval width is minimized if α is the Rayleigh quotient of $\mathbf{v}$.*

Proof. It readily follows that $\beta^2 = \| \sum_i (\lambda_i - \alpha)\mathbf{u}_i\mathbf{u}_i^*\mathbf{v}\|^2$; the latter expression is equal to $\sum_i |\lambda_i - \alpha|^2 |\mathbf{u}_i^*\mathbf{v}|^2$, which in turn is greater than or equal to $\min_i |\lambda_i - \alpha|^2$. □

10.3.2 The simple vector iteration (Power method)

Given is the diagonalizable matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Let $\mathbf{v}^{(0)}$ be any vector with $\|\mathbf{v}^{(0)}\| = 1$. Repeat the following steps:

$$\mathbf{x} = \mathbf{A}\mathbf{v}^{(k)} \text{ and } \mathbf{v}^{(k+1)} = \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

Notice that an approximate eigenvalue can be obtained at each step by means of the Rayleigh quotient. Let the eigenvalue decomposition of $\mathbf{A}$ be $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$, where the eigenvalues are ordered in decreasing magnitude $|\lambda_i| \geq |\lambda_{i+1}|$.

Proposition 10.9. *Assume that $\mathbf{A}$ has a simple largest eigenvalue λ_1 and let $\mathbf{v}_1$ be the corresponding eigenvector. Let $\mathbf{v}^{(0)}$ be any vector that has a nonzero component in the direction of $\mathbf{v}_1$. The simple iteration described above converges toward the dominant eigenvector and the angle between the kth iterate $\mathbf{v}^{(k)}$, and $\mathbf{v}_1$ is of the order $\mathcal{O}\left(\frac{|\lambda_2|^k}{|\lambda_1|^k}\right)$.*

Remark 10.3.1. **(a)** The smaller the ratio $\frac{|\lambda_2|}{|\lambda_1|}$, the faster the convergence.

(b) The result holds even if the matrix is not diagonalizable but $\lambda_1 \neq \lambda_2$.

(c) If the initial vector $\mathbf{v}^{(0)}$ does not have a component in the direction of $\mathbf{v}_1$, convergence is toward $\mathbf{v}_2$, assuming that $|\lambda_2| \neq |\lambda_3|$.

(d) The algorithm does not converge if $|\lambda_1| = |\lambda_2|$, but $\lambda_1 \neq \lambda_2$. A counterexample is given by $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

10.3.3 The inverse vector iteration

Often a good approximation of an eigenvalue is given, $\lambda \approx \lambda_k$. Then, if $\mathbf{A}$ in the simple vector iteration is replaced by $\mathbf{B} = (\mathbf{A} - \lambda\mathbf{I})^{-1}$, the convergence is in general fast, because the largest eigenvalue $\frac{1}{\lambda_k - \lambda}$ of **B** is (much) bigger than the other eigenvalues $\frac{1}{\lambda_j - \lambda}$, $j \neq k$. Again, let $\mathbf{v}^{(0)}$ be any vector with $\|\mathbf{v}^{(0)}\| = 1$. Formally, the following steps are repeated:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{v}^{(k)} \text{ and } \mathbf{v}^{(k+1)} = \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

Remark 10.3.2. **(a)** The solution $\mathbf{x}$ of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{v}^{(k)}$ can be determined by means of the LU factorization or iteratively.

(b) λ is called the *spectral shift*. Since a system of equations has to be solved at each step, this is also called the *shift-and-invert* method.

(c) In general, this inverse iteration is more expensive than the simple iteration discussed earlier. Therefore, in any particular case there is a trade-off between the increased amount of work versus the increased speed of convergence.

10.3.4 Rayleigh quotient iteration

This method is obtained by changing the shift in the inverse iteration method at each step, to become the Rayleigh quotient of the most recent eigenvector estimate.

As before, let $\mathbf{v}^{(0)}$ be any vector with $\|\mathbf{v}^{(0)}\| = 1$. Repeat the following steps:

$$\sigma_k = \rho(\mathbf{v}^{(k)}), \text{ solve } (\mathbf{A} - \sigma_k\mathbf{I})\mathbf{x} = \mathbf{v}^{(k)}, \text{ and set } \mathbf{v}^{(k+1)} = \frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

In this case, convergence is quadratic and in the symmetric case even cubic. Usually, factoring a matrix is required at each step. Thus this method is usually applied only to symmetric tridiagonal matrices [259].

10.3.5 Subspace iteration

Instead of iterating with only one vector, we can iterate with more than one. However, one has to prevent these vectors from converging to the same eigenvector. This is accomplished by orthogonalizing them at each step.

Given is $\mathbf{V}^{(0)} \in \mathbb{R}^{n\times q}$, having full column rank. The following steps are repeated:

$$\text{solve } (\mathbf{A} - \sigma\mathbf{I})\mathbf{X} = \mathbf{V}^{(k)} \text{ and perform a QR factorization } \mathbf{V}^{(k+1)}\mathbf{R}^{(k+1)} = \mathbf{X}. \qquad (10.6)$$

The following result holds.

Proposition 10.10. *Consider the partial (block) Schur decomposition of* **A**,

$$\mathbf{AQ} = \mathbf{Q}\Delta = [\mathbf{Q}_1 \ \mathbf{Q}_2] \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \mathbf{0} & \Delta_{22} \end{pmatrix}, \qquad \Delta \in \mathbb{R}^{k\times k},$$

where $\mathbf{Q} = [\mathbf{q}_1 \cdots \mathbf{q}_k]$ *and the corresponding eigenvalues are*

$$|\lambda_1| \geq \cdots \geq |\lambda_k| > |\lambda_{k+1}| \geq \cdots \geq |\lambda_n|.$$

Let $\mathbf{V}^{(0)} \in \mathbb{R}^{n\times k}$ *be such that* $\mathbf{Q}_1^*\mathbf{V}^{(0)} \in \mathbb{R}^{k\times k}$ *is nonsingular. Then the simple subspace iteration,* (10.6), *converges and the angle between the subspaces spanned by the columns of* $\mathbf{V}^{(k)}$ *and* $\mathbf{Q}_1$ *at the mth step is of the order* $\mathcal{O}\left(\frac{|\lambda_{k+1}|^m}{|\lambda_k|^m}\right)$.

Remark 10.3.3. **(a)** The angle between two subspaces is the maximum over all angles between two vectors, each lying in one subspace. Given two subspaces spanned by the columns of $\mathbf{A}$, $\mathbf{B}$, respectively, let $\mathbf{A} = \mathbf{Q_A R_A}$, $\mathbf{B} = \mathbf{Q_B R_B}$ be the QR-factorizations. The angle between the two subspaces is the largest singular value of $\mathbf{Q_A^* Q_B}$. (The singular values of this product are the principal angles between the two subspaces.)

(b) To speed up convergence, the Rayleigh–Ritz procedure can be incorporated in the subspace iteration algorithm. In this procedure, a few $p \le k$ vectors are sought that converge much faster to eigenvectors of the original matrix.

(c) The *advantages* of subspace iteration are that it is stable, it is simple to implement, it has low storage requirements, and knowledge of good starting vectors can be exploited. The *disadvantage* is it has *slow* convergence.

10.4 Krylov methods

In the iteration methods described above, a sequence of vectors $\mathbf{v}^{(k)}$ is constructed that converges (under mild assumptions) to the eigenvector corresponding to the largest (in modulus) eigenvalue. The convergence rate can be slow. However, notice that at each step only the *last* vector of the sequence $\mathbf{v}^{(k)}$ is used. The approach described in this section attempts to *keep* the information provided by the whole sequence and make good use of it.

As stated earlier, our goal is to determine *approximate* invariant subspaces that satisfy (10.3); the *approximate eigenvalues* are eigenvalues of the projected $\mathbf{A}$ given by (10.4). We first discuss the case where $\mathbf{W} = \mathbf{V}$ and motivate the particular choice of $\mathbf{V}$.

10.4.1 A motivation for Krylov methods

First we attempt to provide a motivation for the use of Krylov methods by concentrating on symmetric matrices $\mathbf{A} = \mathbf{A}^* \in \mathbb{R}^{n \times n}$. Consider the Rayleigh quotient $\rho(\mathbf{x}) = \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}$. The partial derivative of ρ with respect to $\mathbf{x}_i$ is

$$\frac{\partial \rho}{\partial x_i} = \frac{\mathbf{e}_i^*\mathbf{A}\mathbf{x} + \mathbf{x}^*\mathbf{A}\mathbf{e}_i}{\mathbf{x}^*\mathbf{x}} - \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}\,(\mathbf{e}_i^*\mathbf{x} + \mathbf{x}^*\mathbf{e}_i)}{\mathbf{x}^*\mathbf{x}\,\mathbf{x}^*\mathbf{x}} = \frac{2}{\mathbf{x}^*\mathbf{x}}\left[(\mathbf{e}_i^*\mathbf{A}\mathbf{x}) - \frac{\mathbf{x}^*\mathbf{A}\mathbf{x}}{\mathbf{x}^*\mathbf{x}}(\mathbf{e}_i^*\mathbf{x})\right].$$

Thus, the gradient $\nabla\rho = \left[\frac{\partial\rho}{\partial x_1} \cdots \frac{\partial\rho}{\partial x_n}\right]^*$ is

$$\frac{1}{2}\nabla\rho(\mathbf{x}) = \frac{\mathbf{A}\mathbf{x} - \rho(\mathbf{x})\mathbf{x}}{\mathbf{x}^*\mathbf{x}},$$

which shows that the stationary points of the Rayleigh quotient are the eigenvectors of $\mathbf{A}$.

We follow the reasoning of Golub and Van Loan [144]. Consider a matrix $\mathbf{V}$ whose columns are orthonormal, and let $\mathbf{V}_j$ denote the submatrix of $\mathbf{V}$ consisting of its first j columns. It follows from the Cauchy interlacing theorem that the following set of inequalities holds:

$$\lambda_{\max}(\mathbf{V}_1^*\mathbf{A}\mathbf{V}_1) \le \lambda_{\max}(\mathbf{V}_2^*\mathbf{A}\mathbf{V}_2) \le \cdots \le \lambda_{\max}(\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k) \le \lambda_{\max}(\mathbf{A}).$$

The Lanczos algorithm can be derived by imposing the requirement that given $\mathbf{v}_i$, $i = 1, \ldots, j$, the next vector in the sequence, $\mathbf{v}_{j+1}$, be chosen so as to MAXIMIZE $\lambda_{\max}(\mathbf{V}_{j+1}^*\mathbf{A}\mathbf{V}_{j+1})$. The quantities at our disposal for achieving this are $\mathbf{v}_1, \ldots, \mathbf{v}_j$ and the matrix $\mathbf{A}$.

To achieve this goal, let $\mathbf{r}_j \in \mathbb{R}^n$ be such that $\rho(\mathbf{r}_j) = \lambda_{\max}(\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j)$; i.e., the Rayleigh quotient of $\mathbf{r}_j$ is equal to the largest eigenvalue of $\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j$. Let $\mathbf{z} \in \mathbb{R}^k$ be a unit vector such that $\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j\mathbf{z} = \lambda_{\max}(\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j)\mathbf{z}$; it follows that $\mathbf{z}^*\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j\mathbf{z} = \lambda_{\max}(\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j)$. This shows that $\mathbf{r}_j = \mathbf{V}_j\mathbf{z} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_j\}$. The direction of largest increase of the Rayleigh quotient evaluated at $\mathbf{r}_j$ is given by its gradient, namely, $\nabla\rho(\mathbf{r}_j) = \frac{2}{\mathbf{r}_j^*\mathbf{r}_j}(\mathbf{A}\mathbf{r}_j - \rho(\mathbf{r}_j)\mathbf{r}_j)$.

Now the condition that has to be imposed on the choice of the $(j+1)$st column $\mathbf{v}_{j+1}$ is that the span of the columns of the matrix $\mathbf{V}_{j+1}$ *contains* the gradient of the Rayleigh quotient evaluated at $\mathbf{r}_j$:

$$\nabla\rho(\mathbf{r}_j) = \frac{2}{\mathbf{r}_j^*\mathbf{r}_j}(\mathbf{A}\mathbf{r}_j - \rho(\mathbf{r}_j)\mathbf{r}_j) \in \text{span col}\, \mathbf{V}_{j+1} = \text{span col}\, \left[\mathbf{v}_1 \cdots \mathbf{v}_j \;\; \mathbf{v}_{j+1}\right],$$

where $\mathbf{r}_j = \sum_{i=1}^{j} \alpha_i \mathbf{v}_i$. Solving these constraints successively for $j = 1, 2, \ldots, k$, we obtain $\mathbf{v}_k = \mathbf{A}^{k-1}\mathbf{v}_1$. Furthermore, a similar argument shows that this choice of $\mathbf{v}_{j+1}$ minimizes the smallest eigenvalue of $\mathbf{V}_{j+1}^*\mathbf{A}\mathbf{V}_{j+1}$, given the smallest eigenvalue of $\mathbf{V}_j^*\mathbf{A}\mathbf{V}_j$. Therefore, since $\mathbf{V}_k^*\mathbf{V}_k = \mathbf{I}_k$,

$$\text{span col}\, [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k] = \text{span col}\, \left[\mathbf{v}_1, \mathbf{A}\mathbf{v}_1, \ldots, \mathbf{A}^{k-1}\mathbf{v}_1\right] = \text{span col}\, \mathcal{R}_k(\mathbf{A}, \mathbf{v}_1). \tag{10.7}$$

This leads to the problem of computing orthonormal bases for reachability/Krylov subspaces.

See (4.26) for the definition of $\mathcal{R}_k(\mathbf{A}, \mathbf{v}_1)$; it should be mentioned that in the numerical linear algebra community, the resulting spaces are referred to as *Krylov subspaces* and denoted by $\mathcal{K}_k(\mathbf{A}, \mathbf{v}_1)$.

Consequently, conditions (10.3) and (10.7) form the basis for *Krylov* methods.

10.4.2 The Lanczos method

Consider the symmetric $n \times n$ matrix $\mathbf{A}$ together with a sequence of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ that form an orthonormal basis for the reachability (Krylov) subspace $\mathcal{K}_k(\mathbf{A}, \mathbf{v}_1)$. Following the reasoning of the previous section, we need to orthogonalize $\mathbf{A}\mathbf{v}_k$ with respect to the $\mathbf{v}_i$, $i = 1, \ldots, k$. This can be done by applying the Gram–Schmidt procedure to the new direction, namely, $\mathbf{A}\mathbf{v}_k$. The component $\mathbf{r}_k$ of $\mathbf{A}\mathbf{v}_k$ orthogonal to the span of the columns of $\mathcal{R}_k(\mathbf{A}, \mathbf{v}_1)$ is given by

$$\mathbf{r}_k = \mathbf{A}\mathbf{v}_k - \sum_{i=1}^{k}(\mathbf{v}_i^*\mathbf{A}\mathbf{v}_k)\,\mathbf{v}_i = \mathbf{A}\mathbf{v}_k - \mathbf{V}_k\left[\mathbf{V}_k^*\mathbf{A}\mathbf{v}_k\right] = \left[\mathbf{I} - \mathbf{V}_k\mathbf{V}_k^*\right]\mathbf{A}\mathbf{v}_k.$$

Thus the new vector in the sequence is

$$\mathbf{v}_{k+1} = \frac{\mathbf{r}_k}{\|\mathbf{r}_k\|}.$$

Therefore,

$$\mathbf{A}\mathbf{v}_k = \sum_{i=1}^{k+1} \alpha_{i,k}\mathbf{v}_i, \quad \text{where } \alpha_{i,j} = \mathbf{v}_i^* \mathbf{A}\mathbf{v}_j \in \mathbb{R}.$$

Since $\mathbf{v}_j$ depends only on $\mathbf{v}_i$, $i < j$, this procedure has the following property.

Proposition 10.11. *For any choice of* $\mathbf{A} \in \mathbb{R}^{n \times n}$, *the coefficients satisfy* $\alpha_{i,j} = 0$ *for* $i > j + 1$. *Furthermore, since* $\mathbf{A}$ *is symmetric,* $\mathbf{A} = \mathbf{A}^*$, *we have* $\alpha_{i,j} = \alpha_{j,i}$, *and the coefficients are zero for* $j > i + 1$ *as well.*

The above equations can also be written compactly as follows:

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{r}_k\mathbf{e}_k^*, \quad \text{where } \mathbf{H}_k = \mathbf{V}_k^*\mathbf{A}\mathbf{V}_k. \tag{10.8}$$

Several conclusions can be drawn from the above relations. First, the matrix $\mathbf{H}_k$, which is the projection of $\mathbf{A}$ onto $\mathcal{V} = \mathcal{K}(\mathbf{A}, \mathbf{b})$, is *tridiagonal*. For simplicity of notation, let $\alpha_i = \alpha_{i,i}$ and $\beta_{i+1} = \alpha_{i,i+1}$; then

$$\mathbf{H}_k = \begin{bmatrix} \alpha_1 & \beta_2 & & & & \\ \beta_2 & \alpha_2 & \beta_3 & & & \\ & \beta_3 & \alpha_3 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \alpha_{k-1} & \beta_k \\ & & & & \beta_k & \alpha_k \end{bmatrix}. \tag{10.9}$$

This matrix shows that the vectors in the Lanczos procedure satisfy a *three-term recurrence* relationship

$$\mathbf{A}\mathbf{v}_i = \beta_{i+1}\mathbf{v}_{i+1} + \alpha_i\mathbf{v}_i + \beta_i\mathbf{v}_{i-1}, \qquad i = 1, 2, \ldots, k-1.$$

If the remainder $\mathbf{r}_k = 0$, $\mathbf{r}_{k+1}$ cannot be constructed and the Lanczos procedure terminates, in which case if $(\lambda, \mathbf{x})$ is an eigenpair of $\mathbf{H}_k$, $(\lambda, \mathbf{V}_k\mathbf{x})$ is an eigenpair of $\mathbf{A}$ (since $\mathbf{H}_k\mathbf{x} = \lambda\mathbf{x}$ implies $\mathbf{A}\mathbf{V}_k\mathbf{x} = \mathbf{V}_k\mathbf{H}_k\mathbf{x} = \lambda\mathbf{V}_k\mathbf{x}$). However, if $\mathbf{r}_k \neq 0$, we can apply the Rayleigh–Ritz procedure, where $\mathbf{V}_k\mathbf{x}$ is the corresponding Ritz vector. Then the Rayleigh quotient $\rho(\mathbf{V}_k\mathbf{x}) = \frac{\mathbf{x}^*\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k\mathbf{x}}{\mathbf{x}^*\mathbf{x}} = \lambda$ implies that $\mathbf{A}$ has an eigenvalue in the interval $[\lambda + \mu, \lambda - \mu]$, where $\mu = \|\mathbf{A}\mathbf{x} - \lambda\mathbf{x}\|$; moreover, by interlacing, we also have $\lambda_k(\mathbf{A}) \leq \lambda \leq \lambda_1(\mathbf{A})$.

10.4.3 Convergence of the Lanczos method

How close is the estimate of the largest eigenvalue after k steps of the Lanczos procedure? There are several results in this direction. We quote here one of them due to Kaniel and Paige; for details, see [144].

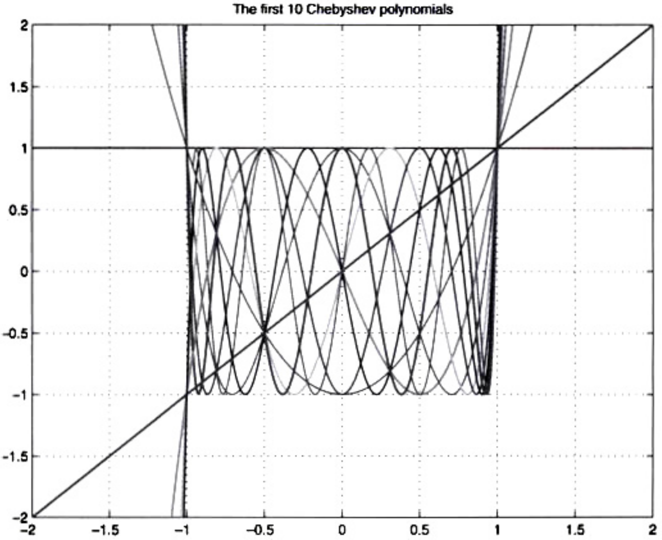


Figure 10.4. *The first* 10 *Chebyshev polynomials.*

Proposition 10.12. *Let* $(\lambda_i, \mathbf{w}_i)$ *be the* i*th eigenpair of the symmetric matrix* $\mathbf{A} \in \mathbb{R}^{n\times n}$*; we assume that* $\lambda_i \geq \lambda_{i+1}$ *and that the eigenvectors are orthonormal. Let* μ_j*,* $j = 1, \ldots, k$*, be the eigenvalues of the tridiagonal* $\mathbf{H}_k$ *obtained after* k *steps of the Lanczos algorithm. Let the starting vector be* $\mathbf{v}_1$*. Then*

$$\lambda_1 \geq \mu_1 \geq \lambda_1 - (\lambda_1 - \lambda_n)\,\nu_1^2, \quad \textit{where} \quad \nu_1 = \frac{\tan\phi_1}{c_{k-1}(1+2\rho_1)},$$

$\cos\phi_1 = |v_1^* w_1|$, $\rho_1 = \frac{\lambda_1-\lambda_2}{\lambda_2-\lambda_n}$, *and* c_i *denotes the Chebyshev polynomial of degree* i*. Furthermore,*

$$\lambda_n \leq \mu_k \leq \lambda_n + (\lambda_1 - \lambda_n)\,\nu_n^2, \quad \textit{where} \quad \nu_n = \frac{\tan\phi_n}{\sqrt{c_{k-1}}(1+2\rho_n)},$$

$\rho_n = \frac{\lambda_{n-1}-\lambda_n}{\lambda_1-\lambda_{n-1}}$, *and* $\cos\phi_n = |v_n^* w_n|$.

Recall that the *Chebyshev polynomials* are generated recursively using the formula $c_j(\xi) = 2\xi c_{j-1}(\xi) - c_{j-2}(\xi)$, $j = 2, 3, \ldots$, where $c_0(\xi) = 1$, $c_1(\xi) = \xi$. Chebyshev polynomials (shown in Figure 10.4) lie within the square bounded by ±1 in both axes and increase rapidly outside the interval $[-1, 1]$. (Since the denominators of ν_1 and ν_n are the values of such a polynomial outside this interval, they are likely to be large.) Thus if the angle between the starting vector and the corresponding eigenvectors of A is not large, ν_1 and ν_n are small.

Figure 10.5 shows the convergence of the Lanczos procedure. Given is a 20×20 symmetric matrix whose eigenvalues lie in the interval $[-15, 15]$. The abscissa shows the values of the various eigenvalues and their approximants, while the rows show the

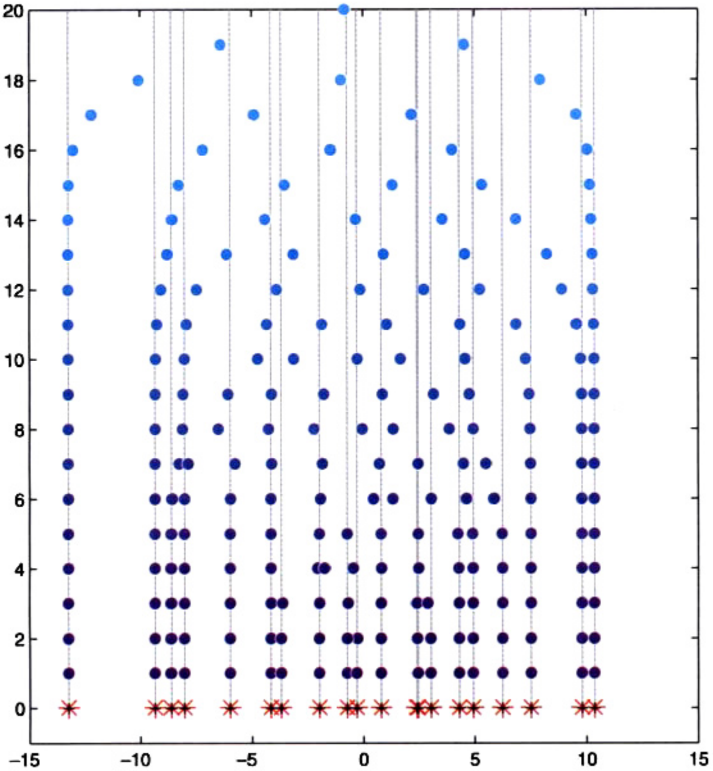


Figure 10.5. *Plot of the eigenvalues of a symmetric* 20 × 20 *matrix using the symmetric Lanczos procedure.*

approximants after a certain number of steps. For instance, the row labeled 20 shows the eigenvalue, obtained after one step of Lanczos, the one labeled 10 shows the eigenvalues obtained at the 11th step, and the one labeled 1, the eigenvalues obtained after 20 steps, which coincide with the exact eigenvalues shown on the row labeled 0. Notice that the eigenvalue obtained at the first step is approximately in the middle of the range of values, while after six steps the two extreme (largest and smallest) eigenvalues are approximated well, and it takes another five steps for the second largest and second smallest eigenvalues to be well approximated. The conclusion is that the spectrum is approximated starting with the extremes. Figure 10.5 was generated by means of a MATLAB GUI due to Sorensen, which can be downloaded from

http://www.caam.rice.edu/~caam551/MatlabCode/matlabcode.html

10.4.4 The Arnoldi method

The procedure described above can also be applied to nonsymmetric matrices **A**. The resulting process is known as the *Arnoldi method*, pictorially depicted in Figure 10.6. The

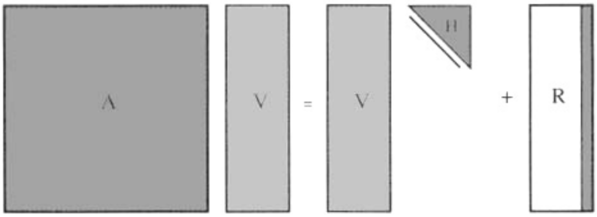


Figure 10.6. *The Arnoldi process: given* $\mathbf{A} \in \mathbb{R}^{n\times n}$*, construct* $\mathbf{V} \in \mathbb{R}^{n\times k}$ *with orthonormal columns such that* $\mathbf{H} \in \mathbb{R}^{k\times k}$ *is an upper Hessenberg matrix and only the last column of the residual* $\mathbf{R} \in \mathbb{R}^{n\times k}$ *is nonzero.*

difference is that in Proposition 10.11, $\alpha_{i,j} \neq \alpha_{j,i}$, and therefore the projected matrix $\mathbf{H}_k$ loses its symmetry and becomes a *Hessenberg matrix*; in this case, we denote the entries by $h_{i,j} = \alpha_{i,j}$:

$$\mathbf{H}_k = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & h_{1,k-1} & h_{1,k} \\ h_{2,1} & h_{2,2} & h_{2,3} & \cdots & h_{2,k-1} & h_{2,k} \\ & h_{3,2} & h_{3,3} & & h_{3,k-1} & h_{3,k} \\ & & & \ddots & \vdots & \vdots \\ & & & & h_{k-1,k-1} & h_{k-1,k} \\ & & & & h_{k,k-1} & h_{k,k} \end{bmatrix}. \qquad (10.10)$$

A key consequence of this lack of tridiagonal structure is that *long recurrences* are now needed to construct $\mathbf{v}_{k+1}$. This is in contrast to the Lanczos procedure, where only three-term recurrences are necessary. The stage is thus set for *restarting* methods for the Arnoldi procedure, which is briefly discussed in section 10.4.11.

10.4.5 Properties of the Arnoldi method

Given is $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $\mathbf{b} \in \mathbb{R}^n$. Let $\mathcal{R}_k(\mathbf{A}, \mathbf{b}) \in \mathbb{R}^{n\times k}$ be the reachability or Krylov matrix defined by (4.26). It is *assumed* that $\mathcal{R}_k$ has full column rank equal to k, which is true, if $(\mathbf{A}, \mathbf{b})$ is reachable.

Problem 10.4.1. *Devise a process which is iterative and at the kth step gives*

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{R}_k, \qquad \mathbf{V}_k,\ \mathbf{R}_k \in \mathbb{R}^{n\times k},\ \mathbf{H}_k \in \mathbb{R}^{k\times k},\ k = 1, 2, \ldots, n$$

These quantities have to satisfy the following conditions at each step:

1. *The columns of* $\mathbf{V}_k$ *are orthonormal:* $\mathbf{V}_k^*\mathbf{V}_k = \mathbf{I}_k,\ k = 1, 2, \ldots, n.$

2. *The residual* $\mathbf{R}_k$ *is orthogonal to the columns of* $\mathbf{V}_k$*, that is, it satisfies the Galerkin condition:* $\mathbf{V}_k^*\mathbf{R}_k = \mathbf{0},\ k = 1, 2, \ldots, n.$

3. span col $\mathbf{V}_k$ = span col $\mathcal{R}_k(\mathbf{A}, \mathbf{b}),\ k = 1, 2, \ldots, n.$

This problem leads to the Arnoldi procedure. The solution has the following structure.

Theorem 10.13.

1. $\mathbf{H}_k$ *is obtained by projecting* $\mathbf{A}$ *onto the span of the columns of* $\mathbf{V}_k$*:* $\mathbf{H}_k = \mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$.

2. *The remainder* $\mathbf{R}_k$ *has rank one and can be written as* $\mathbf{R}_k = \mathbf{r}_k\mathbf{e}_k^*$, *where* $\mathbf{e}_k$ *is the kth unit vector; thus* $\mathbf{r}_k \perp \mathcal{R}_k$.

3. *This further implies that* $\mathbf{v}_{k+1} = \frac{\mathbf{r}_k}{\|\mathbf{r}_k\|}$, *where* $\mathbf{v}_{k+1}$ *is the* $(k+1)$*st column of* $\mathbf{V}$. *Consequently,* $\mathbf{H}_k$ *is an upper Hessenberg matrix.*

4. *Let* $\mathbf{p}_k(\lambda) = \det(\lambda\mathbf{I}_k - \mathbf{H}_k)$ *be the characteristic polynomial of* $\mathbf{H}_k$. *This monic polynomial is the solution of the following minimization problem:*

   $$\mathbf{p}_k = \arg\min \|\mathbf{p}(\mathbf{A})\mathbf{b}\|_2,$$

   where the minimum is taken over all monic polynomials $\mathbf{p}$ *of degree k. Since* $\mathbf{p}_k(\mathbf{A})\mathbf{b} = \mathbf{A}^k\mathbf{b} + \mathcal{R}_k \cdot \underline{\mathbf{p}}$, *where* $\underline{\mathbf{p}}_{i+1}$ *is the coefficient of* λ^i *of the polynomial* $\mathbf{p}_k$, *it also follows that the coefficients of* $\mathbf{p}_k$ *provide the least squares fit between* $\mathbf{A}^k\mathbf{b}$ *and the columns of* $\mathcal{R}_k$.

5. *There holds*

   $$\mathbf{r}_k = \frac{1}{\|\mathbf{p}_{k-1}(\mathbf{A})\mathbf{b}\|}\mathbf{p}_k(\mathbf{A})\mathbf{b}, \quad \mathbf{H}_{k,k-1} = \frac{\|\mathbf{p}_k(\mathbf{A})\mathbf{b}\|}{\|\mathbf{p}_{k-1}(\mathbf{A})\mathbf{b}\|}.$$

Proof. Given the fact that $\mathbf{V}_k^*\mathbf{V}_k = \mathbf{I}_k$, the first item is equivalent to the second item of the problem formulation. Furthermore, the second item is equivalent to the second condition of the above problem. The proof of the first three items is similar to the corresponding properties of the Lanczos procedure (Proposition 10.11). For the remaining items, see [306]. □

The above theorem is based on the following *fundamental* lemma of the Arnoldi procedure. Its proof is left to the reader. (See Problem 42 in Chapter 15.)

Lemma 10.14. *Let* $\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{H} + \mathbf{f}\mathbf{e}_k^*$ *with* $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{H} \in \mathbb{R}^{k\times k}$ *upper Hessenberg,* $\mathbf{V} \in \mathbb{R}^{n\times k}$, $\mathbf{V}^*\mathbf{V} = \mathbf{I}_k$, *and* $\mathbf{V}\mathbf{e}_1 = \mathbf{v}_1$. *There holds*

$$\mathbf{A}^j\mathbf{v}_1 = \mathbf{V}\mathbf{H}^j\mathbf{e}_1 \text{ for } 0 \le j < k,$$

and in addition, for any polynomial ϕ *of degree less than k,*

$$\phi(\mathbf{A})\mathbf{v}_1 = \mathbf{V}\phi(\mathbf{H})\mathbf{e}_1.$$

For $j = k$ *we have*

$$\mathbf{A}^k\mathbf{v}_1 = \mathbf{V}\mathbf{H}^k\mathbf{e}_1 + \nu\mathbf{f}, \quad \nu = \Pi_{i=1}^{k-1}h_{i+1,i} \in \mathbb{R},$$

that is, $\nu = \mathbf{e}_k^*\mathbf{H}^k\mathbf{e}_1$ is the product of the entries of the subdiagonal of **H**. *Furthermore, for any polynomial ϕ of degree k, there holds*

$$\phi(\mathbf{A})\mathbf{v}_1 = \mathbf{V}\phi(\mathbf{H})\mathbf{e}_1 + \nu\alpha_k\mathbf{f},$$

where α_k is the coefficient of the highest power ξ^k, of $\phi(\xi)$.

Remark 10.4.1. FOM, CG, GMRES, MINRES. There are certain iterative methods for solving the set of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, that are closely related to the Arnoldi method. All these methods seek approximate solutions at the kth step, denoted by $\mathbf{x}^{(k)}$, that belong to the Krylov (reachability) subspace spanned by the columns of $\mathcal{R}_k(\mathbf{A}, \mathbf{b}) = [\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}]$. Thus according to the preceding analysis, we must have $\mathbf{x}^{(k)} = \mathbf{V}_k\mathbf{y}$ for some $\mathbf{y} \in \mathbb{R}^k$. (Actually, $\mathbf{x}^{(k)} \in \mathbf{x}^{(0)} + \mathcal{R}_k(\mathbf{A}, \mathbf{b})$; often $\mathbf{x}^{(0)}$ is taken to be zero.)

The *full orthogonalization method* (*FOM*) requires a *Ritz–Galerkin* condition, namely, that the residual be orthogonal to $\mathcal{R}_k(\mathbf{A}, \mathbf{b})$:

$$\mathbf{V}_k^*\left(\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\right) = \mathbf{0}.$$

Thus $\mathbf{y}$ is obtained by solving $\mathbf{H}_{k,k}\mathbf{y} = \|\mathbf{b}\|\mathbf{e}_1$. If $\mathbf{A} = \mathbf{A}^* > 0$ is positive definite, the *conjugate gradient* (*CG*) method results.

The *generalized minimal residual method* (*GMRES*) requires that the residual be minimized,

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|_2 \ \text{ minimal, where } \ \mathbf{x}^{(k)} \in \mathcal{R}_k(\mathbf{A}, \mathbf{b}).$$

Since $\mathbf{A}\mathbf{V}_k = \mathbf{V}_{k+1}\mathbf{H}_{k+1,k}$ and $\mathbf{x}^{(k)} = \mathbf{V}_k\mathbf{y}$, there holds

$$\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|_2 = \|\mathbf{b} - \mathbf{A}\mathbf{V}_k\mathbf{y}\|_2 = \|\mathbf{V}_{k+1}\|\mathbf{b}\|\mathbf{e}_1 - \mathbf{V}_{k+1}\mathbf{H}_{k+1,k}y\|_2 = \|\|\mathbf{b}\|\mathbf{e}_1 - \mathbf{H}_{k+1,k}\mathbf{y}\|_2.$$

Thus $\mathbf{y}$ is the least squares solution of the above problem, namely,

$$\mathbf{y} = \|\mathbf{b}\|\left[\mathbf{H}_{k+1,k}^*\mathbf{H}_{k+1,k}\right]^{-1}\mathbf{H}_{k+1,k}^*\mathbf{e}_1.$$

Finally, if $\mathbf{A} = \mathbf{A}^*$ is symmetric, the method reduces to the *minimal residual method* (*MINRES*).

The residual in both FOM and GMRES can be expressed in terms of a polynomial of degree k: $\psi(s) = \alpha_k s^k + \cdots + \alpha_1 s + \alpha_0$. In the former case, the polynomial should be such that the norm of the residual $\mathbf{r} = \psi(\mathbf{A})\mathbf{b}$ is minimized subject to $\alpha_k = 1$; in the latter case, we seek to minimize the norm of the same residual under the assumption $\alpha_0 = 1$.

10.4.6 An alternative way to look at Arnoldi

Consider a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, a starting vector $\mathbf{b} \in \mathbb{R}^n$, and the corresponding reachability matrix $\mathcal{R}_n = [\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{n-1}\mathbf{b}]$. The following relationship holds true:

$$\mathbf{A}\mathcal{R}_n = \mathcal{R}_n\mathbf{F}, \quad \text{where } \mathbf{F} = \begin{pmatrix} 0 & 0 & \cdots & 0 & -\alpha_0 \\ 1 & 0 & \cdots & 0 & -\alpha_1 \\ 0 & 1 & \cdots & 0 & -\alpha_2 \\ & & \vdots & & \\ 0 & 0 & \cdots & 1 & -\alpha_{n-1} \end{pmatrix},$$

and $\chi_{\mathbf{A}}(s) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$ is the characteristic polynomial of **A**. Compute the QR factorization of $\mathcal{R}_n$:

$$\mathcal{R}_n = \mathbf{VU},$$

where **V** is orthogonal and **U** is upper triangular. It follows that

$$\mathbf{AVU} = \mathbf{VUF} \;\Rightarrow\; \mathbf{AV} = \mathbf{V}\underbrace{\mathbf{UFU}^{-1}}_{\bar{\mathbf{A}}} \;\Rightarrow\; \mathbf{AV} = \mathbf{V}\bar{\mathbf{A}}.$$

Since **U** is upper triangular, so is $\mathbf{U}^{-1}$; furthermore, **F** is upper Hessenberg. Therefore, $\bar{\mathbf{A}}$ being the product of an upper triangular times an upper Hessenberg times an upper triangular matrix, is *upper Hessenberg*. The k-step Arnoldi factorization can now be obtained by considering the first k columns of the above relationship:

$$[\mathbf{AV}]_k = \left[\mathbf{V}\bar{\mathbf{A}}\right]_k \;\Rightarrow\; \mathbf{A}[\mathbf{V}]_k = [\mathbf{V}]_k\bar{\mathbf{A}}_{kk} + \mathbf{f}\mathbf{e}_k^*,$$

where **f** is a multiple of the $(k+1)$st column of **V**. Notice that $\bar{\mathbf{A}}_{kk}$ contains the first k columns and k rows of **A** and is therefore still upper Hessenberg; the columns of $[\mathbf{V}]_k$ provide an orthonormal basis for the space spanned by the first k columns of the reachability matrix $\mathcal{R}_n$.

10.4.7 Properties of the Lanczos method

If **A** is symmetric, we can define an inner product on the space of polynomials as follows:

$$\langle \mathbf{p}(\lambda), \mathbf{q}(\lambda)\rangle = \langle \mathbf{p}(\mathbf{A})\mathbf{b}, \mathbf{q}(\mathbf{A})\mathbf{b}\rangle = \mathbf{b}^*\mathbf{p}(\mathbf{A}^*)\mathbf{q}(\mathbf{A})\mathbf{b} = \mathbf{b}^*\,\mathbf{p}(\mathbf{A})\cdot\mathbf{q}(\mathbf{A})\,\mathbf{b},$$

where the latter is the usual inner product in $\mathbb{R}^n$. In the symmetric case, the Arnoldi process becomes the *Lanczos process*, and the corresponding characteristic polynomials of $\mathbf{H}_k$ are the *Lanczos polynomials*. The Lanczos process applied to symmetric matrices produces $2n-1$ numbers, namely, $\alpha_i = \mathbf{H}_{i,i}$, $i = 1, \ldots, n$, and $\beta_j = \mathbf{H}_{j,j+1} = \mathbf{H}_{j+1,j}$, $j = 1, \ldots, n-1$ (see also (10.9)). The following additional results hold.

Corollary 10.15. *If* $\mathbf{A} = \mathbf{A}^* \in \mathbb{R}^{n\times n}$, *then*

- $\mathbf{H}_k$ *is symmetric and tridiagonal.*

- *the polynomials* $\mathbf{p}_k$, $k = 0, 1, \ldots, n-1$, $\mathbf{p}_0 = 1$, *are orthogonal:* $\langle \mathbf{p}_i, \mathbf{p}_j\rangle = 0$, *for* $i \neq j$, *and* $\beta_k\|\mathbf{p}_{k-1}(\mathbf{A})\mathbf{b}\| = \|\mathbf{p}_k(\mathbf{A})\mathbf{b}\|$. *Therefore, the normalized polynomials are* $\tilde{\mathbf{p}}_j = (\beta_1\beta_2\cdots\beta_j)^{-1}\mathbf{p}_j$.

- *the columns of* $\mathbf{V}_k$ *and the Lanczos polynomials satisfy the following three-term recurrences:*

$$\begin{aligned}
\beta_{k+1}\mathbf{v}_{k+1} &= (\mathbf{A} - \alpha_k\mathbf{I})\mathbf{v}_k &&- \beta_k\mathbf{v}_{k-1},\\
\beta_{k+1}\tilde{\mathbf{p}}_{k+1}(\lambda) &= (\lambda - \alpha_k)\tilde{\mathbf{p}}_k(\lambda) &&- \beta_k\tilde{\mathbf{p}}_{k-1}(\lambda),\\
\mathbf{p}_{k+1}(\lambda) &= (\lambda - \alpha_k)\mathbf{p}_k(\lambda) &&- \beta_k^2\mathbf{p}_{k-1}(\lambda).
\end{aligned}$$

10.4.8 Two-sided Lanczos

If a matrix $\mathbf{A}$ has to be transformed to tridiagonal form but is not symmetric, a modified Lanczos procedure can be applied. This can be seen as an alternative to the Arnoldi method; it provides *short recurrences but no orthogonality* versus *long recurrences with orthogonality.*

Problem 10.4.2. *Given* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and two vectors* $\mathbf{b}, \mathbf{c}^* \in \mathbb{R}^n$, *devise an iterative process that gives at the kth step,*

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{T}_k + \mathbf{R}_k, \;\; \mathbf{A}^*\mathbf{W}_k = \mathbf{W}_k\mathbf{T}_k^* + \mathbf{S}_k.$$

- *Biorthogonality:* $\mathbf{W}_k^*\mathbf{V}_k = \mathbf{I}_k,\;\; k = 1, 2, \ldots, n.$
- span col $\mathbf{V}_k$ = span col $\mathcal{R}_k(\mathbf{A}, \mathbf{b})$, *and* span col $\mathbf{W}_k$ = span col $\mathcal{R}_k(\mathbf{A}^*, \mathbf{c}^*)$, $k = 1, 2, \ldots, n.$
- *Galerkin conditions:* $\mathbf{V}_k^*\mathbf{S}_k = \mathbf{0}$, $\mathbf{W}_k^*\mathbf{R}_k = \mathbf{0}$, $k = 1, 2, \ldots, n.$

This problem is solved by the two-sided Lanczos procedure. Notice that the second condition of the second item above can also be expressed as

$$\text{span rows } \mathbf{W}_k^* = \text{span rows } \mathcal{O}_k(\mathbf{c}, \mathbf{A}),$$

where $\mathcal{O}_k$ is the observability matrix of the pair $(\mathbf{c}, \mathbf{A})$ defined by (4.38). The *assumption* for the solvability of this problem is that $\det(\mathcal{O}_k(\mathbf{c}, \mathbf{A})\mathcal{R}_k(\mathbf{A}, \mathbf{b})) \neq 0$ for all $k = 1, 2, \ldots, n$. The associated *Lanczos polynomials* are defined as before, namely, $\mathbf{p}_k(\lambda) = \det(\lambda\mathbf{I}_k - \mathbf{T}_k)$. In this case, however, the inner product is defined in a different way, namely,

$$\langle \mathbf{p}(\lambda), \mathbf{q}(\lambda)\rangle = \langle \mathbf{p}(\mathbf{A}^*)\mathbf{c}^*, \mathbf{q}(\mathbf{A})\mathbf{b}\rangle = \mathbf{c}\,\mathbf{p}(\mathbf{A}) \cdot \mathbf{q}(\mathbf{A})\,\mathbf{b}.$$

Lemma 10.16.

- $\mathbf{T}_k$ *is obtained by projecting* $\mathbf{A}$ *as follows:* $\mathbf{T}_k = \mathbf{W}_k^*\mathbf{A}\mathbf{V}_k$.
- *The remainders* $\mathbf{R}_k$, $\mathbf{S}_k$ *have rank one and can be written as* $\mathbf{R}_k = \mathbf{r}_k\mathbf{e}_k^*$, $\mathbf{S}_k = \mathbf{q}_k\mathbf{e}_k^*$.
- *This further implies that* $\mathbf{v}_{k+1}$, $\mathbf{w}_{k+1}$ *are scaled versions of* $\mathbf{r}_k$, $\mathbf{q}_k$, *respectively. Consequently,* $\mathbf{T}_k$ *is a tridiagonal matrix, having* $3n-2$ *entries:* α_i, $1 \le i \le n$, β_i, γ_i, $1 \le i \le n-1$.
- *The generalized Lanczos polynomials* $\mathbf{p}_k(\lambda) = \det(\lambda\mathbf{I}_k - \mathbf{T}_k)$, $k = 0, 1, \ldots, n-1$, $\mathbf{p}_0 = 1$, *are orthogonal:* $\langle \mathbf{p}_i, \mathbf{p}_j\rangle = 0$ *for* $i \neq j$.
- *The columns of* $\mathbf{V}_k$, $\mathbf{W}_k$ *and the Lanczos polynomials satisfy the following three-term recurrences:*

$$\begin{array}{lclcl}
\gamma_k\mathbf{v}_{k+1} & = & (\mathbf{A} - \alpha_k\mathbf{I})\mathbf{v}_k & - & \beta_{k-1}\mathbf{v}_{k-1}, \\
\beta_k\mathbf{w}_{k+1} & = & (\mathbf{A}^* - \alpha_k\mathbf{I})\mathbf{w}_k & - & \gamma_{k-1}\mathbf{w}_{k-1}, \\
\gamma_k\mathbf{p}_{k+1}(\lambda) & = & (\lambda - \alpha_k)\mathbf{p}_k(\lambda) & - & \beta_{k-1}\mathbf{p}_{k-1}(\lambda), \\
\beta_k\mathbf{q}_{k+1}(\lambda) & = & (\lambda - \alpha_k)\mathbf{q}_k(\lambda) & - & \gamma_{k-1}\mathbf{q}_{k-1}(\lambda).
\end{array}$$

10.4.9 The Arnoldi and Lanczos algorithms

Conceptually, the Arnoldi procedure, given $\mathbf{A} \in \mathbb{R}^{n\times n}$ and the starting vector $\mathbf{b} \in \mathbb{R}^n$, first sets $\mathbf{V}_1 = \frac{\mathbf{b}}{\|\mathbf{b}\|}$. Then $\mathbf{H}_1 = \mathbf{V}_1^*\mathbf{A}\mathbf{V}_1$. This implies $\mathbf{f}_1 = \mathbf{A}\mathbf{V}_1 - \mathbf{V}_1\mathbf{H}_1 \perp \mathbf{V}_1$. Thus $\mathbf{V}_2 = [\mathbf{V}_1\ \frac{\mathbf{f}_1}{\|\mathbf{f}_1\|}]$; this in turn implies that $\mathbf{H}_2 = \mathbf{V}_2^*\mathbf{A}\mathbf{V}_2$ and $\mathbf{F}_2 = \mathbf{A}\mathbf{V}_2 - \mathbf{V}_2\mathbf{H}_2 = [\mathbf{0}\ \mathbf{f}_2] = \mathbf{f}_2\mathbf{e}_2^*$. Thus the third vector of $\mathbf{V}_3$ is $\frac{\mathbf{f}_2}{\|\mathbf{f}_2\|}$: $\mathbf{V}_3 = [\mathbf{V}_2\ \frac{\mathbf{f}_2}{\|\mathbf{f}_2\|}]$, $\mathbf{H}_3 = \mathbf{V}_3^*\mathbf{A}\mathbf{V}_3$, and so on. Thus at the kth step, the kth column of the projector $\mathbf{V}_k$ is constructed and, simultaneously, the entries of the kth row and column of the projected $\mathbf{A}$-matrix, namely, $\mathbf{H}_k$, follow. Computationally, forming the product $\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$ is expensive. Therefore, the fact that only the last column changes from step to step, plus the $(k+1,k)$th entry of $\mathbf{H}_{k+1}$, is exploited.

The two-sided Lanczos algorithm is executed in a similar way. Now in addition to $\mathbf{A} \in \mathbb{R}^{n\times n}$, we are given two starting vectors, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{c}^* \in \mathbb{R}^n$. First the inner product $\gamma_1 = \mathbf{c}\mathbf{b}$ is computed. Then $\mathbf{V}_1 = \frac{\mathbf{b}}{\sqrt{|\gamma_1|}}$ and $\mathbf{W}_1 = \frac{\pm\mathbf{c}^*}{|\gamma_1|}$ (where the plus/minus sign depends on whether γ_1 is positive or negative); it follows that $\mathbf{T}_1 = \mathbf{W}_1^*\mathbf{A}\mathbf{V}_1$. The second step involves the computation of the remainders: $\mathbf{f}_1 = \mathbf{A}\mathbf{V}_1 - \mathbf{V}_1\mathbf{T}_1$ and $\mathbf{r}_1 = \mathbf{A}^*\mathbf{W}_1 - \mathbf{W}_1\mathbf{T}_1^*$. Subsequently, $\gamma_2 = \mathbf{r}_1^*\mathbf{f}_1$, $\mathbf{V}_2 = [\mathbf{V}_1\ \frac{\mathbf{f}_1}{\sqrt{|\gamma_2|}}]$, $\mathbf{W}_2 = [\mathbf{W}_1\ \frac{\pm\mathbf{r}_1}{\sqrt{|\gamma_2|}}]$, and hence $\mathbf{T}_2 = \mathbf{W}_2^*\mathbf{A}\mathbf{V}_2$. At the third step, we compute the remainders $\mathbf{F}_2 = \mathbf{A}\mathbf{V}_2 - \mathbf{V}_2\mathbf{T}_2$, $\mathbf{R}_2 = \mathbf{A}^*\mathbf{W}_2 - \mathbf{W}_2\mathbf{T}_2^*$; it turns out that they are both rank one: $\mathbf{F}_2 = [\mathbf{0}\ \mathbf{f}_2] = \mathbf{f}_2\mathbf{e}_2^*$, $\mathbf{R}_2 = [\mathbf{0}\ \mathbf{r}_2] = \mathbf{r}_2\mathbf{e}_2^*$, the third column of the two projectors is then obtained by appropriate normalization of $\mathbf{f}_2$ and $\mathbf{r}_2$, and so on. Again, as mentioned earlier, one would not compute, e.g., $\mathbf{A}\mathbf{V}_k$ anew at each iteration.

The Lanczos algorithm: Recursive implementation

Given: the triple $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b}, \mathbf{c}^* \in \mathbb{R}^n$
Find: $\mathbf{V}_k, \mathbf{W}_k \in \mathbb{R}^{n\times k}$, $\mathbf{f}_k, \mathbf{g}_k \in \mathbb{R}^n$, and $\mathbf{T}_k \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{T}_k + \mathbf{f}_k\mathbf{e}_k^*, \quad \mathbf{A}^*\mathbf{W}_k = \mathbf{W}_k\mathbf{T}_k^* + \mathbf{g}_k\mathbf{e}_k^*, \quad \text{where} \tag{10.11}$$

$$\mathbf{T}_k = \mathbf{W}_k^*\mathbf{A}\mathbf{V}_k, \quad \mathbf{V}_k^*\mathbf{W}_k = \mathbf{I}_k, \quad \mathbf{W}_k^*\mathbf{f}_k = \mathbf{0}, \quad \mathbf{V}_k^*\mathbf{g}_k = \mathbf{0}, \tag{10.12}$$

where $\mathbf{e}_k$ denotes the kth unit vector in $\mathbb{R}^n$.

Two-sided Lanczos algorithm

1. $\beta_1 = \sqrt{|\mathbf{b}^*\mathbf{c}^*|}$, $\gamma_1 = \operatorname{sign}(\mathbf{b}^*\mathbf{c}^*)\beta_1$,
   $\mathbf{v}_1 = \mathbf{b}/\beta_1$, $\mathbf{w}_1 = \mathbf{c}^*/\gamma_1$.
2. For $j = 1, \ldots, k$,
   (a) $\alpha_j = \mathbf{w}_j^*\mathbf{A}\mathbf{v}_j$,
   (b) $\mathbf{r}_j = \mathbf{A}\mathbf{v}_j - \alpha_j\mathbf{v}_j - \gamma_j\mathbf{v}_{j-1}$, $\mathbf{q}_j = \mathbf{A}^*\mathbf{w}_j - \alpha_j\mathbf{w}_j - \beta_j\mathbf{w}_{j-1}$,
   (c) $\beta_{j+1} = \sqrt{|\mathbf{r}_j^*\mathbf{q}_j|}$, $\gamma_{j+1} = \operatorname{sign}(\mathbf{r}_j^*\mathbf{q}_j)\beta_{j+1}$,
   (d) $\mathbf{v}_{j+1} = \mathbf{r}_j/\beta_{j+1}$, $\mathbf{w}_{j+1} = \mathbf{q}_j/\gamma_{j+1}$.

The following relationships hold true:

$$\mathbf{V}_k = (\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_k), \quad \mathbf{W}_k = (\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_k);$$

satisfying

$$\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{T}_k + \beta_{k+1}\mathbf{v}_{k+1}\mathbf{e}_k^*, \quad \mathbf{A}^*\mathbf{W}_k = \mathbf{W}_k\mathbf{T}_k^* + \gamma_{k+1}\mathbf{w}_{k+1}\mathbf{e}_k^*,$$

$$\mathbf{T}_k = \begin{pmatrix} \alpha_1 & \gamma_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_k \\ & & \beta_k & \alpha_k \end{pmatrix}, \text{ and } \mathbf{r}_k \in \mathcal{R}_{k+1}(\mathbf{A}, \mathbf{b}), \ \mathbf{q}_k^* \in \mathcal{O}_{k+1}(\mathbf{c}, \mathbf{A}).$$

The Arnoldi algorithm: Recursive implementation

Given: the pair $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$
Find: $\mathbf{V} \in \mathbb{R}^{n\times k}$, $\mathbf{f} \in \mathbb{R}^n$, and $\mathbf{H} \in \mathbb{R}^{k\times k}$ such that

$$\mathbf{A}\mathbf{V} = \mathbf{V}\mathbf{H} + \mathbf{f}\mathbf{e}_k^*, \text{ where} \tag{10.13}$$

$$\mathbf{H} = \mathbf{V}^*\mathbf{A}\mathbf{V}, \ \mathbf{V}^*\mathbf{V} = \mathbf{I}_k, \ \mathbf{V}^*\mathbf{f} = \mathbf{0}, \tag{10.14}$$

where $\mathbf{H}$ is in *upper Hessenberg* form (as before, $\mathbf{e}_k$ denotes the kth unit vector in $\mathbb{R}^n$).

The Arnoldi algorithm

```
1. $\mathbf{v}_1 = \frac{\mathbf{b}}{\|\mathbf{b}\|}$, $\mathbf{w} = \mathbf{A}\mathbf{v}_1$; $\alpha_1 = \mathbf{v}_1^*\mathbf{w}$,
   $\mathbf{f}_1 = \mathbf{w} - \mathbf{v}_1\alpha_1$; $\mathbf{V}_1 = (\mathbf{v}_1)$; $\mathbf{H}_1 = (\alpha_1)$.

2. For $j = 1, 2, \ldots, k-1$,
   $\beta_j = \| \mathbf{f}_j \|$, $\mathbf{v}_{j+1} = \frac{\mathbf{f}_j}{\beta_j}$,
   $\mathbf{V}_{j+1} = \begin{pmatrix}\mathbf{V}_j & \mathbf{v}_{j+1}\end{pmatrix}$, $\hat{\mathbf{H}}_j = \begin{pmatrix}\mathbf{H}_j \\ \beta_j\mathbf{e}_j^*\end{pmatrix}$,
   $\mathbf{w} = \mathbf{A}\mathbf{v}_{j+1}$, $\mathbf{h} = \mathbf{V}_{j+1}^*\mathbf{w}$, $\mathbf{f}_{j+1} = \mathbf{w} - \mathbf{V}_{j+1}\mathbf{h}$,
   $\mathbf{H}_{j+1} = \begin{pmatrix}\hat{\mathbf{H}}_j & \mathbf{h}\end{pmatrix}$.
```

Remark 10.4.2. **(a)** The residual is $\mathbf{f}_j = \mathbf{A}\mathbf{v}_j - \mathbf{V}_j\mathbf{h}_j$, where $\mathbf{h}_j$ is chosen so that the norm $\| \mathbf{f}_j \|$ is minimized. It turns out that $\mathbf{V}_j^*\mathbf{h}_j = \mathbf{0}$, and $\mathbf{h}_j = \mathbf{V}_j^*\mathbf{A}\mathbf{v}_j$, that is, $\mathbf{f}_j = (\mathbf{I} - \mathbf{V}_j\mathbf{V}_j^*)\mathbf{A}\mathbf{v}_j$.

(b) If $\mathbf{A}$ is symmetric, $\mathbf{H}_j$ is tridiagonal, and the Arnoldi algorithm coincides with the Lanczos algorithm.

(c) A high-level code for the Arnoldi and Lanczos procedures is summarized in Figure 10.7.

10.4.10 Rational Krylov methods

To accelerate convergence of the Krylov methods, one can apply a *shift-invert* strategy as shown in section 10.3.3 for the single vector iteration. Therein $\mathbf{A}$ is replaced by $(\mathbf{A} - \lambda\mathbf{I})^{-1}$, where λ is the shift which is close to the eigenvalue of interest. This leads to the family of *rational Krylov* methods. As pointed out by Ruhe [281], a further improvement can be obtained by using *several shifts* $\lambda_1, \ldots, \lambda_k$. The implementation of rational Krylov methods

```
Algorithm: The k-step Arnoldi factorization.
Data: $\mathbf{A} \in \mathbb{R}^{n \times n}$, starting vector $\mathbf{v} \in \mathbb{R}^n$.

$\mathbf{v}_1 = \mathbf{v}/\|\mathbf{v}\|$;
$\mathbf{w} = \mathbf{A}\mathbf{v}_1$; $\alpha_1 = \mathbf{v}_1^*\mathbf{w}$;
$\mathbf{f}_1 \leftarrow \mathbf{w} - \alpha_1\mathbf{v}_1$;
$\mathbf{V}_1 \leftarrow (\mathbf{v}_1)$; $\mathbf{H}_1 \leftarrow (\alpha_1)$;

for $j = 1, 2, \dots, k-1$,

    $\beta_j = \|\mathbf{f}_j\|$; $\mathbf{v}_{j+1} \leftarrow \mathbf{f}_j/\beta_j$;
    $\mathbf{V}_{j+1} \leftarrow (\mathbf{V}_j,\ \mathbf{v}_{j+1})$;
    $\hat{\mathbf{H}}_j \leftarrow \begin{bmatrix} \mathbf{H}_j \\ \beta_j\mathbf{e}_j^* \end{bmatrix}$
    $\mathbf{w} \leftarrow \mathbf{A}\mathbf{v}_{j+1}$;
    $\mathbf{h} \leftarrow \mathbf{V}_{j+1}^*\mathbf{w}$; $\mathbf{f}_{j+1} \leftarrow \mathbf{w} - \mathbf{V}_{j+1}\mathbf{h}$
    $\mathbf{H}_{j+1} \leftarrow (\hat{\mathbf{H}}_j,\ \mathbf{h})$;
end
```

```
Algorithm: The k-step two-sided Lanczos process.
Data: $\mathbf{A} \in \mathbb{R}^{n \times n}$, starting vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

$\beta_1 = \sqrt{|\mathbf{b}^*\mathbf{c}^*|}$; $\gamma_1 = \mathrm{sign}(\mathbf{b}^*\mathbf{c}^*)\beta_1$; $\mathbf{v}_1 = \mathbf{v}/\beta_1$; $\mathbf{w}_1 = \mathbf{w}/\gamma_1$;
$\mathbf{f} = \mathbf{A}\mathbf{v}_1$; $\mathbf{g} = \mathbf{A}^*\mathbf{w}_1$;
$\alpha_1 = \mathbf{w}_1^*\mathbf{f}$;
$\mathbf{f}_1 \leftarrow \mathbf{f} - \alpha_1\mathbf{v}_1$; $\mathbf{g}_1 \leftarrow \mathbf{g} - \alpha_1\mathbf{w}_1$;

for $j = 1, 2, \dots, k-1$,

    $\beta_j = \sqrt{|\mathbf{g}_j^*\mathbf{f}_j|}$; $\gamma_j = \mathrm{sign}\,(\mathbf{g}_j^*\mathbf{f}_j)\beta_j$;
    $\mathbf{v}_{j+1} \leftarrow \mathbf{f}_j/\beta_j$; $\mathbf{w}_{j+1} \leftarrow \mathbf{g}_j/\gamma_j$;

    $\mathbf{f} \leftarrow \mathbf{A}\mathbf{v}_{j+1} - \gamma_j\mathbf{v}_j$; $\mathbf{g} \leftarrow \mathbf{A}^*\mathbf{w}_{j+1} - \beta_j\mathbf{w}_j$;
    $\alpha_{j+1} \leftarrow \mathbf{w}_{j+1}^*\mathbf{f}$;
    $\mathbf{f}_{j+1} \leftarrow \mathbf{f} - \alpha_{j+1}\mathbf{v}_{j+1}$; $\mathbf{g}_{j+1} \leftarrow \mathbf{g} - \alpha_{j+1}\mathbf{w}_{j+1}$
end
```

Figure 10.7. *The Arnoldi and two-sided Lanczos algorithms.*

requires the computation of (direct) LU factorizations of $\mathbf{A} - \lambda_i\mathbf{I}$. These methods will be discussed in more detail in the next chapter.

The rational Arnoldi algorithm

1. Choose shift μ.
   Solve $(\mathbf{A} - \mu\mathbf{I})\mathbf{v}_1 = \mathbf{b}$, normalize $\mathbf{v}_1 \leftarrow \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$;
   Solve $(\mathbf{A} - \mu\mathbf{I})\mathbf{w} = \mathbf{v}_1$; $\alpha_1 = \mathbf{v}_1^*\mathbf{w}$;
   $\mathbf{f}_1 = \mathbf{w} - \mathbf{v}_1\alpha_1$; $\mathbf{V}_1 = (\mathbf{v}_1)$; $\mathbf{H}_1 = (\alpha_1)$.

2. For $j = 1, 2, \dots, k-1$,
   $\beta_j = \| \mathbf{f}_j \|$, $\mathbf{v}_{j+1} = \frac{\mathbf{f}_j}{\beta_j}$.
   $\mathbf{V}_{j+1} = \begin{pmatrix}\mathbf{V}_j & \mathbf{v}_{j+1}\end{pmatrix}$, $\hat{\mathbf{H}}_j = \begin{pmatrix} \mathbf{H}_j \\ \beta_j\mathbf{e}_j^* \end{pmatrix}$,
   Solve $(\mathbf{A} - \mu\mathbf{I})\mathbf{w} = \mathbf{v}_{j+1}$, $\mathbf{h} = \mathbf{V}_{j+1}^*\mathbf{w}$, $\mathbf{f}_{j+1} = \mathbf{w} - \mathbf{V}_{j+1}\mathbf{h}$,
   $\mathbf{H}_{j+1} = \begin{pmatrix}\hat{\mathbf{H}}_j & \mathbf{h}\end{pmatrix}$.

10.4.11 Implicitly restarted Arnoldi and Lanczos methods

The goal of *restarting* the Lanczos and Arnoldi factorizations is to get a better approximation to some desired set of preferred eigenvalues, for example, those eigenvalues that have

- largest modulus,
- largest real part, or
- positive or negative real part.

Since the Arnoldi recurrence gets increasingly expensive as the number of iterations m gets large, one hopes to obtain accurate estimates to the eigenvalues of $\mathbf{A}$ when $m = \dim \mathbf{H}_m \ll \dim \mathbf{A}$. In this case, we hope for the eigenvalues of the projected matrix to capture some salient subset of the spectrum of $\mathbf{A}$, for example, the rightmost eigenvalues of $\mathbf{A}$.

One way to encourage such convergence is to enhance the starting vector $\mathbf{b}$ in the direction of the eigenvalues (or invariant subspace) associated with the desired eigenvalues via a polynomial transformation $\mathbf{b} \leftarrow \mathbf{p}(\mathbf{A})\mathbf{b}$, where $\mathbf{p}$ is a polynomial with roots near unwanted eigenvalues of $\mathbf{A}$. One gets some idea for these roots by performing a basic Arnoldi iteration and computing the spectrum of the projected matrix $\mathbf{H}_m$. An excellent choice for the roots of $\mathbf{p}$ is the set of those eigenvalues of $\mathbf{H}_m$ most unlike the eigenvalues of $\mathbf{A}$ that are sought (if one want the rightmost eigenvalues of $\mathbf{A}$, take as roots of $\mathbf{p}$ the leftmost eigenvalues of $\mathbf{H}_m$). Then one begins the Arnoldi iteration anew with starting vector $\mathbf{p}(\mathbf{A})\mathbf{b}$. This can be accomplished in a numerically stable way via *implicit restarting*.

Suppose, for example, that while $\mathbf{A}$ has eigenvalues in the left half plane, the approximant $\mathbf{H}$ obtained through one of the two methods has one eigenvalue which has positive real part, say, μ, where $\mathbf{V}_m^* \mathbf{A} \mathbf{V}_m \mathbf{x} = \mu \mathbf{x}$, for some $\mathbf{x}$, and

$$\mathbf{A}\mathbf{V}_m = \mathbf{V}_m \mathbf{H}_m +_{\mathbf{m}} \mathbf{e}_{\mathbf{m}}^*.$$

To eliminate this unwanted eigenvalue, the reduced-order matrix obtained at the mth step is projected onto an $(m-1)$st-order system. This is done as follows. First, compute the QR-factorization of $\mathbf{H}_m - \mu \mathbf{I}_m = \mathbf{Q}_m \mathbf{R}_m$. It follows that

$$\mathbf{A}\bar{\mathbf{V}}_m = \bar{\mathbf{V}}_m \bar{\mathbf{H}}_m +_{\mathbf{m}} \mathbf{e}_{\mathbf{m}}^* \mathbf{Q}_{\mathbf{m}}, \quad \text{where } \bar{\mathbf{V}}_{\mathbf{m}} = \mathbf{V}_{\mathbf{m}} \mathbf{Q}_{\mathbf{m}} \text{ and } \bar{\mathbf{H}}_{\mathbf{m}} = \mathbf{Q}_{\mathbf{m}}^* \mathbf{H}_{\mathbf{m}} \mathbf{Q}_{\mathbf{m}}.$$

We now truncate the above relationship to contain $m - 1$ columns; let $\bar{\mathbf{H}}_{m-1}$ denote the principal submatrix of $\bar{\mathbf{H}}_m$, containing the leading $m - 1$ rows and columns.

Theorem 10.17. *Given the above set-up,* $\bar{\mathbf{H}}_{m-1}$ *can be obtained through an* $(m-1)$*-step Arnoldi process with* $\mathbf{A}$ *and starting vector* $\bar{\mathbf{b}} = (\mu \mathbf{I}_n - \mathbf{A})\mathbf{b}$*:*

$$\mathbf{A}\bar{\mathbf{V}}_{m-1} = \bar{\mathbf{V}}_{m-1} \bar{\mathbf{H}}_{m-1} + \bar{\mathbf{f}} \mathbf{e}_{m-1}^*.$$

This process can be repeated to eliminate other unwanted eigenvalues from the reduced-order matrix.

Remark 10.4.3. **(a)** Implicit restarting was introduced by Sorensen [305]. This paper offers two new ideas, namely, implicit restarting (i.e., restarting through the QR decomposition, without explicitly forming the new starting vector) and extra shifts, i.e., using unwanted eigenvalues of $\mathbf{H}_m$ as roots of $\mathbf{p}$. The combination has proved especially effective. This has been implemented in ARPACK [227]. For additional eigenvalue algorithms, see [35]. For a recent study on restarted Lanczos algorithms, see [258]. A high-level code for IRAM is given in Figure 10.8.

(b) Krylov methods and implicit restarting have been worked out for special classes of matrices. For the Hamiltonian case, see [62], [111].

(c) Convergence of IRAM. The convergence of IRAMs has been studied in some detail. See [46], [307].

```
Algorithm: IRAM.

1. Compute an $m$-step Arnoldi factorization:
   $\mathbf{AV}_m = \mathbf{V}_m\mathbf{H}_m + \mathbf{f}_m\mathbf{e}_m^*,\ m = p + k.$

2. Repeat until convergence:

       Compute $\Lambda(\mathbf{H}_m)$ and select $p$ shifts $\mu_1, \ldots, \mu_p$.
       $\mathbf{q}^* \leftarrow \mathbf{e}_m^*$
       for $j = 1, \ldots, p$
           Factor $[\mathbf{Q}, \mathbf{R}] = \mathrm{qr}(\mathbf{H}_m - \mu_j\mathbf{I})$;
           $\mathbf{H}_m \leftarrow \mathbf{Q}^*\mathbf{H}_m\mathbf{Q}$; $\mathbf{V}_m \leftarrow \mathbf{V}_m\mathbf{Q}$;
           $\mathbf{q}^* \leftarrow \mathbf{q}^*\mathbf{Q}$;
       end
       $\mathbf{f}_k \leftarrow \mathbf{v}_{k+1}\hat{\beta}_k + \mathbf{f}_m\sigma_k$;
       $\mathbf{V}_k \leftarrow \mathbf{V}_m(1:n, 1:k)$; $\mathbf{H}_k \leftarrow \mathbf{H}_m(1:k, 1:k)$;
       Beginning with the $k$-step Arnoldi factorization
           $\mathbf{AV}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{f}_k\mathbf{e}_k^*$,
       apply $p$ additional steps to obtain a new $m$-step
       Arnoldi factorization
           $\mathbf{AV}_m = \mathbf{V}_m\mathbf{H}_m + \mathbf{f}_m\mathbf{e}_m^*$.

   end
```

Figure 10.8. *The implicitly restarted Arnoldi algorithm.*

We now present some examples that illustrate the Krylov methods.

Example 10.18. Consider the following symmetric matrix $\mathbf{A}$ and starting vector $\mathbf{b}$:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 2 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The symmetric Lanczos or Arnoldi procedure for steps $k = 1, 2, 3, 4$ yields

$$\mathbf{V}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{H}_1 = [2], \quad \mathbf{R}_1 = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix},$$

$$\mathbf{V}_2 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{\sqrt{6}} \\ 0 & \frac{2}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{6}} \end{bmatrix}, \quad \mathbf{H}_2 = \begin{bmatrix} 2 & \sqrt{6} \\ \sqrt{6} & \frac{8}{3} \end{bmatrix}, \quad \mathbf{R}_2 = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\sqrt{54}} \\ 0 & \frac{-1}{\sqrt{54}} \\ 0 & \frac{1}{\sqrt{54}} \end{bmatrix},$$

$$\mathbf{V}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}, \quad \mathbf{H}_3 = \begin{bmatrix} 2 & \sqrt{6} & 0 \\ \sqrt{6} & \frac{8}{3} & \frac{1}{\sqrt{18}} \\ 0 & \frac{1}{\sqrt{18}} & \frac{4}{3} \end{bmatrix}, \quad \mathbf{R}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{3}}{2} \\ 0 & 0 & 0 \\ 0 & 0 & \frac{-\sqrt{3}}{2} \end{bmatrix},$$

$$\mathbf{V}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{3}} & 0 \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} \end{bmatrix}, \quad \mathbf{H}_4 = \begin{bmatrix} 2 & \sqrt{6} & 0 & 0 \\ \sqrt{6} & \frac{8}{3} & \frac{1}{\sqrt{18}} & 0 \\ 0 & \frac{1}{\sqrt{18}} & \frac{4}{3} & \frac{\sqrt{3}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{3}}{\sqrt{2}} & 0 \end{bmatrix}, \quad \mathbf{R}_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

These matrices satisfy $\mathbf{AV}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{R}_k$, which, due to the orthogonality of the columns of $\mathbf{V}_k$, and $\mathbf{R}_k$, implies $\mathbf{H}_k = \mathbf{V}_k^*\mathbf{AV}_k$, $k = 1, 2, 3, 4$. The associated orthogonal polynomials are the characteristic polynomials of H_k:

$$\begin{aligned} \mathbf{p}_0(x) &= 1, \\ \mathbf{p}_1(x) &= \det(x\mathbf{I}_1 - \mathbf{H}_1) = x - 2, \\ \mathbf{p}_2(x) &= \det(x\mathbf{I}_2 - \mathbf{H}_2) = x^2 - 14x/3 - 2/3, \\ \mathbf{p}_3(x) &= \det(x\mathbf{I}_3 - \mathbf{H}_3) = x^3 - 6x^2 + 11x/2 + 1. \end{aligned}$$

Finally, the characteristic polynomial at the fourth step, $\mathbf{p}_4(x) = \det(x\mathbf{I}_4 - \mathbf{H}_4) = x^4 - 6x^3 + 4x^2 + 8x + 1$, is the same as the characteristic polynomial of $\mathbf{A}$, and therefore in the associated inner product, it corresponds to the zero polynomial. To normalize these polynomials, we compute $\beta_1 = \mathbf{H}_4(1, 2)$, $\beta_2 = \mathbf{H}_4(2, 3)$, $\beta_3 = \mathbf{H}_4(3, 4)$; thus the orthonormal polynomials are

$$\begin{aligned} \tilde{\mathbf{p}}_0(x) &= 1, \\ \tilde{\mathbf{p}}_1(x) &= p_1(x)/\beta_1 = \tfrac{1}{\sqrt{6}}(x - 2), \\ \tilde{\mathbf{p}}_2(x) &= p_2(x)/\beta_1/\beta_2 = \sqrt{3}(3x^2 - 14x - 2), \\ \tilde{\mathbf{p}}_3(x) &= p_3(x)/\beta_1/\beta_2/\beta_3 = \sqrt{2}(2x^3 - 12x^2 + 11x + 2). \end{aligned}$$

The weights that define the Lanczos inner product turn out to be $\gamma_1^2 = 0.4294$, $\gamma_2^2 = 0.0067$, $\gamma_3^2 = 0.5560$, $\gamma_4^2 = 0.0079$ corresponding to the eigenvalues $\lambda_1 = 4.8154$, $\lambda_2 = 2.0607$, $\lambda_3 = -0.1362$, $\lambda_4 = -0.7398$. It is readily checked that

$$\sum_{i=1}^{4} \gamma_i^2 \cdot \tilde{\mathbf{p}}_k(\lambda_i) \cdot \tilde{\mathbf{p}}_\ell(\lambda_i) = \delta_{k,\ell}, \qquad k, \ell = 0, 1, 2, 3,$$

where $\delta_{k,\ell}$ is the Kronecker delta symbol. The eigenvalues of successive approximants $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$, $\mathbf{H}_4$ are

$$2, \begin{pmatrix} -0.1387 \\ 4.8054 \end{pmatrix}, \begin{pmatrix} -0.1550 \\ 1.3405 \\ 4.8145 \end{pmatrix}, \begin{pmatrix} -0.7399 \\ -0.1362 \\ 2.0607 \\ 4.8153 \end{pmatrix}.$$

Example 10.19. Here we consider

$$\mathbf{A} = \frac{1}{4}\begin{bmatrix} 23 & -15 & -3 & 3 \\ 3 & -11 & 1 & -1 \\ 3 & 5 & 1 & -1 \\ 7 & 1 & -19 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

The Arnoldi procedure $\mathbf{AV}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{f}_k\mathbf{e}_k^*$ yields the following matrices for $k = 1, 2, 3, 4$:

$$\mathbf{V}_1 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{H}_1 = [0], \quad \mathbf{f}_1 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix},$$

$$\mathbf{V}_2 = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{H}_2 = \begin{bmatrix} 0 & 2 \\ 2 & 2 \end{bmatrix}, \quad \mathbf{f}_2 = \begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix},$$

$$\mathbf{V}_3 = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{bmatrix}, \quad \mathbf{H}_3 = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 2 & 0 \\ 0 & 4 & -2 \end{bmatrix}, \quad \mathbf{f}_3 = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix},$$

$$\mathbf{V}_4 = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \quad \mathbf{H}_4 = \begin{bmatrix} 0 & 2 & 2 & 5 \\ 2 & 2 & 0 & 0 \\ 0 & 4 & -2 & 2 \\ 0 & 0 & 2 & 4 \end{bmatrix}, \quad \mathbf{f}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The eigenvalues of the successive approximants $\mathbf{H}_1, \mathbf{H}_2, \mathbf{H}_3, \mathbf{H}_4$ are

$$0, \quad \begin{pmatrix} 3.2361 \\ -1.2361 \end{pmatrix}, \quad \begin{pmatrix} 3.7866 \\ -1.8933 + 1.6594i \\ -1.8933 - 1.6594i \end{pmatrix}, \quad \begin{pmatrix} 5.4641 \\ 2 \\ -1.4641 \\ -2 \end{pmatrix}.$$

Example 10.20. In this example we will examine the effects of different starting vectors. To start with, let

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

In this case, $\mathbf{b}$ is actually equal to $\mathbf{e}_1$ and $\mathbf{A}$ can be brought to upper Hessenberg form by a simple permutation of the columns and rows. Thus

$$\mathbf{V}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \Rightarrow \mathbf{H}_4 = \mathbf{V}_4^*\mathbf{AV}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Finally, the eigenvalues of $\mathbf{H}_k$ for $k = 1, 2, 3, 4$ are $\{1\}, \{2, 0\}, \{2.2056, -0.1028 \pm 0.6655i\}$, $\{2.3247, 0, 0.3376 \pm 0.5623i\}$.

If the starting vector is chosen as $\mathbf{b} = [1 \;\; 1 \;\; 0 \;\; 0]^*$, the Arnoldi procedure terminates after three steps, as the starting vector belongs to the eigenspace of $\mathbf{A}$ spanned by the eigenvectors corresponding to the eigenvalues $\{2.3247, 0.3376 \pm 0.5623i\}$. Thus

$$\mathbf{V}_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} & 0 \\ 0 & \frac{2}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{H}_3 = \mathbf{V}_3^*\mathbf{A}\mathbf{V}_3 = \frac{1}{6}\begin{bmatrix} 9 & \sqrt{3} & 6\sqrt{2} \\ 3\sqrt{3} & 9 & 0 \\ 0 & 2\sqrt{6} & 0 \end{bmatrix}.$$

The eigenvalues of $\mathbf{H}_2$ are $\{2, 1\}$. Finally, for the starting vector $\mathbf{b} = [0 \;\; 0 \;\; 1 \;\; 1]^*$ we get

$$\mathbf{V}_4 = \begin{bmatrix} 0 & \frac{2}{\sqrt{5}} & \frac{\sqrt{55}}{55} & \frac{\sqrt{22}}{11} \\ 0 & \frac{1}{\sqrt{5}} & -\frac{2\sqrt{55}}{55} & -\frac{2\sqrt{22}}{11} \\ \frac{1}{\sqrt{2}} & 0 & \frac{\sqrt{55}}{11} & -\frac{\sqrt{22}}{22} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{\sqrt{55}}{11} & \frac{\sqrt{22}}{22} \end{bmatrix}, \quad \mathbf{H}_4 = \mathbf{V}_4^*\mathbf{A}\mathbf{V}_4 = \begin{bmatrix} 1 & \frac{\sqrt{10}}{5} & \frac{\sqrt{110}}{10} & 0 \\ \frac{\sqrt{10}}{2} & \frac{7}{5} & -\frac{9\sqrt{11}}{55} & -\frac{7\sqrt{110}}{110} \\ 0 & \frac{\sqrt{11}}{5} & \frac{18}{55} & \frac{7\sqrt{10}}{55} \\ 0 & 0 & \frac{5\sqrt{10}}{22} & \frac{3}{11} \end{bmatrix}.$$

The eigenvalues of the projected matrices are $\{1\}$; $\{0.1802, 2.2198\}$; $\{2.3580, .1846 \pm .8004i\}$; $\{2.3247, 0, 0.3376 \pm 0.5623i\}$.

10.5 Chapter summary

The *SVD-based* approximation methods presented in the preceding three chapters require dense computations of the order $\mathcal{O}(n^3)$ and storage of the order $\mathcal{O}(n^2)$, where n is the dimension of the system to be approximated. Thus, their compelling properties notwithstanding, they cannot be applied to large problems. What is needed in such cases are *iterative methods* which can be implemented by means of vector-matrix multiplications exclusively. This leads to a different class of approximation methods, the origin of which lies in *eigenvalue estimation*, that is, estimation of a few eigenvalues of a given operator (matrix). The preceding chapter was dedicated to the presentation of these iterative methods, known collectively as *Krylov methods*.

Eigenvalue estimation methods are divided into two parts: the *single vector iteration* methods and the *subspace iteration* methods. The latter category includes *Krylov subspace methods*, which consist of the *Arnoldi* and the *Lanczos* procedures. It is worth noting at this stage that the *Krylov subspaces* are, in the system theoretic language, *reachability* and *observability* subspaces. This connection will be exploited in the next chapter, where we will show that Krylov subspace iterations are intimately connected with model reduction.

This page intentionally left blank

Chapter 11

Model Reduction Using Krylov Methods

In this chapter, we turn our attention to the third use of the Krylov iteration as described at the beginning of the previous chapter (page 314), that is, *approximation by moment matching*. If the to-be-approximated system $\mathbf{\Sigma}$ is described in state space terms (4.13), this problem can be solved iteratively and in a numerically efficient manner by means of the Arnoldi and Lanczos factorizations.

There is an extended literature on this topic. First, the paper by Gragg and Lindquist [146] discusses many connections between system theory and the realization problem on the one hand and numerical analysis, Padé approximation, and Lanczos methods on the other. See also the course notes of Van Dooren [333], the work of Gutknecht [161], and the book of Komzsik [209], as well as [154], [132], [236], [37].

Approximation by moment matching

A linear discrete- or continuous-time system defined by a convolution sum or integral is uniquely determined by its impulse response $\mathbf{h}$. Equivalently, the Laplace transform of the impulse response $\mathbf{H}$, called the *transfer function* of the system, can be used. For the systems considered here, $\mathbf{H}$ is a rational function. In this case, the complexity of the system turns out to be the rank of the Hankel operator $\mathcal{H}$ defined in section 5.4, or the McMillan degree of $\mathbf{H}$ (see page 96). Therefore, one way to approximate a system is to approximate its transfer function $\mathbf{H}$ by a rational function of lower degree. This can be done by matching some terms of the Laurent series expansion of $\mathbf{H}$ at various points of the complex plane. Often, the expansion around infinity is used:

$$\mathbf{H}(s) = \sum_{i\geq 0} \mathbf{h}_i s^{-i}.$$

The coefficients $\mathbf{h}_t$ are known as *Markov parameters* of the system. We now seek a rational function,

$$\hat{\mathbf{H}}(s) = \sum_{i\geq 0} \hat{\mathbf{h}}_i s^{-i},$$

which matches, for instance, the first $\ell < n$ Markov parameters of the original system, that is, $\mathbf{h}_i = \hat{\mathbf{h}}_i$ for $i = 1, 2, \ldots, \ell$. This is the *partial realization problem* discussed in section 4.4.4.

The Lanczos and Arnoldi methods

In the case in which the system is given in terms of state space data, the partial realization problem can be solved in a numerically efficient way. Two approaches fall into this category: the *Lanczos* and the *Arnoldi* methods. Let $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ be given, where $\mathbf{B}$ and $\mathbf{C}^*$ are column vectors. In this case, the starting vectors are not arbitrary (as in the case of eigenvalue approximation) but are determined by $\mathbf{B}$ and $\mathbf{C}$.

The k-step Arnoldi iteration applied to the pair $(\mathbf{A}, \mathbf{B})$ produces a matrix with orthonormal columns, $\mathbf{V}_k$. A reduced-order system $\hat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \end{array}\right)$ can be constructed by applying the *Galerkin projection* $\Pi = \mathbf{V}\mathbf{V}^*$ to $\boldsymbol{\Sigma}$:

$$\hat{\mathbf{A}} = \mathbf{V}_k^*\mathbf{A}\mathbf{V}_k, \;\; \hat{\mathbf{B}} = \mathbf{V}_k^*\mathbf{B}, \;\; \hat{\mathbf{C}} = \mathbf{C}\mathbf{V}_k.$$

The main property of the reduced system $\hat{\boldsymbol{\Sigma}}$ is that it matches k Markov parameters of $\boldsymbol{\Sigma}$, namely,

$$\hat{\mathbf{h}}_j = \hat{\mathbf{C}}\hat{\mathbf{A}}^{j-1}\hat{\mathbf{B}} = \mathbf{C}\mathbf{A}^{j-1}\mathbf{B} = \mathbf{h}_j, \qquad j = 1, \ldots, k.$$

Similarly, the two-sided Lanczos iteration uses $\mathbf{A}$ together with two starting vectors, namely, $\mathbf{B}$ and $\mathbf{C}^*$, to iteratively produce two matrices $\mathbf{V}_k$ and $\mathbf{W}_k$ with k columns each, which are no longer orthonormal but are *biorthogonal*, that is, $\mathbf{W}_k^*\mathbf{V}_k = \mathbf{I}_k$. A reduced-order system $\hat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \end{array}\right)$ is now constructed by applying the Petrov–Galerkin *projection* defined by $\Pi = \mathbf{V}_k\mathbf{W}_k^*$ to $\boldsymbol{\Sigma}$:

$$\hat{\mathbf{A}} = \mathbf{W}_k^*\mathbf{A}\mathbf{V}_k, \;\; \hat{\mathbf{B}} = \mathbf{W}_k^*\mathbf{B}, \;\; \hat{\mathbf{C}} = \mathbf{C}\mathbf{V}_k.$$

This reduced system $\hat{\boldsymbol{\Sigma}}$ has the property that it matches $2k$ moments of $\boldsymbol{\Sigma}$:

$$\hat{\mathbf{h}}_j = \hat{\mathbf{C}}\hat{\mathbf{A}}^{j-1}\hat{\mathbf{B}} = \mathbf{C}\mathbf{A}^{j-1}\mathbf{B} = \mathbf{h}_j, \qquad j = 1, \ldots, 2k.$$

Put in a different way, both the *Lanczos* and the *Arnoldi* methods correspond at the nth step to the computation of a canonical form of the triple $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. In the *generic case*, these two canonical forms are as follows. For the former, $\mathbf{A}$ is in *upper Hessenberg* form, $\mathbf{B}$ is a multiple of the unit vector $\mathbf{e}_1$ and has $\mathbf{C}$ arbitrary entries. The canonical form in the *Lanczos* method corresponds to $\mathbf{A}$ being in *tridiagonal form*, while $\mathbf{B}$ and $\mathbf{C}^*$ are multiples of the first unit vector $\mathbf{e}_1$. In both cases, the reduced-order system is obtained by truncating the state $\mathbf{x} = (x_1 \;\, x_2 \cdots x_n)^*$ to $\hat{\mathbf{x}} = (x_1 \;\, x_2 \;\cdots\; x_k)^*$, where $k < n$. Notice that the resulting $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ retain the same structure as the full-order $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Furthermore, as already mentioned, the so-constructed reduced-order models match k, $2k$, Markov parameters of the original system, respectively. Furthermore, by applying *rational Krylov* methods, reduced-order systems are obtained that match moments at different points in the complex plane.

Finally, it should be stressed that these methods are useful because they can be implemented iteratively, i.e., $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ can be computed without computing the corresponding canonical forms of **A**, **B**, **C** first and *then* truncating. As a consequence, the Lanczos and Arnoldi procedures have reliable numerical implementations, which involve matrix-vector multiplications, exclusively. Their range of applicability is for large n; they are competitive with dense methods for $n \approx 400$ (i.e., $\mathcal{O}(10^2)$) or higher, depending on the problem.

11.1 The moments of a function

Given a matrix-valued function of time $\mathbf{h} : \mathbb{R} \to \mathbb{R}^{p\times m}$, its kth *moment* is defined as

$$\eta_k = \int_0^\infty t^k \mathbf{h}(t)\, dt, \qquad k = 0, 1, 2, \ldots. \tag{11.1}$$

If this function has a *Laplace transform* defined by $\mathbf{H}(s) = \mathcal{L}(\mathbf{h}) = \int_0^\infty \mathbf{h}(t)e^{-st}dt$, the kth moment of $\mathbf{h}$ is the kth derivative of $\mathbf{H}$ evaluated at $s = 0$:

$$\eta_k = (-1)^k \left.\frac{d^k}{ds^k}\mathbf{H}(s)\right|_{s=0} \in \mathbb{R}^{p\times m}, \qquad k \geq 0. \tag{11.2}$$

For our purposes, we also make use of a generalized notion of moments, namely, the *moments of* **h** *around the* (arbitrary) *point* $s_0 \in \mathbb{C}$:

$$\eta_k(s_0) = \int_0^\infty t^k \mathbf{h}(t) e^{-s_0 t}\, dt. \tag{11.3}$$

These generalized moments turn out to be the derivatives of $\mathbf{H}(s)$ evaluated at $s = s_0$:

$$\eta_k(s_0) = (-1)^k \left.\frac{d^k}{ds^k}\mathbf{H}(s)\right|_{s=s_0} \in \mathbb{R}^{p\times m}, \qquad k \geq 0. \tag{11.4}$$

In our context, **h** is the *impulse response* of a linear system $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, i.e., $\mathbf{h}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B} + \delta(t)\mathbf{D}$, $t \geq 0$. Thus, since $\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$, the moments at $s_0 = 0$ are

$$\eta_0(0) = \mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B}, \quad \eta_k(0) = \mathbf{C}\mathbf{A}^{-(k+1)}\mathbf{B}, \qquad k > 0,$$

and those at s_0 are

$$\eta_0(s_0) = \mathbf{D} + \mathbf{C}(s_0\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}, \quad \eta_k(s_0) = \mathbf{C}(s_0\mathbf{I} - \mathbf{A})^{-(k+1)}\mathbf{B}, \qquad k > 0.$$

It should be noted that the moments determine the coefficients of the Laurent series expansion of the transfer function $\mathbf{H}(s)$ in the neighborhood of $s_0 \in \mathbb{C}$; in particular,

$$\begin{aligned}\mathbf{H}(s) &= \mathbf{H}(s_0) + \mathbf{H}^{(1)}(s_0)\frac{(s - s_0)}{1!} + \cdots + \mathbf{H}^{(k)}(s_0)\frac{(s - s_0)^k}{k!} + \cdots \\ &= \eta_0(s_0) + \eta_1(s_0)\frac{(s - s_0)}{1!} + \cdots + \eta_k(s_0)\frac{(s - s_0)^k}{k!} + \cdots.\end{aligned} \tag{11.5}$$

In a similar way, we may expand $\mathbf{H}(s)$ in a Laurent series around infinity:

$$\mathbf{H}(s) = \mathbf{D} + \mathbf{C}\mathbf{B}\,s^{-1} + \mathbf{C}\mathbf{A}\mathbf{B}\,s^{-2} + \cdots + \mathbf{C}\mathbf{A}^k\mathbf{B}\,s^{-(k+1)} + \cdots. \tag{11.6}$$

The quantities

$$\eta_0(\infty) = \mathbf{D}, \quad \eta_k(\infty) = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B}, \qquad k > 0,$$

are the *Markov parameters* $\mathbf{h}_k$ of $\boldsymbol{\Sigma}$ defined in (4.7).

11.2 Approximation by moment matching

Given $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, expand the transfer function around s_0 as in (11.5), where the moments at s_0 are η_j, $j \geq 0$. The approximation problem consists of finding $\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \hat{\mathbf{D}} \end{array}\right)$, where

$$\hat{\mathbf{H}}(s) = \hat{\eta}_0 + \hat{\eta}_1 \frac{(s - s_0)}{1!} + \hat{\eta}_2 \frac{(s - s_0)^2}{2!} + \hat{\eta}_3 \frac{(s - s_0)^3}{3!} + \cdots$$

such that for appropriate ℓ,

$$\boxed{\eta_j = \hat{\eta}_j, \qquad j = 1, 2, \ldots, \ell.}$$

As already mentioned, when the expansion is around infinity, the moments are called *Markov parameters*, and the resulting problem is known as *partial realization*. If the moments are chosen at zero, the corresponding problem is known as *Padé approximation*. In the general case, the problem is known as *rational interpolation*.

Consider the system where $\mathbf{A}$ is stable and the eigenvalue of largest modulus is λ. The Markov parameters of this system grow as $|\lambda|^k$ (while other moments grow exponentially in $\max_\lambda |s_0 - \lambda|^{-1}$). Thus in many cases the computation of the moments is numerically problematic. Given $(\mathbf{C}, \mathbf{A}, \mathbf{B})$, the following are *key properties* of the ensuing algorithms:

- moment matching is achieved *without* computation of moments, and
- the procedure is implemented *iteratively*.

Their consequence is numerical reliability. The algorithms that achieve this are called *Krylov* methods and, in particular, *Lanczos* and *Arnoldi* methods. In the general case, they are known as *rational Krylov* methods.

11.2.1 Two-sided Lanczos and moment matching

Given is the scalar system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{b} \\ \hline \mathbf{c} & \end{array}\right)$, where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b}, \mathbf{c}^* \in \mathbb{R}^n$. The goal is to find $\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{b}} \\ \hline \hat{\mathbf{c}} & \end{array}\right)$, where $\hat{\mathbf{A}} \in \mathbb{R}^{k \times k}$, $\hat{\mathbf{b}}, \hat{\mathbf{c}}^* \in \mathbb{R}^k$, $k < n$, so that the first $2k$ *Markov parameters* $\mathbf{h}_i = \mathbf{c}\mathbf{A}^{i-1}\mathbf{b}$, of Σ, and $\hat{\mathbf{h}}_i = \hat{\mathbf{c}}\hat{\mathbf{A}}^{i-1}\hat{\mathbf{b}}$, of $\hat{\Sigma}$, are matched:

$$\mathbf{h}_i = \hat{\mathbf{h}}_i, \qquad i = 1, \ldots, 2k. \tag{11.7}$$

We proceed by defining the following four quantities associated with $\mathbf{\Sigma}$:
(i) the $k \times n$ *observability* matrix $\mathcal{O}_k$,
(ii) the $n \times k$ *reachability* matrix $\mathcal{R}_k$,

$$\mathcal{O}_k = \begin{bmatrix} \mathbf{c} \\ \mathbf{c}\mathbf{A} \\ \vdots \\ \mathbf{c}\mathbf{A}^{k-1} \end{bmatrix} \in \mathbb{R}^{k \times n}, \quad \mathcal{R}_k = \begin{bmatrix} \mathbf{b} & \mathbf{A}\mathbf{b} & \cdots & \mathbf{A}^{k-1}\mathbf{b} \end{bmatrix} \in \mathbb{R}^{n \times k},$$

(iii) the $k \times k$ Hankel matrix $\mathcal{H}_k$, and
(iv) its shift $\sigma\mathcal{H}_k$,

$$\mathcal{H}_k = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_k \\ \mathbf{h}_2 & \mathbf{h}_3 & \cdots & \mathbf{h}_{k+1} \\ \vdots & & \ddots & \\ \mathbf{h}_k & \mathbf{h}_{k+1} & \cdots & \mathbf{h}_{2k-1} \end{bmatrix}, \ \sigma\mathcal{H}_k = \begin{bmatrix} \mathbf{h}_2 & \mathbf{h}_3 & \cdots & \mathbf{h}_{k+1} \\ \mathbf{h}_3 & \mathbf{h}_4 & \cdots & \mathbf{h}_{k+2} \\ \vdots & & \ddots & \\ \mathbf{h}_{k+1} & \mathbf{h}_{k+2} & \cdots & \mathbf{h}_{2k} \end{bmatrix}.$$

Therefore, $\mathcal{H}_k = \mathcal{O}_k\mathcal{R}_k$ and $\sigma\mathcal{H}_k = \mathcal{O}_k\mathbf{A}\mathcal{R}_k$. Next comes the key assumption of this procedure, namely, that $\det \mathcal{H}_i \neq 0$, $i = 1, \ldots, k$. This allows the computation of the LU factorization of $\mathcal{H}_k$:

$$\mathcal{H}_k = \mathbf{LU}, \ \mathbf{L}(i, j) = 0, \ i < j, \ \text{ and } \mathbf{U}(i, j) = 0, i > j.$$

Furthermore, $\mathbf{L}$ and $\mathbf{U}$ are chosen so that the absolute values of their diagonal entries are equal:

$$\mathbf{L}(i, i) = \pm\mathbf{U}(i, i).$$

Define the maps

$$\Pi_\mathbf{L} = \mathbf{L}^{-1}\mathcal{O}_k \ \text{ and } \ \Pi_\mathbf{U} = \mathcal{R}_k\mathbf{U}^{-1}.$$

It follows trivially that these maps satisfy (a) $\Pi_\mathbf{L}\Pi_\mathbf{U} = \mathbf{I}$, and thus (b) $\Pi = \Pi_\mathbf{U}\Pi_\mathbf{L}$ is an oblique projection. We are now ready to define a reduced-order system:

$$\boxed{\bar{\Sigma} = \left(\begin{array}{c|c} \bar{\mathbf{A}} & \bar{\mathbf{b}} \\ \hline \bar{\mathbf{c}} & \end{array}\right), \quad \text{where} \quad \bar{\mathbf{A}} = \Pi_\mathbf{L}\mathbf{A}\Pi_\mathbf{U}, \ \ \bar{\mathbf{b}} = \Pi_\mathbf{L}\mathbf{b}, \ \ \bar{\mathbf{c}} = \mathbf{c}\Pi_\mathbf{U}.} \qquad (11.8)$$

Theorem 11.1. $\bar{\Sigma}$ *as defined above satisfies the equality* (11.7) *of the Markov parameters. Furthermore,* $\bar{\mathbf{A}}$ *is tridiagonal, and* $\bar{\mathbf{b}}$, $\bar{\mathbf{c}}^*$ *are multiples of the unit vector* $\mathbf{e}_1$.

Proof. First notice that we can write

$$\Pi_\mathbf{U}\Pi_\mathbf{L} = \mathcal{R}_k\mathbf{U}^{-1}\mathbf{L}^{-1}\mathcal{O}_k = \mathcal{R}_k\left(\mathcal{O}_k\mathcal{R}_k\right)^{-1}\mathcal{O}_k.$$

Therefore

$$\mathcal{O}_k\Pi_\mathbf{U}\Pi_\mathbf{L} = \mathcal{O}_k \ \text{ and } \ \Pi_\mathbf{U}\Pi_\mathbf{L}\mathcal{R}_k = \mathcal{R}_k,$$

which shows that $\Pi_\mathbf{U}\Pi_\mathbf{L}$ is a projection along the rows of $\mathcal{O}_k$ onto the span of the columns of $\mathcal{R}_k$. Using the first projection relationship, we obtain

$$\left.\begin{array}{lcll} \bar{\mathbf{c}} & = & & = \mathbf{c}\Pi_\mathbf{U} \\ \bar{\mathbf{c}}\bar{\mathbf{A}} & = & \mathbf{c}\Pi_\mathbf{U}\Pi_\mathbf{L}\mathbf{A}\Pi_\mathbf{U} & = \mathbf{c}\mathbf{A}\Pi_\mathbf{U} \\ & \vdots & & \\ \bar{\mathbf{c}}\bar{\mathbf{A}}^{k-1} & = & \mathbf{c}\Pi_\mathbf{U}\underbrace{\Pi_\mathbf{L}\mathbf{A}\Pi_\mathbf{U}\cdots\Pi_\mathbf{L}\mathbf{A}\Pi_\mathbf{U}}_{k-1 \text{ times}} & = \mathbf{c}\mathbf{A}^{k-1}\Pi_\mathbf{U} \end{array}\right\} \Rightarrow \bar{\mathcal{O}}_k = \mathcal{O}_k\Pi_\mathbf{U}.$$

Similarly, we have $\bar{\mathcal{R}}_k = \Pi_{\mathbf{L}}\mathcal{R}_k$. Combining the last two equalities, we get

$$\bar{\mathcal{O}}_k\bar{\mathcal{R}}_k = \underbrace{[\mathcal{O}_k\Pi_{\mathbf{U}}\Pi_{\mathbf{L}}]}_{\mathcal{O}_k}\mathcal{R}_k = \mathcal{O}_k\mathcal{R}_k = \mathcal{H}_k.$$

Furthermore,

$$\bar{\mathcal{O}}_k\bar{\mathbf{A}}\bar{\mathcal{R}}_k = \underbrace{\mathcal{O}_k\Pi_{\mathbf{U}}}_{\bar{\mathcal{O}}_k}\Pi_{\mathbf{L}}\mathbf{A}\Pi_{\mathbf{U}}\underbrace{\Pi_{\mathbf{L}}\mathcal{R}_k}_{\bar{\mathcal{R}}_k} = \underbrace{\mathcal{O}_k\Pi_{\mathbf{U}}\Pi_{\mathbf{L}}}_{\mathcal{O}_k}\mathbf{A}\underbrace{\Pi_{\mathbf{U}}\Pi_{\mathbf{L}}\mathcal{R}_k}_{\mathcal{R}_k} = \mathcal{O}_k\mathbf{A}\mathcal{R}_k = \sigma\mathcal{H}_k.$$

Finally, it readily follows that

$$\bar{\mathcal{R}}_k = \mathbf{U} \text{ and } \bar{\mathcal{O}}_k = \mathbf{L},$$

where

$$\bar{\mathcal{R}}_k = \left[\begin{array}{cccc} \bar{\mathbf{b}} & \bar{\mathbf{A}}\bar{\mathbf{b}} & \cdots & \bar{\mathbf{A}}^{k-1}\bar{\mathbf{b}} \end{array}\right] \in \mathbb{R}^{n\times k} \text{ and } \bar{\mathcal{O}}_t = \left[\begin{array}{c} \bar{\mathbf{c}} \\ \bar{\mathbf{c}}\bar{\mathbf{A}} \\ \vdots \\ \bar{\mathbf{c}}\bar{\mathbf{A}}^{k-1} \end{array}\right] \in \mathbb{R}^{k\times n}.$$

Since $\mathbf{L}$ is lower triangular, $\bar{\mathcal{R}}_k = \mathbf{U}$ implies $\bar{\mathbf{b}} = \mathbf{U}(1,1)\mathbf{e}_1$ and $\bar{\mathbf{A}}(i,j) = 0$ for $i > j$, while, because $\mathbf{U}$ is upper triangular, $\bar{\mathcal{O}}_k = \mathbf{L}$ implies that $\bar{\mathbf{c}} = \mathbf{L}(1,1)\mathbf{e}_1$ and $\bar{\mathbf{A}}(i,j) = 0$ for $i < j$. Thus in the basis formed by the columns of $\Pi_{\mathbf{U}}$, $\bar{\mathbf{b}}$, $\bar{\mathbf{c}}$ are multiples of the unit vector $\mathbf{e}_1$ and that $\bar{\mathbf{A}}$ is *tridiagonal.*

This completes the proof of the theorem. □

Corollary 11.2. *The maps* Π_L *and* Π_U *defined above are equal to* $\mathbf{W}_k^*$, $\mathbf{V}_k$, *respectively, of the two-sided Lanczos procedure defined in section* 10.4.8.

The Lanczos procedure leads to the following canonical form for SISO linear systems:

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \\ \hline \mathbf{c} & \end{array}\right] = \left[\begin{array}{cccccc|c} \alpha_1 & \gamma_2 & & & & & \beta_1 \\ \beta_2 & \alpha_2 & \gamma_3 & & & & \\ & \beta_3 & \alpha_3 & \ddots & & & \\ & & \ddots & \ddots & \gamma_{n-1} & & \\ & & & \beta_{n-1} & \alpha_{n-1} & \gamma_n & \\ & & & & \beta_n & \alpha_n & \\ \hline \gamma_1 & & & & & & \end{array}\right].$$

11.2.2 Arnoldi and moment matching

The Arnoldi factorization can be used for model reduction as follows. Recall the QR factorization of the reachability matrix $\mathcal{R}_k \in \mathbb{R}^{n\times k}$ (recall also that $\mathcal{H}_k = \mathcal{O}_k\mathcal{R}_k$); an *orthogonal* projection $\Pi = \mathbf{V}\mathbf{V}^*$ can then be attached to this factorization, where from section 10.4.6

$$\mathcal{R}_k = \mathbf{VU} \Rightarrow \mathbf{V} = \mathcal{R}_k\mathbf{U}^{-1},$$

where $\mathbf{V} \in \mathbb{R}^{n\times k}$, $\mathbf{V}^*\mathbf{V} = \mathbf{I}_k$, and $\mathbf{U}$ is upper triangular. We are now ready to define a reduced-order system:

$$\boxed{\hat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{b}} \\ \hline \hat{\mathbf{c}} & \end{array}\right), \quad \text{where} \quad \hat{\mathbf{A}} = \mathbf{V}^*\mathbf{A}\mathbf{V}, \;\; \hat{\mathbf{b}} = \mathbf{V}^*\mathbf{b}, \;\; \hat{\mathbf{c}} = \mathbf{c}\mathbf{V}.} \tag{11.9}$$

Theorem 11.3. *$\hat{\boldsymbol{\Sigma}}$ defined above satisfies the equality of the Markov parameters $\hat{\mathbf{h}}_i = \mathbf{h}_i$, $i = 1, \ldots, k$. Furthermore, $\hat{\mathbf{A}}$ is in Hessenberg form, and $\hat{\mathbf{b}}$ is a multiple of the unit vector* $\mathbf{e}_1$.

Proof. First notice that since $\mathbf{U}$ is upper triangular, $\mathbf{v}_1 = \frac{\mathbf{b}}{\|\mathbf{b}\|}$, and since $\mathbf{V}^*\mathcal{R}_k = \mathbf{U}$, it follows that $\hat{\mathbf{b}} = \mathbf{u}_1 = \|\ \mathbf{b}\ \|\ \mathbf{e}_1$; therefore $\hat{\mathbf{b}} = \mathbf{V}^*\mathbf{b}$, $\mathbf{V}\mathbf{V}^*\mathbf{b} = \mathbf{V}\hat{\mathbf{b}} = \mathbf{b}$, and hence $\hat{\mathbf{A}}\hat{\mathbf{b}} = \mathbf{V}^*\mathbf{A}\mathbf{V}\mathbf{V}^*\mathbf{b} = \mathbf{V}^*\mathbf{A}\mathbf{b}$. In general, since $\mathbf{V}\mathbf{V}^*$ is an orthogonal projection the columns of $\mathcal{R}_k$, we have $\mathbf{V}\mathbf{V}^*\mathcal{R}_k = \mathcal{R}_k$; moreover, $\hat{\mathcal{R}}_k = \mathbf{V}^*\mathcal{R}_k$. Hence

$$(\hat{\mathbf{h}}_1 \;\cdots\; \hat{\mathbf{h}}_k) = \hat{\mathbf{c}}\hat{\mathcal{R}}_k = \mathbf{c}\mathbf{V}\mathbf{V}^*\mathcal{R}_k = \mathbf{c}\mathcal{R}_k = (\mathbf{h}_1 \;\cdots\; \mathbf{h}_k).$$

Finally, the upper triangularity of $\mathbf{U}$ implies that $\mathbf{A}$ is in Hessenberg form. □

Corollary 11.4. *The map* $\mathbf{V}$ *defined above is the same as that of the Arnoldi procedure defined in section* 10.4.5.

The Arnoldi procedure leads to a canonical form for the triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$. From (10.10) follows

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \\ \hline \mathbf{c} & \end{array}\right] = \left[\begin{array}{cccccc|c} h_{11} & h_{12} & h_{13} & \cdots & h_{1,n-1} & h_{1n} & \beta_1 \\ h_{21} & h_{22} & h_{23} & \cdots & h_{2,n-1} & h_{2n} & \\ & h_{32} & h_{33} & & h_{3,n-1} & h_{3,n} & \\ & & & \ddots & \vdots & \vdots & \\ & & & & h_{n-1,n-1} & h_{n-1,n} & \\ & & & & h_{n,n-1} & h_{nn} & \\ \hline \gamma_1 & \gamma_2 & \gamma_3 & \cdots & \gamma_{n-1} & \gamma_n & \end{array}\right].$$

An error expression for the Lanczos model reduction procedure

Making use of formula (5.29), we can derive an exact error expression for approximation by means of the Lanczos procedure. Let $\mathbf{H}$, $\hat{\mathbf{H}}$ be the transfer functions of the original and the reduced systems. Assuming diagonalizability of the $\mathbf{A}$, $\hat{\mathbf{A}}$ matrices, let (c_i, λ_i), $(\hat{c}_i, \hat{\lambda}_i)$ be the residues and poles of the original and reduced systems, respectively. There holds

$$\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^{n} c_i \left[\mathbf{H}(-\lambda_i^*) - \hat{\mathbf{H}}(-\lambda^*)\right] + \sum_{i=1}^{k} \hat{c}_i \left[\mathbf{H}(-\hat{\lambda}_i) - \hat{\mathbf{H}}(-\hat{\lambda}_i^*)\right]. \tag{11.10}$$

For a local error bound, see [32]. This expression points to the following rule of thumb for choosing interpolation points for model reduction:

Interpolate at the mirror image $-\lambda_i^*$ of the poles λ_i of the original system.

Properties of Krylov methods

1. The number of operations needed to compute a reduced system of order k given an order n system using Lanczos or Arnoldi factorizations is $\mathcal{O}(kn^2)$ for dense systems, $\mathcal{O}(k^2n)$ for sparse systems, versus $\mathcal{O}(n^3)$ operations needed for the SVD-based methods. The requirement for memory is $\mathcal{O}(kn)$. A more precise operation count can be found in section 14.1.

2. Only *matrix-vector* multiplications are required—no matrix factorizations or inversions. There is no need to compute the transformed nth-order model and then truncate. This reduces the *ill-conditioning* that arises in SVD methods.

3. Krylov methods can also be applied to MIMO systems, where $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{p\times n}$, m, and/or p are bigger than one, at the expense of increased bookkeeping. For details, see [156] and references therein.

4. Krylov methods have simple derivation and algorithms and, when used for eigenvalue estimation, high convergence rate.

5. Drawbacks:

   - Numerical issue: Arnoldi/Lanczos methods lose orthogonality. This comes from the instability of the classical Gram–Schmidt procedure. The remedy consists of reorthogonalization. Knowledge of good starting vectors cannot be exploited in model reduction since in this case they are fixed.

   - The reduced-order system may not be stable, even if the original system is stable. One remedy is *implicit restart* of Lanczos and Arnoldi, described in section 10.4.11.

   - The Lanczos method breaks down if $\det \mathcal{H}_i = 0$ for *some* $1 \leq i \leq n$. The remedy in this case is provided by *look-ahead* methods. The Arnoldi method breaks down if $\mathcal{R}_i$ does not have full rank; this is a consequence of the lack of reachability of the pair $(\mathbf{A}, \mathbf{b})$ and happens less frequently than the singularity of some principal minor of the Hankel matrix (although $\mathcal{H}_n$ is nonsingular).

   - The resulting reduced-order system tends to approximate the high frequency poles of the original system. Hence the steady-state error may be significant. The remedy is to match the coefficients of the Laurent expansion of the transfer function at frequencies other than infinity. This leads to the *rational Krylov* approach.

   - A drawback of nonsymmetric Lanczos is the need to work with $\mathbf{A}^*$, which can be difficult to use, e.g., on a distributed memory parallel computer.

A remark on the stability/instability of reduced models

With regard to the instability of reduced-order models using Krylov methods, a remark is in order. For this we need to introduce the *numerical range* of the map $\mathbf{A} \in \mathbb{C}^{n\times n}$:

$$\Gamma(\mathbf{A}) = \left\{\mathbf{x}^*\mathbf{A}\mathbf{x} : \|\mathbf{x}\| = 1\right\}.$$

If $\mathbf{A}$ is normal ($\mathbf{AA}^* = \mathbf{A}^*\mathbf{A}$) the numerical range is equal to the convex hull of the spectrum of $\mathbf{A}$. A consequence of this is that for stable $\mathbf{A}$, the projection $\mathbf{V}^*\mathbf{AV}$ can never be unstable. If $\mathbf{A}$ is nonnormal, $\Gamma(\mathbf{A})$ may extend into the right half plane $\mathbb{C}_+$, and, consequently, depending on the starting vector, the projection may turn out to be unstable. This explains the origin of instability in Krylov methods.

It can also be argued that the lack of stability is not necessarily a bad thing, because it may help capture, for example, the initial slope of the impulse or step response of the system; this is based on the fact that the slope of the norm of the matrix exponential $\|e^{\mathbf{A}t}\|$ at the origin is $\max\{\mathbf{x}^*(\mathbf{A}+\mathbf{A}^*)\mathbf{x}\}$, which is one of the extremal points of the numerical range of $\mathbf{A}$.

Lanczos and continued fractions

The Lanczos canonical form of a linear system is related to a continued fraction expansion of its transfer function.

The continued fraction expansion of a rational function, obtained be repeatedly dividing the denominator by the numerator, is always defined. However, the Lanczos canonical form and the associated tridiagonal form of $\mathbf{A}$ does not always exist. It can be shown that Lanczos breaks down precisely when one or more remainders in the above division process have degree greater than one. These and more general issues are the subject of the section on *recursive interpolation* in Chapter 4.

Example 11.5. A connection between the Lanczos procedure and system theory described in [146] is that the coefficients of the Lanczos canonical form can be read off from the *continued fraction* expansion of the transfer function. For developments related to this continued fraction expansion and generalizations, see the section on recursive interpolation mentioned above and [9], [15], and [12]. For a discussion of the connection between tridiagonalization of $\mathbf{A}$ and minimal realizations, see [260].

Consider the system described by the following transfer function in continued fraction form:

$$\mathbf{H}(s) = \cfrac{1}{s+2+\cfrac{4}{s-1-\cfrac{4}{s+2+\frac{12}{s}}}} = \frac{s^3+s^2+6s-12}{s^4+3s^3+12s^2+8s+24}.$$

The three approximants obtained by truncating this continued fraction expansion are

$$\mathbf{H}_1(s) = \frac{1}{s+2},$$

$$\mathbf{H}_2(s) = \cfrac{1}{s+2+\cfrac{4}{s-1}} = \frac{s-1}{s^2+s+2},$$

$$\mathbf{H}_3(s) = \cfrac{1}{s+2+\cfrac{4}{s-1-\cfrac{4}{s+2}}} = \frac{s^2+s-6}{s^3+3s^2-4} = \frac{(s+3)(s-2)}{(s-1)(s+2)^2}.$$

A realization of $\mathbf{H}(s) = \mathbf{c}(s\mathbf{I}_4 - \mathbf{A})^{-1}\mathbf{b}$, with $\mathbf{A}$ in companion form is given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -24 & -8 & -12 & -3 \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{c} = \begin{bmatrix} -12 & 6 & 1 & 1 \end{bmatrix}.$$

The continued fraction expansion of $\mathbf{H}$ induces a different realization, namely,

$$\bar{\mathbf{A}} = \begin{bmatrix} -2 & 2 & 0 & 0 \\ -2 & 2 & 2 & 0 \\ 0 & 2 & -2 & \sqrt{12} \\ 0 & 0 & -\sqrt{12} & 0 \end{bmatrix}, \ \bar{\mathbf{b}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \bar{\mathbf{c}} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}.$$

The latter realization $\left(\begin{array}{c|c} \bar{\mathbf{A}} & \bar{\mathbf{b}} \\ \hline \bar{\mathbf{c}} & \end{array}\right)$ is in *Lanczos* form, since $\bar{\mathbf{A}}$ is *tridiagonal*, and $\bar{\mathbf{b}}$, $\bar{\mathbf{c}}^*$ are multiples of $\mathbf{e}_1$. Indeed, the biorthogonal Lanczos transformation matrices are

$$\mathbf{V} = \begin{bmatrix} 0 & 0 & 0 & \sqrt{192} \\ 0 & 0 & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{2} & \frac{1}{2} & 0 \\ 1 & \frac{1}{2} & 1 & \sqrt{3} \end{bmatrix},$$

$$\mathbf{W}^* = \begin{bmatrix} -12 & 6 & 1 & 1 \\ -24 & -4 & -2 & 0 \\ 0 & -4 & 0 & 0 \\ 8\sqrt{3} & 0 & 0 & 0 \end{bmatrix} \quad \Rightarrow \ \bar{\mathbf{A}} = \mathbf{W}^*\mathbf{A}\mathbf{V}, \ \bar{\mathbf{b}} = \mathbf{W}^*\mathbf{b}, \ \bar{\mathbf{c}} = \mathbf{c}\mathbf{V}.$$

The Arnoldi factorization applied to $\mathbf{A}$ with starting vector $\mathbf{b}$ yields

$$\hat{\mathbf{A}} = \begin{bmatrix} -3 & -12 & -8 & -24 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \ \hat{\mathbf{b}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \hat{\mathbf{c}} = \begin{bmatrix} 1 & 1 & 6 & -12 \end{bmatrix}.$$

Finally, the Arnoldi factorization applied to $\mathbf{A}^*$ with starting vector $\mathbf{c}^*$ yields

$$\tilde{\mathbf{A}} = \begin{bmatrix} 0.8791 & -1.9058 & 7.2276 & 14.8904 \\ 2.1931 & -1.4171 & 8.9876 & 18.6272 \\ 0 & 0.1455 & -3.6603 & -8.0902 \\ 0 & 0 & 1.6608 & 1.1983 \end{bmatrix}, \ \tilde{\mathbf{b}} = \begin{bmatrix} 13.4907 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$
$$\tilde{\mathbf{c}} = \begin{bmatrix} 0.0741 & -0.0973 & 0.4314 & 0.8938 \end{bmatrix},$$

which shows that although both $\mathbf{A}^*$ and $\tilde{\mathbf{A}}$ are in upper Hessenberg form, they are not the same; this is because the starting vector is not a multiple of $\mathbf{e}_1$.

11.3 Partial realization and rational interpolation by projection

In the preceding sections, we saw that rational Krylov methods achieve model reduction by moment matching, which leads to realization and interpolation problems. In this section we will revisit this problem from a more general point of view.

The issue of partial realization and, more generally, rational interpolation by projection has been treated in the work of Skelton and coworkers; see, e.g., [93], [366], [367]. Grimme in his thesis [152] showed how to accomplish this by means of Krylov methods. Recently, Van Dooren, Gallivan, and coworkers have revisited this area and provided connections with the Sylvester equation [128], [129].

Suppose that we are given a system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B}, \mathbf{C}^* \in \mathbb{R}^n$. We wish to find lower-dimensional models $\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \end{array}\right)$, where $\hat{\mathbf{A}} \in \mathbb{R}^{k\times k}$, $\hat{\mathbf{B}}, \hat{\mathbf{C}}^* \in \mathbb{R}^k$, $k < n$, such that $\hat{\mathbf{\Sigma}}$ preserves some properties of the original system, such as stability, norm-boundedness, passivity, etc. We wish to study this problem through appropriate projection methods. In other words, we will seek $\mathbf{V} \in \mathbb{R}^{n\times k}$ and $\mathbf{W} \in \mathbb{R}^{n\times k}$ such that $\mathbf{W}^*\mathbf{V} = \mathbf{I}_k$, and from the introduction, the reduced system should be given by

$$\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{W}^*\mathbf{A}\mathbf{V} & \mathbf{W}^*\mathbf{B} \\ \hline \mathbf{C}\mathbf{V} & \mathbf{D} \end{array}\right). \tag{1.8}$$

Proposition 11.6. *Let* $\mathbf{V} = [\mathbf{B}, \mathbf{AB}, \dots, \mathbf{A}^{k-1}\mathbf{B}] = \mathcal{R}_k(\mathbf{A}, \mathbf{B})$ *and* $\mathbf{W}$ *be any left inverse of* V. *Then* $\hat{\mathbf{\Sigma}}$ *defined by* (1.8) *is a partial realization of* $\mathbf{\Sigma}$ *and matches k Markov parameters.*

From a numerical point of view, one would not use $\mathbf{V}$ as defined above to construct $\hat{\mathbf{\Sigma}}$, since usually the columns of $\mathbf{V}$ are almost linearly dependent. As it turns out, any matrix whose column span is the same as that of $\mathbf{V}$ can be used.

Proof. We have $\hat{\mathbf{C}}\hat{\mathbf{B}} = \mathbf{C}\mathbf{V}\mathbf{W}^*\mathbf{B} = \mathbf{C}\mathcal{R}_k(\mathbf{A}, \mathbf{B})\mathbf{e}_1 = \mathbf{C}\mathbf{B}$; furthermore,

$$\hat{\mathbf{C}}\hat{\mathbf{A}}^j\hat{\mathbf{B}} = \mathbf{C}\mathcal{R}_k(\mathbf{A}, \mathbf{B})\mathbf{W}^*\mathbf{A}^j\mathcal{R}_k(\mathbf{A}, \mathbf{B})\mathbf{e}_1.$$

This expression is equal to $\mathbf{C}\mathcal{R}_k(\mathbf{A}, \mathbf{B})\mathbf{W}^*\mathbf{A}^j\mathbf{B} = \mathbf{C}\mathcal{R}_k(\mathbf{A}, \mathbf{B})\mathbf{e}_{j+1} = \mathbf{C}\mathbf{A}^j\mathbf{B}$, $j = 1, \dots, k-1$. □

Suppose now that we are given k distinct points $s_j \in \mathbb{C}$. $\mathbf{V}$ is defined as the generalized reachability matrix (4.85),

$$\mathbf{V} = \left[(s_1\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}, \dots, (s_k\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}\right],$$

and as before, let $\mathbf{W}$ be any left inverse of $\mathbf{V}$. The next proposition holds.

Proposition 11.7. *$\hat{\boldsymbol{\Sigma}}$ defined by* (1.8) *interpolates the transfer function of $\boldsymbol{\Sigma}$ at the points s_j, that is,*

$$\mathbf{H}(s_j) = \mathbf{C}(s_j\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} = \hat{\mathbf{C}}(s_j\mathbf{I}_k - \hat{\mathbf{A}})^{-1}\hat{\mathbf{B}} = \hat{\mathbf{H}}(s_j), \qquad j = 1, \dots, k.$$

Proof. The following string of equalities leads to the desired result:

$$\begin{aligned}\hat{\mathbf{C}}(s_j\mathbf{I}_k - \hat{\mathbf{A}})^{-1}\hat{\mathbf{B}} &= \mathbf{C}\mathbf{V}(s_j\mathbf{I}_k - \mathbf{W}^*\mathbf{A}\mathbf{V})^{-1}\mathbf{W}^*\mathbf{B}\\ &= \mathbf{C}\left[(s_1\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}, \dots, (s_k\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}\right]\left(\mathbf{W}^*(s_j\mathbf{I}_n - \mathbf{A})\mathbf{V}\right)^{-1}\mathbf{W}^*\mathbf{B}\\ &= \left[\mathbf{C}(s_1\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}, \dots, \mathbf{C}(s_k\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}\right]\left([\cdots\ \mathbf{W}^*\mathbf{B}\ \cdots]\right)^{-1}\mathbf{W}^*\mathbf{B}\\ &= \left[\mathbf{C}(s_1\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}, \dots, \mathbf{C}(s_k\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}\right]\mathbf{e}_j\\ &= \mathbf{C}(s_j\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}. \qquad \square\end{aligned}$$

The next result concerns matching the value of the transfer function at a given point $s_0 \in \mathbb{C}$, together with $k-1$ derivatives. For this we define the *generalized reachability matrix,*

$$\mathbf{V} = \left[(s_0\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}, (s_0\mathbf{I}_n - \mathbf{A})^{-2}\mathbf{B}, \dots, (s_0\mathbf{I}_n - \mathbf{A})^{-k}\mathbf{B}\right], \qquad (11.11)$$

together with any left inverse $\mathbf{W}$ thereof.

Proposition 11.8. *$\hat{\boldsymbol{\Sigma}}$ defined by* (1.8) *interpolates the transfer function of $\boldsymbol{\Sigma}$ at s_0, together with $k-1$ derivatives at the same point:*

$$\begin{aligned}\frac{(-1)^j}{j!}\frac{d^j}{ds^j}\mathbf{H}(s)\bigg|_{s=s_0} &= \mathbf{C}(s_0\mathbf{I}_n - \mathbf{A})^{-(j+1)}\mathbf{B}\\ &= \hat{\mathbf{C}}(s_0\mathbf{I}_k - \hat{\mathbf{A}})^{-(j+1)}\hat{\mathbf{B}} = \frac{(-1)^j}{j!}\frac{d^j}{ds^j}\hat{\mathbf{H}}(s)\bigg|_{s=s_0},\end{aligned}$$

where $j = 0, 1, \dots, k-1$.

Proof. Let $\mathbf{V}$ be as defined in (11.11), and let $\mathbf{W}$ be such that $\mathbf{W}^*\mathbf{V} = \mathbf{I}_r$. It readily follows that the projected matrix $s_0\mathbf{I}_r - \hat{\mathbf{A}}$ is in companion form, and therefore its powers are obtained by shifting its columns to the right:

$$\begin{aligned}s_0\mathbf{I}_k - \hat{\mathbf{A}} &= \mathbf{W}^*(s_0\mathbf{I}_n - \mathbf{A})\mathbf{V} = [\mathbf{W}^*\mathbf{B}, \mathbf{e}_1, \dots, \mathbf{e}_{k-1}] \Rightarrow\\ (s_0\mathbf{I}_k - \hat{\mathbf{A}})^\ell &= [\underbrace{*\cdots*}_{\ell-1}, \mathbf{W}^*\mathbf{B}, \mathbf{e}_1, \dots, \mathbf{e}_{k-\ell}].\end{aligned}$$

Consequently, $[\mathbf{W}^*(s_0\mathbf{I}_n - \mathbf{A})\mathbf{V}]^{-\ell}\mathbf{W}^*\mathbf{B} = \mathbf{e}_\ell$, which finally implies

$$\hat{\mathbf{C}}(s_0\mathbf{I}_k - \hat{\mathbf{A}})^{-\ell}\hat{\mathbf{B}} = \mathbf{C}\mathbf{V}\,[\mathbf{W}^*(s_0\mathbf{I} - \mathbf{A})\mathbf{V}]^{-\ell}\,\mathbf{W}^*\mathbf{B} = \mathbf{C}\mathbf{V}\mathbf{e}_\ell = \mathbf{C}(s_0\mathbf{I}_n - \mathbf{A})^{-\ell}\mathbf{B},$$
$$\ell = 1, 2, \dots, k,$$

completing the proof. $\square$

A moment's reflection shows that any $\bar{\mathbf{V}}$ that spans the same column space as $\mathbf{V}$ achieves the same objective. Furthermore, any projector composed of any combination of the above three cases achieves matching of an appropriate number of Markov parameters and moments. This is formalized next. For this purpose, we need the partial reachability matrix $\mathcal{R}_k(\mathbf{A},\mathbf{B}) = \begin{bmatrix}\mathbf{B} & \mathbf{AB} & \cdots & \mathbf{A}^{k-1}\mathbf{B}\end{bmatrix}$, and the partial generalized reachability matrix:

$$\mathcal{R}_k(\mathbf{A},\mathbf{B};\sigma) = \begin{bmatrix}(\sigma\mathbf{I}_n-\mathbf{A})^{-1}\mathbf{B} & (\sigma\mathbf{I}_n-\mathbf{A})^{-2}\mathbf{B} & \cdots & (\sigma\mathbf{I}_n-\mathbf{A})^{-k}\mathbf{B}\end{bmatrix}.$$

Corollary 11.9. Rational Arnoldi. (a) *If* $\mathbf{V}$ *as defined in the above three cases is replaced by* $\bar{\mathbf{V}} = \mathbf{VR}$, $\mathbf{R}\in\mathbb{R}^{k\times k}$, $\det\mathbf{R}\neq 0$, *and* $\mathbf{W}$ *by* $\bar{\mathbf{W}} = \mathbf{R}^{-1}\mathbf{W}$, *the same matching results hold true.*

(b) *Let* $\mathbf{V}$ *be such that*

$$\text{span col}\,\mathbf{V} = \text{span col}\,\begin{bmatrix}\mathcal{R}_k(\mathbf{A},\mathbf{B}) & \mathcal{R}_{m_1}(\mathbf{A},\mathbf{B};\sigma_1) & \cdots & \mathcal{R}_{m_\ell}(\mathbf{A},\mathbf{B};\sigma_\ell)\end{bmatrix}.$$

Let also $\mathbf{W}$ *be any left inverse of* $\mathbf{V}$. *The reduced system* (1.8) *matches k Markov parameters and* m_i *moments at* $\sigma_i\in\mathbb{C}$, $i=1,\ldots,\ell$.

The above corollary can be interpreted as *rational Arnoldi*. Its numerical implementation is obtained by combining regular Arnoldi with shift-invert Arnoldi for the shifts $\sigma_1,\ldots,\sigma_\ell$.

11.3.1 Two-sided projections

The results just discussed can be strengthened if the row span of the left matrix $\mathbf{W}^*$ is chosen to match the row span of an observability or generalized observability matrix. In such a case, twice as many moments can be matched with a reduced system of the same dimension as in the previous section. However, this is not always possible. Certain nonsingularity conditions have to be satisfied for this to happen.

We will denote by $\mathcal{O}_k(\mathbf{C},\mathbf{A})\in\mathbb{R}^{k\times n}$ the partial observability matrix consisting of the first k rows of $\mathcal{O}_n(\mathbf{C},\mathbf{A})\in\mathbb{R}^{n\times n}$ (see (4.38)). The first case is

$$\mathbf{V} = \mathcal{R}_k(\mathbf{A},\mathbf{B}),\ \ \mathbf{W} = \underbrace{(\mathcal{O}_k(\mathbf{C},\mathbf{A})\mathcal{R}_k(\mathbf{A},\mathbf{B}))}_{\mathcal{H}_k}{}^{-1}\,\mathcal{O}_k(\mathbf{C},\mathbf{A}).$$

Proposition 11.10. *Assuming that* $\det\mathcal{H}_k\neq 0$, $\hat{\boldsymbol{\Sigma}}$ *defined by* (1.8) *is a partial realization of* $\boldsymbol{\Sigma}$ *and matches* $2k$ *Markov parameters.*

Notice that the projection used above is more general that the one used for the two-sided Lanczos procedure and therefore fails less often. (Only one determinant has to be nonsingular for it not to fail.) The key reason for this is that the two-sided Lanczos process forces the resulting matrix $\hat{\mathbf{A}}$ to be *tridiagonal*, which is a rather stringent condition. This in turn is equivalent to the existence of an LU factorization of $\mathcal{H}_k$.

Given $2k$ distinct points $s_1,\ldots,s_{2k}$, we will make use of the following generalized reachability and observability matrices (recall (4.85) and (4.86)):

$$\tilde{\mathbf{V}} = \begin{bmatrix}(s_1\mathbf{I}_n-\mathbf{A})^{-1}\mathbf{B} & \cdots & (s_k\mathbf{I}_n-\mathbf{A})^{-1}\mathbf{B}\end{bmatrix},$$
$$\tilde{\mathbf{W}} = \begin{bmatrix}(s_{k+1}\mathbf{I}_n-\mathbf{A}^*)^{-1}\mathbf{C}^* & \cdots & (s_{2k}\mathbf{I}_n-\mathbf{A}^*)^{-1}\mathbf{C}^*\end{bmatrix}.$$

Proposition 11.11. *Assuming that* $\det \tilde{\mathbf{W}}^*\tilde{\mathbf{V}} \neq 0$*, the transfer function of the projected system* $\hat{\Sigma}$ *defined by* (1.8) *where* $\mathbf{V} = \tilde{\mathbf{V}}$ *and* $\mathbf{W} = \tilde{\mathbf{W}}(\tilde{\mathbf{V}}^*\tilde{\mathbf{W}})^{-1}$ *interpolates the transfer function of* Σ *at the points* s_i, $i = 1, \ldots, 2k$.

Proof. The string of equalities that follows proves the desired result:

$$\begin{aligned}
\hat{\mathbf{C}}(s_i\mathbf{I}_k - \hat{\mathbf{A}})^{-1}\hat{\mathbf{B}} &= \mathbf{C}\tilde{\mathbf{V}}\left(s_i\mathbf{I}_k - (\tilde{\mathbf{W}}^*\tilde{\mathbf{V}})^{-1}\tilde{\mathbf{W}}^*\mathbf{A}\tilde{\mathbf{V}}\right)^{-1}(\tilde{\mathbf{W}}^*\tilde{\mathbf{V}})^{-1}\tilde{\mathbf{W}}^*\mathbf{B}\\
&= \mathbf{C}\tilde{\mathbf{V}}\left(s_i\tilde{\mathbf{W}}^*\tilde{\mathbf{V}} - \tilde{\mathbf{W}}^*\mathbf{A}\tilde{\mathbf{V}}\right)^{-1}\tilde{\mathbf{W}}^*\mathbf{B}\\
&= \mathbf{C}\tilde{\mathbf{V}}\left(\tilde{\mathbf{W}}^*(s_i\mathbf{I}_n - \mathbf{A})\tilde{\mathbf{V}}\right)^{-1}\tilde{\mathbf{W}}^*\mathbf{B}\\
&= \mathbf{C}\tilde{\mathbf{V}}\left(\tilde{\mathbf{W}}^*[\cdots\ \mathbf{B}\ \cdots]\right)^{-1}\tilde{\mathbf{W}}^*\mathbf{B} = \mathbf{C}\tilde{\mathbf{V}}\mathbf{e}_i = \mathbf{C}(s_i\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B},\\
&\quad \text{for } i = 1, \ldots, k\\
&= \mathbf{C}\tilde{\mathbf{V}}\left(\begin{bmatrix} \vdots \\ \mathbf{C} \\ \vdots \end{bmatrix}\tilde{\mathbf{V}}\right)^{-1}\tilde{\mathbf{W}}^*\mathbf{B} = \mathbf{e}_i^*\tilde{\mathbf{W}}^*\mathbf{B} = \mathbf{C}(s_i\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}.
\end{aligned}$$

These relationships are valid for $i = k + 1, \ldots, 2k$. □

Remark 11.3.1. **(a)** The same procedure as above can be used to approximate *implicit systems*, i.e., systems that are given in a generalized form $\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$, $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)$, where $\mathbf{E}$ may be singular. The reduced system is given by

$$\hat{\mathbf{E}} = \mathbf{W}^*\mathbf{E}\mathbf{V},\ \hat{\mathbf{A}} = \mathbf{W}^*\mathbf{A}\mathbf{V},\ \hat{\mathbf{B}} = \mathbf{W}^*\mathbf{B},\ \hat{\mathbf{C}} = \mathbf{C}\mathbf{V},$$

where the projection $\mathbf{V}\mathbf{W}^*$ is given by

$$\begin{aligned}
\mathbf{W}^* &= \begin{bmatrix} \mathbf{C}(s_{k+1}\mathbf{E} - \mathbf{A})^{-1} \\ \vdots \\ \mathbf{C}(s_{2k}\ \mathbf{E} - \mathbf{A})^{-1} \end{bmatrix} \in \mathbb{R}^{k\times n},\\
\mathbf{V} &= \left[(s_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\ \cdots\ (s_k\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\right] \in \mathbb{R}^{n\times k}.
\end{aligned}$$

(b) The general case of model reduction of MIMO systems by means of tangential interpolation is analyzed in [131].

Remark 11.3.2. *The Sylvester equation and projectors.* An important connection between rational interpolation and the Sylvester equation follows from Theorem 6.5 and in particular formula (6.15). This shows that the solution of an appropriate Sylvester equation $\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{H} + \mathbf{B}\mathbf{G} = \mathbf{0}$ provides a projector that interpolates the original system $\mathbf{C}$, $\mathbf{A}$, $\mathbf{B}$ at minus the eigenvalues of $\mathbf{H}$. Therefore, the projectors discussed in the previous section can be obtained by solving Sylvester equations. This result was first observed in [128], [129]. See also [310].

11.3.2 How general is model reduction by rational Krylov?

Consider a SISO system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ of order n. Its transfer function is a scalar proper rational function $\mathbf{H} = \mathbf{n}/\mathbf{d}$ (i.e., the degree of the numerator is less than the degree of the denominator). Assume that the reduced system $\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \hat{\mathbf{D}} \end{array}\right)$ is also SISO of order k and described by the rational transfer function $\hat{\mathbf{H}} = \hat{\mathbf{n}}/\hat{\mathbf{d}}$. The question is whether there exists a projection $\Pi = \mathbf{V}\mathbf{W}^*$ (in general, oblique) where $\mathbf{V}$, $\mathbf{W}$ are obtained by means of the *rational Krylov procedure*, such that

$$\hat{\mathbf{A}} = \mathbf{W}^*\mathbf{A}\mathbf{V}, \quad \hat{\mathbf{B}} = \mathbf{W}^*\mathbf{B}, \quad \hat{\mathbf{C}} = \mathbf{C}\mathbf{V}.$$

Since $\mathbf{D}$, $\hat{\mathbf{D}}$ are not involved in the above projection, we will assume that $\mathbf{D} = \hat{\mathbf{D}} = \mathbf{0}$. The answer to this question is *affirmative* in the *generic case*. To see this, consider the difference of the transfer functions

$$\mathbf{H} - \hat{\mathbf{H}} = \frac{\mathbf{n}}{\mathbf{d}} - \frac{\hat{\mathbf{n}}}{\hat{\mathbf{d}}} = \frac{\mathbf{n}\hat{\mathbf{d}} - \hat{\mathbf{n}}\mathbf{d}}{\mathbf{d}\hat{\mathbf{d}}}.$$

This difference has $n+k$ zeros (including those at infinity): $\mu_i, i = 1, \ldots, n+k$. Therefore, $\hat{\mathbf{H}}$ can be constructed by interpolating $\mathbf{H}$ at $2k+1$ of these zeros. For $k = n-1$, all μ_i have to be used, while for $k < n-1$ there are several choices of the interpolation points.

In the *nongeneric case*, that is, when $\mathbf{d}$ and $\hat{\mathbf{d}}$ have a common factor of degree, say, ℓ, there is a restriction on the degree of $\hat{\mathbf{H}}$ which can be obtained by interpolation, namely,

$$k \leq n - \ell - 1. \tag{11.12}$$

The situation, however, is not as simple for MIMO systems, that is, systems described by nonscalar proper rational matrix transfer functions. See [138], [130] for details. Here is a simple example that illustrates the SISO case.

Example 11.12. Let $\mathbf{\Sigma}$ be given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -1 & -5 & -10 & -10 & -5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{C} = [1 \;\; 8 \;\; 26 \;\; 40 \;\; 5], \quad \mathbf{D} = 0,$$

which implies $\mathbf{H}(s) = \frac{5s^4+40s^3+26s^2+8s+1}{(s+1)^5}$. The reduced-order system $\hat{\mathbf{\Sigma}}$ is

$$\hat{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -3 & 3 \end{bmatrix}, \quad \hat{\mathbf{B}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \hat{\mathbf{C}} = [-1 \;\; 0 \;\; 5], \quad \hat{\mathbf{D}} = 0,$$

which implies $\hat{\mathbf{H}}(s) = \frac{5s^2-1}{(s-1)^3}$. It readily follows that

$$\mathbf{H}(s) - \hat{\mathbf{H}}(s) = \frac{-128\,s^5}{(s+1)^5\,(s-1)^3},$$

which shows that this expression has five zeros at 0 and three at infinity. As interpolation points, we pick all five zeros at 0 and one at infinity. Thus we have

$$\mathbf{V} = \begin{bmatrix} \mathbf{A}^{-1}\mathbf{B} & \mathbf{A}^{-2}\mathbf{B} & \mathbf{A}^{-3}\mathbf{B} \end{bmatrix} = \begin{bmatrix} -1 & 5 & -15 \\ 0 & -1 & 5 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\bar{\mathbf{W}}^* = \begin{bmatrix} \mathbf{C}\mathbf{A}^{-1} \\ \mathbf{C}\mathbf{A}^{-2} \\ \mathbf{C} \end{bmatrix} = \begin{bmatrix} 3 & 16 & 30 & 0 & -1 \\ 1 & 0 & -30 & -16 & -3 \\ 1 & 8 & 26 & 40 & 5 \end{bmatrix}.$$

Therefore,

$$\bar{\mathbf{W}}^*\mathbf{V} = \begin{bmatrix} -3 & -1 & 5 \\ -1 & 5 & 15 \\ -1 & -3 & -1 \end{bmatrix} \Rightarrow \mathbf{W}^* = \left(\bar{\mathbf{W}}^*\mathbf{V}\right)^{-1}\bar{\mathbf{W}}^* = \begin{bmatrix} -1 & -5 & -10 & 21 & 3 \\ 0 & -1 & -5 & -23 & -3 \\ 0 & 0 & -1 & 8 & 1 \end{bmatrix}.$$

Finally,

$$\tilde{\mathbf{A}} = \mathbf{W}^*\mathbf{A}\mathbf{V} = \begin{bmatrix} 3 & 1 & 0 \\ -3 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \tilde{\mathbf{B}} = \mathbf{W}^*\mathbf{B} = \begin{bmatrix} 3 \\ -3 \\ 1 \end{bmatrix}, \tilde{\mathbf{C}} = \mathbf{C}\mathbf{V} = \begin{bmatrix} -1 & -3 & -1 \end{bmatrix}.$$

It is readily checked that the triples $(\hat{\mathbf{C}}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ and $(\tilde{\mathbf{C}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ are equivalent and therefore $\hat{\mathbf{H}}(s) = \tilde{\mathbf{C}}(s\mathbf{I} - \tilde{\mathbf{A}})^{-1}\tilde{\mathbf{B}}$.

11.3.3 Model reduction with preservation of stability and passivity

Recall the definitions of passivity and positive realness discussed in section 5.9.1. Given is a passive system described by $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$. This means that its transfer function $\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$ is positive real.

The following result is shown in [27]; see also [234]. For a result derived from the one presented below, see [309]. For different approaches to the problem, see [33], [117], [205], [118], [352].

Lemma 11.13. *Let λ_i be such that $\mathbf{H}^*(-\lambda_i) + \mathbf{H}(\lambda_i) = \mathbf{0}$. If the frequencies (interpolation points) in Proposition* 11.10 *are chosen so that s_j, $j = 1, \ldots, k$, are (stable) spectral zeros, and $s_{j+k} = -s_j$, $j = 1, \ldots, k$, i.e., as zeros of the spectral factors and their mirror images, the projected system is both stable and passive.*

Recall that the zeros of the square system $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$, where $\mathbf{D}$ is nonsingular, are the eigenvalues of the matrix $\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$. If $\mathbf{D}$ is singular, the zeros are the finite eigenvalues of a generalized eigenvalue problem, namely, all finite λ for which

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} - \lambda \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

loses rank. Thus when $\mathbf{D}+\mathbf{D}^*$ is nonsingular, the zeros of the spectral factors are the eigenvalues of the following *Hamiltonian* matrix:

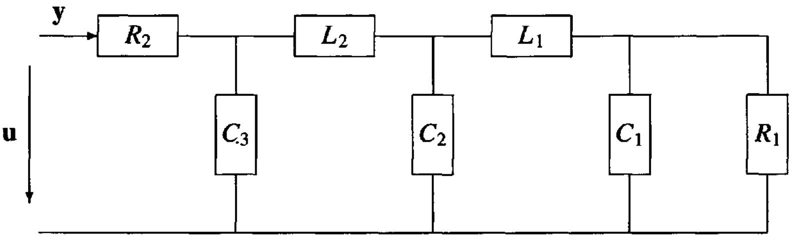
$$\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\mathbf{A}^* \end{pmatrix} - \begin{pmatrix} \mathbf{B} \\ -\mathbf{C}^* \end{pmatrix} (\mathbf{D}+\mathbf{D}^*)^{-1} \begin{pmatrix} \mathbf{C} & \mathbf{B}^* \end{pmatrix}.$$

In particular, if the reduced-order system has dimension k, we need to compute k eigenvalues $\lambda_i \in \mathbb{C}_-$. If the nonsingularity condition is not satisfied, we end up with a generalized eigenvalue problem of the pair $\mathcal{A}$, $\mathcal{E}$, where

$$\mathcal{A} = \begin{pmatrix} \mathbf{A} & \mathbf{0} & \mathbf{B} \\ \mathbf{0} & -\mathbf{A}^* & -\mathbf{C}^* \\ \mathbf{C} & \mathbf{B}^* & \mathbf{D}+\mathbf{D}^* \end{pmatrix}, \quad \mathcal{E} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Remark 11.3.3. A further result in passivity preserving model reduction was developed by Sorensen [309], who showed that the invariant of spaces $(\mathcal{A}, \mathcal{E})$ provide the desired passivity preserving projector $\mathbf{VW}^*$. Therefore, the problem of model reduction with preservation of passivity has been transformed into a *structured eigenvalue problem* and can thus be solved using ARPACK software.

Example 11.14. We consider the following Resistor-Inductor-Capacitor (RLC) ladder network. The state variables are as follows: x_1, the voltage across C_1; x_2, the current through L_1; x_3, the voltage across C_2; x_4, the current through L_2; and x_5, the voltage across C_3. The input is the voltage $\mathbf{u}$, and the output is the current $\mathbf{y}$, as shown in the figure below. We assume that all the capacitors and inductors have unit value, while $R_1 = \frac{1}{2}\,\Omega$, $R_2 = \frac{1}{5}\,\Omega$.



$$\mathbf{A} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -5 \end{bmatrix}, \ \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 2 \end{bmatrix}, \ \mathbf{C} = [0\ 0\ 0\ 0\ -2], \ \mathbf{D} = 1,$$

$$\mathbf{H}(s) = \frac{s^5+3s^4+6s^3+9s^2+7s+3}{s^5+7s^4+14s^3+21s^2+23s+7},$$

$$\mathbf{H}(s)+\mathbf{H}(-s) = \frac{2(s^{10}-s^8-12s^6+5s^4+35s^2-21)}{(s^5+7s^4+14s^3+21s^2+23s+7)(s^5-7s^4+14s^3-21s^2+23s-7)}.$$

The zeros of the stable spectral factor are $-.1833 \pm 1.5430i$, $-.7943$, -1.3018, -1.8355; these can also be computed as the square roots in the left half plane of the eigenvalues of

$$\mathbf{A}^2 - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\mathbf{A} = \begin{bmatrix} 3 & -2 & 1 & 0 & 0 \\ 2 & -2 & 0 & 1 & 0 \\ 1 & 0 & -2 & 0 & 1 \\ 0 & 1 & 0 & -2 & -5 \\ 0 & 0 & 1 & 1 & 4 \end{bmatrix}.$$

According to Lemma 11.13, the interpolation points are therefore $s_1 = 1.8355$, $s_2 = -1.8355$, $s_3 = 1.3018$, $s_4 = -1.3018$:

$$\bar{\mathbf{V}} = [(s_1\mathbf{I}_5 - \mathbf{A})^{-1}\mathbf{B} \quad (s_2\mathbf{I}_5 - \mathbf{A})^{-1}\mathbf{B}] = \begin{bmatrix} .0065 & .2640 \\ .0250 & .0434 \\ .0524 & .1842 \\ .1212 & -.2947 \\ .2749 & .7251 \end{bmatrix},$$

$$\bar{\mathbf{W}} = \begin{pmatrix} \mathbf{C}(s_3\mathbf{I}_5 - \mathbf{A})^{-1} \\ \mathbf{C}(s_4\mathbf{I}_5 - \mathbf{A})^{-1} \end{pmatrix} = \begin{pmatrix} -.0157 & .0519 & -.0833 & .1603 & -.2919 \\ 1.0671 & -.7451 & .0972 & -.6186 & -.7081 \end{pmatrix},$$

$$\bar{\mathbf{W}}\bar{\mathbf{V}} = -\begin{pmatrix} .0634 & .2762 \\ .2762 & .0640 \end{pmatrix} = \mathbf{LU} \Rightarrow \mathbf{L} = \begin{pmatrix} .2317 & 1 \\ 1 & 0 \end{pmatrix}, \mathbf{U} = -\begin{pmatrix} .2762 & .0640 \\ 0 & .2613 \end{pmatrix},$$

$$\mathbf{V} = \bar{\mathbf{V}}\mathbf{U}^{-1} = \begin{bmatrix} -.0236 & -1.0040 \\ -.0905 & -0.1439 \\ -.1897 & -0.6584 \\ -.4388 & 1.2351 \\ -.9953 & -2.5310 \end{bmatrix},$$

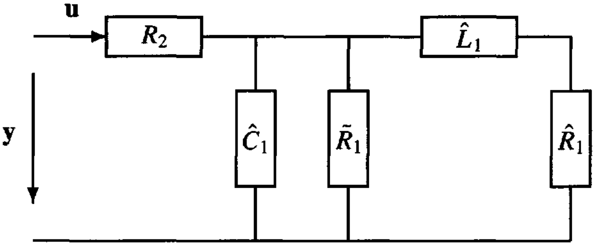
$$\mathbf{W} = \mathbf{L}^{-1}\bar{\mathbf{W}} = \begin{bmatrix} 1.0671 & -.7451 & .0972 & -.6186 & -.7081 \\ -.2630 & .2245 & -.1058 & .3036 & -.1279 \end{bmatrix}.$$

The reduced system is

$$\hat{\mathbf{A}} = \mathbf{WAV} = -\begin{pmatrix} 3.2923 & 5.0620 \\ 0.9261 & 2.5874 \end{pmatrix}, \ \hat{\mathbf{B}} = \mathbf{WB} = -\begin{pmatrix} 1.4161 \\ 0.2560 \end{pmatrix},$$

$\hat{\mathbf{C}} = \mathbf{CV} = [1.9905 \;\; 5.0620]$, and $\hat{\mathbf{D}} = \mathbf{D}$. It readily follows that the zeros of the spectral factors are s_1, s_2, s_3, and s_4, thus showing that the reduced system is passive. Furthermore, $\hat{\mathbf{H}}(s_i) = \mathbf{H}(s_i)$, $i = 1, 2, 3, 4$. In addition,

$$\hat{\mathbf{H}}(s) = \mathbf{D} + \hat{\mathbf{C}}(s\mathbf{I}_2 - \hat{\mathbf{A}})^{-1}\mathbf{B} = 1 + \frac{4.1152s + 2.3419}{s^2 + 1.7652s + 1.4891}.$$



The values of the above elements are $\hat{R}_1 = 2.8969\,\Omega$, $\hat{L}_1 = 5.0904\,H$, $\tilde{R}_1 = 3.4404\,\Omega$, $\hat{C}_1 = 0.2430\,F$, $R_2 = 1\,\Omega$.

11.4 Chapter summary

In this chapter we discussed Krylov methods applied to model reduction. First we showed that the Lanczos and Arnoldi procedures yield reduced-order models that match a certain number of Markov parameters (i.e., moments at infinity) of the original model. Thus while Arnoldi matches a number of moments equal to the order of the reduced model, Lanczos matches twice as many moments.

In the second part, we examined model reduction by moment matching from a more general point of view. It was shown that rational Krylov methods match moments at preassigned points in the complex plane, that is, they provide an iterative solution of the *rational interpolation* problem. Furthermore, the question of the generality of the rational Krylov procedure has been briefly investigated. Finally, the reduction with preservation of passivity was shown to be solvable using rational Krylov methods, by choosing the interpolation points as a subset of the spectral zeros of the system. The solution of this problem can thus be reduced to the solution of a structured eigenvalue problem.

This page intentionally left blank

Part V

SVD–Krylov Methods and Case Studies

This page intentionally left blank

Chapter 12

SVD–Krylov Methods

The basic ingredient of SVD-based methods consists of certain gramians $\mathcal{P}$, $\mathcal{Q}$, usually referred to as reachability, observability gramians, respectively. These are positive (semi) definite, can be simultaneously diagonalized, and are used to construct a projector onto the dominant eigenspace of the product $\mathcal{P}\mathcal{Q}$. These gramians are solutions to Lyapunov equations.

Gramians are also involved in Krylov-based methods, but these are given in factored form $\mathcal{P} = \mathcal{R}\mathcal{R}^*$, $\mathcal{Q} = \mathcal{O}^*\mathcal{O}$, where $\mathcal{R}$, $\mathcal{O}$ are reachability, observability or generalized reachability, generalized observability matrices; these matrices are used to project the original model so that interpolation (moment matching) conditions are satisfied. Consequently, there is no need for solving Lyapunov equations in this case. A further aspect of Krylov methods is that they can be implemented in an *iterative* way, using well-known procedures.

12.1 Connection between SVD and Krylov methods

This chapter is dedicated to a discussion of various connections between these two approximation methods. There are four such connections:

1. Reachability and generalized reachability matrices can be obtained as solutions of Sylvester equations, the latter being a general form of Lyapunov equations. The same holds for observability and generalized observability matrices.

2. By appropriate choice of the weightings, weighted balanced truncation methods can be reduced to Krylov methods.

3. There are methods that combine attributes of both approaches, for instance, model reduction by least squares. In this case, instead of two, only one gramian (the observability gramian $\mathcal{Q}$) is computed, while at the same time, moment matching takes place. The former is the SVD part while the latter is the Krylov part of this method.

4. The bottleneck in applying SVD-based methods (balanced truncation) to large-scale systems is the fact that the solution of Lyapunov equations requires $O(n^3)$ operations, where n is the dimension of the original system. One remedy to this situation consists of using iterative methods for solving these Lyapunov equation approximately. This leads to approximately balancing transformations and to reduced systems obtained by approximately balanced truncation.

The first three items will be discussed in the three sections that follow. The fourth item will be discussed in section 12.2.

12.1.1 Krylov methods and the Sylvester equation

Recall from Chapter 6 that the solution to the Sylvester equation $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$ can be written according to (6.14) as

$$\mathbf{X} = \underbrace{\left[\, (\mu_1\mathbf{I} + \mathbf{A})^{-1}\mathbf{c}_1 \ \cdots \ (\mu_k\mathbf{I} + \mathbf{A})^{-1}\mathbf{c}_1 \,\right]}_{\mathcal{R}(\mathbf{A},\mathbf{c}_1)} \underbrace{\begin{bmatrix} \mathbf{c}_2^*\hat{\mathbf{w}}_1 & & \\ & \ddots & \\ & & \mathbf{c}_2^*\hat{\mathbf{w}}_n \end{bmatrix}}_{\tilde{\mathbf{W}}^*} \mathbf{W}^*,$$

where $\mathbf{W}^*\mathbf{B} = \mathbf{M}\mathbf{W}^*$, $\mathbf{M} = \operatorname{diag}(\mu_1, \ldots, \mu_k)$, is the eigenvalue decomposition of $\mathbf{B}$.

We recognize the matrix $\mathcal{R}(\mathbf{A}, \mathbf{c}_1)$ as a generalized reachability matrix (see (4.85)), and according to Proposition 11.7, if we chose any left inverse $\mathbf{W}^*$ thereof, the two matrices lead to a projected system which interpolates at $-\mu_i^*$, i.e., at the mirror image the eigenvalues of $\mathbf{B}$.

Thus, if we want to interpolate at given points λ_i, choose a matrix $\mathbf{B}$ whose eigenvalues are $\mu_i = -\lambda_i^*$, choose $\mathbf{c}_1 = \mathbf{b}$ and $\mathbf{c}_2$ such that the pair $(\mathbf{B}, \mathbf{c}_2^*)$ is reachable, and solve the resulting Sylvester equation. The solution provides one part of the required projector.

12.1.2 From weighted balancing to Krylov

Recall the *weighted reachability* gramian,

$$\mathcal{P} = \frac{1}{2\pi}\int_{-\infty}^{\infty} (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\,\mathbf{W}(i\omega)\mathbf{W}^*(-i\omega)\,\mathbf{B}^*(-i\omega\mathbf{I} - \mathbf{A}^*)^{-1}\,d\omega,$$

defined by (7.33). Therein, $\mathbf{W}$ is the input weight to the system which will be assumed for simplicity to be a scalar function. Let $\mathbb{I}$ denote the unit step function ($\mathbb{I}(t) = 1, t \geq 0$, and $\mathbb{I}(t) = 0$, otherwise). We define the weight

$$\begin{aligned}\mathbf{W}_\epsilon(i\omega)\mathbf{W}_\epsilon(-i\omega) &= \frac{1}{2\epsilon}\left[\mathbb{I}(\omega - (\omega_0 - \epsilon)) - \mathbb{I}(\omega - (\omega_0 + \epsilon))\right] \\ &\quad + \frac{1}{2\epsilon}\left[\mathbb{I}(-\omega - (\omega_0 - \epsilon)) - \mathbb{I}(-\omega - (\omega_0 + \epsilon))\right].\end{aligned}$$

As $\epsilon \to 0$, the above expression tends to the sum of two impulses $\delta(\omega - \omega_0) + \delta(-\omega - \omega_0)$. In this case the gramian becomes

$$\mathcal{P} = (i\omega_0\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(-i\omega_0\mathbf{I} - \mathbf{A}^*)^{-1} + (-i\omega_0\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{B}^*(i\omega_0\mathbf{I} - \mathbf{A}^*)^{-1}.$$

Therefore,

$$\mathcal{P} = \underbrace{\left[(i\omega_0\mathbf{I} - \mathbf{A})^{-1}\mathbf{B},\ \ (-i\omega_0\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\right]}_{\mathbf{V}} \begin{bmatrix} \mathbf{B}^*(-i\omega_0\mathbf{I} - \mathbf{A}^*)^{-1} \\ \mathbf{B}^*(\ \ i\omega_0\mathbf{I} - \mathbf{A}^*)^{-1} \end{bmatrix}.$$

Thus projection with $\mathbf{VW}^*$, where $\mathbf{W}^*\mathbf{V} = \mathbf{I}_2$, yields a reduced system which interpolates the original system at $\pm i\omega_0$. Thus, with appropriate choice of the weights, weighted balanced truncation reduces to a Krylov method (rational interpolation) and matches the original transfer function at points on the imaginary axis.

Example 12.1. The purpose of this example is to show yet another way to combine SVD and Krylov methods. Recall Example 11.5. Given the reachability, observability matrices $\mathcal{R}_4(\mathbf{A},\mathbf{B})$, $\mathcal{O}_4(\mathbf{C},\mathbf{A})$, the gramians $\mathcal{P} = \mathcal{R}_4(\mathbf{A},\mathbf{B})\mathcal{R}_4^*(\mathbf{A},\mathbf{B})$ and $\mathcal{Q} = \mathcal{O}_4^*(\mathbf{C},\mathbf{A})\mathcal{O}_4(\mathbf{C},\mathbf{A})$ can be assigned. The balancing method would now provide transformations which simultaneously diagonalize $\mathcal{P}$ and $\mathcal{Q}$. In this new basis, model reduction would amount to truncation of the states which correspond to small eigenvalues of the product $\mathcal{PQ}$. For this choice of gramians, notice that the square roots of these eigenvalues are the absolute values of the eigenvalues of the 4 × 4 Hankel matrix associated with the triple $(\mathbf{C},\mathbf{A},\mathbf{B})$.

We will now compute a balancing transformation that diagonalizes the gramians simultaneously. For this purpose, let

$$\mathcal{R}_4^*(\mathbf{A},\mathbf{B})\mathcal{O}_4^*(\mathbf{C},\mathbf{A}) = \mathcal{H}_4 = \mathbf{USV}^*.$$

Then

$$\mathbf{T} = \mathbf{S}^{-1/2}\mathbf{V}^*\mathcal{O}_4(\mathbf{C},\mathbf{A}) \ \text{ and } \ \mathbf{T}^{-1} = \mathcal{R}_4(\mathbf{A},\mathbf{B})\mathbf{U}\mathbf{S}^{-1/2}.$$

It follows that

$$\mathbf{U} = \begin{bmatrix} -0.0522 & -0.0279 & -0.7458 & -0.6635 \\ 0.2013 & 0.2255 & 0.6209 & -0.7233 \\ -0.5292 & -0.7921 & 0.2366 & -0.1911 \\ -0.8226 & 0.5665 & 0.0470 & -0.0119 \end{bmatrix}, \ \mathbf{V} = \mathbf{U}\,\mathrm{diag}\,(1,-1,1,-1),$$

and $\mathbf{S} = \mathrm{diag}\,(71.7459, 64.0506, 2.4129, 1.1082)$. Thus the balancing transformation is

$$\mathbf{T} = \begin{bmatrix} -3.4952 & -4.6741 & 0.3783 & -0.4422 \\ 5.3851 & -3.6051 & 1.1347 & -0.2233 \\ 3.4811 & -10.6415 & -2.5112 & -1.1585 \\ -15.3406 & -10.8670 & -2.8568 & -0.6985 \end{bmatrix} \Rightarrow \mathbf{T}\mathcal{P}\mathbf{T}^* = \mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1} = \mathbf{S},$$

$$\mathbf{F} = \mathbf{TAT}^{-1} = \begin{bmatrix} -6.6419 & 5.6305 & -1.5031 & 2.4570 \\ -5.6305 & 3.1458 & -1.0941 & 1.7895 \\ -1.5031 & 1.0941 & 0.1628 & -1.0490 \\ -2.4570 & 1.7895 & 1.0490 & 0.3333 \end{bmatrix}, \ \mathbf{G} = \mathbf{TB} = \begin{bmatrix} -0.4422 \\ -0.2233 \\ -1.1585 \\ -0.6985 \end{bmatrix},$$

and $\mathbf{H} = \mathbf{CT}^{-1} = \mathbf{G}^*\,\mathrm{diag}\,(1,-1,1,-1)$. Model reduction now proceeds by retaining the leading parts of $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$.

Thus given the gramians $\mathcal{P} = \mathcal{R}\mathcal{R}^*$ and $\mathcal{Q} = \mathcal{O}^*\mathcal{O}$, depending on how we proceed, different reduced-order models are obtained. For instance, if we apply an SVD-type method, the gramians are simultaneously diagonalized. If we choose a Krylov method, on the other hand, the square root factors $\mathcal{R}$ and $\mathcal{O}$ are directly used to construct the projector. Notice that the former methods loose their interpolation properties.

To complete the example, we also quote the result obtained by applying the Arnoldi method (11.10):

$$\tilde{\mathbf{F}}_1 = [-2],\ \tilde{\mathbf{G}}_1 = [1],\ \tilde{\mathbf{H}}_1 = [1].$$

$$\tilde{\mathbf{F}}_2 = \begin{bmatrix} 0 & -2 \\ 1 & -1 \end{bmatrix},\ \tilde{\mathbf{G}}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ \tilde{\mathbf{H}}_2 = [1,\ \ -2].$$

$$\tilde{\mathbf{F}}_3 = \begin{bmatrix} 0 & 0 & 4 \\ 1 & 0 & 0 \\ 0 & 1 & -3 \end{bmatrix},\ \tilde{\mathbf{G}}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},\ \tilde{\mathbf{H}}_3 = [1,\ \ -2,\ 0].$$

$$\tilde{\mathbf{F}} = \begin{bmatrix} 0 & 0 & 0 & -24 \\ 1 & 0 & 0 & -8 \\ 0 & 1 & 0 & -12 \\ 0 & 0 & 1 & -3 \end{bmatrix},\ \tilde{\mathbf{G}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix},\ \tilde{\mathbf{H}} = [1,\ \ -2,\ 0,\ 4].$$

In this case, according to Proposition 11.10, the truncated model of order k matches $2k$ Markov parameters.

12.1.3 Approximation by least squares

In this section, we discuss the approximation of a stable discrete-time system in the *least squares sense*. It turns out that this model reduction method matches *moments*, i.e., Markov parameters, and guarantees stability of the reduced-order system. This section follows [158]. Other contributions along these lines are [93] and [178].

Consider the discrete-time system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{b} \\ \hline \mathbf{c} & \end{array}\right)$, $|\lambda_i(\mathbf{A})| < 1$. Recall the Hankel matrix (4.63). For this section, we use the notation

$$\mathcal{H}_k = \begin{bmatrix} h_1 & h_2 & \cdots & h_k \\ h_2 & h_3 & \cdots & h_{k+1} \\ \vdots & \vdots & & \vdots \\ h_k & h_{k+1} & \cdots & h_{2k-1} \\ \vdots & \vdots & & \vdots \end{bmatrix},\quad \mathbf{h}_{k+1} = \begin{bmatrix} h_{k+1} \\ h_{k+2} \\ \vdots \\ h_{2k} \\ \vdots \end{bmatrix}.$$

We denote by $\mathcal{O}$ the infinite observability matrix and by $\mathcal{R}_k$ the reachability matrix containing k terms. Thus $\mathcal{H}_k = \mathcal{O}\mathcal{R}_k$ and $\mathbf{h}_{k+1} = \mathcal{O}A^k\mathbf{b}$. Furthermore, $\mathcal{Q} = \mathcal{O}^*\mathcal{O}$ is the observability *gramian* of the system.

A reduced system is obtained by computing the *least squares* fit of the $(k+1)$st column $\mathbf{h}_{k+1}$ of $\mathcal{H}$ on the preceding k columns of $\mathcal{H}$, i.e., the columns of $\mathcal{H}_k$:

$$\mathcal{H}_k\mathbf{x}_{LS} = \mathbf{h}_{k+1} + \mathbf{e}_{LS},$$

where $\mathbf{e}_{LS}$ is the least squares error vector. From the standard theory of least squares it follows that

$$\mathbf{x}_{LS} = \left(\mathcal{H}_k^*\mathcal{H}_k\right)^{-1}\mathcal{H}_k^*\mathbf{h}_{k+1} = \left(\mathcal{R}_k^*\mathcal{Q}\mathcal{R}_k\right)^{-1}\mathcal{R}_k^*\mathcal{Q}\mathbf{A}^k\mathbf{b}.$$

The characteristic polynomial of the resulting system is

$$\chi_{LS}(\sigma) = (1 \ \ \sigma \ \ \cdots \ \ \sigma^k)\begin{pmatrix} -\mathbf{x}_{LS} \\ 1 \end{pmatrix}.$$

The corresponding matrices are

$$\mathbf{A}_{LS} = \Pi_L\mathbf{A}\Pi_R,\ \mathbf{b}_{LS} = \Pi_L\mathbf{b},\ \mathbf{c}_{LS} = \mathbf{c}\Pi_R,$$

where the projection $\Pi_R\Pi_L$ is defined as

$$\Pi_L = \left(\mathcal{R}_k^*\mathcal{Q}\mathcal{R}_k\right)^{-1}\mathcal{R}_k^*\mathcal{Q},\ \Pi_R = \mathcal{R}_k \ \Rightarrow\ \Pi_L\Pi_R = \mathbf{I}_k.$$

If we let $\mathbf{x}_{LS}^* = (\alpha_0 \ \cdots \ \alpha_{k-1})$, the reduced triple has the following form:

$$\mathbf{\Sigma}_{LS} = \left(\begin{array}{c|c} \mathbf{A}_{LS} & \mathbf{b}_{LS} \\ \hline \mathbf{c}_{LS} & \end{array}\right) = \left(\begin{array}{ccccc|c} 0 & 0 & \cdots & 0 & \alpha_0 & 1 \\ 1 & 0 & \cdots & 0 & \alpha_1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \alpha_{k-2} & 0 \\ 0 & 0 & \cdots & 1 & \alpha_{k-1} & 0 \\ \hline h_1 & h_2 & \cdots & h_{k-1} & h_k & \end{array}\right). \tag{12.1}$$

Lemma 12.2. Stability of the least squares approximant. *The least squares approximant $\mathbf{\Sigma}_{LS}$ of the stable system $\mathbf{\Sigma}$ is stable.*

Before giving the proof of this result, we transform the approximant to a new basis. Let $\mathcal{Q} = \mathbf{R}^*\mathbf{R}$ be the Cholesky factorization of $\mathcal{Q}$, and let

$$\mathcal{R}_k^*\mathbf{R}^*\mathbf{R}\mathcal{R}_k = \mathbf{W}^*\mathbf{W}, \qquad \mathbf{W} \in \mathbb{R}^{k\times k},\ \det\mathbf{W} \neq 0,$$

be the Cholesky factorization of $\mathcal{R}_k^*\mathcal{Q}\mathcal{R}_k$. We define

$$\mathbf{T} = \mathbf{R}\mathcal{R}_k\mathbf{W}^{-1} \in \mathbb{R}^{n\times k} \ \Rightarrow\ \mathbf{T}^*\mathbf{T} = \mathbf{I}_k.$$

It follows that

$$\bar{\mathbf{A}}_{LS} = \mathbf{W}\mathbf{A}_{LS}\mathbf{W}^{-1} = \mathbf{T}^*\mathbf{A}\mathbf{T}.$$

Thus by similarity, $\mathbf{A}_{LS}$ and $\bar{\mathbf{A}}_{LS}$ have the same characteristic polynomial. We now show that $\bar{\mathbf{A}}_{LS}$ has eigenvalues inside the unit circle.

Proposition 12.3. *If* $|\lambda(\mathbf{A})| < 1$, *it follows that* $|\lambda(\bar{\mathbf{A}}_{LS})| \leq 1$.

Proof. Recall that $\mathbf{A}^*\mathcal{Q}\mathbf{A} + \mathbf{c}^*\mathbf{c} = \mathcal{Q}$. Assuming without loss of generality that $\mathcal{Q} = \mathbf{I}$, it follows that $\mathbf{A}^*\mathbf{A} \leq \mathbf{I}_n$; multiplying this relationship by $\mathbf{T}^*$ and $\mathbf{T}$ on the left and right, respectively, we have $\mathbf{T^*A^*AT} \leq \mathbf{I}_k$. Since $\begin{pmatrix} \mathbf{I}_n & \mathbf{L}^* \\ \mathbf{L} & \mathbf{I}_m \end{pmatrix} > \mathbf{0}$ is equivalent to $\mathbf{I}_n - \mathbf{L^*L} > \mathbf{0}$ or $\mathbf{I}_m - \mathbf{LL}^*$ for $\mathbf{L} \in \mathbb{R}^{m\times n}$, the latter inequality is equivalent to $\mathbf{ATT^*A^*} \leq \mathbf{I}_n$, which in turn implies $\bar{\mathbf{A}}_{LS}\bar{\mathbf{A}}_{LS}^* = \mathbf{T^*ATT^*A^*T} \leq \mathbf{I}_k$; this concludes the proof of the proposition. □

Proof. We will now prove Lemma 12.2. The above proposition implies that the least squares approximant may have poles on the unit circle. To show that this cannot happen, we complete the columns of $\mathbf{T}$ to obtain a unitary matrix: $\mathbf{U} = [\mathbf{T}\ \ \mathbf{S}]$, $\mathbf{UU}^* = \mathbf{U^*U} = \mathbf{I}_n$. Then $\mathbf{I}_n - \mathbf{A^*A} = \mathbf{c^*c}$ implies $\mathbf{I}_n - \mathbf{U^*A^*UU^*AU} = \mathbf{U^*c^*cU}$ which, written explicitly, gives

$$\begin{pmatrix} \mathbf{I}_k & \\ & \mathbf{I}_{n-k} \end{pmatrix} - \begin{pmatrix} \mathbf{T^*A^*T} & \mathbf{T^*A^*S} \\ \mathbf{S^*A^*T} & \mathbf{S^*A^*S} \end{pmatrix}\begin{pmatrix} \mathbf{T^*AT} & \mathbf{T^*AS} \\ \mathbf{S^*AT} & \mathbf{S^*AS} \end{pmatrix} = \begin{pmatrix} \mathbf{T^*c^*cT} & \mathbf{T^*c^*cS} \\ \mathbf{S^*c^*cT} & \mathbf{S^*c^*cS} \end{pmatrix}.$$

The $(1, 1)$ block of this equation is

$$\mathbf{I}_k - \mathbf{T^*A^*TT^*AT} - \mathbf{T^*A^*SS^*AT} = \mathbf{T^*c^*cT}.$$

Assume now that the reduced system has a pole of magnitude one, i.e., there exists $\mathbf{x} \in \mathbb{C}^k$ such that

$$\mathbf{T^*ATx} = \lambda\mathbf{x} \quad \text{and} \quad |\lambda| = 1.$$

Multiplying the above equation with $\mathbf{x}^*$ on the left and $\mathbf{x}$ on the right, we obtain

$$\mathbf{x^*T^*A^*SS^*ATx} + \mathbf{x^*T^*c^*cTx} = \|\mathbf{S^*ATx}\|^2 + \|\mathbf{cTx}\|^2 = 0.$$

This implies $\mathbf{cTx} = \mathbf{0}$ and $\mathbf{S^*ATx} = \mathbf{0}$, which means that the reduced system is not observable if it has a pole on the unit circle. Now let $\mathbf{y} = \mathbf{ATx}$; then $\mathbf{U^*y} = \begin{pmatrix} \mathbf{T}^* \\ \mathbf{S}^* \end{pmatrix}\mathbf{y} = \begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix} \Rightarrow \mathbf{y} = (\mathbf{T}\ \ \mathbf{S})\begin{pmatrix} \mathbf{x} \\ \mathbf{0} \end{pmatrix} \Rightarrow \mathbf{y} = \mathbf{Tx}$; hence $\mathbf{Tx}$ is an eigenvector of $\mathbf{A}$ with eigenvalue 1, which is in the null space of $\mathbf{c}$; this is a contradiction of the fact that $\mathbf{c}, \mathbf{A}$ is observable. This concludes the proof. □

From (12.1) it follows that the least squares approximant matches the first k Markov parameters of the original system.

The above considerations are summarized next.

Theorem 12.4. *Given the discrete-time stable system $\mathbf{\Sigma}$, let $\mathbf{\Sigma}_{LS}$ be the kth-order approximant obtained by means of the least squares fit of the $(k+1)$st column of the Hankel matrix of $\mathbf{\Sigma}$ to the preceding k columns. $\mathbf{\Sigma}_{LS}$ is given by* (12.1) *and enjoys the following properties:*

- *$\mathbf{\Sigma}_{LS}$ is stable.*
- *k Markov parameters are matched:* $\mathbf{cA}^{i-1}\mathbf{b} = \bar{\mathbf{c}}\bar{\mathbf{A}}^{i-1}\bar{\mathbf{b}}$, $i = 1, \dots, k$.

Remark 12.1.1. *Connection with Prony's method.* The least squares method just presented is related to Prony's method. For details on this and other issues, e.g., generalization to continuous-time systems, MIMO systems, and an error bound, see [158]. A different way to obtain stable reduced-order systems by projection is described in [207].

12.2 Introduction to iterative methods

Recently, there has been renewed interest in iterative projection methods for model reduction. Three leading efforts in this area are Padé via Lanczos (PVL) [114], multipoint rational interpolation [152], and implicitly restarted dual Arnoldi [188].

The PVL approach exploits the deep connection between the (nonsymmetric) Lanczos process and classic moment matching techniques; for details see [333], [336]. The multipoint rational interpolation approach utilizes the rational Krylov method of Ruhe [281] to provide moment matching of the transfer function at selected frequencies and hence to obtain enhanced approximation of the transfer function over a broad frequency range. These techniques have proved to be very effective. PVL has enjoyed considerable success in circuit simulation applications. Rational interpolation achieves remarkable approximation of the transfer function with very-low-order models. Nevertheless, there are shortcomings to both approaches. In particular, since the methods are local, it is difficult to establish rigorous error bounds. Heuristics have been developed that appear to work, but no global results exist. Second, the rational interpolation method requires selection of interpolation points. This is not an automated process and relies on ad hoc specification by the user.

The dual Arnoldi method runs two separate Arnoldi processes, one for the reachability subspace and one for the observability subspace, and then constructs an oblique projection from the two orthogonal Arnoldi basis sets. The basis sets and the reduced model are updated using a generalized notion of implicit restarting. The updating process is designed to iteratively improve the approximation properties of the model. Essentially, the reduced model is reduced further, keeping the best features, and then expanded via the dual Arnoldi processes to include new information. The goal is to achieve approximation properties related to balanced realizations. Other related approaches [68], [228], [194], [267] work directly with projected forms of the two Lyapunov equations to obtain low rank approximations to the system gramians. An overview of similar model reduction methods can be found in [336]. See also [92] and [210].

In the following, we describe two approaches to iterative projection methods. The first uses the cross gramian introduced in section 4.3.2. The second is based on Smith-related iterative methods. These two sections closely follow [310] and [157], respectively.

12.3 An iterative method for approximate balanced reduction

Computational techniques are well known for producing a balancing transformation $\mathbf{T}$ for small- to medium-scale problems [167], [224]. Such methods rely on an initial Schur decomposition of $\mathbf{A}$ followed by additional factorization schemes of dense linear algebra. The computational complexity involves $O(n^3)$ arithmetic operations and the storage of several dense matrices of order n, i.e., $O(n^2)$ storage. For large state space systems, this approach for obtaining a reduced model is clearly intractable. Yet, computational experiments indicate that such systems are representable with very-low-order models. This provides the primary motivation for seeking methods to construct projections of low order.

The *cross gramian* $\mathcal{X}$ of Σ is defined for square systems ($m = p$) by (4.59); it is the solution to the Sylvester equation

$$\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{B}\mathbf{C} = \mathbf{0}. \tag{4.59}$$

Our approach to model reduction as reported in section 12.3.1 consists of constructing low rank k approximate solutions to this matrix by setting $\mathcal{X} = \mathbf{V}\hat{\mathcal{X}}\mathbf{W}^*$ with $\mathbf{W}^*\mathbf{V} = \mathbf{I}_k$ and then projecting using $\mathbf{V}$ together with $\mathbf{W}$. An implicit restart mechanism is presented which allows the computation of an approximation to the best rank k approximation to $\mathcal{X}$. Furthermore, a reduced basis constructed from this procedure has an error estimate in the SISO case.

As mentioned earlier, construction of a balancing transformation has typically relied on solving for the reachability and observability gramians $\mathcal{P}$, $\mathcal{Q}$ and developing a balancing transformation from the EVD of the product $\mathcal{P}\mathcal{Q}$. However, it turns out that in the SISO case, and in the case of symmetric MIMO systems, a balancing transformation can be obtained directly from the eigenvector basis for the cross gramian.

Recall Lemma 5.6, which states that the eigenvalues of the Hankel operator for square systems are given by the eigenvalues of the cross gramian. In addition, the following two results hold.

Lemma 12.5. *Let* $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ *be a stable SISO system that is reachable and observable. There is a nonsingular symmetric matrix* $\mathbf{J}$ *such that* $\mathbf{AJ} = \mathbf{JA}^*$, *and* $\mathbf{CJ} = \mathbf{B}^*$. *It follows that the three gramians are related as follows:* $\mathcal{P} = \mathcal{X}\mathbf{J}$ *and* $\mathcal{Q} = \mathbf{J}^{-1}\mathcal{X}$.

Proof. Let $\mathcal{R} = [\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B}]$, $\mathcal{O} = [\mathbf{C}^*, \mathbf{A}^*\mathbf{C}^*, \ldots, (\mathbf{A}^*)^{n-1}\mathbf{C}^*]^*$, and define $\mathcal{H}_k = \mathcal{O}\mathbf{A}^k\mathcal{R}$. The hypothesis implies that both $\mathcal{R}$ and $\mathcal{O}$ are nonsingular, and it is easily shown that the Hankel matrix $\mathcal{H}_k$ is symmetric. Define $\mathbf{J} = \mathcal{R}\mathcal{O}^{-*}$. Note $\mathbf{J} = \mathcal{O}^{-1}\mathcal{H}_0\mathcal{O}^{-*}$ so that $\mathbf{J} = \mathbf{J}^*$. Moreover,

$$\mathbf{CJ} = \mathbf{e}_1^*\mathcal{O}\mathbf{J}^* = \mathbf{e}_1^*\mathcal{O}\mathcal{O}^{-1}\mathcal{R}^* = \mathbf{e}_1^*\mathcal{R}^* = \mathbf{B}^*.$$

To complete the proof, we note that $\mathbf{AJ} = \mathbf{A}\mathcal{R}\mathcal{O}^{-*} = \mathcal{O}^{-1}\mathcal{H}_1\mathcal{O}^{-*}$, and hence $\mathbf{AJ} = (\mathbf{AJ})^* = \mathbf{JA}^*$. It follows that $\mathbf{A}(\mathcal{X}\mathbf{J}) + (\mathcal{X}\mathbf{J})\mathbf{A}^* + \mathbf{BB}^* = \mathbf{0}$, which implies $\mathcal{P} = \mathcal{X}\mathbf{J}$. Similarly, multiplying the cross gramian equation on the left by $\mathbf{J}^{-1}$ we obtain $\mathbf{A}^*\mathbf{J}^{-1}\mathcal{X} + \mathbf{J}^{-1}\mathcal{X}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$, which in turn implies the desired $\mathcal{Q} = \mathbf{J}^{-1}\mathcal{X}$. □

Lemma 12.6. *Let* $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$ *be a stable SISO system that is reachable and observable. Suppose that* (4.59) *is satisfied. Then* $\mathcal{X}$ *is diagonalizable with* $\mathcal{X}\mathbf{Z} = \mathbf{ZD}$, *where* $\mathbf{Z}$ *is nonsingular and* $\mathbf{D}$ *is real and diagonal. Moreover, up to a diagonal scaling of its columns,* $\mathbf{Z}$ *is a balancing transformation for the system.*

Proof. Since $\mathcal{P} = \mathcal{X}\mathbf{J}$ is symmetric positive definite, it has a Cholesky factorization $\mathcal{X}\mathbf{J} = \mathbf{LL}^*$ with $\mathbf{L}$ lower triangular and nonsingular. This implies $\mathbf{L}^{-1}\mathcal{X}\mathbf{L} = \mathbf{L}^*\mathbf{J}^{-1}\mathbf{L} = \mathbf{QDQ}^*$, where $\mathbf{Q}$ is orthogonal and $\mathbf{D}$ is real and diagonal since $\mathbf{L}^*\mathbf{J}^{-1}\mathbf{L}$ is symmetric. Thus,

$$\mathcal{X} = \mathbf{ZDZ}^{-1} \text{ and } \mathbf{J} = \mathbf{ZD}^{-1}\mathbf{Z}^* \text{ with } \mathbf{Z} = \mathbf{LQ}.$$

Since $|\mathbf{D}|^{1/2}\mathbf{D} = \mathbf{D}|\mathbf{D}|^{1/2}$, we may replace $\mathbf{Z}$ by $\mathbf{Z}|\mathbf{D}|^{-1/2}$ to obtain

$$\mathbf{J} = \mathbf{ZD}_{\mathbf{J}}\mathbf{Z}^* \text{ and } \mathcal{X} = \mathbf{ZDZ}^{-1} \text{ with } \mathbf{D}_{\mathbf{J}} = |\mathbf{D}|\mathbf{D}^{-1} = \operatorname{diag}(\pm 1).$$

It follows that $\mathcal{X}\mathbf{J} = \mathbf{Z}(\mathbf{D}\mathbf{D_J})\mathbf{Z}^*$ and $\mathbf{J}^{-1}\mathcal{X} = \mathbf{Z}^{-*}(\mathbf{D_J}\mathbf{D})\mathbf{Z}^{-1}$, since $\mathbf{D_J} = \text{diag}\,(\pm 1)$ implies $\mathbf{D_J} = \mathbf{D_J}^{-1}$. If we let $\mathbf{S} = \mathbf{D}\mathbf{D_J} = |\mathbf{D}|$, we note $\mathbf{S}$ is a diagonal matrix with positive diagonal elements, and the above discussion together with (4.59) gives

$$(\mathbf{Z}^{-1}\mathbf{A}\mathbf{Z})\mathbf{S} + \mathbf{S}(\mathbf{Z}^*\mathbf{A}^*\mathbf{Z}^{-*}) + \mathbf{Z}^{-1}\mathbf{B}\mathbf{B}^*\mathbf{Z}^{-*} = \mathbf{0}.$$

After some straightforward manipulations, we find that $(\mathbf{Z}^{-1}\mathbf{A}\mathbf{Z})^*\mathbf{S} + \mathbf{S}(\mathbf{Z}^{-1}\mathbf{A}\mathbf{Z}) + \mathbf{Z}^*\mathbf{C}^*\mathbf{C}\mathbf{Z} = \mathbf{0}$. To conclude the proof, we note that $\mathbf{C}\mathbf{J} = \mathbf{B}^*$ implies $\mathbf{C}\mathbf{Z}\mathbf{D_J} = \mathbf{B}^*\mathbf{Z}^{-*}$, and hence the system transformed by $\mathbf{Z}$ is indeed balanced. □

A corollary of this result is that if the cross gramian is diagonal, then the system is essentially balanced. Since $\mathbf{Z}$ can be taken to be a diagonal matrix such that $\mathbf{C}\mathbf{Z}\mathbf{e}_j = \pm\mathbf{B}^*\mathbf{Z}^{-*}\mathbf{e}_j$, a diagonal linear transformation balances the system. This diagonal transformation is constructed trivially from the entries of **B** and **C**.

As mentioned earlier, in the SISO case the absolute values of the diagonal entries of **D** are the Hankel singular values of the system. If we assume that the diagonal entries of **D** have been ordered in decreasing order of magnitude, then the $n \times k$ matrix $\mathbf{Z}_k$ consisting of the leading k columns of **Z** provides a truncated balanced realization with all of the desired stability and error properties.

*The question, then, is how to compute a reasonable approximation to $\mathbf{Z}_k$ directly. We wish to avoid computing all of **Z** first, followed by truncation, especially in the large-scale setting.*

12.3.1 Approximate balancing through low rank approximation of the cross gramian

We now consider the problem of computing the best rank k approximation $\mathcal{X}_k$ to the cross gramian $\mathcal{X}$. We seek a restarting mechanism analogous to implicit restarting for eigenvalue computations that enable us to compute $\mathcal{X}_k$ directly instead of computing all of $\mathcal{X}$ and truncating.

Motivation. Suppose $\mathcal{X} = \mathbf{U}\mathbf{S}\mathbf{V}^*$ is the SVD of $\mathcal{X}$. Let $\mathcal{X} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^* + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^*$, where $\mathbf{U} = [\mathbf{U}_1,\ \mathbf{U}_2]$ and $\mathbf{V} = [\mathbf{V}_1,\ \mathbf{V}_2]$. Projecting the cross gramian equation on the right with $\mathbf{V}_1$ we get

$$(\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{B}\mathbf{C})\mathbf{V}_1 = \mathbf{0} \ \Rightarrow\ \mathbf{A}\mathbf{U}_1\mathbf{S}_1 + \mathbf{U}_1\mathbf{S}_1(\mathbf{V}_1^*\mathbf{A}\mathbf{V}_1) + \mathbf{B}\mathbf{C}\mathbf{V}_1 = \mathbf{E},$$

where $\mathbf{E} = -\mathbf{U}_2\mathbf{S}_1\mathbf{V}_1^*\mathbf{A}\mathbf{V}_1$. Observe that $\|\mathbf{E}\|_2 = O(\sigma_{k+1})\|\mathbf{A}\|$, where σ_{k+1} is the first neglected singular value, that is, the $(k+1)$st singular value of $\mathcal{X}$.

Procedure. This suggests the following type of iteration. First obtain a projection of **A**,

$$\mathbf{A}\mathbf{V} + \mathbf{V}\mathbf{H} = \mathbf{F} \ \text{with}\ \mathbf{V}^*\mathbf{V} = \mathbf{I}, \qquad \mathbf{V}^*\mathbf{F} = \mathbf{0},$$

where $\mathbf{V} \in \mathbb{R}^{n\times m}$ and $\mathbf{H} \in \mathbb{R}^{m\times m}$ with $k < m \ll n$. We require that this projection yield a stable matrix **H**. If not, **V** must be modified. Using the technique of implicit restarting as developed in [153], we can cast out the unstable eigenvalues. This would result in a **V** and an **H** of smaller dimension. As long as this dimension remains greater than k, we can execute the remaining steps described below. Should the dimension fall below k, then some method must be used to expand the basis set. As an example, one could use a block form of Arnoldi with the remaining $\mathbf{V}$ as a starting block of vectors.

Once $\mathbf{V}$ and $\mathbf{H}$ have been computed, solve the Sylvester equation

$$\mathbf{AW} + \mathbf{WH} + \mathbf{BCV} = \mathbf{0} \tag{12.2}$$

for $\mathbf{W} \in \mathbb{R}^{n\times m}$. We now compute the SVD of $\mathbf{W}$:

$$\mathbf{W} = \mathbf{QSY}^* = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{pmatrix} \begin{pmatrix} \mathbf{S}_1 & \\ & \mathbf{S}_2 \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1^* \\ \mathbf{Y}_2^* \end{pmatrix},$$

$$\text{where } \mathbf{Q}_1 \in \mathbb{R}^{n\times k},\ \mathbf{S}_1 \in \mathbb{R}^{k\times k},\ \mathbf{Y}_1 \in \mathbb{R}^{m\times k},$$

and $\mathbf{S}_1$ contains the k most significant singular values of $\mathbf{W}$. If we now project (12.2) on the right with $\mathbf{Y}_1$, we obtain

$$\mathbf{AU}_1\mathbf{S}_1 + \mathbf{U}_1\mathbf{S}_1\mathbf{H}_1 + \mathbf{BCV}_1 = -\mathbf{Q}_2\mathbf{S}_2\mathbf{Y}_2^*\mathbf{HY}_1, \quad \text{where } \mathbf{H}_1 = \mathbf{Y}_1^*\mathbf{HY}_1.$$

At this point the projection has been updated from $\mathbf{V} \in \mathbb{R}^{n\times m}$ to $\mathbf{V}_1 = \mathbf{VY}_1 \in \mathbb{R}^{n\times k}$. We must now have a mechanism to expand the subspace generated by $\mathbf{V}_1$ so that the above process can be repeated. One possibility is as follows. First notice that the expression $\mathcal{X}' = \mathbf{Q}_1\mathbf{S}_1\mathbf{Y}_1^*\mathbf{V}^* \in \mathbb{R}^{n\times n}$ can be considered as an approximate solution of the equation $\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{BC} = \mathbf{0}$; therefore, we project the expression $\mathbf{A}\mathcal{X}' + \mathcal{X}'\mathbf{A} + \mathbf{BC}$ on the left with $\mathbf{Q}_1^*$. We call the result $\mathbf{E}'$,

$$\mathbf{E}' = (\mathbf{Q}_1^*\mathbf{AQ}_1)(\mathbf{S}_1\mathbf{Y}_1^*\mathbf{V}^*) + (\mathbf{S}_1\mathbf{Y}_1^*\mathbf{V}^*)\mathbf{A} + \mathbf{Q}_1^*\mathbf{BC},$$

and then solve the equation $\mathbf{HZ} + \mathbf{ZA} + \mathbf{E}' = \mathbf{0}$ for $\mathbf{Z} \in \mathbb{R}^{k\times n}$. This represents a residual correction to the solution of the Sylvester equation $\mathbf{A}\mathcal{X} + \mathcal{X}\mathbf{A} + \mathbf{BC} = \mathbf{0}$ when projected on the left. However, instead of adding this correction to the basis set $\mathbf{V}_1$, we simply adjoin the columns of $\mathbf{Z}^*$ to the subspace spanned by the columns of $\mathbf{V}_1$ and project $\mathbf{A}$ onto this space. Let

$$\begin{pmatrix} \mathbf{VY}_1 & \mathbf{Z}^* \end{pmatrix} = \begin{pmatrix} \mathbf{V}_1' & \mathbf{V}_2' \end{pmatrix} \begin{pmatrix} \mathbf{S}_1' & \\ & \mathbf{S}_2' \end{pmatrix} \begin{pmatrix} \mathbf{U}_1'^* \\ \mathbf{U}_2'^* \end{pmatrix},$$

$$\mathbf{V}_1 \in \mathbb{R}^{n\times k},\ \mathbf{S}_1' \in \mathbb{R}^{k\times k},\ \mathbf{U}_1' \in \mathbb{R}^{2k\times k}.$$

Thus the updated projector at this stage is $\mathbf{V}_1'$, the updated projection of $\mathbf{A}$ is $\mathbf{H}' = \mathbf{V}_1'^*\mathbf{AV}_1'$, and the procedure can go to (12.2).

The latter portion of the iteration is analogous to the Davidson part of the Jacobi–Davidson algorithm for eigenvalue calculation proposed by Sleijpen and van der Vorst [300]. These ideas are summarized in the algorithm sketched in Figure 12.1 for the complete iteration.

There have been many ideas for the numerical solution of Lyapunov and Sylvester equations [176], [350], [184], [284], [297], [267], [228], [194]. The approach described here nearly gives the best rank k approximation directly. However, the best rank k approximation is not a fixed point of this iteration. A slightly modified correction equation is needed to achieve this. Note that the iteration of Hodel, Poola, and Tenison [176] for the Lyapunov equation also suffers from this difficulty.

```
(Implicit) Restarting Algorithm
1. $\mathbf{AV} + \mathbf{VH} = \mathbf{F}$ with $\mathbf{V}^*\mathbf{V} = \mathbf{I}$ and $\mathbf{V}^*\mathbf{F} = \mathbf{0}$
2. while (not_converged),
    2.1 Solve Sylvester equation projected in $\mathcal{R}(\mathbf{A}, \mathbf{B})$:
        Solve $\mathbf{AW} + \mathbf{WH} + \mathbf{BCV} = \mathbf{0}$
        $[\mathbf{Q}, \mathbf{S}, \mathbf{Y}] = svd(\mathbf{W})$;

    2.2 Contract the space (keep largest singular values):
        $\mathbf{S}_1 = \mathbf{S}(1 : k, 1 : k)$;
        $\mathbf{Q}_1 = \mathbf{Q}(:, 1 : k), \mathbf{Y}_1 = \mathbf{Y}(:, 1 : k)$;
        $\mathbf{V} \leftarrow \mathbf{VY}_1$;   $\mathbf{H} \leftarrow \mathbf{Y}_1^*\mathbf{HY}_1$;

    2.3 Correct Sylvester equation projected in $\mathcal{O}(\mathbf{C}, \mathbf{A})$:
        Form $\mathbf{E}' = (\mathbf{Q}_1^*\mathbf{AQ}_1)(\mathbf{S}_1\mathbf{V}^*) + (\mathbf{S}_1\mathbf{V}^*)\mathbf{A} + \mathbf{Q}_1^*\mathbf{BC}$;
        Solve $\mathbf{HZ} + \mathbf{ZA} + \mathbf{E}' = \mathbf{0}$

    2.4 Expand the space – Adjoin correction and project:
        $[\mathbf{V}, \mathbf{S}, \mathbf{U}] = svd([\mathbf{V},\ \mathbf{Z}])$
        New projector: $\mathbf{V} \leftarrow \mathbf{V}(:, 1 : k)$
        New projected A: $\mathbf{H} \leftarrow \mathbf{V}^*\mathbf{AV}$;
```

Figure 12.1. *An implicitly restarted method for the cross gramian.*

We make a final remark on this method. The equation $\mathbf{AW} + \mathbf{WH} + \mathbf{BCV} = \mathbf{0}$ introduces in the projector directions, which lie in the *reachability* or *Krylov space* spanned by the columns of $\mathcal{R}(\mathbf{A}, \mathbf{B})$, while the correction equation $\mathbf{HZ} + \mathbf{ZA} + \mathbf{E}' = \mathbf{0}$ introduces directions which are orthogonal to the unobservable space $\ker \mathcal{O}(\mathbf{C}, \mathbf{A})$.

A special Sylvester equation

Efficient solution of a special Sylvester equation provides the key to steps 2.1 and 2.3 in the algorithm in Figure 12.1. Both steps result in an equation of the form $\mathbf{AZ} + \mathbf{ZH} + \mathbf{M} = \mathbf{0}$, where $\mathbf{H}$ is a $k \times k$ stable matrix and $\mathbf{M}$ is an $n \times k$ matrix with $k \ll n$. The special feature of this Sylvester equation is that $\mathbf{A}$ is much larger in dimension than $\mathbf{H}$. The first method we propose here is the one described in section 6.1.5.

Since the eigenvalues of $\mathbf{A}$ are assumed to be in the open left half plane and the eigenvalues of $-\mathbf{H}^*$ are in the open right half plane, the k eigenvalues of largest real part for the block upper triangular matrix in (6.21) are the desired eigenvalues. When k is small, it is possible to compute the eigenvalues of $\mathbf{H}$ in advance of the computation of the partial Schur decomposition in (6.21). Within this framework, the implicitly restarted Arnoldi method [305] (implemented in ARPACK [227]) can be used effectively to compute this partial Schur decomposition. If there is a reasonable gap between the eigenvalues of $\mathbf{H}$ and the imaginary axis, then the implicitly restarted Arnoldi method (IRAM) is successful in computing the k eigenvalues of largest real part using only matrix vector products. In any case, exact knowledge of the desired eigenvalues provides several opportunities to enhance convergence. One possibility is to use a single Cayley transformation (costing one sparse direct factorization) to map the eigenvalues of $\mathbf{A}$ to the interior and the eigenvalues of $\mathbf{H}$ to the exterior of the unit disk.

An alternative would be to construct a Schur decomposition $\mathbf{H} = \mathbf{QRQ}^*$ and transform the equation to $\mathbf{A(ZU)} + \mathbf{(ZU)R} + \mathbf{(MU)} = \mathbf{0}$, where $\mathbf{R}$ is (quasi) upper triangular, and then solve for the columns of $\hat{\mathbf{Z}} = \mathbf{ZU}$ from left to right via

$$(\mathbf{A} - \rho_{jj}\mathbf{I})\hat{\mathbf{z}}_j = -\mathbf{m}_j - \sum_{i=1}^{j-1} \hat{\mathbf{z}}_i \rho_{ij} \text{ for } j = 1, 2, \ldots, k,$$

where ρ_{ij} are the elements of the upper triangular matrix $\mathbf{R}$, and $\hat{\mathbf{z}}_j$, $\hat{\mathbf{m}}_j$ are the columns of $\hat{\mathbf{Z}}$, $\hat{\mathbf{M}} = \mathbf{MU}$. This would require a separate sparse direct factorization of a large $n \times n$ complex matrix at each step j. That would be k such factorizations, and each of these would have a potentially different sparsity pattern for $\mathbf{L}$ and $\mathbf{U}$ due to pivoting for stability. Staying in real arithmetic would require working with quadratic factors involving $\mathbf{A}$ and hence would destroy sparsity.

Approximate balancing transformation

Let the SVD of the cross gramian be $\mathcal{X} = \mathbf{USV}^* = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^* + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^*$, where $\mathbf{S}_1$ contains the significant singular values. It follows that $\mathbf{U}_1$ is an orthonormal basis for a subspace spanned by the columns of the reachability subspace $\mathcal{R}(\mathbf{A}, \mathbf{B})$, while $\mathbf{V}_1$ provides an orthonormal basis for a subspace orthogonal to the unobservable space ker $\mathcal{O}(\mathbf{C}, \mathbf{A})$. Our goal in this section is to obtain an *approximate* balancing transformation through the eigenvectors of $\mathcal{X}_1 = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1$ corresponding to nonzero eigenvalues.

In the SISO case, we know that $\mathcal{X}$ has real eigenvalues. If $\mathcal{X}_1$ were obtained as $\mathcal{X}_1 = \mathbf{Z}_1\mathbf{D}_1\mathbf{W}_1^*$, where the diagonal elements of $\mathbf{D}_1$ are the eigenvalues of largest magnitude (i.e., the dominant Hankel singular values) and $\mathbf{W}_1^*\mathbf{Z}_1 = \mathbf{I}_1$, then, as discussed previously, $\mathbf{Z}_1$ would provide a balancing transformation. Instead, we have the best rank k approximation to $\mathcal{X}$ in $\mathcal{X}_1$. Therefore, we shall attempt to approximate the relevant eigenvector basis for $\mathcal{X}$ with an eigenvector basis for $\mathcal{X}_1$. It is easily seen that any eigenvector of $\mathcal{X}_1$ corresponding to a nonzero eigenvalue must be in the range of $\mathbf{U}_1$. In fact, we see that

$$\mathcal{X}_1\mathbf{U}_1\mathbf{S}_1^{1/2} = \mathbf{U}_1\mathbf{S}_1^{1/2}\mathbf{G},$$

where $\mathbf{G} = \mathbf{S}_1^{1/2}\mathbf{V}_1^*\mathbf{U}_1\mathbf{S}_1^{1/2}$. If $\mathbf{GZ} = \mathbf{ZD}_1$ with $\mathbf{D}_1$ real and diagonal, then taking

$$\mathbf{Z}_1 = \mathbf{U}_1\mathbf{S}_1^{1/2}\mathbf{Z}|\mathbf{D}_1|^{-1/2} \text{ and } \mathbf{W}_1 = \mathbf{V}_1\mathbf{S}_1^{1/2}\mathbf{Z}^{-*}|\mathbf{D}_1|^{-1/2}$$

provides $\mathcal{X}_1 = \mathbf{Z}_1\mathbf{D}_1\mathbf{W}_1^*$ with $\mathbf{W}_1^*\mathbf{Z}_1 = \mathbf{I}_1$. Note also that $\mathcal{X}\mathbf{Z}_1 = \mathcal{X}_1\mathbf{Z}_1 + (\mathcal{X} - \mathcal{X}_1)\mathbf{Z}_1 = \mathbf{Z}_1\mathbf{D}_1 + O(\sigma_{k+1})$, and thus we have an approximate balancing transformation represented by this $\mathbf{Z}_1$. The resulting reduced model is

$$\mathbf{A}_1 = \mathbf{W}_1^*\mathbf{AZ}_1, \quad \mathbf{B}_1 = \mathbf{W}_1^*\mathbf{B}, \quad \mathbf{C}_1 = \mathbf{CZ}_1;$$

this projected system is approximately balanced.

Remark 12.3.1. *Stability of the reduced model.* Our methods pertain to stable systems and therefore it is important in many applications that the reduced model be stable as well. We must have the eigenvalues of the reduced matrix $\mathbf{A}_1$ in the left half plane. The reduced model obtained through the algorithm shown in Figure 12.1 is almost always stable in practice, but occasionally it might be unstable. One approach to achieving a stable reduced model is to apply the techniques developed in [153] to the projected quantities. That would amount to applying the implicit restarting mechanism to rid the projected matrix of unstable modes.

The preceding discussion was primarily concerned with the SISO case. For MIMO systems, we propose the idea of embedding the system in a symmetric system; this is explored next.

Extension of theory and computation to MIMO systems

In the large-scale setting, there is a clear advantage to working with the cross gramian instead of with the two gramians related to reachability and observability. In addition to the fact that only one Sylvester equation need be solved, there is the question of compatibility that arises when working with the pair of gramians. Since two separate projections must be computed, one cannot be certain that the two subspaces are the same as the ones that would have been achieved through computing the full gramians and then truncating.

The crucial property of the three gramians in the SISO case is $\mathcal{X}^2 = \mathcal{P}\mathcal{Q}$. It is easy to see that this relationship holds true for MIMO systems which are *symmetric*, i.e., the transfer function is symmetric. Of course, this is not generally the case. To make use of this property, we propose to embed the given system into a symmetric system with more inputs and outputs but the same number of state variables. Given the m-input, p-output system $\boldsymbol{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \end{array}\right)$, we seek $\tilde{\mathbf{B}} \in \mathbb{R}^{n\times p}$ and $\tilde{\mathbf{C}} \in \mathbb{R}^{m\times n}$ such that the augmented system

$$\hat{\boldsymbol{\Sigma}} = \left(\begin{array}{c|c} \hat{\mathbf{A}} & \hat{\mathbf{B}} \\ \hline \hat{\mathbf{C}} & \end{array}\right) = \left(\begin{array}{c|cc} \mathbf{A} & \tilde{\mathbf{B}} & \mathbf{B} \\ \hline \mathbf{C} & & \\ \tilde{\mathbf{C}} & & \end{array}\right) \in \mathbb{R}^{(n+m+p)\times(n+p+m)}$$

is square and symmetric, i.e., the Markov parameters $\hat{\mathbf{C}}\hat{\mathbf{A}}^{\ell}\hat{\mathbf{B}} \in \mathbb{R}^{(m+p)\times(m+p)}$ are symmetric for all $\ell \geq 0$. That this can be done follows readily from properties of system realizations. The important aspect of this embedding is that the complexity (McMillan degree or number of states) of the system has not increased. Therefore, the norm ($\mathcal{H}_2$ or $\mathcal{H}_\infty$) of the original system is bounded from above by that of the augmented system.

MIMO systems and the symmetrizer

Let $\mathbf{J} = \mathbf{J}^*$ be a symmetrizer for $\mathbf{A}$, i.e., $\mathbf{AJ} = \mathbf{JA}^*$. The following quantities are defined:

$$\tilde{\mathbf{B}} = \mathbf{JC}^* \text{ and } \tilde{\mathbf{C}} = \mathbf{B}^*\mathbf{J}^{-1} \;\Rightarrow\; \hat{\mathbf{B}} = \begin{bmatrix}\tilde{\mathbf{B}} & \mathbf{B}\end{bmatrix},\; \hat{\mathbf{C}} = \begin{bmatrix}\mathbf{C} \\ \tilde{\mathbf{C}}\end{bmatrix} \;\Rightarrow\; \hat{\mathbf{B}} = \mathbf{J}\hat{\mathbf{C}}^*.$$

The *augmented system* is $\hat{\boldsymbol{\Sigma}}$, where $\hat{\mathbf{A}} = \mathbf{A}$ and $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ are as defined above. This system has the property that its Hankel operator is symmetric. Therefore, using the same tools as above, we can show that

$$\hat{\mathcal{P}} = \hat{\mathcal{X}}\mathbf{J}^{-1},\ \hat{\mathcal{Q}} = \mathbf{J}\hat{\mathcal{X}} \ \Rightarrow\ \hat{\mathcal{X}}^2 = \hat{\mathcal{P}}\hat{\mathcal{Q}}.$$

A straightforward calculation shows that the gramians of the augmented and the original systems are related as follows:

$$\hat{\mathcal{P}} = \mathcal{P} + \mathbf{J}\mathcal{Q}\mathbf{J},\quad \hat{\mathcal{Q}} = \mathcal{Q} + \mathbf{J}^{-1}\mathcal{P}\mathbf{J}^{-1},\quad \hat{\mathcal{X}} = \mathbf{J}\hat{\mathcal{Q}} = \hat{\mathcal{P}}\mathbf{J}^{-1}.$$

The choice of symmetrizer

At this stage, the symmetrizer is any symmetric matrix that satisfies $\mathbf{AJ} = \mathbf{JA}^*$. For simplicity, let us assume that $\mathbf{A}$ is *diagonalizable* and has been transformed to the basis where $\mathbf{A}$ is diagonal. In this case, the symmetrizer $\mathbf{J}$ is an arbitrary diagonal matrix. The question arises of how to best choose the diagonal entries of $\mathbf{J}$.

The criterion that is chosen is that the *Hankel operator of the augmented system be close to that of the original system.* To address this issue, we make use of the variational characterization of balancing as developed in [268], [244].

Let $\mathbf{T}$ be a basis change in the state space. Then, as already mentioned, the gramians are transformed as follows: $\mathcal{P} \rightarrow \mathbf{T}\mathcal{P}\mathbf{T}^*$, $\mathcal{Q} \rightarrow \mathbf{T}^{-*}\mathcal{P}\mathbf{T}^{-1}$. Consider the following criterion:

$$\mathcal{J}(\mathbf{T}) = \text{trace}\,(\mathbf{T}\mathcal{P}\mathbf{T}^*) + \text{trace}\,(\mathbf{T}^{-*}\mathcal{Q}\mathbf{T}^{-1}).$$

For a fixed state space basis, the above quantity is equal to the sum of the eigenvalues of the reachability and of the observability gramians. The question that arises is to find the minimum of $\mathcal{J}(\mathbf{T})$ as a function of all nonsingular transformations $\mathbf{T}$. First notice that $\mathcal{J}$ can be expressed in terms of the positive definite matrix $\Phi = \mathbf{T}^*\mathbf{T}$:

$$\mathcal{J} = \text{trace}\,\left[\mathcal{P}\Phi + \mathcal{Q}\Phi^{-1}\right], \qquad \Phi = \mathbf{T}^*\mathbf{T} > \mathbf{0}.$$

The following result is due to Helmke and Moore.

Proposition 12.7. *The minimum of* $\mathcal{J}$ *is* $\mathcal{J}_* = \min_{\Phi>0} \mathcal{J} = 2\sum_{k=1}^{n} \sigma_k$, *and the minimizer is* $\Phi_* = \mathcal{P}^{-1/2}\left(\mathcal{P}^{1/2}\mathcal{Q}\mathcal{P}^{1/2}\right)^{1/2}\mathcal{P}^{-1/2}$.

We should remark that the first part of the above proposition can be proved using elementary means, as in section 7.1. It readily follows that with the eigenvalue decomposition $\mathcal{P}^{1/2}\mathcal{Q}\mathcal{P}^{1/2} = \mathbf{U}\Sigma^2\mathbf{U}^*$, a resulting balancing transformation is $\mathbf{T}_b = \Sigma^{1/2}\mathbf{U}^*\mathcal{P}^{-1/2}$. In other words, $\mathbf{T}_b\mathcal{P}\mathbf{T}_b^* = \mathbf{T}_b^{-*}\mathcal{Q}\mathbf{T}_b^{-1} = \mathbf{P}$. The transformation $\mathbf{T}_b$ is unique up to orthogonal similarity ($\mathbf{P}$ need not be diagonal). In our case, we wish to compute an appropriate symmetrizer. The criterion (the sum of the traces of the two gramians) for the augmented system is as follows:

$$\begin{aligned}\mathcal{J}(\mathbf{J}) &= \text{trace}\,(\mathcal{P} + \mathbf{J}\mathcal{Q}\mathbf{J}) + \text{trace}\,(\mathcal{Q} + \mathbf{J}^{-1}\mathcal{P}\mathbf{J}^{-1})\\ &= \text{trace}\,(\mathcal{P} + \mathcal{Q}) + \underbrace{\text{trace}\,(\mathbf{J}\mathcal{Q}\mathbf{J} + \mathbf{J}^{-1}\mathcal{P}\mathbf{J}^{-1})}_{\mathcal{J}_1(\mathbf{J})}.\end{aligned}$$

We compute the diagonal $\mathbf{J} = \text{diag}(j_1, \ldots, j_n)$ so that the above trace is minimized. The first summand does not depend on $\mathbf{J}$. The second is

$$\mathcal{J}_1(\mathbf{J}) = \sum_{i=1}^{n} \left[p_{ii} j_i^2 + q_{ii} \frac{1}{j_i^2} \right].$$

The minimum of $\mathcal{J}_1$ is achieved for $j_i^2 = \sqrt{\frac{q_{ii}}{p_{ii}}}$:

$$\min \mathcal{J}_1 = 2 \sum_{i=1}^{n} \sqrt{p_{ii} q_{ii}} \quad \Rightarrow \quad \min \mathcal{J} = \left(\sqrt{p_{ii}} + \sqrt{q_{ii}} \right)^2.$$

This should be compared with twice the sum of the trace of the two gramians, namely, $2 \sum_{i=1}^{n} (p_{ii} + q_{ii})$. The difference of the two traces is $\sum_{i=1}^{n} (\sqrt{p_{ii}} - \sqrt{q_{ii}})^2$.

The above computation was carried through under the assumption that $\mathbf{A}$ is diagonal. Let in particular

$$\mathbf{A} = \text{diag}(-\lambda_1, \ldots, -\lambda_n), \ \mathbf{B} = \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix}^*, \ \mathbf{C} = \begin{pmatrix} c_1 & \cdots & c_n \end{pmatrix},$$

$$\text{where } b_i \in \mathbb{R}^m, \ c_i \in \mathbb{R}^p.$$

In this representation, the diagonal entries of the two gramians are

$$p_{ii} = \frac{b_i^* b_i}{\lambda_i + \lambda_i^*}, \quad q_{ii} = \frac{c_i c_i^*}{\lambda_i + \lambda_i^*}.$$

Furthermore, by applying the state space transformation $\mathbf{T}$, which is diagonal with entries $t_{ii} = \sqrt{\frac{c_i c_i^*}{b_i^* b_i}}$, we can ensure that $p_{ii} = q_{ii}$. Therefore, if $\mathbf{A}$ is diagonalizable, there exists a symmetrizer that guarantees that the sum of the singular values of the augmented system is twice that of the original.

Computational efficiency

The method we have proposed appears to be computationally expensive. However, it seems to converge quite rapidly. In fact, our test examples indicate that for medium-scale problems ($n \approx 400$), it is already competitive with existing dense methods. Moreover, it can provide balanced realizations where most other methods fail.

The proposed implicit restarting approach involves a great deal of work associated with solving the required special Sylvester equations. However, the iterative method is based on adjoining residual corrections to the current approximate solutions to the cross gramian equations. We observe very rapid convergence in practice. Usually three major iterations are sufficient. Nevertheless, there is no proof of convergence, and this seems to be a general difficulty with projection approaches of this type (see, e.g., [176]).

Finally, we would like to point to (i) a preconditioned subspace method for solving the Lyapunov equation [175] and (ii) a recursive way of computing dominant singular subspaces [83]. Numerical experiments with iterative methods for solving the Lyapunov equation can be found in [73].

12.4 Iterative Smith-type methods for model reduction

Let $\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) \in \mathbb{R}^{(n+p)\times(n+m)}$ be the model to be reduced. Closely related to this system are the continuous-time Lyapunov equations (6.1). Under the assumptions of stability, reachability, and observability, the above equations have unique symmetric positive definite solutions $\mathcal{P}$, $\mathcal{Q} \in \mathbb{R}^{n\times n}$, called the *reachability, observability gramians*, respectively. The square roots of the eigenvalues of the product $\mathcal{P}\mathcal{Q}$ are the so-called Hankel singular values $\sigma_i(\mathbf{\Sigma})$ of the system $\mathbf{\Sigma}$ which are basis independent. As discussed in section 9.4, in many cases, the eigenvalues of $\mathcal{P}$, $\mathcal{Q}$ as well as the Hankel singular values $\sigma_i(\mathbf{\Sigma})$ decay very rapidly. This motivates the development of low rank approximate solution schemes and leads to model reduction by truncation.

Recall from section 7.3 the factorizations $\mathcal{P} = \mathbf{U}\mathbf{U}^*$ and $\mathcal{Q} = \mathbf{L}\mathbf{L}^*$ where $\mathbf{U}$ and $\mathbf{L}$ are the *square roots* of the gramians, respectively. Given the singular value decomposition (SVD) $\mathbf{U}^*\mathbf{L} = \mathbf{W}\Sigma\mathbf{V}^*$, according to (7.16) the maps $\mathbf{T}_1$ and $\mathbf{T}_{i1}$ satisfy $\mathbf{T}_{i1}\mathbf{T}_1 = \mathbf{I}_k$ and hence $\mathbf{T}_1\mathbf{T}_{i1}$ is an oblique projector. Then according to (7.17), a reduced model is obtained by letting $\mathbf{A}_1 = \mathbf{T}_1\mathbf{A}\mathbf{T}_{i1}$, $\mathbf{B}_1 = \mathbf{T}_1\mathbf{B}$, $\mathbf{C}_1 = \mathbf{C}\mathbf{T}_{i1}$. The reduced model $\mathbf{\Sigma}_1$ is balanced and asymptotically stable, and the $\mathcal{H}_\infty$-norm of the error system is bounded from above by twice the sum of the neglected singular values.

As mentioned, the above formulation for balanced truncation requires the knowledge of the square roots $\mathbf{U}$ and $\mathbf{L}$. The Bartels–Stewart method [43] as modified by Hammarling [164] is the standard direct method for the solution of Lyapunov equations of small to moderate size. Since this method requires the computation of a Schur decomposition, it is not appropriate for large-scale problems. Moreover, as previously explained, $\mathcal{P}$ and $\mathcal{Q}$ are often found to have *numerically* low rank compared to n. This *low rank phenomenon* leads to the idea of approximating the gramians with low rank solutions. Next we briefly summarize some approaches that aim at obtaining low rank approximants for the gramians $\mathcal{P}$ and $\mathcal{Q}$. In section 12.4.1, we mention the alternating direction implicit (ADI), Smith, and cyclic Smith(l) iterations that lead to an approximate solution to $\mathcal{P}$ and hence have the storage requirement of $\mathcal{O}(n^2)$. Section 12.4.2 summarizes the low rank versions of ADI and cyclic Smith(l) iterations which compute low rank approximants to $\mathbf{U}$ and $\mathbf{L}$, instead of $\mathcal{P}$ and $\mathcal{Q}$, and hence reduce the memory complexity to $\mathcal{O}(nk)$. For a detailed analysis of these algorithms, see [267], [263], [303] and references therein.

For large n and especially for slowly converging iterations, the number of columns of the solution of the low rank ADI iteration can easily exceed manageable memory capacity. Below, we introduce and analyze a modified low rank Smith method that essentially retains the convergence properties but overcomes the memory requirements. The development follows [157].

12.4.1 ADI, Smith, and cyclic Smith(l) iterations

In the following discussion, we focus on the approximate solution of a single Lyapunov equation $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is stable and diagonalizable, and $\mathbf{B} \in \mathbb{R}^{n\times m}$. This applies equally to the computation of low rank approximants to the observability gramian. In all of these methods, the idea is to transform, using spectral mappings of the type $\omega(\lambda) = \frac{\mu^*-\lambda}{\mu+\lambda}$, where $\mu \in \mathbb{C}_-$, a continuous-time Lyapunov equation into a discrete-time Lyapunov or Stein equation, for which the solution is obtained by an infinite summation. The parameter μ is the *shift*. The ADI method uses a different shift at each step and obtains a series of Stein equations. The Smith method uses a single shift parameter and hence is a special case of the ADI iteration. The Smith(l) method is also a special case of the ADI iteration, where l shifts are used in a cyclic manner.

The ADI iteration

The ADI iteration was first introduced by Peaceman and Rachford [263] for solving linear systems of equations arising from the discretization of elliptic boundary value problems. In general, the ADI iteration is applied to linear systems $\mathbf{My} = \mathbf{b}$, where $\mathbf{M}$ is symmetric positive definite and can be split into the sum of two symmetric positive definite matrices $\mathbf{M} = \mathbf{M}_1 + \mathbf{M}_2$ for which the following iteration is efficient:

$$\begin{aligned}\mathbf{y}_0 &= \mathbf{0},\\ (\mathbf{M}_1 + \mu_j\mathbf{I})\mathbf{y}_{j-1/2} &= \mathbf{b} - (\mathbf{M}_2 - \mu_j\mathbf{I})\mathbf{y}_{j-1},\\ (\mathbf{M}_2 + \eta_j\mathbf{I})\mathbf{y}_j &= \mathbf{b} - (\mathbf{M}_1 - \eta_j\mathbf{I})\mathbf{y}_{j-1/2} \text{ for } j = 1, 2, \ldots, J.\end{aligned}$$

The ADI shift parameters μ_j and η_j are determined from spectral bounds on $\mathbf{M}_1$ and $\mathbf{M}_2$ to increase the convergence rate. Application to the Lyapunov equation is obtained through the following iteration step:

$$\mathcal{P}_i^{\mathrm{A}} = (\mathbf{A}-\mu_i^*\mathbf{I})(\mathbf{A}+\mu_i\mathbf{I})^{-1}\mathcal{P}_{i-1}^{\mathrm{A}}[(\mathbf{A}-\mu_i^*\mathbf{I})(\mathbf{A}+\mu_i\mathbf{I})^{-1}]^* - 2\rho_i(\mathbf{A}+\mu_i\mathbf{I})^{-1}\mathbf{BB}^*(\mathbf{A}+\mu_i\mathbf{I})^{-*}, \tag{12.3}$$

where $\mathcal{P}_0^{\mathrm{A}} = \mathbf{0}$, the shift parameters $\{\mu_1, \mu_2, \mu_3, \ldots\}$ are elements of $\mathbb{C}_-$, and $\rho_i = \mathcal{R}e\,(\mu_i)$.

The spectral radius $\rho_{\mathrm{ADI}} = \rho(\prod_{i=1}^{l}(\mathbf{A} - \mu_i^*\mathbf{I})(\mathbf{A} + \mu_i\mathbf{I})^{-1})$ determines the rate of convergence where l is the number of shifts used. The minimization of ρ_{ADI} with respect to shift parameters μ_i is called the ADI min-max problem:

$$\{\mu_1, \mu_2, \ldots, \mu_l\} = \arg\min_{\mu_i \in \mathbb{C}_-}\max_{\lambda\in\sigma(A)} \frac{|\,(\lambda - \mu_1^*)\cdots(\lambda - \mu_l^*)\,|}{|\,(\lambda + \mu_1)\cdots(\lambda + \mu_l)\,|}.$$

See [104], [313], and [350] for contributions to the solution of the ADI min-max problem. It can be shown that if $\mathbf{A}$ is diagonalizable, the lth iterate satisfies the inequality

$$\|\mathcal{P} - \mathcal{P}_l^{\mathrm{A}}\|_F \leq \|\mathcal{X}\|^2\|\mathcal{X}^{-1}\|^2\,\rho_{\mathrm{ADI}}^2\,\|\mathcal{P}\|_F,$$

where $\mathcal{X}$ is the matrix of eigenvectors of $\mathbf{A}$.

Smith's method

For every real scalar $\mu < 0$, $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{BB}^* = \mathbf{0}$ is equivalent to the Stein equation

$$\mathcal{P} - \mathbf{A}_\mu\mathcal{P}\mathbf{A}_\mu^* = \mathbf{B}_\mu\mathbf{B}_\mu^*, \tag{12.4}$$

where

$$\mathbf{A}_\mu = (\mathbf{A} - \mu\mathbf{I})(\mathbf{A} + \mu\mathbf{I})^{-1}, \quad \mathbf{B}_\mu = \sqrt{-2\mu}\,(\mathbf{A} + \mu I)^{-1}\mathbf{B}. \tag{12.5}$$

Hence using the bilinear transformation $\omega = \frac{\mu-\lambda}{\mu+\lambda}$, the continuous-time problem has been transformed into a discrete-time one. Then according to Proposition 4.35, the Stein equation

(12.4) has the same solution as its continuous-time counterpart. The solution to the Stein equation is $\mathcal{P} = -2\mu \sum_{j=0}^{\infty} \mathbf{A}_{\mu}^{j}\mathbf{B}_{\mu}\mathbf{B}_{\mu}^{*}(\mathbf{A}_{\mu}^{*})^{j}$. The low rank Smith method amounts to truncating the series at an appropriate step, while the modified low rank Smith method updates a low rank SVD approximation to $\mathcal{P}$ by updating and truncating the SVD as each term $\mathbf{A}_{\mu}^{j}\mathbf{B}_{\mu}\mathbf{B}_{\mu}^{*}(\mathbf{A}_{\mu}^{*})^{j}$ is added in.

Since $\mathbf{A}$ is stable, $\rho(\mathbf{A}_{\mu}) < 1$ and the sequence $\{\mathcal{P}_i^S\}_{i=0}^{\infty}$ generated by the iteration $\mathcal{P}_0^S = \mathbf{0}$ and $\mathcal{P}_{i+1}^S = \mathbf{B}_{\mu}\mathbf{B}_{\mu}^{*} + \mathbf{A}_{\mu}\mathcal{P}_i^S\mathbf{A}_{\mu}^{*}$ converges to the solution $\mathcal{P}$. Thus the Smith iterates can be written as

$$\mathcal{P}_i^S = \sum_{j=1}^{i} \mathbf{A}_{\mu}^{j-1}\mathbf{B}_{\mu}\mathbf{B}_{\mu}^{*}(\mathbf{A}_{\mu}^{j-1})^{*}.$$

Notice that the Smith method is the same as the ADI method, when all the shifts are equal.

Formulas for multiple shifts μ_j and for complex shifts are fairly straightforward to derive [157].

The Smith(l) iteration

Penzl observed in [267] that in many cases the ADI method converges very slowly when only a single shift (Smith iteration) is used and a moderate increase of the number of shifts l accelerates the convergence. But he also observed that the speed of convergence is hardly improved by a further increase of l. These observations lead to the idea of cyclic Smith(l) iteration, which is a special case of ADI where l different shifts are used in a cyclic manner, i.e., $\mu_{i+jl} = \mu_i$ for $j = 1, 2, \ldots$. It is easy to show that the Smith(l) iterates are generated as follows:

$$\mathcal{P}_i^{Sl} = \sum_{j=1}^{i} \mathbf{A}_d^{i-1}\mathbf{T}(\mathbf{A}_d^{i-1})^{*}, \quad \text{where } \mathbf{A}_d = \prod_{i=1}^{l}(\mathbf{A} - \mu_i\mathbf{I})(\mathbf{A} + \mu_i\mathbf{I})^{-1}, \quad \mathbf{T} = \mathbf{P}_l^A.$$

As in Smith's method, $\mathcal{P} - \mathbf{A}_d\mathcal{P}\mathbf{A}_d^{*} = \mathbf{T}$ is equivalent to an appropriately defined continuous-time Lyapunov equation.

12.4.2 Low rank ADI and low rank Smith(l) iterations

Since the ADI, Smith, and Smith(l) iterations outlined in section 12.4.1 compute the solution $\mathcal{P}$ explicitly, the storage requirement is $O(n^2)$. But one should notice that in many cases the storage requirement is the limiting factor rather than the amount of computation. The remedy is the low rank methods. Instead of explicitly forming the solution $\mathcal{P}$, the low rank methods compute and store the low rank approximate square root factors reducing the storage requirement to $O(nr)$, where r is the numerical rank of $\mathcal{P}$.

The key idea in the low rank versions of Smith(l) and ADI methods is to write

$$\mathcal{P}_i^{Sl} = \mathbf{Z}_i^{Sl}(\mathbf{Z}_i^{Sl})^{*} \text{ and } \mathcal{P}_i^{A} = \mathbf{Z}_i^{A}(\mathbf{Z}_i^{A})^{*}.$$

The low rank ADI (LR-ADI) is based on (12.3). Combining the above relationships, (12.3) can be rewritten in terms of $\mathbf{Z}_i^A$ as

$$\mathbf{Z}_i^A = [\ (\mathbf{A} - \mu_i\mathbf{I})(\mathbf{A} + \mu_i\mathbf{I})^{-1}\mathbf{Z}_{i-1}^A, \quad \sqrt{-2\mu_i}\,(\mathbf{A} + \mu_i\mathbf{I})^{-1}\mathbf{B}\ ],$$

where $\mathbf{Z}_1^A = \sqrt{-2\mu_1}(\mathbf{A} + \mu_1 I)^{-1}\mathbf{B}$. When the number of shift parameters is limited, the cyclic low rank Smith method (LR-Smith(l)) is a more efficient alternative to the LR-ADI. It consists of two steps. First, the iterate $\mathbf{Z}_1^{Sl}$ is obtained by an l step low rank ADI iteration; i.e., the LR-Smith(l) is initialized by

$$\mathbf{Z}_1^{Sl} = \mathbf{B}_d = \mathbf{Z}_l^A, \tag{12.6}$$

where $\mathbf{B}_d$ is defined in (12.5). It then follows that the kth step LR-Smith(l) iterate is given by

$$\mathbf{Z}_k^{Sl} = [\mathbf{B}_d \;\; \mathbf{A}_d\mathbf{B}_d \;\; \mathbf{A}_d^2\mathbf{B}_d \;\; \cdots \;\; \mathbf{A}_d^{k-1}\mathbf{B}_d]. \tag{12.7}$$

One should notice that while a k-step LR-ADI iteration requires k matrix factorizations, a k-step LR-Smith(l) iteration computes only l matrix factorization. Moreover, if the shifts $\{\mu_1, \ldots, \mu_l\}$ are used in a cyclic manner, the cyclic LR-Smith(l) iteration is equivalent to the LR-ADI iteration.

We note that at the ith step $\mathbf{Z}_i^A$ and $\mathbf{Z}_i^{Sl}$ has $m \times i$ and $m \times l \times i$ columns, respectively. Hence, as discussed in [157], *when m is large and/or the convergence is slow*, i.e., $\rho(\mathbf{A}_d)$ is close to 1, the number of columns of $\mathbf{Z}_k^A$ and $\mathbf{Z}_k^{Sl}$ easily reaches an unmanageable level of memory requirements, and these two methods fail. In section 12.4.3, we introduce a modified LR-Smith(l) iteration to overcome this problem and to retain the low rank structure.

Next we present some convergence results for the LR-Smith(l) iteration without proofs. The proofs of these result can be found in [157].

Proposition 12.8. *Define $\mathbf{E}_{kp} = \mathcal{P} - \mathcal{P}_k$ and $\mathbf{E}_{kq} = \mathcal{Q} - \mathcal{Q}_k$, and let $\mathbf{A} = \mathcal{X}(\Lambda)\mathcal{X}^{-1}$ be the eigenvalue decomposition of $\mathbf{A}$. The k-step LR-Smith(l) iterates satisfy*

$$0 \le \operatorname{trace}(\mathbf{E}_{kp}) = \operatorname{trace}(\mathcal{P} - \mathcal{P}_k) \le K\, m\, l\, (\rho(\mathbf{A}_d))^{2k} \operatorname{trace}(\mathcal{P}),$$
$$0 \le \operatorname{trace}(\mathbf{E}_{kq}) = \operatorname{trace}(\mathcal{Q} - \mathcal{Q}_k) \le K\, p\, l\, (\rho(\mathbf{A}_d))^{2k} \operatorname{trace}(\mathcal{Q}),$$

where $K = \kappa(\mathcal{X})^2$, and $\kappa(\mathcal{X})$ denotes the condition number of $\mathcal{X}$.

Let σ_i and $\hat{\sigma}_i$ denote the Hankel singular values resulting from the full rank exact gramians and the low rank approximate gramian, respectively, i.e., $\sigma_i^2 = \lambda_i(\mathcal{P}\mathcal{Q})$ and $\hat{\sigma}_i^2 = \lambda_i(\mathcal{P}_k\mathcal{Q}_k)$. The following holds true:

Corollary 12.9. *Let σ_i and $\hat{\sigma}_i$ be defined as above. Define $\hat{n} = kl\min(m, p)$. Then,*

$$0 \le \sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{\hat{n}} \hat{\sigma}_i^2 \le K\, l\, (\rho(\mathbf{A}_d))^{2k} \Bigg(K \min(m, p)(\rho(\mathbf{A}_d))^{2k} \operatorname{trace}(\mathcal{P}) \operatorname{trace}(\mathcal{Q})$$
$$+ m \operatorname{trace}(\mathcal{P}) \sum_{i=0}^{k-1} \|\mathbf{C}_d\mathbf{A}_d^i\|_2^2 + p \operatorname{trace}(\mathcal{Q}) \sum_{i=0}^{k-1} \|\mathbf{A}_d^i\mathbf{B}_d\|_2^2 \Bigg),$$

where K is the square of the condition number of $\mathcal{X}$.

12.4.3 The modified LR-Smith(l) iteration

As discussed in section 12.4.2, the LR-Smith(l) iteration has the drawback that when applied to the usual Lyapunov equation $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$, the number of columns of the approximate square root increases by $m \times l$ at each step, where m is the number of inputs and l is the number of cyclic shifts applied. Hence in the case of large m and slow convergence, this causes storage problems. In what follows, we propose a modified LR-Smith(l) iteration so that the number of columns in the low rank square root factor does not increase unnecessarily at each step. The key idea of the proposed method is to compute the SVD of the iterate at each step and, given a tolerance τ, to replace the iterate with its best low rank approximation. However, we do not recompute the SVD; instead we update it after each step to include the new information and then truncate given the tolerance τ.

The proposed algorithm

As stated in section 12.4.1, the continuous-time Lyapunov equation is equivalent to the Stein equation

$$\mathcal{P} - \mathbf{A}_d\mathcal{P}\mathbf{A}_d^* + \mathbf{B}_d\mathbf{B}_d^* = \mathbf{0}, \tag{12.8}$$

where $\mathbf{A}_d$ and $\mathbf{B}_d$ are defined in (12.5) and (12.6), respectively. Let $\mathbf{Z}_k^{Sl}$ be the kth LR-Smith(l) iterate as defined in (12.7). Then the approximate low rank gramian at the kth step is

$$\mathcal{P}_k^{Sl} = \mathbf{Z}_k^{Sl}(\mathbf{Z}_k^{Sl})^* = \sum_{i=0}^{k-1} \mathbf{A}_d^i\mathbf{B}_d\mathbf{B}_d^*(\mathbf{A}_d^i)^*. \tag{12.9}$$

Let the short SVD (S-SVD) of $\mathbf{Z}_k^{Sl}$ be $\mathbf{Z}_k^{Sl} = \mathbf{V}\Sigma\mathbf{W}^*$, where $\mathbf{V} \in \mathbb{R}^{n\times(mlk)}$, $\Sigma \in \mathbb{R}^{(mlk)\times(mlk)}$, and $\mathbf{W} \in \mathbb{R}^{(mlk)\times(mlk)}$. Consequently, the S-SVD of $\mathcal{P}_k^{Sl}$ is given by $\mathcal{P}_k^{Sl} = \mathbf{V}\Sigma^2\mathbf{V}^*$, and it is enough to store $\mathbf{V}$ and Σ only. In other words, $\tilde{\mathbf{Z}}_k = \mathbf{V}\Sigma$ is also a low rank square root factor for $\mathcal{P}_k^{Sl}$.

Let $\tau > 0$ be a prespecified tolerance value. Assume that until the kth step of the algorithm all the iterates $\mathbf{Z}_i^{Sl}$ satisfy $\frac{\sigma_{\min}(\mathbf{Z}_i^{Sl})}{\sigma_{\max}(\mathbf{Z}_i^{Sl})} > \tau$ for $i = 1, \ldots, k$. At the $(k+1)$st step, the approximants $\mathbf{Z}_{k+1}^{Sl}$ and $\mathcal{P}_{k+1}^{Sl}$ are readily computed as

$$\mathbf{Z}_{k+1}^{Sl} = [\mathbf{Z}_k^{Sl} \;\; \mathbf{A}_d^k\mathbf{B}_d] \text{ and } \mathcal{P}_{k+1}^{Sl} = \mathcal{P}_k^{Sl} + \mathbf{A}_d^k\mathbf{B}_d\mathbf{B}_d^*(\mathbf{A}_d^k)^*.$$

Define $\mathbf{B}_{(k)} = \mathbf{A}_d^k\mathbf{B}_d$ and decompose

$$\mathbf{B}_{(k)} = \mathbf{V}\Gamma + \hat{\mathbf{V}}\Theta,$$

where $\Gamma \in \mathbb{R}^{(mlk)\times(ml)}$, $\Theta \in \mathbb{R}^{(ml)\times(ml)}$, $\mathbf{V}^*\hat{\mathbf{V}} = \mathbf{0}$, and $\hat{\mathbf{V}}^*\hat{\mathbf{V}} = \mathbf{I}_{ml}$. In view of the above decomposition, we define the matrix

$$\hat{\mathbf{Z}}_{k+1} = [\mathbf{V} \;\; \hat{\mathbf{V}}]\underbrace{\begin{bmatrix} \Sigma & \Gamma \\ \mathbf{0} & \Theta \end{bmatrix}}_{\hat{\mathbf{S}}}.$$

Let $\hat{\mathbf{S}}$ have the following SVD: $\hat{\mathbf{S}} = \mathbf{T}\hat{\Sigma}\mathbf{Y}^*$. We note that since $\hat{\mathbf{S}} \in \mathbb{R}^{(k+1)ml \times (k+1)ml}$, taking the SVD of $\hat{\mathbf{S}}$ is inexpensive. Then it readily follows that $\tilde{\mathbf{Z}}_{k+1}$ is given by

$$\tilde{\mathbf{Z}}_{k+1} = \tilde{\mathbf{V}}\hat{\Sigma}, \quad \text{where } \tilde{\mathbf{V}} = [\mathbf{V}, \ \hat{\mathbf{V}}]\mathbf{T},$$

and $\tilde{\mathbf{V}} \in \mathbb{R}^{n \times (k+1)ml}$, $\hat{\Sigma} \in \mathbb{R}^{(k+1)ml \times (k+1)ml}$. Again one should notice that $\tilde{\mathbf{Z}}_{k+1}$ is simply obtained from $\tilde{\mathbf{Z}}_k$, which is already available, and from the SVD of $\hat{\mathbf{S}}$, which is easy to compute. Next, we partition $\hat{\Sigma}$ and $\tilde{\mathbf{V}}$ conformally:

$$\tilde{\mathbf{Z}}_{k+1} = [\tilde{\mathbf{V}}_1 \ \ \tilde{\mathbf{V}}_2] \begin{bmatrix} \hat{\Sigma}_1 & \\ & \hat{\Sigma}_2 \end{bmatrix}, \quad \text{where } \frac{\hat{\Sigma}_2(1,1)}{\hat{\Sigma}_1(1,1)} < \tau. \tag{12.10}$$

The $(k+1)$st low rank square root factor is approximated by

$$\tilde{\mathbf{Z}}_{k+1} \approx \tilde{\mathbf{V}}_1\hat{\Sigma}_1. \tag{12.11}$$

Hence the $(k+1)$st step low rank gramian is computed as $\tilde{\mathcal{P}}_{k+1} = \tilde{\mathbf{Z}}_{k+1}(\tilde{\mathbf{Z}}_{k+1})^*$; this means that we simply ignore the singular values which are less than the given tolerance. Hence, from the kth to the $(k+1)$st step, the number of columns of $\tilde{\mathbf{Z}}_{k+1}$ generally does not increase. An increase occurs only if more than r singular values of $\tilde{\mathbf{Z}}_{k+1}$ fall above the tolerance $\tau\sigma_1$. In any case, there can be at most ml additional columns added at any step which is the same as the original LR-Smith(l) iteration discussed in section 12.4.2.

Convergence properties of the modified LR-Smith(l) iteration

In this section, we present some convergence results for the k-step approximate solutions $\tilde{\mathcal{P}}_k$ and $\tilde{\mathcal{Q}}_k$ which are, respectively, the modified LR-Smith(l) solutions to the two Lyapunov equations $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$, $\mathbf{A}^*\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{C}^*\mathbf{C} = \mathbf{0}$, where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $\mathbf{C} \in \mathbb{R}^{p\times n}$. For the proofs, see [157]. We introduce the set

$$\mathcal{I}_{\mathcal{P}} = \{i : \text{such that in (12.10) } \hat{\Sigma}_2 \neq \mathbf{0} \text{ for } \tilde{\mathbf{Z}}_i, \ i = 1, 2, \ldots, k\}.$$

In other words, this is the set of indices for which some columns have been eliminated from the ith approximant. Then for each $i \in \mathcal{I}_{\mathcal{P}}$, we denote by $n_i^{\mathcal{P}}$ the number of neglected singular values. $\mathcal{I}_{\mathcal{Q}}$ and $n_i^{\mathcal{Q}}$ are defined similarly.

Proposition 12.10. *Let $\tilde{\mathbf{Z}}_k$ be the kth step modified LR-Smith(l) iterate corresponding to* (12.8), *and let $\tilde{\mathcal{P}}_k = \tilde{\mathbf{Z}}_k(\tilde{\mathbf{Z}}_k)^*$, $\mathcal{P}_k$ be the kth step LR-Smith(l) iterates given by* (12.9). *Define $\Delta_{kp} = \mathcal{P} - \tilde{\mathcal{P}}_k$. It follows that $\tilde{\mathcal{P}}_k$ and Δ_{kp} satisfy*

$$\|\Delta_{kp}\| = \|\mathcal{P}_k^{SL} - \tilde{\mathcal{P}}_k\| \leq \tau^2 \sum_{i\in\mathcal{I}_{\mathcal{P}}} (\sigma_{max}(\tilde{\mathbf{Z}}_i))^2, \tag{12.12}$$

$$0 \leq \operatorname{trace}(\Delta_{kp}) = \operatorname{trace}(\mathcal{P}_k^{SL} - \tilde{\mathcal{P}}_k) \leq \tau^2 \sum_{i\in\mathcal{I}_{\mathcal{P}}} n_i^{\mathcal{P}} (\sigma_{max}(\tilde{\mathbf{Z}}_i))^2, \tag{12.13}$$

where τ is the tolerance value of the modified LR-Smith(l) algorithm.

In view of Proposition 12.8, the following result holds.

Proposition 12.11. *Let $\tilde{\mathbf{Z}}_k$ be the kth step modified LR-Smith(l) iterate corresponding to (12.8), $\tilde{\mathcal{P}}_k = \tilde{\mathbf{Z}}_k(\tilde{\mathbf{Z}}_k)^*$, $\mathbf{E}_{kp} = \mathcal{P} - \tilde{\mathcal{P}}_k$. Let $\mathbf{A} = \mathcal{X}\Lambda\mathcal{X}^{-1}$ be the eigenvalue decomposition of $\mathbf{A}$. The k-step modified LR-Smith(l) iterates satisfy*

$$0 \leq \operatorname{trace}(\mathbf{E}_{kp}) = \operatorname{trace}(\mathcal{P} - \tilde{\mathcal{P}}_k) \leq K\, m\, l\, \rho(\mathbf{A}_d)^{2k} \operatorname{trace}(\mathcal{P}) + \tau^2 \sum_{i \in \mathcal{I}_\mathcal{P}} n_i^\mathcal{P}\ (\sigma_{max}(\tilde{\mathbf{Z}}_i))^2,$$

where K, as before, is the square of the condition number of $\mathcal{X}$.

We note that the bounds for the traces of the errors in Propositions 12.8 and 12.11 differ only by an order of $\mathcal{O}(\tau^2)$. The next result introduces a convergence result for the computed Hankel singular values similar to the one in Corollary 12.9.

Corollary 12.12. *Let σ_i and $\tilde{\sigma}_i$ denote Hankel singular values resulting from the full rank exact gramians $\mathcal{P}$ and $\mathcal{Q}$ and from the modified LR-Smith(l) approximants $\tilde{\mathcal{P}}_k$ and $\tilde{\mathcal{Q}}_k$, respectively: ${\sigma_i}^2 = \lambda_i(\mathcal{P}\mathcal{Q})$ and $\tilde{\sigma}_i^2 = \lambda_i(\tilde{\mathcal{P}}_k\tilde{\mathcal{Q}}_k)$. Define $\hat{n} = kl \min(m, p)$. Then,*

$$\begin{aligned}0 \leq \sum_{i=1}^{n} \sigma_i^2 - \sum_{i=1}^{\hat{n}} \tilde{\sigma}_i^2 \leq K\, l\, (\rho(\mathbf{A}_d))^{2k} \Bigg[& K \min(m, p)(\rho(\mathbf{A}_d))^{2k} \operatorname{trace}(\mathcal{P}) \operatorname{trace}(\mathcal{Q}) \\ & + m \ \operatorname{trace}(\mathcal{P}) \sum_{i=0}^{k-1} \|\mathbf{C}_d\mathbf{A}_d^i\|_2^2 + p \ \operatorname{trace}(\mathcal{Q}) \sum_{i=0}^{k-1} \|\mathbf{A}_d^i\mathbf{B}_d\|_2^2 \\ & + \tau_\mathcal{P}^2 \|\mathcal{Q}_k^{SL}\| \sum_{i \in \mathcal{I}_\mathcal{P}} n_i^\mathcal{P}\ (\sigma_{max}(\tilde{\mathbf{Z}}_i))^2 + \tau_\mathcal{Q}^2 \|\mathcal{P}_k^{SL}\| \sum_{i \in \mathcal{I}_\mathcal{Q}} n_i^\mathcal{Q}\ (\sigma_{max}(\tilde{\mathbf{Y}}_i))^2 \\ & + \tau_\mathcal{P}^2\tau_\mathcal{Q}^2 \sum_{i \in \mathcal{I}_\mathcal{P}} (\sigma_{max}(\tilde{\mathbf{Z}}_i))^2 \sum_{i \in \mathcal{I}_\mathcal{Q}} n_i^\mathcal{Q}\ (\sigma_{max}(\tilde{\mathbf{Y}}_i))^2 \Bigg],\end{aligned}$$

where $\tau_\mathcal{P}$, $\tau_\mathcal{Q}$ are the given tolerance values, and $K = \kappa(\mathcal{X})^2$.

Once again the bounds for the error in the computed Hankel singular values in Corollaries 12.9 and 12.12 differ by only the summation of terms, which are $\mathcal{O}(\tau_\mathcal{P}^2)$, $\mathcal{O}(\tau_\mathcal{Q}^2)$, and $\mathcal{O}(\tau_\mathcal{P}^2\tau_\mathcal{P}^2)$.

12.4.4 A discussion on the approximately balanced reduced system

Let $\boldsymbol{\Sigma}_k = \left(\begin{array}{c|c} \mathbf{A}_k & \mathbf{B}_k \\ \hline \mathbf{C}_k & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{W}_k^*\mathbf{A}\mathbf{V}_k & \mathbf{W}_k^*\mathbf{B} \\ \hline \mathbf{C}\mathbf{V}_k & \mathbf{D} \end{array}\right)$ be the kth-order reduced system obtained by exact balanced truncation. Similarly, let $\hat{\boldsymbol{\Sigma}}_k = \left(\begin{array}{c|c} \hat{\mathbf{A}}_k & \hat{\mathbf{B}}_k \\ \hline \hat{\mathbf{C}}_k & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} \hat{\mathbf{W}}_k^*\mathbf{A}\hat{\mathbf{V}}_k & \hat{\mathbf{W}}_k^*\mathbf{B} \\ \hline \mathbf{C}\hat{\mathbf{V}}_k & \mathbf{D} \end{array}\right)$ be the kth-order reduced system obtained by approximate balanced truncation, where the approximate low rank square roots $\mathbf{Z}_r^{Sl}$ and $\mathbf{Y}_r^{Sl}$ are used instead of the exact square roots in computing $\hat{\mathbf{W}}_k$ and $\hat{\mathbf{V}}_k$. The following equation is easily derived:

$$\hat{\mathbf{A}}_k\hat{\mathbf{S}}_k + \hat{\mathbf{S}}_k\hat{\mathbf{A}}_k^* + \hat{\mathbf{B}}_k\hat{\mathbf{B}}_k^* = \hat{\mathbf{W}}_k^*(\mathbf{A}\Delta + \Delta\mathbf{A}^*)\hat{\mathbf{W}}_k,$$

where Δ is the error in $\mathcal{P}$, i.e., $\Delta = \mathcal{P} - \mathcal{P}_r^{SI}$ and $\hat{\mathbf{S}}_k$ is the diagonal matrix with the approximate Hankel singular values as diagonal entries. Stability of the reduced system is not always guaranteed, but if the approximate singular values are close enough to the original ones, then it is.

Next we examine how close $\boldsymbol{\Sigma}_k$ is to $\hat{\boldsymbol{\Sigma}}_k$. Toward this goal, define $\Delta_{\mathbf{V}} = \mathbf{V}_k - \hat{\mathbf{V}}_k$ and $\Delta_{\mathbf{W}} = \mathbf{W}_k - \hat{\mathbf{W}}_k$ and let $\|\Delta_{\mathbf{V}}\| \leq \tau$ and $\|\Delta_{\mathbf{W}}\| \leq \tau$, where τ is a small number; in other words, we assume that $\hat{\mathbf{V}}_k$ and $\hat{\mathbf{W}}_k$ are close to $\mathbf{V}_k$ and $\mathbf{W}_k$. It can be shown that $\Delta_{\mathbf{A}} = \mathbf{A}_k - \hat{\mathbf{A}}_k$, $\Delta_{\mathbf{B}} = \mathbf{B}_k - \hat{\mathbf{B}}_k$, $\Delta_{\mathbf{C}} = \mathbf{C}_k - \hat{\mathbf{C}}_k$ satisfy

$$\|\Delta_{\mathbf{A}}\| \leq \tau\|\mathbf{A}\|(\|\mathbf{W}_k\| + \|\mathbf{V}_k\|) + \tau^2\|\mathbf{A}\|, \quad \|\Delta_{\mathbf{B}}\| \leq \tau\|\mathbf{B}_k\|, \quad \text{and} \quad \|\Delta_{\mathbf{C}}\| \leq \tau\|\mathbf{C}_k\|.$$

Thus, the closer $\hat{\mathbf{V}}_k$, $\hat{\mathbf{W}}_k$ are to $\mathbf{V}_k$, $\mathbf{W}_k$, that is, the smaller τ, the closer $\hat{\boldsymbol{\Sigma}}_k$ will be to $\boldsymbol{\Sigma}_k$.

12.4.5 Relation of the Smith method to the trapezoidal rule

We consider the system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, $\mathbf{y} = \mathbf{c}\mathbf{x}$ and recall that for $\mathbf{u}(\tau) = \delta(\tau)$ and $\mathbf{x}(0) = \mathbf{0}$, $\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{B}$. Moreover, the solution of the Lyapunov equation $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$ is

$$\mathcal{P} = \int_0^\infty \mathbf{x}(\tau)\mathbf{x}(\tau)^* \mathrm{d}\tau = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^*t}\mathrm{d}\tau.$$

The Smith method derives from the equivalence of the Lyapunov equation to a Stein equation (12.4) where (12.5) holds.

There is a direct relation due to Sorensen between a Smith method and the well-known trapezoidal rule (Crank–Nicholson) method for integrating the differential equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$. We consider $\mathbf{x}_j \approx \mathbf{x}(jh)$, $j = 0, 1, \ldots$, to be computed by the trapezoidal rule, and we recall that $\mathbf{x}_0 = \mathbf{0}$ and that $\mathbf{u}(\tau) = \delta(\tau)$. For any positive step size h,

$$\mathbf{x}((j+1)h) - \mathbf{x}(jh) = \int_{jh}^{(j+1)h} \dot{\mathbf{x}}(\tau)\mathrm{d}\tau = \int_{jh}^{(j+1)h} \mathbf{A}\mathbf{x}(\tau)\mathrm{d}\tau + \mathbf{B}\int_{jh}^{(j+1)h} \delta(\tau)\mathrm{d}\tau$$

for each interval $[jh, (j+1)h)$. A straightforward application of the trapezoidal rule along with the conditions $\mathbf{x}(0) = \mathbf{0}$ and $\int_0^h \delta(\tau)\mathrm{d}\tau = 1$ gives $\mathbf{x}_1 = (\mathbf{I} - \frac{h}{2}\mathbf{A})^{-1}\mathbf{B}$ and $\mathbf{x}_{j+1} = (\mathbf{I} - \frac{h}{2}\mathbf{A})^{-1}(\mathbf{I} + \frac{h}{2}\mathbf{A})\mathbf{x}_j$ for the remaining intervals. Hence, $\mathbf{x}_{j+1} = \left[(\mathbf{I} - \frac{h}{2}\mathbf{A})^{-1}(\mathbf{I} + \frac{h}{2}\mathbf{A})\right]^j (\mathbf{I} - \frac{h}{2}\mathbf{A})^{-1}\mathbf{B}$, for $j = 0, 1, 2, \ldots$. Now, putting $\mu = \frac{2}{h}$ gives

$$\mathcal{P} = \int_0^\infty \mathbf{x}(\tau)\mathbf{x}(\tau)^*\mathrm{d}\tau = h\sum_{j=0}^{\infty} \mathbf{x}_{j+1}\mathbf{x}_{j+1}^* = 2\mu\sum_{j=0}^{\infty} \mathbf{A}_\mu^j \mathbf{B}_\mu \mathbf{B}_\mu^* (\mathbf{A}_\mu^*)^j,$$

where $\mathbf{A}_\mu = (\mathbf{A} - \mu\mathbf{I})^{-1}(\mathbf{A} + \mu\mathbf{I})$ and $\mathbf{B}_\mu = (\mathbf{A} - \mu\mathbf{I})^{-1}\mathbf{B}$ as before, since $h\mu^2 = \frac{2}{\mu}\mu^2 = 2\mu$.

This establishes the direct connection between the low rank Smith method (in simplest form) and the technique of constructing an approximate trajectory $\{\mathbf{x}_j\}$ via the trapezoidal rule and then developing an approximation to $\mathcal{P}$ directly by applying the simplest quadrature rule $\mathcal{P} = h\sum_{j=0}^{\infty} \mathbf{x}_{j+1}\mathbf{x}_{j+1}^*$. The result is easily generalized to multiple inputs by linearity and superposition. It is interesting to note that quadrature rules have been proposed previously [155] in this context for directly approximating the integral formula $\mathcal{P} = \int_0^\infty e^{\mathbf{A}t}\mathbf{B}\mathbf{B}^* e^{\mathbf{A}^*t}d\tau$ for the gramian.

12.5 Chapter summary

The two basic families of approximation methods are the SVD-based and the Krylov-based methods. As mentioned, however, each has its own set of advantages and disadvantages. The former preserve important properties like stability and have a global error bound; in return, they can be applied to systems of low to moderate dimension. The latter are moment matching methods which have iterative implementations. Therefore, they can be applied to large-scale systems. But since moment matching methods are local in frequency, stability is in general, not guaranteed.

In the preceding chapter, we established connections between these two families of approximation methods. First, it was shown that reachability and generalized reachability matrices can be obtained as solutions of Sylvester equations, which are a general form of Lyapunov equations. The same holds for observability and generalized observability matrices. Second, it was argued that by appropriate choice of weightings, weighted balanced truncation methods can be reduced to Krylov methods. Third, the *least squares* approximation method was introduced and was shown to combine some of the attributes of each of the basic approximation methods. Finally, two iterative methods for solving Lyapunov equations have been presented, leading to approximation methods which combine the best attributes of SVD-based and Krylov-based methods.

Chapter 13

Case Studies

The results presented in earlier chapters will now be illustrated by means of several case studies. The first section deals with the approximation of three images, which can be considered as static systems, using the SVD. The next three sections deal with dynamical systems. Section 13.2 examines the reduction of five systems of relatively low order ($n = 48$ to $n = 348$); the reduced models obtained using a variety of methods are compared with each other. Section 13.3 investigates the reduction of models of two modules of the ISS, namely, module 1R with $n = 270$ states and module 12A with $n = 1412$ states. A variety of model reduction methods are compared with the *modal gain factor* method, which is a modified modal approximation method commonly used for the reduction of structural systems. Finally, in section 13.4, iterative Smith-type methods are applied to the CD player model used in section 13.2 and to three further models with dimension $n = 1006$, $n = 3468$, and $n = 20{,}736$ states.

13.1 Approximation of images

Here we apply the SVD discussed in section 3.2 to the approximation of three black-and-white images: a clown from MATLAB and the pictures of galaxies NGC6782 and NGC6822 taken by the Hubble telescope.

A picture can be described as a rectangular matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, obtained by a two-dimensional sampling and quantization of the intensity at each of the pixels. Coding of the picture consists of reducing the number of bits which are necessary for each pixel, without, however, compromising the quality of the reconstructed picture from the coded data.

The compression is as follows. The original picture of size $n \times m$ requires storage of size $n \times m$. The SVD compressed picture of rank r requires storage of size $(n + m + 1) \cdot r$. Thus the compression ratio in storage is $\frac{(n+m+1)\cdot r}{n\cdot m}$; assuming that $n \approx m$, this ratio becomes approximately $\frac{2\cdot r}{n}$. If k bits of information are required per pixel, then the reduction ratio is still the same. The SDD, which is an approximate SVD method discussed briefly in section 3.2.6, achieves a compression ratio $\frac{r\cdot(3n+3m+k)}{n\cdot m\cdot k}$ which, for large $m \approx n$, yields the compression ratio $\frac{3\cdot r}{n\cdot k}$.

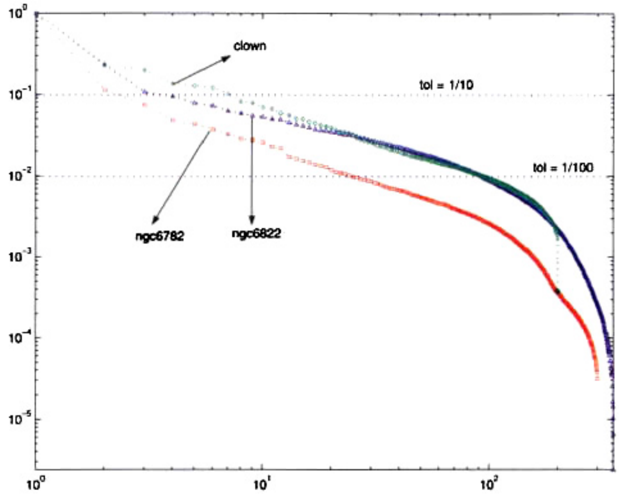


Figure 13.1. *The normalized singular values of the clown and galaxies NGC6822 and NGC6782.*

First we compute and plot the singular values of the three images (see Figure 13.1). These plots provide the *trade-off between achievable accuracy and desired complexity*. For instance if we wish to approximate the images with an accuracy of 1%, the plots show that the resulting complexity of the `ngc6782` galaxy approximant is about 30, while that of the approximants of the other two images is about 100; thus the latter two images are more difficult to approximate than the former. Figure 13.2 shows the original pictures together with 10% and 5% approximants.

13.2 Model reduction algorithms applied to low-order systems

In this section, we apply both SVD-based and Krylov-based model reduction algorithms to five different dynamical systems: a building model, a heat transfer model, a model of a CD player, a model of a clamped beam, and a low-pass Butterworth filter. The order k of the reduced models will be determined by a given *tolerance*, namely, $\tau = 1 \times 10^{-3}$. Given the Hankel singular values σ_i, $i = 1, \ldots, n$, of each system, k will be the smallest positive integer such that

$$\sigma_{k+1} < \tau\sigma_1.$$

The following table shows the order of the original system n, the number of inputs m, and outputs p, as well as the order of the reduced system, k:

	n	m	p	k
Building model	48	1	1	31
Heat model	197	2	2	5
CD player model	120	1	1	12
Clamped beam model	348	1	1	13
Butterworth filter	100	1	1	35

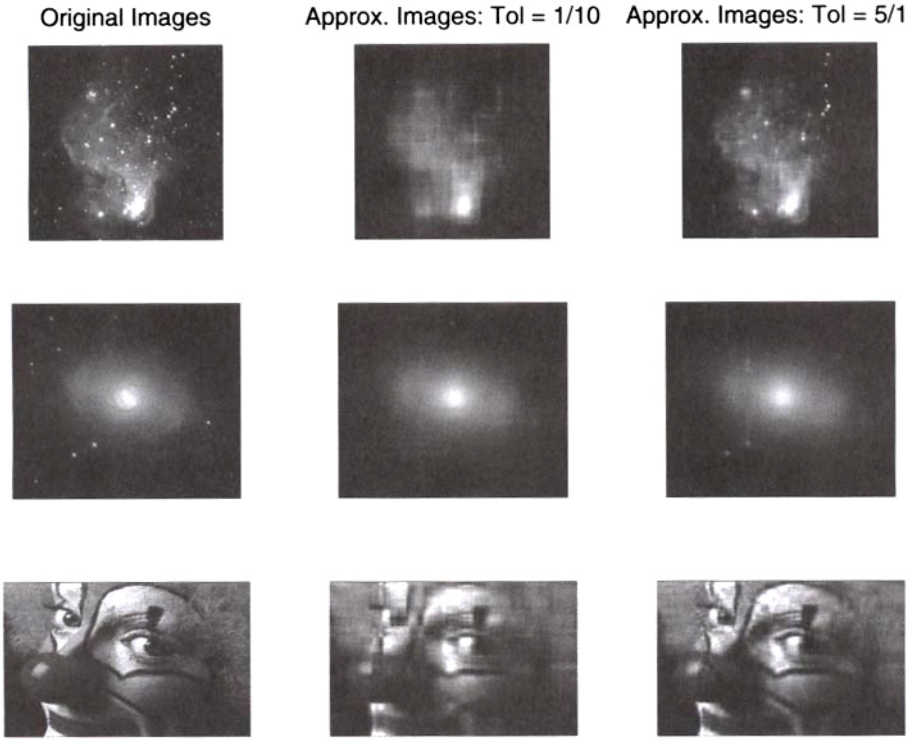


Figure 13.2. *The original images together with the 10% and 5% approximants.*

The Hankel singular values of each model are depicted in Figure 13.3, top. For comparison, the highest Hankel singular values are normalized to 1. The bottom of Figure 13.3 shows the relative degree reduction $\frac{k}{n}$ versus the error tolerance $\tau = \frac{\sigma_k}{\sigma_1}$; this shows how much the order can be reduced for a given tolerance: the lower the curve, the easier it is to approximate the system. It thus follows that among all models, for any fixed tolerance smaller than 1.0×10^{-1}, the building model is the hardest to approximate.

It should be stressed that the tolerance is a *user-specified* quantity that determines the trade-off between accuracy and complexity. Once specified, it determines completely the reduced-order models for SVD-based methods. The eigenvalues (poles) of the reduced system, for instance, are placed automatically. On the other hand, to apply the Krylov-based methods, one has to choose the interpolation points and their multiplicities.

The model reduction approaches used are three SVD-based methods, namely, balanced truncation, optimal Hankel-norm approximation, and singular perturbation approximation; three Krylov-based methods, namely, the Arnoldi procedure, the Lanczos procedure, and the rational Krylov procedure; and two SVD–Krylov-based methods, namely, the least squares method described in section 12.1.3 and the iterative method described in section 12.3, which will be referred to as *approximate balanced reduction.*

In the subsections that follow, each system is briefly described and the amplitude Bode plots (more precisely, the largest singular value of the frequency responses) of the full- and reduced-order models as well as of the corresponding error systems are plotted. Moreover,

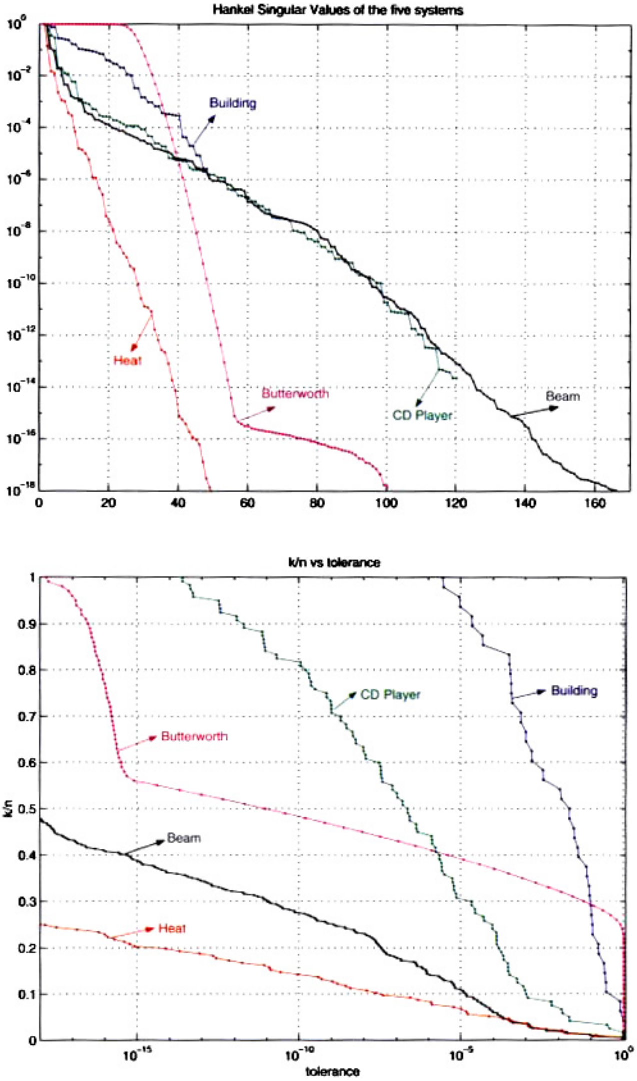


Figure 13.3. *Top pane: normalized Hankel singular values of the heat model, the Butterworth filter, the clamped beam model, the building example, the CD player model. Bottom pane: relative degree reduction $\frac{k}{n}$ versus error tolerance $\frac{\sigma_k}{\sigma_1}$.*

the relative $\mathcal{H}_\infty$- and $\mathcal{H}_2$-norms of the error systems are tabulated. Since the approximants obtained by balanced reduction and approximate balanced reduction are (almost) indistinguishable for all but the heat model, we plot and tabulate results for one of these approximants.

13.2.1 Building model

The full-order model is that of a building (Los Angeles University Hospital) with 8 floors, each of which has 3 degrees of freedom, namely, displacements in the x and y directions,

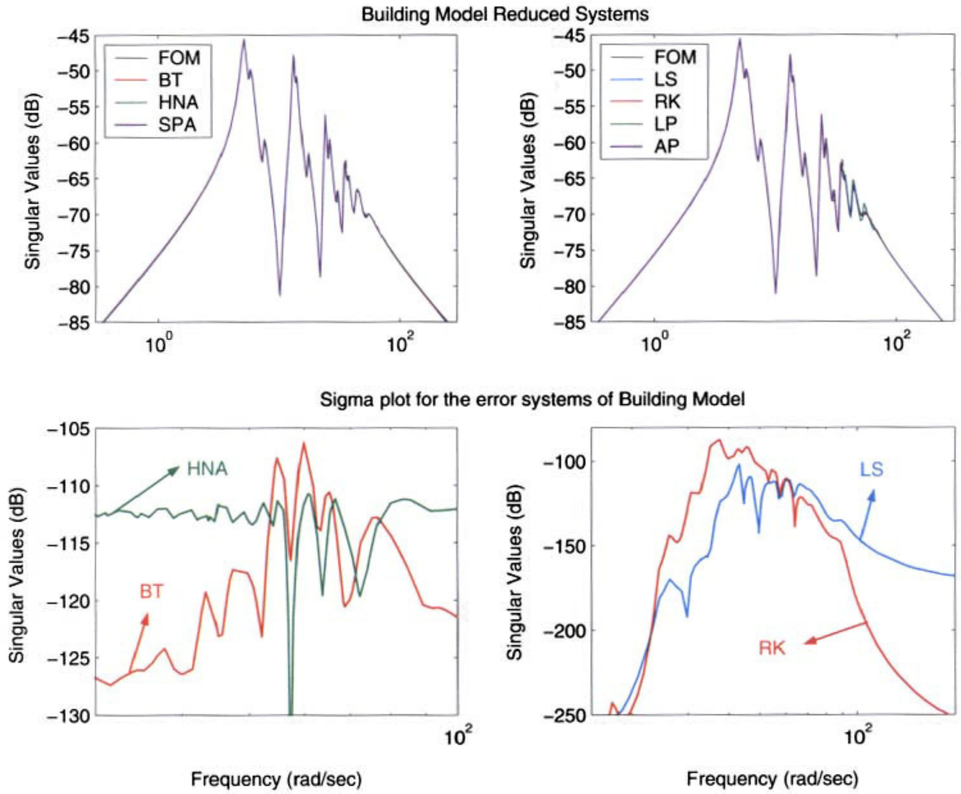


Figure 13.4. *Building model. Top: Bode plot of original and reduced systems. Bottom: error Bode plots.*

and rotation. The 24 resulting variables $\mathbf{q}_i$, $i = 1, \ldots, 24$, and their derivatives $\dot{\mathbf{q}}_i$, $i = 1, \ldots, 24$, satisfy a vector second-order differential equation of the form $\mathbf{M}\ddot{\mathbf{q}}(t) + \mathbf{D}\dot{\mathbf{q}}(t) + \mathbf{K}\mathbf{q}(t) = \mathbf{v}\,\mathbf{u}(t)$, where $\mathbf{u}(t)$ is the input. This equation can be written in state space form by defining the state $\mathbf{x}^* = [\ \mathbf{q}^*\ \ \dot{\mathbf{q}}^*\ ]^* \in \mathbb{R}^{48}$:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{D} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} \mathbf{0} \\ \mathbf{M}^{-1}\mathbf{v} \end{bmatrix} \mathbf{u}(t).$$

We will assume that the input affects the first coordinate $\mathbf{q}_1(t)$, that is, $\mathbf{v} = [1\ \ 0\ \ \cdots\ \ 0]^*$, and the output is $\mathbf{y}(t) = \dot{\mathbf{q}}_1(t) = \mathbf{x}_{25}(t)$.

The state space model has order 48 and is SISO. For this example, the pole closest to the imaginary axis has real part equal to -2.62×10^{-1}. We approximate the system with a model of order 31. The largest singular value $\sigma_{\max}$ of the frequency response of the reduced-order and of the error systems is shown in Figure 13.4, top and bottom, respectively. Since the expansion of transfer function $\mathbf{H}(s)$ around ∞ results in unstable reduced systems for Arnoldi and Lanczos, the shifted version of these two methods with $s_0 = 1$ was used instead. The effect of choosing s_0 as a low frequency point is observed in the right panels of the figure, as Arnoldi and Lanczos yield good approximants for the low frequency range.

The same holds for the rational Krylov method, since the interpolation points were chosen in the 1 to 100 rad/sec region. When compared to SVD-based methods, the moments matching–based methods are better for low frequencies. Among the SVD-based methods, singular perturbation, balanced reduction methods, are the best for low, high, frequencies, respectively. When we consider the whole frequency range, balancing and singular perturbation are closer to the original model. But in terms of the relative $\mathcal{H}_\infty$ error norm, Hankel-norm approximation is the best. As expected, rational Krylov, Arnoldi, and Lanczos result in high relative errors since they are local. The error norms are displayed next.

	$\mathcal{H}_\infty$-norm of error	$\mathcal{H}_2$-norm of error
Balanced	9.64×10^{-4}	2.04×10^{-3}
Hankel	5.50×10^{-4}	6.25×10^{-3}
Sing. Pert.	9.65×10^{-4}	$2,42 \times 10^{-2}$
Rat. Krylov	7.51×10^{-3}	1.11×10^{-2}
Lanczos	7.86×10^{-3}	1.26×10^{-2}
Arnoldi	1.93×10^{-2}	3.33×10^{-2}
Least Squares	1.35×10^{-3}	1.95×10^{-3}

13.2.2 Heat diffusion model

The original system is a plate with two heat sources and temperature measurements at two locations. It is described by the heat equation. A model of order 197 is obtained by spatial discretization. The real part of the pole closest to the imaginary axis is -1.52×10^{-2}. It is observed from Figure 13.3 that this system is easy to approximate since its Hankel singular values decay rapidly. We approximate it with models of order 5. Lanczos and Arnoldi have not been used in this case. As expected, due to the low tolerance value, all methods generate satisfactory approximants matching the full-order model through the frequency range (see Figure 13.5). Only the rational Krylov method is inaccurate around the frequency 1 rad/sec, due to the nonautomated choice of interpolation points. The error norms are tabulated below.

	$\mathcal{H}_\infty$-norm of error	$\mathcal{H}_2$-norm of error
Balanced	2.03×10^{-3}	5.26×10^{-2}
Approx. Bal.	4.25×10^{-3}	4.68×10^{-2}
Hankel	1.93×10^{-3}	6.16×10^{-2}
Sing. Pert.	2.39×10^{-3}	7.39×10^{-2}
Rat. Krylov	1.92×10^{-2}	2.01×10^{-1}

13.2.3 The CD player

This model describes the dynamics between the lens actuator and the radial arm position of a portable CD player. The model has 120 states, a single input, and a single output. The pole closest to the imaginary axis has real part equal to -2.43×10^{-2}. Approximants have order 12. The first Markov parameter of the system is zero. Hence, instead of expanding the transfer function around ∞, we expand it around $s_0 = 200$. This overcomes the breakdown in the Lanczos procedure. We also use rational Arnoldi with $s_0 = 200$. Figure 13.6, top, shows the largest singular values of the frequency response of the reduced-order models

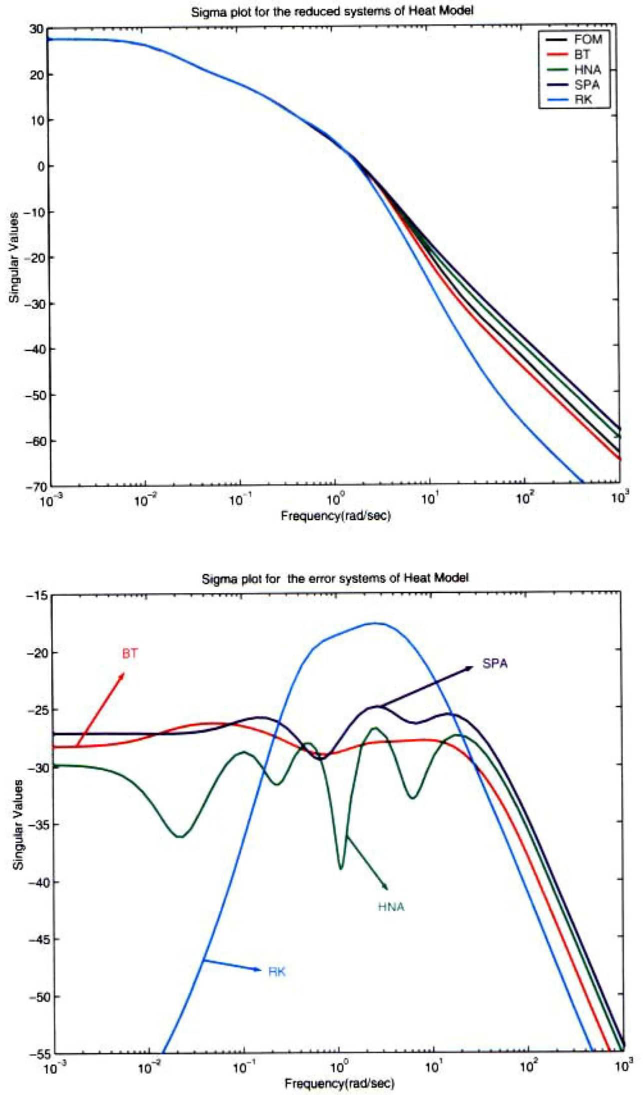
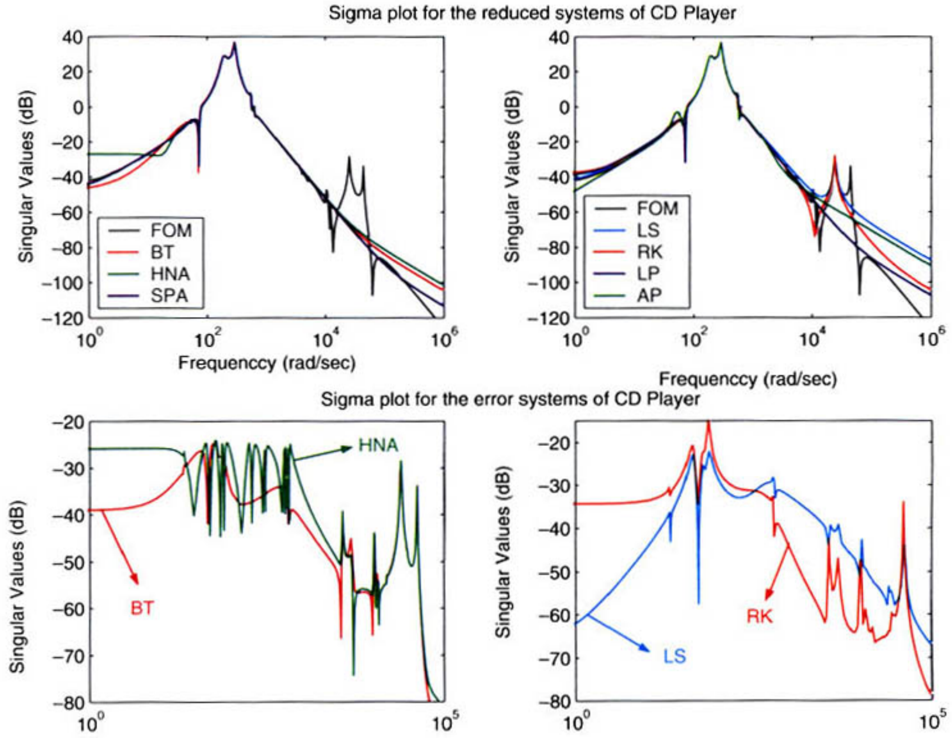


Figure 13.5. *Top: Bode plot of original and reduced systems for the heat diffusion model. Bottom: error Bode plots.*

together with that of the full-order model. It should be noticed that only the rational Krylov method approximates well the peaks around the frequency range 10^4 to 10^5 rad/sec; this is due to the choice of interpolation points in this frequency region. Among the SVD-based ones, Hankel-norm approximation is the worst for both low and high frequencies. The largest singular value of the frequency response of the error systems in Figure 13.6, bottom, reveals that the SVD-based methods are better when we consider the whole frequency range. Despite doing a good job at low and high frequencies, rational Krylov has the highest relative $\mathcal{H}_\infty$ and $\mathcal{H}_2$ error norms, as listed in the table below. But notice that rational



Sigma plot for the reduced systems of CD Player

Sigma plot for the error systems of CD Player

Figure 13.6. *Top: Bode plot of original and reduced systems for the CD player model. Bottom: error Bode plots.*

Krylov is better than Arnoldi and Lanczos except for the frequency range 10^2 to 10^3 rad/sec.

	$\mathcal{H}_\infty$-norm of error	$\mathcal{H}_2$-norm of error
Balanced	9.74×10^{-4}	3.92×10^{-3}
Approx. Bal.	9.74×10^{-4}	3.92×10^{-3}
Hankel	9.01×10^{-4}	4.55×10^{-3}
Sing. Pert.	1.22×10^{-3}	4.16×10^{-3}
Rat. Krylov	5.60×10^{-2}	4.06×10^{-2}
Arnoldi	1.81×10^{-2}	1.84×10^{-2}
Lanczos	1.28×10^{-2}	1.28×10^{-2}
Least Squares	1.14×10^{-3}	3.39×10^{-3}

13.2.4 Clamped beam model

This is the model of a clamped beam with proportional (Rayleigh) damping; it has 348 states and is SISO. It is obtained by spatial discretization of an appropriate partial differential equation. The input is the force applied to the free end of the beam, while the output is the resulting displacement. In this case, the real part of the pole closest to the imaginary axis is -5.05×10^{-3}. We approximate the system with models of order 13. The plots of

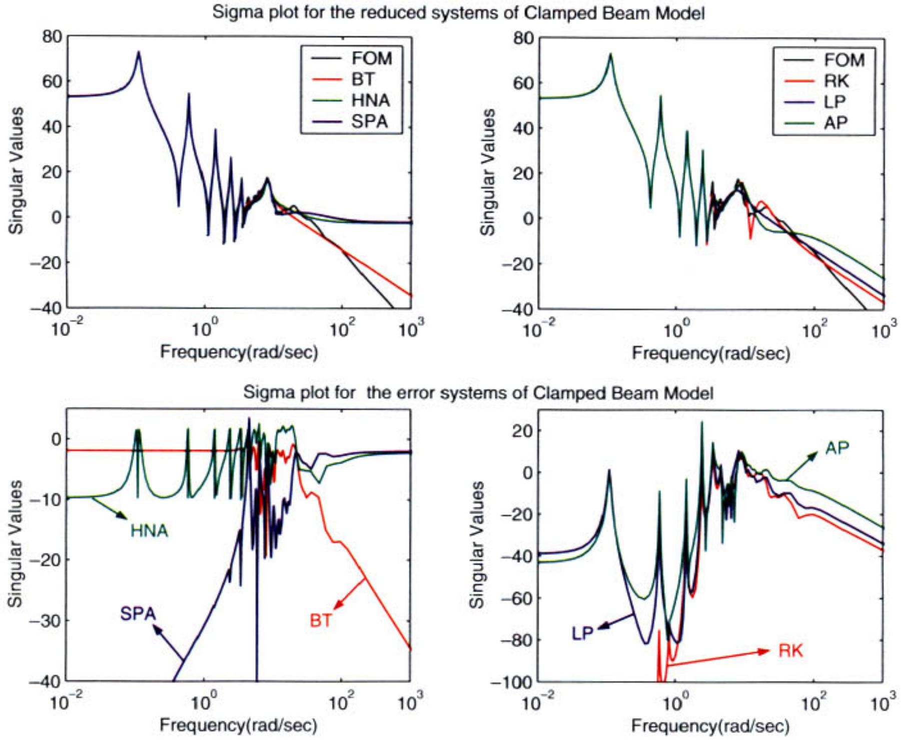


Figure 13.7. *$\sigma_{\max}$ of the frequency response of the reduced systems (upper plots) and of the error systems (lower plots) of the clamped beam.*

the largest singular value of the frequency response of the approximants and error systems is shown in Figure 13.7, left and right, respectively. Since the first Markov parameter is zero, to circumvent the breakdown of Lanczos, rational Lanczos with $s_0 = 1$ is used; we also use rational Arnoldi with the same shift, namely, $s_0 = 1$. For rational Krylov, the interpolation points $s_0 = 1$ and $s_0 = 3$ were used. The ensuing approximant is one of the best approximants. The Lanczos and Arnoldi procedures lead to good approximants for the frequency range 0 to 1 rad/sec, due to the choice of the interpolation point. Balanced model reduction is the best among the SVD methods after 1 rad/sec. In terms of error norms, SVD-based methods are better than moment matching–based methods, but the differences are not as pronounced as for the previous examples. The error norms are tabulated next.

	$\mathcal{H}_\infty$-norm of error	$\mathcal{H}_2$-norm of error
Balanced	2.14×10^{-4}	7.69×10^{-3}
Hankel	2.97×10^{-4}	8.10×10^{-3}
Sing. Pert.	3.28×10^{-4}	4.88×10^{-2}
Rat. Krylov	5.45×10^{-4}	8.88×10^{-3}
Arnoldi	3.72×10^{-3}	1.68×10^{-2}
Lanczos	9.43×10^{-4}	1.67×10^{-2}

13.2.5 Low-pass Butterworth filter

The full-order model in this case is a low-pass Butterworth filter of order 100 with cutoff frequency 1 rad/sec and gain 1 in the pass band. The normalized Hankel singular values are shown in Figure 13.3. It should be noticed that the first 25 Hankel singular values are (almost) equal, and, consequently, this system cannot be reduced to order less than 25 using SVD-based methods. Given the tolerance $\tau = 1 \times 10^{-3}$, the order of the approximants is 35. Since the transfer function in this case has no finite zeros, the rational Arnoldi and Lanczos are used with interpolation point $s_0 = 0.1$. Rational Krylov with interpolation points $s_1 = 1 \times 10^{-5}$, $s_2 = 0.1$, $s_3 = 0.15$, and $s_4 = 0.18$ was also used. As Figure 13.8, top, shows, the approximation around the cutoff frequency provided by the moment matching methods is not good. On the other hand, SVD-based methods produce good approximants for the whole frequency range. Among the SVD-based methods, Hankel-norm approximation is the best in terms of the $\mathcal{H}_\infty$-norm, while balanced reduction is best in terms of the $\mathcal{H}_2$-norm of the error system. Here are the resulting error norms:

	$\mathcal{H}_\infty$-norm of error	$\mathcal{H}_2$-norm of error
Balanced	6.29×10^{-4}	5.19×10^{-4}
Approx. Bal.	6.29×10^{-4}	5.19×10^{-4}
Hankel	5.68×10^{-4}	1.65×10^{-3}
Sing. Pert.	6.33×10^{-4}	5.21×10^{-4}
Rat. Krylov	1.02×10^{0}	4.44×10^{-1}
Arnoldi	1.02×10^{0}	5.38×10^{-1}
Lanczos	1.04×10^{0}	3.68×10^{-1}

Conclusions. A comparative study of several algorithms for model reduction was presented, namely, balanced reduction, approximate balanced reduction, singular perturbation approximation, Hankel-norm approximation, Arnoldi procedure, Lanczos procedure, rational Krylov method, and least squares approximation. The first four make use of Hankel singular values, three are based on matching the moments, i.e., the coefficients of the Laurent expansion of the transfer function around a given point in the complex plane, and one is the SVD-Krylov method described in section 12.3.1. These algorithms have been applied to five dynamical systems.

The results show that balanced reduction and approximate balanced reduction are the best when we consider the whole frequency range. Between these two, approximate balancing has the advantage that it computes a reduced system *iteratively*, and therefore the need to obtain a balanced realization of the full-order system first and subsequently truncate is eliminated. Consequently, the computational cost and storage requirements are reduced. Among the SVD-based methods, Hankel-norm approximation has the lowest $\mathcal{H}_\infty$ error norm in most cases but the highest $\mathcal{H}_2$ error norm, and for low frequencies it gives the worst approximation. Since moment matching methods are local, they usually lead to higher error norms than SVD-based methods; in return, they reduce the computational cost and storage requirements. Among these methods, rational Krylov may lead to better results due to the flexibility of choosing the interpolation points. This selection, however, is not an automated process and has to be specified by the user. In contrast, for SVD-based methods, the specification of an error tolerance determines the reduced model.

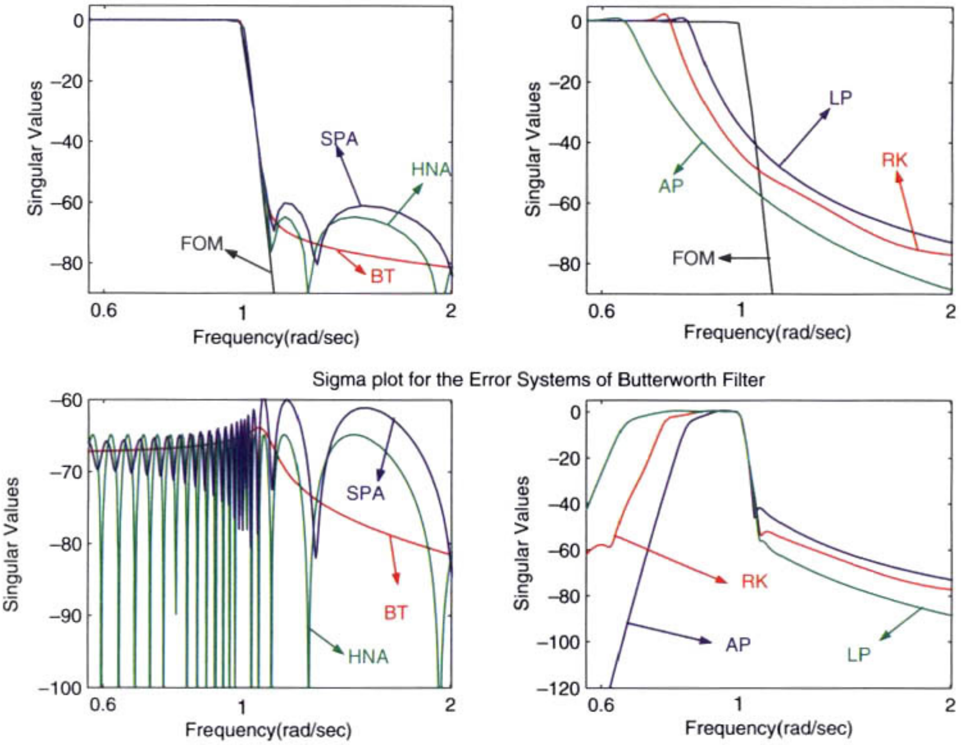


Figure 13.8. *Top: Bode plots of original and reduced systems for the Butterworth filter. Bottom: error Bode plots.*

13.3 Approximation of the ISS 1R and 12A flex models

The model reduction problem for two structural systems will be examined next, namely, modules 1R and 12A of the ISS (International Space Station).

The assembly and operation of the ISS poses unique control challenges due to its complex, variable, and flexible structure as well as the variety of operational modes and control systems. It is estimated that more than 40 space shuttle flights will be required to complete the assembly. For the program to be successful, mission requirements must be met using a variety of control systems from various international partners. Hence, it is critically important that an integrated assessment of the robust stability and performance of the *guidance navigation and control* (GN&C) system be carried out to certify the vehicle for flight readiness.

The integrated control system flex structure assessment and flight readiness certification process must identify the potential for dynamic interaction between the flexible structure and the control systems. As the assembly progresses from early, nearly symmetric stages to late-stage complex configurations, the structural frequency spectrum exhibits progressively more complex characteristics, such as densely packed modes, lower frequency modes, and directionally coupled dominant modes. The structural flexibility during assembly is shown in Figure 2.7, Chapter 2. The figure shows particular stages in the space station assembly sequence and a frequency response plot from thruster commands (roll, pitch, yaw) to filtered rate gyro sensors. It is evident that the complexity and size of the flex models grow as new structural elements are added. Hence, to perform controller flex structure dynamic interaction assessment, it becomes necessary to reduce the flex models to complete the analysis in a timely manner that meets the assembly schedule.

We compare the following reduction methods: balanced reduction, approximate balanced reduction, weighted balanced reduction, the Arnoldi and Lanczos procedures, and the *modal gain factor* (MGF) method.

The multivariable MGF method is a generalization of the standard SISO *modal approximation* method, which weighs the flex modes by frequency. First, it is assumed that the flex model is in modal form, that is,

$$\Sigma = \left[\begin{array}{cc|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{B}_1 \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{B}_2 \\ \hline \mathbf{C}_1 & \mathbf{C}_2 & \end{array}\right] = \left[\begin{array}{cc|c} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ -\mathrm{diag}\,(\omega_i^2) & \mathrm{diag}\,(2\zeta_i) & \Phi_1 \\ \hline \Phi_0^x & \Phi_0^{\dot{x}} & \end{array}\right],$$

where Φ_0^x is the mode shape for displacement or position outputs, $\Phi_0^{\dot{x}}$ is the mode shape for velocity outputs, Φ_1 is the mode shape for the inputs, $\omega_i =\mid \lambda_i(\mathbf{A}) \mid$ and $\zeta_i = \mathcal{R}e(\lambda_i(\mathbf{A}))$, where $\lambda_{i+1}(\mathbf{A}) = [\lambda_i(\mathbf{A})]^*$, is the complex conjugate of $\lambda_i(\mathbf{A})$, for $i = 1, 3, 5, \ldots, n-1$. The MGF corresponding to ω_i is defined as follows:

$$\mathrm{MGF}_i = \left\|[\mathbf{C}_i(:,i) + \mathbf{C}_2(:,i)\omega_i]\mathbf{B}_2(i,:)/\omega_i^2\right\|_2.$$

Note that this expression averages the displacement and velocity outputs. After computing the modal gain factors, given a tolerance τ, reduction is obtained simply by truncating the modes with $\mathrm{MGF}_i < \tau$. A more general formulation using arbitrary input/output pairs is

$$\mathrm{MGF}_i = \left\|[\mathbf{C}_i(j:k,i) + \mathbf{C}_2(:,i)\omega_i]\mathbf{B}_2(i,l:m)/\omega_i^2\right\|_2,$$

where $[l,\ldots,m]$ is the input set and $[j,\ldots,k]$ is the output set.

In addition to the MGF method, we use the Smith and cyclic Smith(l) methods to compute approximate reachability, observability, and cross gramians (see section 12.4). Approximate balanced truncation is then used to obtain reduced-order models.

13.3.1 Stage 1R

This is the flex model of the Russian Service Module 1R of the ISS. It has 270 states, 3 inputs, and 3 outputs. Figure 13.9, top, depicts the Hankel singular values of the system, where the largest has been normalized to one. The system will be approximated with reduced models of order 26; this corresponds to a tolerance of $\tau = 8.4 \times 10^{-3}$, i.e., $\frac{\sigma_{k+1}}{\sigma_1} \leq \tau$.

We apply balanced, approximate balanced, and weighted balanced reductions, as well as reductions based on the Smith and Smith(l) iterative methods. For weighted balanced reduction, the input and output weights are 3-input, 3-output systems which are equal. Three types of weights are considered. In the first case, the system from the ith input to the jth output is a band-pass filter over the frequency range 0.5 to 100 rad/sec. In the second

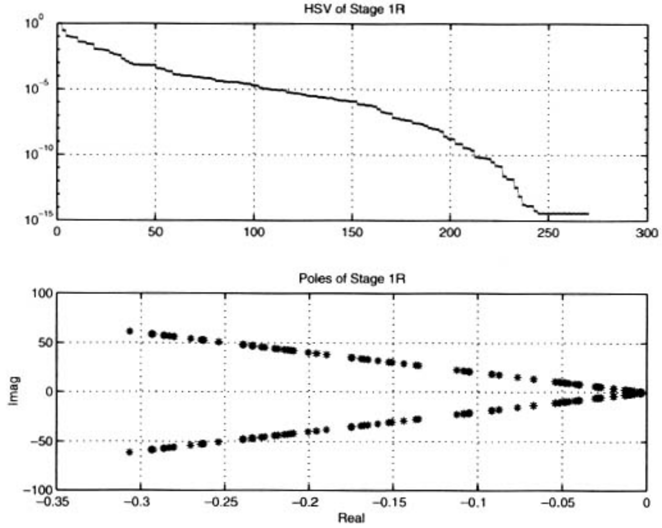


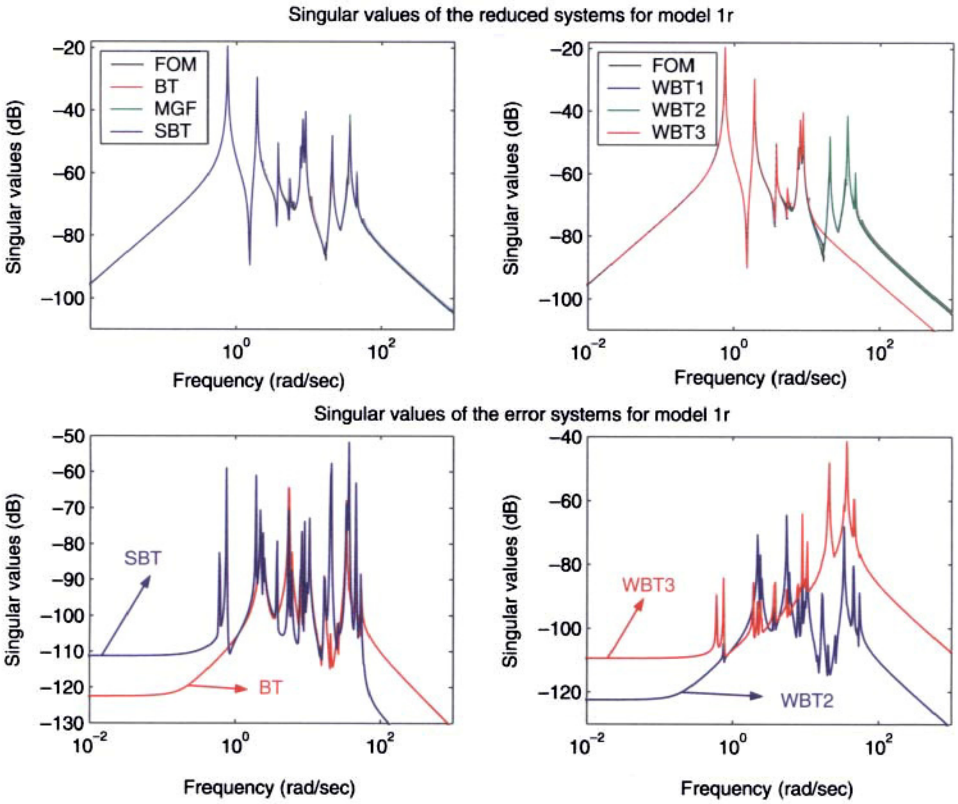
Figure 13.9. *Hankel singular values and poles of the* 1*R ISS model.*

case, diagonal weighting is used, i.e., only the system from the ith input to the ith output is nonzero. The last case is similar to the first case, except that the frequency range is 0.5 to 10 rad/sec. Furthermore, using Smith and Smith(l) methods, we compute low rank approximations of the reachability ($\mathcal{P}$), observability ($\mathcal{Q}$), and cross ($\mathcal{X}$) gramians of the 1R model. Smith and Smith(l) methods transform the system into a discrete-time system using single and multiple shifts; thus the solution is obtained as an infinite sum. The spectral radius of the resulting discrete-time system matrix $\mathbf{A}_d$ determines the convergence rate and the effectiveness of this method. Due to the large spread of the eigenvalues of $\mathbf{A}$ as depicted in the lower pane of Figure 13.9, Smith and Smith(l) methods have slow convergence rates (the spectral radius $\rho(\mathbf{A}_d)$ is close to 1) and show poor results as far as the computation of the approximate gramians is concerned. For the 1R model, Smith's method, using a single shift, results in a spectral radius of 0.9991, while using $l = 20$ shifts, the spectral radius $\rho(\mathbf{A}_d)$ could not be reduced to less than 0.9973. Therefore, only the Smith method is considered. The iteration was run for 50 steps, and the approximants $\bar{\mathcal{P}}$, $\bar{\mathcal{Q}}$, and $\bar{\mathcal{X}}$ were obtained. Their relative error norms are high due to slow convergence:

$$\frac{\|\mathcal{P}-\bar{\mathcal{P}}\|}{\|\mathcal{P}\|} = 8.83 \times 10^{-1}, \quad \frac{\|\mathcal{Q}-\bar{\mathcal{Q}}\|}{\|\mathcal{Q}\|} = 8.51 \times 10^{-1}, \quad \frac{\|\mathcal{X}-\bar{\mathcal{X}}\|}{\|\mathcal{X}\|} = 8.85 \times 10^{-1}.$$

After obtaining the low rank approximate gramians, balanced reduction and approximate balanced reduction are applied using $\bar{\mathcal{P}}$, $\bar{\mathcal{Q}}$, and $\bar{\mathcal{X}}$. These approximants will be referred to as *Smith-balanced* and *Smith-approximate-balanced* reduced models, respectively.

Figure 13.10 shows the Bode plots of the resulting reduced systems as well as of the error systems. Since balanced and approximately balanced reduction yield almost the same approximant, only one of them is shown. These figures show that all reduced models



Singular values of the reduced systems for model 1r

Singular values of the error systems for model 1r

Figure 13.10. *Model reduction plots of the 1R ISS model.*

approximate the peaks of the frequency response well. The exception is the third weighted balanced approximant, which deviates after 10 rad/sec. This is expected because the weighting was chosen over the range 0.5 to 10 rad/sec. The same figure also shows that among the weighted balanced approximants, the second (corresponding to diagonal weighting) is the best.

It is interesting to see that although the errors in the computed gramians and Hankel singular values are significant, Smith-balanced and Smith-approximate balanced models work well. This means that although $\bar{\mathcal{P}}$, $\bar{\mathcal{Q}}$, and $\bar{\mathcal{X}}$ are not good approximants of $\mathcal{P}$, $\mathcal{Q}$, and $\mathcal{X}$, the *dominant eigenspaces* of $\bar{\mathcal{P}}\bar{\mathcal{Q}}$, and $\bar{\mathcal{X}}$ are good approximations of the *dominant eigenspaces* of $\mathcal{P}\mathcal{Q}$ and $\mathcal{X}$, respectively.

The MGF approximant is better for low frequencies, while balanced and approximate balanced approximants have a smaller error at moderate and high frequencies. The Smith-balanced and Smith-approximate-balanced models are comparable with the other approximants, especially at moderate frequencies. Indeed, although they have the highest error considering the whole frequency range, for high frequencies they are the best approximants. Notice that Smith-balanced and Smith-approximate-balanced models yield almost the same approximants.

The next table lists the relative $\mathcal{H}_\infty$-norm of the error systems (quotient of the error system norm and the original system norm). Balanced and approximate-balanced reductions have the lowest error norm, while the modal gain factor method yields a slightly higher error norm. The error norms of both Smith approximants are high. The worst among all is the third weighted balanced approximant.

	Balancing	App. Bal.	Smith Bal.	Smith App. Bal.
H_∞-norms	5.71×10^{-3}	5.71×10^{-3}	2.44×10^{-2}	2.42×10^{-2}
	MGF	Weighted-1	Weighted-2	Weighted-3
H_∞-norms	5.75×10^{-3}	2.73×10^{-2}	5.71×10^{-3}	8.04×10^{-2}

Let $\mathbf{H}(s)$ and $\bar{\mathbf{H}}(s)$ denote the transfer functions of the FOM and ROMs, respectively. Furthermore, let $\omega_{\max}$ be the frequency at which the FOM attains its $\mathcal{H}_\infty$-norm, that is, $\sigma_{\max}\mathbf{H}(i\omega_{\max})$ is the $\mathcal{H}_\infty$-norm of the FOM. The pair of left, right, singular vectors of $\mathbf{H}(i\omega_{\max})$ will be denoted by $\mathbf{u}_i$, $\mathbf{v}_i$, and that of $\bar{\mathbf{H}}(i\omega_{\max})$ by $\bar{\mathbf{u}}_i$, $\bar{\mathbf{v}}_i$. The table below lists the relative errors $\mathcal{E}(\mathbf{u}_i) = \frac{\|\mathbf{u}_i - \bar{\mathbf{u}}_i\|}{\|\mathbf{u}_i\|}$, $\mathcal{E}(\mathbf{v}_i) = \frac{\|\mathbf{v}_i - \bar{\mathbf{v}}_i\|}{\|\mathbf{v}_i\|}$ for the various ROMs. Notice that while all reduced models match the first pair of singular vectors, this is not the case for the second and third pairs of singular vectors. The model gain factor method is the best in this respect.

	Balancing	App. Bal.	Smith Bal.	Smith App. Bal.
$\mathcal{E}(\mathbf{u}_1)$	1.03×10^{-5}	8.99×10^{-6}	8.66×10^{-4}	8.67×10^{-4}
$\mathcal{E}(\mathbf{u}_2)$	1.33×10^{0}	1.33×10^{0}	2.07×10^{-1}	1.28×10^{0}
$\mathcal{E}(\mathbf{u}_3)$	5.89×10^{-2}	5.69×10^{-2}	2.11×10^{-2}	2.98×10^{-2}
$\mathcal{E}(\mathbf{v}_1)$	2.11×10^{-5}	2.19×10^{-5}	4.99×10^{-4}	4.76×10^{-4}
$\mathcal{E}(\mathbf{v}_2)$	1.35×10^{0}	1.35×10^{0}	2.66×10^{-1}	1.24×10^{0}
$\mathcal{E}(\mathbf{v}_3)$	2.51×10^{-2}	2.46×10^{-2}	5.19×10^{-3}	1.38×10^{-2}
	MGF	Weighted-1	Weighted-2	Weighted-3
$\mathcal{E}(\mathbf{u}_1)$	2.02×10^{-5}	3.54×10^{-5}	2.79×10^{-5}	5.34×10^{-4}
$\mathcal{E}(\mathbf{u}_2)$	5.15×10^{-2}	1.43×10^{0}	1.32×10^{0}	1.36×10^{0}
$\mathcal{E}(\mathbf{u}_3)$	1.00×10^{-3}	2.61×10^{-1}	5.89×10^{-2}	2.97×10^{-1}
$\mathcal{E}(\mathbf{v}_1)$	2.07×10^{-5}	4.85×10^{-5}	2.22×10^{-5}	1.32×10^{-4}
$\mathcal{E}(\mathbf{v}_2)$	5.34×10^{-2}	1.48×10^{0}	1.35×10^{0}	1.33×10^{0}
$\mathcal{E}(\mathbf{v}_3)$	3.07×10^{-4}	2.48×10^{-1}	2.49×10^{-2}	2.07×10^{-2}

13.3.2 Stage 12A

This is the flex model of stage 12A of the ISS. It has 1412 states, one input, and one output. The leading 200 normalized Hankel singular values of the system are shown in Figure 13.11, top; we see that their decay is slower than that in the case of model 1R. The system is approximated with reduced models of order 226, which corresponds to a tolerance of $\tau = 2.7 \times 10^{-4}$.

We apply the following approximation methods: balanced reduction, approximate-balanced reduction, Arnoldi, Lanczos, Smith, and Smith(l). Since the expansion of the transfer function around $s_0 = \infty$ results in an unstable reduced system, we use the shifted versions of the Arnoldi and Lanczos procedures with $s_0 = 6$ and $s_0 = 4$, respectively. Furthermore, using the Smith and Smith(l) methods, we compute $\bar{\mathcal{P}}$, $\bar{\mathcal{Q}}$, $\bar{\mathcal{X}}$, which lead to

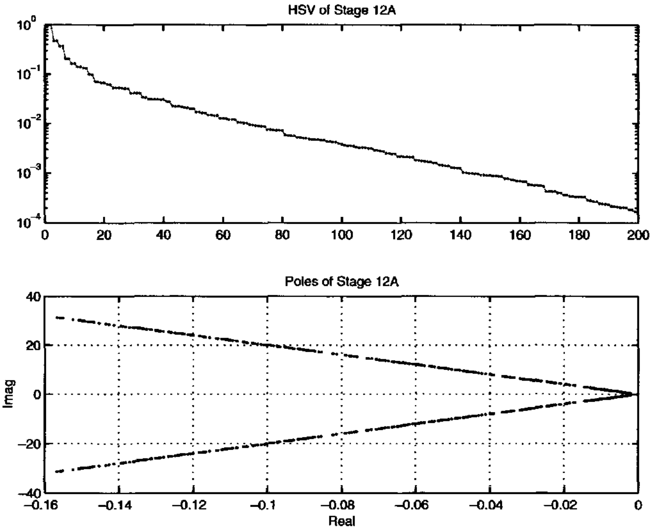


Figure 13.11. *The leading Hankel singular values (top) and the poles (bottom) of the* 12A *ISS model.*

Smith-balanced and Smith-approximate-balanced approximants. The eigenvalues (poles) of model 12A shown in Figure 13.11, bottom, exhibit the same pattern as those of model 1R. Hence, multiple shifts do not help convergence, and consequently only one shift is used. The resulting spectral radius is $\rho(\mathbf{A}_d) = 0.9989$. The iteration is run for 500 steps and the approximate gramians $\bar{\mathcal{P}}$, $\bar{\mathcal{Q}}$, and $\bar{\mathcal{X}}$ have the following error norms:

$$\frac{\|\mathcal{P} - \bar{\mathcal{P}}\|}{\|\mathcal{P}\|} = 2.04 \times 10^0, \quad \frac{\|\mathcal{Q} - \bar{\mathcal{Q}}\|}{\|\mathcal{Q}\|} = 1.56 \times 10^0, \quad \frac{\|\mathcal{X} - \bar{\mathcal{X}}\|}{\|\mathcal{X}\|} = 4.16 \times 10^0.$$

As for 1R, the relative errors in the approximate gramians are high. But the error in the approximate Hankel singular values is much lower in this case, and the leading $\hat{\sigma}_i$ are good approximants of the corresponding σ_i.

Figure 13.12 shows the Bode plots of the reduced-order and error systems. As before, balanced, approximate balanced, reduction, on the one hand, and Smith-balanced, Smith-approximate-balanced, reduction, on the other yield (almost) the same approximants; hence only the results for the balanced and Smith-balanced approximants are depicted. As seen from the same figure, all reduced models work quite well. MGF approximants show some deviation around 5 rad/sec and around ∞. As in the 1R case, although the errors in the computed gramians are significant, Smith-balanced and Smith-approximate-balanced models are good approximants. The plots of the largest singular value σ_{max} of the frequency response of the error systems show that Lanczos and Arnoldi outperform the other methods for low and high frequencies.

For moderate frequencies, except for Lanczos, the remaining methods yield comparable results. Finally, while the MGF method is better for low frequencies, balanced truncation is better for high frequencies. The relative $\mathcal{H}_\infty$-norms of the error systems are given in the following table:

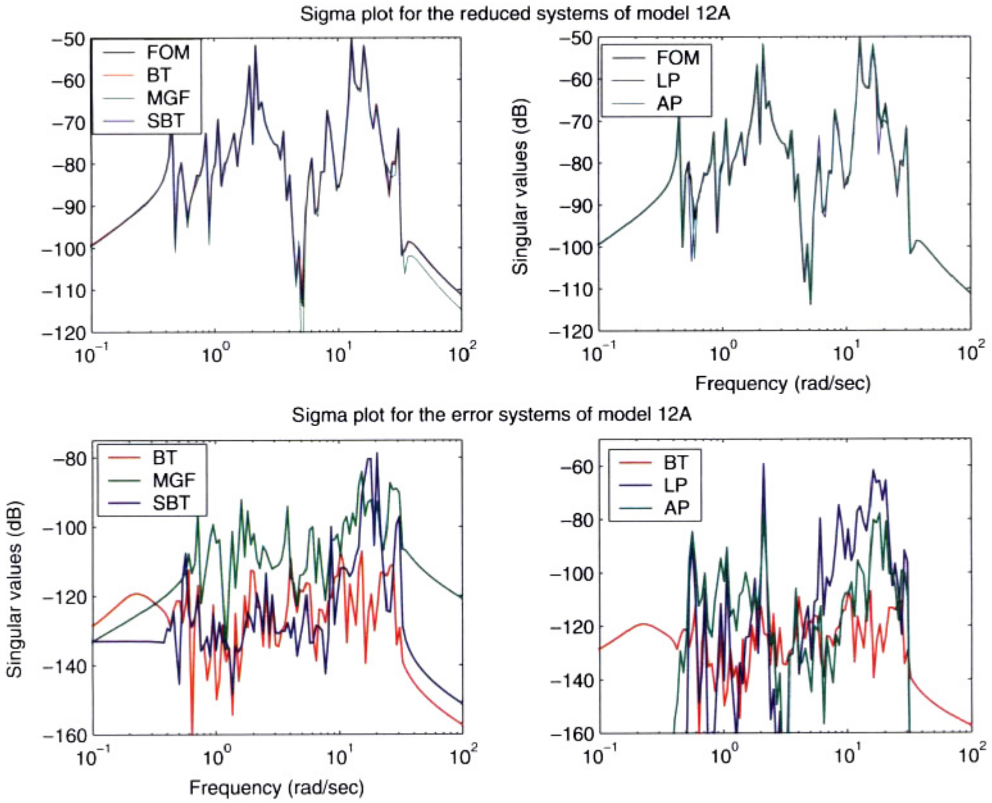


Figure 13.12. *Model reduction plots of the* 12*A ISS model.*

	Balancing	App. Bal.	Smith Bal.	Smith App. Bal.
$\mathcal{H}_\infty$-norms	1.43×10^{-3}	1.43×10^{-3}	3.74×10^{-2}	3.74×10^{-2}

	MGF	Arnoldi	Lanczos	
$\mathcal{H}_\infty$-norms	2.05×10^{-2}	4.11×10^{-2}	3.49×10^{-1}	

Thus, balanced and approximate balanced reduced models have the lowest error norm. The error of the MFG method in this case is much higher. The Lanczos and Arnoldi procedures yield high error norms as well. The Smith method lies between MGF and Arnoldi.

Conclusions. Several model reduction methods have been applied to the flex models of stage 1R and stage 12A of the ISS. While the MGF method yields better results than balanced reduction for very low frequencies, the latter is preferable for high frequencies. Furthermore, balanced and approximate balanced reduction yield the lowest error norms. The Smith iterative method yields high errors in approximating the gramians, but nevertheless the reduced models are satisfactory. Finally, Lanczos and Arnoldi yield good approximants close to the chosen interpolation points, but the resulting overall error norms are high.

13.4 Iterative Smith-type methods

Next, the results presented in section 12.4 will be illustrated by means of numerical examples. In each case, both the LR-Smith(l) iterates $\mathcal{P}_k^{SI}$, $\mathcal{Q}_k^{SI}$, as well as the modified LR-Smith(l) iterates $\tilde{\mathcal{P}}_k$, $\tilde{\mathcal{Q}}_k$, will be computed, together with the error bounds introduced in section 12.4.3. For comparison purposes, balanced reduction is also applied using the full rank gramians $\mathcal{P}$, $\mathcal{Q}$, the approximate gramians $\mathcal{P}_k^{SI}$, $\mathcal{Q}_k^{SI}$, and $\tilde{\mathcal{P}}_k$, $\tilde{\mathcal{Q}}_k$. The resulting reduced-order systems and the Hankel singular values are compared.

13.4.1 The CD player

Here we will consider the same model as in section 13.2.3. It turns out (see, e.g., [157]) that the eigenvalues of $\mathbf{A}$ have a wide spread, that is, they have both big and small real and imaginary parts. This makes it hard to obtain a low spectral radius $\rho(\mathbf{A}_d)$. A single shift results in $\rho(\mathbf{A}_d) = 0.99985$, while with $l = 40$ shifts $\rho(\mathbf{A}_d)$ could not be reduced below 0.98. Hence only a single shift is considered. LR-Smith(l) and the modified LR-Smith(l) iterations are run for $k = 70$ iterations. The tolerance values are chosen to be $\tau_{\mathcal{P}} = 1 \times 10^{-6}$ for $\tilde{\mathcal{P}}_k$ and $\tau_{\mathcal{Q}} = 8 \times 10^{-6}$ for $\tilde{\mathcal{Q}}_k$. The resulting low rank LR-Smith(l) square root factors $\mathbf{Z}_k^{SI}$ and $\mathbf{Y}_k^{SI}$ have 70 columns, while the modified version of the algorithm yields low rank square root factors $\tilde{\mathbf{Z}}_k$ and $\tilde{\mathbf{Y}}_k$ which have only 25 columns. The relative errors between the computed gramians are

$$\frac{\|\mathcal{P}_k^{SI} - \tilde{\mathcal{P}}_k\|}{\|\mathcal{P}_k^{SI}\|} = 4.13 \times 10^{-10} \text{ and } \frac{\|\mathcal{Q}_k^{SI} - \tilde{\mathcal{Q}}_k\|}{\|\mathcal{Q}_k^{SI}\|} = 2.33 \times 10^{-10}.$$

These numbers show the effectiveness of the modified algorithm. The errors between the exact and the computed gramians are

$$\frac{\|\mathcal{P} - \mathcal{P}_k^{SI}\|}{\|\mathcal{P}\|} = \frac{\|\mathcal{P} - \tilde{\mathcal{P}}_k\|}{\|\mathcal{P}\|} = 3.95 \times 10^{-3} \text{ and}$$

$$\frac{\|\mathcal{Q} - \mathcal{Q}_k^{SI}\|}{\|\mathcal{Q}\|} = \frac{\|\mathcal{Q} - \tilde{\mathcal{Q}}_k\|}{\|\mathcal{Q}\|} = 8.24 \times 10^{-1}.$$

Figure 13.13, top, depicts the normalized Hankel singular values of the FOM, σ_i; the Hankel singular values resulting from $\mathcal{P}_k^{SI}$ and $\mathcal{Q}_k^{SI}$, $\hat{\sigma}_i$; and the Hankel singular values resulting from $\tilde{\mathcal{P}}_k$ and $\tilde{\mathcal{Q}}_k$, $\tilde{\sigma}_i$. As the figure shows, the leading $\hat{\sigma}_i$, $i = 1, \ldots, 14$, are well approximated. Moreover,

$$\sum_{i=1}^{30} \hat{\sigma}_i - \sum_{i=1}^{30} \tilde{\sigma}_i = 1.64 \times 10^{-3}.$$

The errors and the corresponding error bounds are tabulated below. One should notice that the error bounds (12.12) for the norm and (12.13) for the trace are tight, as they are of the order $\mathcal{O}(\tau^2)$. Also, as stated after Corollary 12.12, the upper bound for the error between $\hat{\sigma}_i$ and $\tilde{\sigma}_i$ is small. The bounds (12.9) and (12.14) are loose because of the slow convergence ($\rho(A_d) = 0.99985$). (See [157] for more details.)

Exercise 3-6-6. Convert the following problem into standard form using Scheme I and solve using the simplex method:

$$\begin{array}{lrcrcrcl} \min & x + 2y + 3z \\ \text{subject to} & x & - & y & + & 3z & \le & 3, \\ & 4x & + & y & & & \ge & 1, \\ & & & & & z & \ge & 0. \end{array}$$

Exercise 3-6-7. Convert the following problem to standard form and solve using the two-phase simplex method:

$$\begin{array}{lrcrcrcrcr} \max & -2x_1 - x_2 - x_3 - 2x_4 \\ \text{subject to} & x_1 & - & x_2 & + & x_3 & - & x_4 & = & -1, \\ & -x_1 & - & x_2 & - & x_3 & - & x_4 & = & -3, \\ & -x_1 & + & x_2 & - & x_3 & + & x_4 & \le & 1, \\ & x_1 & + & x_2 & - & x_3 & - & x_4 & \le & -1, \\ & -x_1 & - & x_2 & + & x_3 & + & x_4 & \le & 1, \\ & -x_1 & + & x_2 & + & x_3 & - & x_4 & \le & -2, \\ & & & & & & & x_1, x_4 & \ge & 0. \end{array}$$

(Note that the variables x_2, x_3 are free.)

In Example 3-6-5, we added only one equation and one variable to the problem during conversion to standard form. When there are multiple free variables and multiple equations, we end up increasing the size of the problem significantly, as seen in Exercise 3-6-7. By using a variant on the Scheme I technique, we can generate a standard-form problem by just adding one extra variable and one extra constraint. The role of the extra variable is to absorb the maximum negativity of the free variables y. We replace y by a set of nonnegative variables $\hat{y}$ and the extra variable η as follows:

$$y \text{ free} \iff y = \hat{y} - e\eta, \qquad \hat{y} \ge 0, \qquad \eta \ge 0.$$

For the equality constraints, we make the following substitution:

$$Ex + Fy = g \iff \begin{array}{rcl} Ex + Fy & \ge & g, \\ e'(Ex + Fy - g) & \le & 0. \end{array}$$

Here $e = (1, 1, \dots, 1)'$ is a vector of ones of appropriate dimension. By making these two substitutions into the general form (3.22), we obtain the following standard-form linear program:

$$\begin{array}{lrcrcrcr} \min\limits_{x,\hat{y},\eta} & p'x + q'(\hat{y} - e\eta) \\ \text{subject to} & Bx & + & C(\hat{y} & - & e\eta) & \ge & d, \\ & Ex & + & F(\hat{y} & - & e\eta) & \ge & g, \\ & -e'Ex & - & e'F(\hat{y} & - & e\eta) & \ge & -e'g, \\ & & & & & x, \hat{y}, \eta & \ge & 0. \end{array}$$

Exercise 3-6-8. Use the above approach to solve the problem given in Exercise 3-6-7.

Σ_k^{Sl}, and $\tilde{\Sigma}_k$ denote the 12th-order reduced systems obtained through balanced reduction using the exact square root factors $\mathbf{Z}$ and $\mathbf{Y}$, the LR-Smith(l) iterates $\mathbf{Z}_k^{Sl}$ and $\mathbf{Y}_k^{Sl}$, and the modified LR-Smith(l) iterates $\tilde{\mathbf{Z}}_k$ and $\tilde{\mathbf{Y}}_k$, respectively. Also let Σ denote the FOM. Figure 13.13, bottom left, depicts the amplitude Bode plots of the FOM Σ and the reduced systems Σ_k, Σ_k^{Sl}, and $\tilde{\Sigma}_k$; thus although the approximate gramians are not very good approximations of the exact gramians, Σ_k^{Sl} and $\tilde{\Sigma}_k$ are close to Σ_k. The amplitude Bode plots of the error systems $\Sigma - \Sigma_k$, $\Sigma - \Sigma_k^{Sl}$, and $\Sigma - \tilde{\Sigma}_k$ are shown in Figure 13.13, bottom right. Thus Σ_k^{Sl} and $\tilde{\Sigma}_k$ are almost equal as expected since the errors between $\tilde{\mathcal{P}}_k$ and $\mathcal{P}_k^{Sl}$ and $\tilde{\mathcal{Q}}_k$ and $\mathcal{Q}_k^{Sl}$ are small; see Figure 13.13, top right.

13.4.2 A system of order $n = 1006$

This example is from [265]. The FOM is a dynamical system of order 1006. The state space matrices are given by $\mathbf{A} = \text{diag}\,[\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4]$, where

$$\mathbf{A}_1 = \begin{bmatrix} -1 & 100 \\ -100 & -1 \end{bmatrix}, \mathbf{A}_2 = \begin{bmatrix} -1 & 200 \\ -200 & -1 \end{bmatrix}, \mathbf{A}_3 = \begin{bmatrix} -1 & 400 \\ -400 & -1 \end{bmatrix},$$

$$\mathbf{A}_4 = \text{diag}\,[-1, -2, \ldots, -1000], \quad \mathbf{B}^* = \mathbf{C} = [\ \underbrace{10 \cdots 10}_{6}\ \ \underbrace{1 \cdots 1}_{1000}\ ].$$

Using $l = 10$ shifts, the spectral radius of $\mathbf{A}_d$ is reduced to $\rho(\mathbf{A}_d) = 0.7623$, which results in fast convergence. It is easy to see that the spectrum of $\mathbf{A}$ is

$$\{-1, -2, \ldots, -1000, -1 \pm j100, -1 \pm j200, -1 \pm j400\}.$$

Thus six of the shifts are chosen so that the six complex eigenvalues of $\mathbf{A}$ are eliminated. We run the LR-Smith(l) and the modified LR-Smith(l) methods for $k = 30$ iterations. The tolerance values are $\tau_{\mathcal{P}} = \tau_{\mathcal{Q}} = 3 \times 10^{-5}$. The resulting low rank LR-Smith(l) square root factors $\mathbf{Z}_k^{Sl}$ and $\mathbf{Y}_k^{Sl}$ have 300 columns. On the other hand, the modified low rank Smith method yields low rank square root factors $\tilde{\mathbf{Z}}_k$ and $\tilde{\mathbf{Y}}_k$ which have only 19 columns. The relative errors between the computed gramians are small, namely,

$$\frac{\|\mathcal{P}_k^{Sl} - \tilde{\mathcal{P}}_k\|}{\|\mathcal{P}_k^{Sl}\|} = 2.75 \times 10^{-8} \text{ and } \frac{\|\mathcal{Q}_k^{Sl} - \tilde{\mathcal{Q}}_k\|}{\|\mathcal{Q}_k^{Sl}\|} = 2.32 \times 10^{-8}.$$

The errors between the exact and computed gramians are

$$\frac{\|\mathcal{P} - \mathcal{P}_k^{Sl}\|}{\|\mathcal{P}\|} = 4.98 \times 10^{-5}, \quad \frac{\|\mathcal{P} - \tilde{\mathcal{P}}_k\|}{\|\mathcal{P}\|} = 2.71 \times 10^{-8},$$

$$\frac{\|\mathcal{Q} - \mathcal{Q}_k^{Sl}\|}{\|\mathcal{Q}\|} = 4.98 \times 10^{-5}, \quad \frac{\|\mathcal{Q} - \tilde{\mathcal{Q}}_k\|}{\|\mathcal{Q}\|} = 2.31 \times 10^{-8}.$$

The normalized HSVs σ_i, $\hat{\sigma}_i$ and $\tilde{\sigma}_i$, $i = 1, \ldots, 19$, are depicted in Figure 13.14, top left. There holds

$$\sum_{i=1}^{19} \hat{\sigma}_i - \sum_{i=1}^{19} \tilde{\sigma}_i = 2.76 \times 10^{-6}.$$

All the errors and the corresponding error bounds are tabulated below.

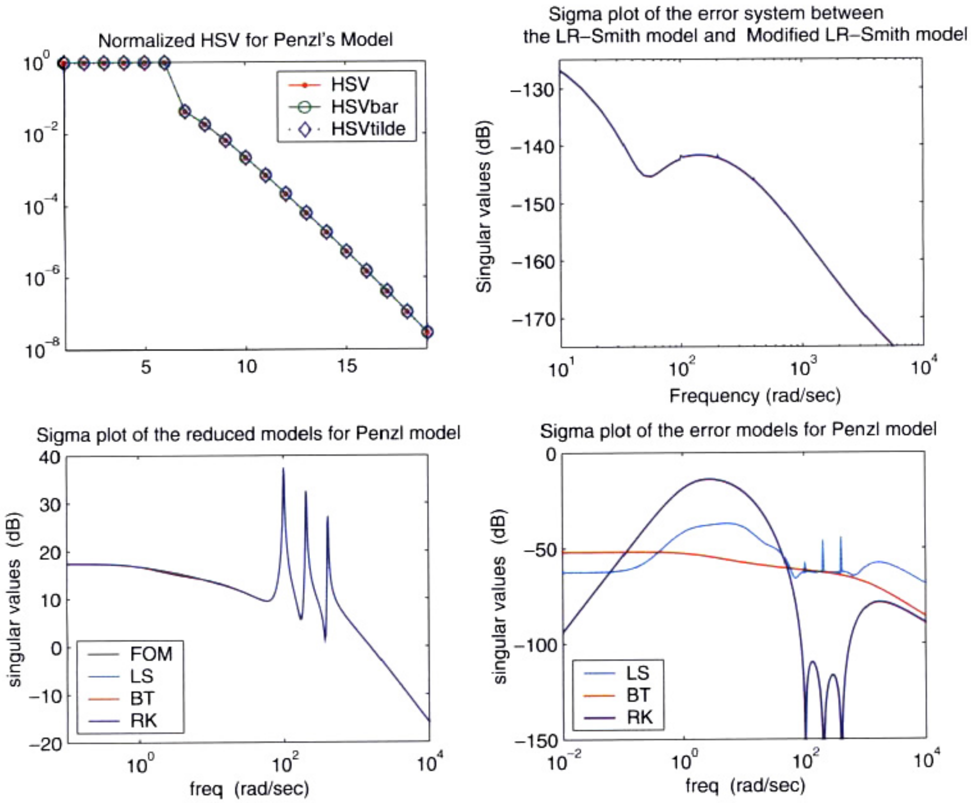


Figure 13.14. *Top left: the normalized Hankel singular values σ_i, $\hat{\sigma}_i$, and $\tilde{\sigma}_i$. Top right: the amplitude Bode plot of error system $\mathbf{\Sigma}_k^{SI} - \tilde{\mathbf{\Sigma}}_k$ for the 1006 example. Bottom left: amplitude Bode plots of the FOM $\mathbf{\Sigma}$ and the reduced systems $\mathbf{\Sigma}_k$, $\mathbf{\Sigma}_k^{SI}$, and $\tilde{\mathbf{\Sigma}}_k$. Bottom right: amplitude Bode plots of the error systems $\mathbf{\Sigma} - \mathbf{\Sigma}_k$, $\mathbf{\Sigma} - \mathbf{\Sigma}_k^{SI}$, and $\mathbf{\Sigma} - \tilde{\mathbf{\Sigma}}_k$ for the $n = 1006$ example.*

$\|\mathcal{P}_k^{SI} - \tilde{\mathcal{P}}_k\|$	upper bound	trace $(\mathcal{P}_k^{SI} - \tilde{\mathcal{P}}_k)$	upper bound	$\sum\sigma_i^2 - \sum\hat{\sigma}_i^2$	upper bound
1.12×10^{-6}	1.35×10^{-6}	2.52×10^{-6}	1.30×10^{-5}	$2.99. \times 10^{-8}$	2.69×10^{-2}

$\|\mathcal{Q}_k^{SI} - \tilde{\mathcal{Q}}_k\|$	upper bound	trace $(\mathcal{Q}_k^{SI} - \tilde{\mathcal{Q}}_k)$	upper bound	$\sum\sigma_i^2 - \sum\tilde{\sigma}_i^2$	upper bound
1.20×10^{-6}	1.35×10^{-6}	3.00×10^{-6}	1.30×10^{-5}	4.19×10^{-5}	$2.69 \times 10^{-2} + 1.33 \times 10^{-3}$

$\|\Sigma - \Sigma_k\|_\infty$	$\|\Sigma - \Sigma_k^{SI}\|_\infty$	$\|\Sigma - \tilde{\Sigma}_k\|_\infty$	$\|\Sigma_k^{SI} - \tilde{\Sigma}_k\|_\infty$	$\|\Sigma_k - \Sigma_k^{SI}\|_\infty$	$\|\Sigma_k - \tilde{\Sigma}_k\|_\infty$
1.47×10^{-4}	1.47×10^{-4}	1.47×10^{-4}	2.40×10^{-9}	7.25×10^{-11}	7.25×10^{-11}

The error bounds (12.12) and (12.13) are again tight, while the bounds (12.9) and (12.14) are tighter compared to the CD example. This is due to the normality of $\mathbf{A}$ and to the fast convergence rate, i.e., the small value of $\rho(\mathbf{A}_d)$.

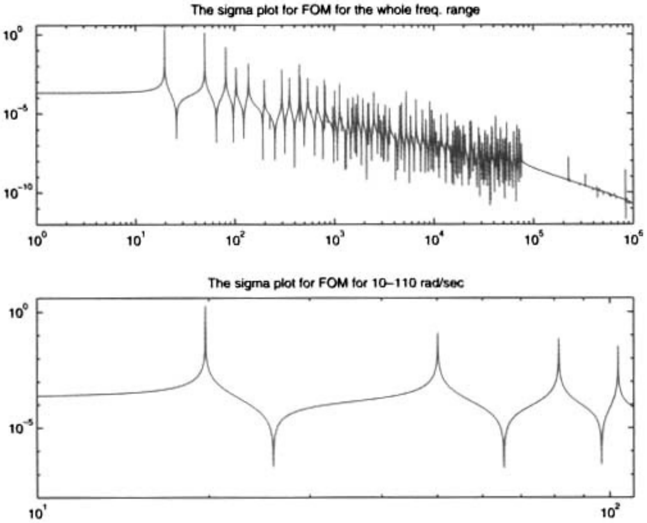


Figure 13.15. *Frequency response of FOM.*

Using balanced reduction, a reduced model of order 11 is obtained. Figure 13.14, bottom left depicts the amplitude Bode plots of the FOM Σ and the reduced systems Σ_k, Σ_k^{SI}, and $\tilde{\Sigma}_k$. The amplitude Bode plots of the error systems $\Sigma - \Sigma_k$, $\Sigma - \Sigma_k^{SI}$, and $\Sigma - \tilde{\Sigma}_k$ are shown in the bottom right of this figure. Again, Σ_k^{SI} and $\tilde{\Sigma}_k$ are almost identical. The Bode plots of the error $\Sigma_k^{SI} - \tilde{\Sigma}_k$ are shown in the top right of the figure. Recall that Σ_k^{SI} has been obtained using a square root factor having 300 columns, while $\tilde{\Sigma}_k$ has been obtained using a square root factor having only 19 columns, which indicates that the rank of the LR-Smith(l) iterate is low.

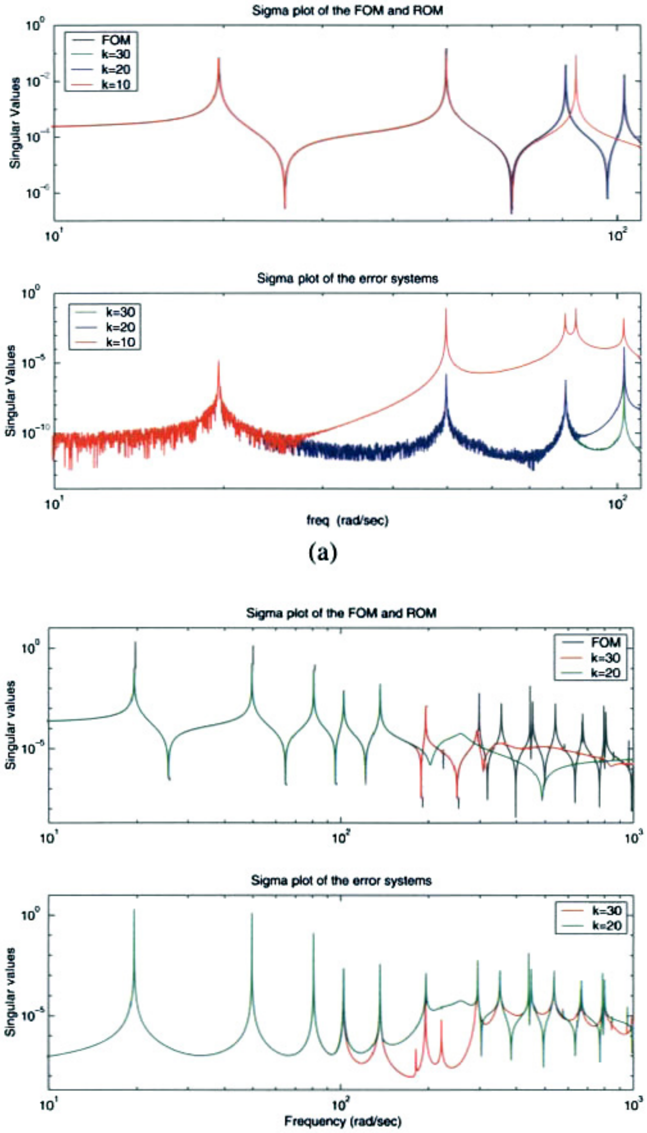
13.4.3 Aluminum plate $n = 3468$

This is the structural model of an aluminum plate of $0.5m$ by $0.5m$, Young modulus $7.0\,10^{10}N/m^2$, thickness $0.001m$, Poisson ratio 0.33, and density $2700kg/m^3$ and no structural damping. It was discretized by a grid of 16×16 solid shell elements. The plate is subjected to a unit point force in the coordinate $(0.125m, 0.125m)$. The goal is to compute the amplitude of the vertical displacement in the same position for the frequency range $\omega \in [10, 110]$ Hz. The dimension of the second-order system is 1734, that is, $n = 3468$.

Its frequency response is shown in Figure 13.15. Figure 13.16 shows the frequency responses of error systems and approximants obtained by rational Krylov methods with $s_0 = 0$ and $s_0 = 100$.

13.4.4 Heat distribution on a plate $n = 20{,}736$

The FOM describes the two-dimensional heat equation on a square plate with adiabatic (no heat flux) boundary conditions. The plate consists of nine subplates. The number of discretization points in the x and y axes is 144; this leads to a FOM of order 20,736. Each

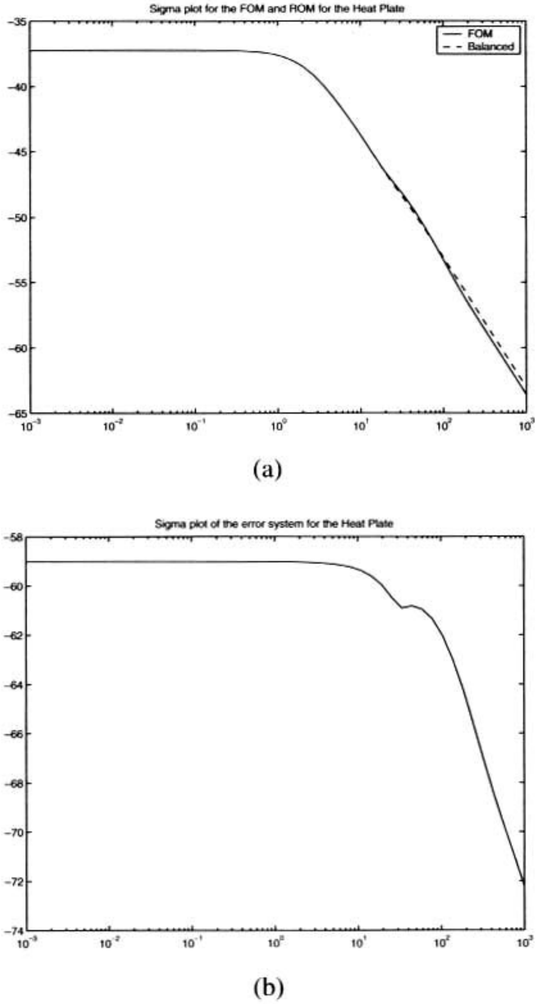


(a)

(b)

Figure 13.16. *Rational Krylov method: expansion point $s_0 = 0$ (a); expansion point $s_0 = 100$ (b).*

subblock is uniformly heated, which results in $m = 9$ inputs. As observations, we choose the middle points of each one of the subblocks. This leads to $p = 9$ outputs. We ran the modified low rank Smith algorithm with tolerance values $\tau_{\mathcal{P}} = \tau_{\mathcal{Q}} = 10^{-6}$ for $k = 40$ steps using $l = 2$ shifts and obtained the low rank square root factors $\tilde{Z}_k$ and $\tilde{Y}_k$ with 85 columns. Note that an exact LR-Smith(l) factor has 720 columns. Using the approximate Cholesky factors $\tilde{Z}_k$ and $\tilde{Y}_k$, the system was reduced to dimension 9 by means of balanced reduction.



(a)

(b)

Figure 13.17. (a) *Amplitude Bode plots of the FOM and ROM.* (b) *Amplitude Bode plots of the error systems for heat example.*

The reduced model is *almost* balanced, as the following numerical results show:

$$\|\mathcal{P}_{red} - \text{diag}(\mathcal{P}_{red})\| = 7.28 \times 10^{-9}, \quad \|\mathcal{Q}_{red} - \text{diag}(\mathcal{Q}_{red})\| = 1.05 \times 10^{-12},$$

$$\|\mathcal{P}_{red} - \mathcal{Q}_{red}\| = 1.04 \times 10^{-9},$$

where $\mathcal{P}_{red}$, $\mathcal{Q}_{red}$ denote the gramians for the reduced model. The sigma plots of the FOM and ROM and that of the error model are shown in Figure 13.17. Thus given a FOM of order 20,736, this iterative method produced an approximately balanced approximant of order 9, based on Cholesky factors having only 85 columns.

Chapter 14
Epilogue

The goal for this book is to discuss model reduction issues following the diagram in Figure 1.3. After the preparatory material of Part II, SVD-based approximation methods were investigated in Part III. In particular, approximation by balanced truncation, weighted balanced truncation, and optimal and suboptimal Hankel-norm approximation were studied. The closely related POD methods were also briefly discussed. This is the *left branch* of the diagram in Figure 1.3. In Part IV, Krylov-based approximation methods were discussed. Since they are based on eigenvalue estimation methods, some effort was spent discussing this topic. This was followed by an account of the Krylov or moment matching methods for model reduction, namely, the Arnoldi and Lanczos procedures, as well as their rational counterparts. This is the *right branch* of Figure 1.3. Finally, in Part V, ways of merging the left and right branches of this diagram were proposed; this leads to the lower part of the diagram, which is concerned with iterative SVD-based methods. As pointed out, weighted balanced methods provide a direct link between SVD-based and Krylov-based approximation methods. Finally, numerical experiments illustrating the features of the various methods were presented.

14.1 Projectors, computational complexity, and software

The unifying feature of all model reduction methods presented in this book is that they involve *projections*. Let $\Pi = \mathbf{V}\mathbf{W}^*$ be a projection, i.e., $\Pi^2 = \Pi$. The corresponding reduced-order model $\hat{\mathbf{\Sigma}}$ is obtained by means of formula (1.8):

$$\frac{d}{dt}\hat{\mathbf{x}} = (\mathbf{W}^*\mathbf{A}\mathbf{V})\hat{\mathbf{x}} + (\mathbf{W}^*\mathbf{B})\mathbf{u}, \quad \hat{\mathbf{y}} = (\mathbf{C}\mathbf{V})\hat{\mathbf{x}} + \mathbf{D}\mathbf{u}.$$

The quality of the approximant is mostly measured in terms of the frequency response $\mathbf{H}(i\omega) = \mathbf{D} + \mathbf{C}(i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ and, in particular, its *peak*, called the $\mathcal{H}_\infty$*-norm*, or its *2-norm* (energy), called the $\mathcal{H}_2$*-norm*.

Approximation by balanced truncation emphasizes the energy of the gramians expressed either in the time domain (4.43), (4.44) or in the frequency domain (4.51), (4.52). Krylov methods adapt to the frequency response and emphasize relative contributions of $\mathbf{H}$. The *choices of projectors* for some basic model reduction methods are as follows:

1. *Balanced truncation.* The key step involves solving for the gramians $\mathcal{P}$, $\mathcal{Q}$, which satisfy the Lyapunov equations: $\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^*=\mathbf{0}$, $\mathbf{A}^*\mathcal{Q}+\mathcal{Q}\mathbf{A}+\mathbf{C}^*\mathbf{C}=\mathbf{0}$. The reduced-order system is obtained by *projection* of the full-order system onto the dominant eigenspace of $\mathcal{P}\mathcal{Q}$.

2. *Optimal and suboptimal Hankel-norm approximation.* The first step is to solve for the gramians as above. Subsequently, the given system must be embedded in an all-pass (unitary) system followed by projection onto its stable eigensystem.

3. *Krylov-based approximation: Arnoldi and Lanczos procedures.* The key step here is the iterative computation of the low-order system obtained by projection onto subspaces of the reachability, observability, as well as generalized reachability, observability spaces. This leads to moment matching.

The main attributes of the two main model reduction methods are summarized next.

- **SVD-based methods:**

  1. preservation of stability,
  2. global error bound,
  3. applicability up to $n \approx 1000$, depending on computer speed and memory.

- **Krylov-based methods:**

  1. numerical efficiency,
  2. applicability: $n \gg 1000$.

It is clear from the above list that the two main model reduction methods have disjoint sets of advantages and disadvantages. Therefore, one tries to develop methods that combine the best attributes of each approach. This leads to the third class of methods, which we refer to as *SVD-Krylov methods*, and should satisfy as many of the following properties as possible.

- **SVD-Krylov methods:**

  1. preservation of stability,
  2. global error bound,
  3. numerical efficiency,
  4. applicability: $n \gg 1000$.

Complexity

The *complexity* of these methods is defined as the number of *flops* (floating point operations) required to compute the reduced system. Let n denote the dimension of the *original* system, k the dimension of the *reduced* system, and α the average number of nonzero elements per row/column of $\mathbf{A}$.

1. *Balanced truncation.*

   • Using *dense* decompositions and taking into account only the dominant terms of the total cost, approximately $70n^3$ flops are required to compute the gramians and approximately $30n^3$ flops to perform the balancing (eigen-decomposition).

   • Using *approximate* and/or *sparse* decompositions, it costs approximately αmn, $m \gg k$, flops per iteration to compute the gramians, while balancing costs approximately k^2n flops.

2. *Optimal Hankel-norm approximation.*

   • Using *dense* decompositions and taking into account only the dominant terms of the total cost, approximately $70n^3$ flops are required to compute the gramians and approximately $60n^3$ flops to perform the balancing and the embedding.

   • The *complexity* of *optimal Hankel-norm approximation* using *approximate* and/or *sparse* decompositions is as follows. The computation of gramians costs approximately αmn, $m \gg k$, flops per iteration, while the embedding costs of the order n^3 flops.

3. *Krylov approximation* requires in general approximately $k\alpha n$ flops, if sparse operations are used. Consequently, for dense operations approximately kn^2 flops are required.

   In more detail, at the jth step, the Arnoldi procedure requires αn flops for matrix-vector multiplication $\mathbf{w} = \mathbf{A}\mathbf{v}$, plus $2jn$ flops for orthogonalization. Thus k steps of Arnoldi require

$$k\alpha n + 2\frac{k(k+1)}{2}n = (\alpha + k + 1)kn \quad \text{flops.}$$

   The Lanczos procedure at the jth step requires $2\alpha n$ flops for matrix-vector multiplication $\mathbf{w} = \mathbf{A}\mathbf{v}$, plus $6n$ flops for the three-term recurrence. Thus k steps of Lanczos require

$$2k\alpha n + 6kn = 2(\alpha + 3)kn \quad \text{flops.}$$

Software

The goal with respect to model reduction is to produce software that provides reliable solutions of large-dimensional numerical problems encountered in complex applications. Turning numerically reliable algorithms into high performance, portable, and robust numerical software relies on several areas of expertise: numerical linear algebra, computer science, computer hardware, etc. For an overview of these issues, see [347].

A number of software packages are available for numerical linear algebra and model reduction purposes. We list a few (see also [91] for more general purpose software):

1. General purpose software (including model reduction):
   MATLAB: http://www.mathworks.com
   SCILAB: http://www.scilab.org
   MATRIXX: http://www.ni.com/matrixx/

2. Model reduction software: SLICOT. For details, see [59], [345], [339]. This software package can be downloaded from http://www.win.tue.nl/niconet.

3. Parallel model reduction software: PSLICOT; see [53].

4. Linear algebra package: LAPACK; see section 3.3.5 and [7].

5. For large-scale eigenvalue problems: ARPACK and its parallel implementation P_ARPACK. See [227] for details. These software packages are available at http://www.caam.rice.edu/software/ARPACK.

6. Some specialized software packages:
   SUGAR MEMS simulation: http://www-bsac.eecs.berkeley.edu/cadtools/sugar/sugar/
   MOZART ozone propagation: http://www.acd.ucar.edu/science/mozart/mozart/index.php
   ANSYS® Simulation of systems described by PDEs: http://www.ansys.com/

14.2 Open problems

Numerous open problems exist in the area of model reduction for large-scale systems. We conclude this book with a partial list. Many more problems, related, for instance, to *data assimilation* (which is of importance in model reduction for weather prediction and air quality simulations as mentioned in section 2.2.1), are omitted.

Decay rate of the Hankel singular values: General case

In section 9.4, upper bounds on the rates of decay as well as approximate rates of decay of the eigenvalues of a single Lyapunov equation were presented. Numerical examples show the usefulness of these results. Furthermore, lower bounds for the decay of the Hankel singular values (which are the squares roots of the eigenvalues of the product of the solutions of two Lyapunov equations) have also been derived. Computationally, both results require the knowledge of all eigenvalues (poles) of the system. At first, one would need to extend the upper bound to the eigenvalues of the product of two gramians. The next step would be to develop bounds on the rates of decay when only partial information about the eigenvalues is available (like an inclusion domain in the complex plane). In this regard, see Figures 9.9, 9.10, 9.11, and 9.12.

Choice of expansion points for rational Krylov methods

It section 11.3.2, we showed that given two scalar proper rational functions $\mathbf{H}_1$, $\mathbf{H}_2$ of McMillan degree $n_1 > n_2$, respectively, the latter can be obtained almost always from the former through interpolation, which can be implemented iteratively by means of the rational Krylov procedure. Therefore, in the generic SISO case, *any* reduced-order model is attainable through appropriate choice of interpolation points. The question arises of *how* to choose the interpolation points so as to minimize a desirable error criterion. Krylov and rational Krylov methods provide no guidance in that respect. In our considerations, the formula for the $\mathcal{H}_2$-norm of the system derived in section 5.5.2 suggests that to keep this norm small, the interpolation points should be chosen as the *mirror image* of some of the poles of the original system. (See the considerations in section 11.2.2.) The first problem that arises in this regard is to determine *which* poles of the original system to use. In other words, which choice of mirror images will minimize, for instance, the $\mathcal{H}_2$-norm of the error system? More generally, are there other rules for choosing the interpolation points so that a desirable norm of the error system is minimized?

Choice of spectral zeros in passive model reduction by interpolation

In section 11.3.3, it is shown that by choosing the interpolation points as spectral zeros of the original passive system, the reduced system will be automatically passive and stable. As in the previous problem, the issue that arises is to choose those spectral zeros that will reduce or minimize some appropriate norm of the error system. For an illustration of this issue, see Example 11.14.

Weighted SVD and Krylov-based model reduction

Weighted SVD methods have been presented in section 7.6. One unexplored issue is the influence of frequency weights on the sensitivity of the reduced system; furthermore, ways of computing the logarithm of the resolvent in large-scale settings are lacking. In contrast to unweighted methods, stability and error bounds are issues that have not been satisfactorily resolved and hence require further investigation. Furthermore, the problem of weighted Krylov methods is less well understood; one approach is given in [133]. Again, stability and error bounds are important aspects of this problem that need further study.

Iterative SVD model reduction methods

In sections 12.3 and 12.4, iterative methods for computing SVD-type approximants were proposed. The open problems are convergence of the first algorithm presented as well as stability. The second algorithm is guaranteed to converge to a reduced system which is balanced; however, it is not clear how many steps are needed for the reduced system, which is approximately balanced, to be stable. The balanced canonical form discussed in section 7.4 may turn out to be important in this case. Furthermore, iterative methods for weighted balanced reduction are missing.

Model reduction of second-order systems

In many cases, e.g., when mechanical systems modeled as mass-spring-damper are involved, the reduction should respect the second-order structure (the state is composed of positions and the corresponding velocities). Converting the system to first order and applying existing methods for model reduction is an option, but it is done at the expense of destroying the second-order structure. (The reduced system may no longer be representable as a mass-spring-damper system.) For an overview of this problem from an eigenvalue viewpoint, see [325]. From a system theoretic viewpoint, the problem was originally investigated in [239]. Recently, Chahlaoui et al. [84] have revisited the problem. Issues still remaining are stability, error bounds, the choice of gramians, and iterative methods. Approaches to Krylov methods for second-order systems can be found in [44], [36]. Weighted reduction for this class of systems remains an open problem.

Model reduction of differential algebraic equations or descriptor systems

The solution of dynamical systems described by differential algebraic equations has been investigated in [78]. Concerning model reduction of linear descriptor system, all rational Krylov methods can be readily extended to handle such systems; the difference is that in the shift invert part $(s\mathbf{I}-\mathbf{A})^{-1}$ is replaced by $(s\mathbf{E}-\mathbf{A})^{-1}$, where $\mathbf{E}$ is a singular matrix. Many aspects of balanced methods for descriptor systems have been worked out in [317], [318], [319], [320]. A descriptor toolbox is also available [343].

Model reduction for linear time-varying systems

The problem of model reduction of linear, discrete-time, periodically time-varying systems using SVD-type methods has been studied extensively. For an account, see [344] and references therein. Krylov methods for the same class of systems have been investigated in [183]; see also [211]. Balanced truncation for continuous-time, time-varying systems has been worked out in [217] and [286]; the latter reference treats the discrete-time, time-varying case. What needs further investigation is the numerical implementation of these approaches to systems of high complexity.

Model reduction of nonlinear systems

The study of model reduction methods for linear time-varying systems is the first step toward the systematic study of model reduction for nonlinear systems. Currently, the only systematic and widely used method is model reduction by means of POD; see section 9.1 for some details on the advantages and drawbacks of this method and a brief survey of the associated literature. Error bounds, stability, and iterative methods are issues that need to be investigated, perhaps for special classes of nonlinear systems.

Model reduction of structured interconnected systems, with application to MEMS

The difficulty in analyzing and synthesizing MEMS, also known as *microsystems*, stems from the complexity of the underlying structure as well as its heterogeneity; such a system may contain mechanical, electrical, fluidic, thermal, acoustic, and other components. This leads to multidomain and multiphysics problems.

The development of microsystems begins with modeling. The spatial distribution leads typically to PDEs describing the various parts (electrical, mechanical, etc.). Subsequently the interactions between the various parts are modeled (e.g., electromechanical, electrothermal). These equations are then turned into systems of ODEs through discretization. In addition, small signal behavior is often not enough to describe the system, due to nonlinearities, and furthermore there may be several operating points. This leads to long iteration cycles. The procedure from design to manufacturing to testing is time-consuming. Hence rapid prototyping is not feasible.

These issues can be addressed by reduction of the order of the microsystem model. In such a procedure, the various physical systems (electrical, mechanical, etc.) must be reduced separately and the reduced models put together by means of the interconnections. Furthermore, the procedure should produce reduced-order systems which have a physical interpretation as electrical, mechanical, etc. The overall goal is to understand as many effects occurring in microsystems as possible through simulation. The following references can be consulted for details on this challenging problem: [294], [278], [290], [291].

Controller design for high-order systems

Very often a given system must be controlled to achieve desired performance objectives. Roughly speaking, the complexity of the controller is the same as that of the plant. Therefore, if the latter is high, so will be the former. For implementation purposes, however, the controller must usually have low complexity, leading to controller complexity reduction. Another possible approach to the design of low complexity controllers is first to reduce the order of the system to be controlled and then compute a controller based on the reduced system. This can be dangerous, however, as the low-order controller may fail to stabilize the original system. These issues are discussed in detail in [252].

If, however, the order of the given system is very high (in the thousands of state variables or above), such an approach may not be feasible. In this case, some reduction of the original system is required in order to get a start on the computation of the controller. Depending on the underlying equations (e.g., Navier–Stokes, Burgers, heat, etc., and linearity/nonlinearity), such problems have to be investigated individually. Thus, besides computational aspects, issues concerning robustness need to be addressed, in particular when the low-order controller is hooked up with the original system (potentially described by PDEs).

Such problems have been studied (although not exhaustively) in the computational fluid dynamics community. (See section 9.1 for a few references in this direction.) Further work is necessary in connection with problems arising, e.g., in microsystems.

Model reduction of uncertain systems

Sometimes at the modeling stage, one ends up with a nominal system which is linear, together with uncertainties of the modeling parameters involved; these uncertainties can in numerous cases be expressed in terms of linear fractional transformations. The ensuing model reduction problem is solvable in principle, using LMIs. For details on the use of the linear fractional transformation framework and the solution of the associated model reduction problems in terms of LMIs, see [47] and references therein.

As pointed out in the introduction, problems involving LMIs lead to algorithms whose computational complexity is usually between $n^4 \cdots n^6$. Thus they cannot be applied at present to even moderate-size problems. Work is therefore needed to derive methodologies for the reduction of uncertain systems which lead to algorithms of low computational complexity.

This page intentionally left blank

Chapter 15

Problems

Problem 1. Show that the (unilateral) Laplace transform of the formula (4.16) for the matrix exponential is in agreement with formula (4.74), i.e.,

$$\mathcal{L}(\mathbf{f}_i)(s) = \frac{\mathbf{q}^{(i)}(s)}{\mathbf{q}(s)}, \qquad i = 1, \ldots, n,$$

where $\mathbf{q}$ is the characteristic polynomial of $\mathbf{F}$, and $\mathbf{q}^{(i)}$ are the pseudoderivative polynomials of $\mathbf{q}$ defined by (4.72).

Problem 2. Given is

$$\mathbf{F} = \begin{pmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & & 1 \\ -q_0 & -q_1 & -q_2 & \cdots & -q_{n-1} \end{pmatrix} \in \mathbb{R}^{n \times n}.$$

Let $\mathbf{q}(s) = q_0 + q_1 s + \cdots + q_{n-1}s^{n-1} + s^n$ be the characteristic polynomial of $\mathbf{F}$. Show that

$$\mathbf{q}(s)\,(s\mathbf{I}_n - \mathbf{F})^{-1} = \left\{ \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{n-1} \end{bmatrix} \begin{bmatrix} \mathbf{q}^{(1)}(s) & \mathbf{q}^{(2)}(s) & \cdots & \mathbf{q}^{(n-1)}(s) & \mathbf{q}^{(n)}(s) \end{bmatrix} \right\} \operatorname{mod} \mathbf{q}(s).$$

Hence show that (4.74) holds for $\mathbf{F}$ in companion form as given above, and consequently adj $(\mathbf{F}) = -\mathbf{q}^{(1)}(\mathbf{F})$, where $\mathbf{q}^{(1)}(s)$ is the first pseudoderivative of $\mathbf{q}(s)$, defined in (4.72). In addition, prove that (4.71) is indeed a realization of $\frac{\mathbf{p}(s)}{\mathbf{q}(s)}$.

Problem 3. Given is the 2×2 rational matrix:

$$\mathbf{H}(s) = \begin{pmatrix} \frac{(2\alpha+1)s^2+\alpha s+\alpha-1}{s^2(s-1)} & \frac{(2\alpha-3)s+1-\alpha}{s^2} \\ \frac{1}{s} & \frac{1}{s} \end{pmatrix}, \qquad \alpha \in \mathbb{R}.$$

Using Silverman's algorithm, compute a minimal realization $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ of $\mathbf{H}$, as a function of α. For $\alpha = 1$, compute also a minimal realization using MATLAB routines, and show that it is equivalent to the realization obtained via the Silverman algorithm.

Problem 4. Using the Silverman algorithm, find a minimal realization of the sequence of natural numbers where all multiples of 3 have been replaced by zero:

$$\mathcal{S} = \{1, 2, 0, 4, 5, 0, 7, 8, 0, \ldots\}.$$

In the above sequence, assume that $h_0 = 0$. Compute the sum of the series $\sum_{n>0} h_n s^{-n}$. Give the minimal recursion polynomial as well as an expression for *all* recursion polynomials of the sequence $\mathcal{S}$.

Problem 5. Consider the Lyapunov equation

$$\mathbf{AX} + \mathbf{XA}^* = \mathbf{Q},$$

where all matrices are $n \times n$.

(a) Provided that $\mathbf{Q} \geq 0$ (positive semidefinite and no assumption of reachability is made), show that $\pi(\mathbf{A}) = n$ implies $\nu(\mathbf{X}) = 0$ and $\nu(\mathbf{A}) = n$ implies $\pi(\mathbf{X}) = 0$.

(b) If $\delta(\mathbf{A}) = n$ (all eigenvalues of $\mathbf{A}$ are on the imaginary axis), show that $\mathbf{Q} = \mathbf{0}$.

Problem 6. Propose a remedy for the transformation of the continuous-time to the discrete-time Lyapunov equation discussed in section 6.1.7 of the notes, when the matrix $\mathbf{A}$ has an eigenvalue equal to 1.

Problem 7. Let $\mathbf{AV} = \mathbf{VH}$, where $\mathbf{A}$, $\mathbf{V}$, $\mathbf{H}$ are real and square matrices. Prove that $\mathbf{V}$ is nonsingular if and only if $\mathbf{A}$ is similar to $\mathbf{H}$.

Problem 8. Find the SVDs and compute the rank one approximations, which are optimal in the 2-norm, of the matrices listed below. Use paper and pencil.

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix}, \qquad \mathbf{A}_2 = \begin{pmatrix} 6 & 3 \\ -1 & 2 \end{pmatrix},$$

$$\mathbf{A}_3 = \begin{pmatrix} 2 & 3 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{A}_4 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

Problem 9. (a) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\det \mathbf{A} \neq 0$. What is the relationship between the singular values of $\mathbf{A}$ and $\mathbf{A}^{-1}$?

(b) Let $\lambda \in \mathbb{R}$ be an eigenvalue of $\mathbf{A}$. Show that $\sigma_n(\mathbf{A}) \leq |\lambda| \leq \sigma_1(\mathbf{A})$.

(c) Find the singular values of $\mathbf{A} = \begin{pmatrix} p & -q \\ q & p \end{pmatrix}$. Explain your answer geometrically. Find $\mathbf{u} \in \mathbb{R}^2$, $\|\mathbf{u}\|_2 = 1$, such that $\|\mathbf{Au}\|_2 = \sigma_1$. Explain.

Problem 10. (a) Show by direct computation that the least squares solution of $\mathbf{Ax} = \mathbf{b}$ is given by

$$\mathbf{x}_{LS} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}, \qquad \mathbf{A} \in \mathbb{R}^{n\times m},\ n \geq m,\ \operatorname{rank}\mathbf{A} = m.$$

(b) Using the SVD of $\mathbf{A} = \mathbf{U\Sigma V}^*$, show that

$$\mathbf{x}_{LS} = \sum_{i=1}^{m} \frac{\mathbf{u}_i^*\mathbf{b}}{\sigma_i}\mathbf{v}_i \qquad \text{and} \qquad \|\mathbf{Ax}_{LS} - \mathbf{b}\|^2 = \sum_{i=m+1}^{n} (\mathbf{u}_i^*\mathbf{b})^2.$$

What is the geometrical interpretation of this problem? Do $\mathbf{Ax}_{LS}$ and $\mathbf{b} - \mathbf{Ax}_{LS}$ lie in the im(A), ker(A) or in a space perpendicular to im(A)? Justify your results.

(c) Using these formulas, find the least squares solution of $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \frac{1}{10}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 3 & -4 \\ 4 & 3 \end{pmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}.$$

Problem 11. (a) Show that if $\mathbf{A} \in \mathbb{R}^{m\times n}$, then $\|\mathbf{A}\|_2 \leq \sqrt{\|\mathbf{A}\|_1\|\mathbf{A}\|_\infty}$.

(b) Suppose $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^n$. Show that if $\mathbf{E} = \mathbf{uv}^*$ and $\mathbf{v}^*\mathbf{u} = 1$, then

$$\|\mathbf{E}\|_F = \|\mathbf{E}\|_2 = \|\mathbf{I} - \mathbf{E}\|_2 = \|\mathbf{u}\|_2\|\mathbf{v}\|_2.$$

Problem 12. For any A, let E be a matrix such that $\|\mathbf{E}\|_2 < \sigma_{min}(\mathbf{A})$, then prove that rank $(\mathbf{A} + \mathbf{E}) \geq \operatorname{rank}(\mathbf{A})$.

Problem 13. Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^m$ and assume $\mathbf{A} \in \mathbb{R}^{m\times n}$. Show that if $\operatorname{rank}\mathbf{A} = n$, then $\|\mathbf{x}\|_\mathbf{A} = \|\mathbf{Ax}\|$ is a vector norm on $\mathbb{R}^n$.

Problem 14. Given is a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{n\times n}$ and a nonsingular upper triangular matrix $\mathbf{T} = (t_{ij}) \in \mathbb{R}^{n\times n}$. The following relationships are to be shown for $p = 1, 2, \infty$:

(a) $\|\mathbf{A}\|_p \geq |a_{ij}|$.

(b) $\|\mathbf{T}^{-1}\|_p \geq \frac{1}{\min_i |t_{ii}|}$.

(c) Therefore the p-condition number of T satisfies

$$\kappa_p(\mathbf{T}) \geq \frac{\max_i |t_{ii}|}{\min_j |t_{jj}|}.$$

Problem 15. Approximation of the `clown.mat` image using MATLAB: after starting MATLAB, type

```
load clown;
Z = ind2gray(X, map);
[mz,nz] = size(Z);
imshow(Z,64);
mag = 2;
truesize(1, [mz*mag, nz*mag]);\\[-5mm]
```

Compute the SVD of **Z** using the command `[U,S,V] = svd(Z);`

1. Plot the singular values of **Z** on a logarithmic scale.
2. Compute approximant having error less than 10%, 5%, 2% of the largest singular value of **Z**. What is the rank of the corresponding approximants? Also, for each case compute the compression ratio (compression ratio is defined as the number of bytes required to store the approximant divided by the original image size in bytes.)
3. Now tile the image into four equal pieces. For each of the above errors, use SVD to approximate the subimages and then reconstruct the complete image from them. Compute the compression ratio for this image. Compute the 2-norm error of the approximant and then compare the result with the previous one. Which method is better? Which one requires more computations?
4. Attach to your results the original image, the approximant from the first method and second method, for the case in which the error is less than 2%.

Problem 16. Prove that the 1-Schatten norm of a matrix, also known as the *trace norm*,

$$\| \mathbf{M} \|_{\text{trace}} = \sum_{i=1}^{n} \sigma_i(\mathbf{M}), \qquad \mathbf{M} \in \mathbb{R}^{n\times n},$$

satisfies the triangle inequality.

Hint. First prove the following lemma. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ have singular values $\sigma_i(\mathbf{A})$, $i = 1, \ldots, n$, arranged in decreasing order, and let $\mathbf{C} \in \mathbb{R}^{n\times n}$ be a rank k partial isometry (that is, $\mathbf{C} = \mathbf{V}\mathbf{W}^*$, $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n\times k}$, $\mathbf{V}^*\mathbf{V} = \mathbf{I}_k$, and $\mathbf{W}^*\mathbf{W} = \mathbf{I}_k$). Then for each $k = 1, \ldots, n$, we have $\sum_{i=1}^{k} \sigma_i(\mathbf{A}) = \max_{\mathbf{C}} \mid \text{trace}\,(\mathbf{AC}) \mid$.

Problem 17. Find the Frobenius norm of the identity $\mathbf{I} \in \mathbb{R}^{n\times n}$. Show that the p, q induced norm of the identity matrix satisfies $\|\mathbf{I}\|_{p,q} \geq 1$. Compute this induced norm as a function of p, q, n. In particular, show that the $(1, 1)$ induced norm of the identity is equal to 1.

Problem 18. **(a)** Prove that if $\| \cdot \|$ is a vector norm, so is $\gamma \| \cdot \|$ for any fixed $C > 0$.

(b) Prove that if $\|\cdot\|$ is a matrix norm satisfying the submultiplicativity property (3.8), then $\gamma\|\cdot\|$ is also a matrix norm satisfying (3.8) for any fixed $\gamma \geq 1$.

(c) For which γ is the scaled norm $\gamma\|\cdot\|_{(1,2)}$ an induced norm? Answer the same question for $\gamma\|\cdot\|_{(1,1)}$.

Problem 19. Prove that the p, q induced norms defined by (3.4) satisfy the submultiplicativity property (3.8), that is,

$$\|\mathbf{M}\|_{p,q} \leq \|\mathbf{M}\|_{p,r}\,\|\mathbf{M}\|_{r,q}.$$

Problem 20. Prove that a unitarily invariant norm of $\mathbf{M} \in \mathbb{R}^{n\times m}$ (i.e., a norm satisfying $\|\mathbf{UMV}\| = \|\mathbf{M}\|$ for all unitary $\mathbf{U}$, $\mathbf{V}$) is a convex function of its singular values (a function is called convex if $\mathbf{f}(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha\mathbf{f}(\mathbf{x}) + (1-\alpha)\mathbf{f}(\mathbf{y})$ for all $\mathbf{x}$, $\mathbf{y}$, and scalars α).

Problem 21. Prove the equivalence of the following three statements: **(1)** $\mathbf{P} \in \mathbb{R}^{n\times n}$ is an orthogonal projection; **(2)** the projection is symmetric $\mathbf{P} = \mathbf{P}^*$; **(3)** the spectral norm of $\mathbf{P}$ is one: $\|\mathbf{P}\|_2 = 1$.

Problem 22. Consider the class $\mathcal{M}_n$ of all square matrices of size n with real entries. An inner product is defined in $\mathcal{M}_n$ as follows:

$$\langle \mathbf{X}, \mathbf{Y}\rangle = \text{trace}\ (\mathbf{X}^*\mathbf{Y}),$$

where * denotes complex conjugation and transposition. This inner product induces a norm on $\mathcal{M}_n$ which is the *Frobenius norm.*

(a) Find the distance δ_0 between a *fixed* $\mathbf{X}_0 \in \mathcal{M}_n$ and the set of all multiples of $n \times n$ orthogonal matrices:

$$\delta_0 = \inf_{\alpha,\mathbf{U}} \|\mathbf{X}_0 - \alpha\mathbf{U}\|_F,$$

where the infimum is taken over all scalars α and orthogonal matrices $\mathbf{U}$.

(b) Find if possible, α_0 and $\mathbf{U}_0$ such that

$$\delta_0 = \|\mathbf{X}_0 - \alpha_0\mathbf{U}_0\|_F.$$

What is α_0 and $\mathbf{U}_0$ if $\mathbf{X}_0$ is a symmetric matrix?

Problem 23. Prove Proposition 7.5.

Problem 24. Given $\mathbf{A} \in \mathbb{R}^{n\times n}$, prove that the conditions below are equivalent:

- $\mathbf{A} + \mathbf{A}^* \leq \mathbf{0}$,
- $\| e^{\mathbf{A}t} \|_2 \leq 1,\ t > 0,$
- $\exists \mathbf{Q} = \mathbf{Q}^* > 0$ such that $\mathbf{AQ} + \mathbf{QA}^* \leq \mathbf{0}$ and $\mathbf{A}^*\mathbf{Q} + \mathbf{QA} \leq \mathbf{0}$.

Problem 25. A discrete-time system Σ is described by its impulse response

$$\mathbf{h}(k) = 2^{-k},\ k = 0, 1, 2, \ldots, \text{ and } \mathbf{h}(k) = 0,\ k < 0.$$

(a) Compute the singular values of the associated operator $\mathcal{S}$ using the results of section 5.2.

(b) Compute a matrix representation for $\mathcal{S}$ and $\mathcal{S}^*\mathcal{S}$ explicitly. Using your favorite software package, compute successive approximations of the spectrum of this operator by computing the eigenvalues of finite symmetric $n \times n$ submatrices $\mathbf{M}_n$ of $\mathcal{S}^*\mathcal{S}$ for $n = 10, 20, 50, 100, 200, 500, 1000$. Plot the eigenvalues. Compare with the result in the first part of this problem.

Problem 26. Consider the system described by $\mathbf{H}(s) = \frac{1}{s+1}$. Find its $\mathcal{L}_2(0, T)$-norm using (5.26) for $T = 1, 10, 100$. Compare with the $\mathcal{L}_2(0, \infty)$-norm.

Transform this system in discrete time and compute the same norm using the operator $\mathcal{K}$ defined following formula (5.26). Compare the results.

Problem 27. Consider the rational functions:

$$\mathbf{G}(s) = \frac{-44.1s^3 + 334s^2 + 1034s + 390}{s^6 + 20s^5 + 155s^4 + 586s^3 + 1115s^2 + 1034s + 390},$$

$$\mathbf{H}_N(s) = \sum_{i=1}^{N} \frac{2^i}{s + 2^i}.$$

(a) Using MATLAB (or equivalent), compute the *Hankel* singular values of $\mathbf{G}$. Furthermore, compute a realization of $\mathbf{G}$ which is *balanced* (in the Lyapunov sense). Hence determine balanced truncations $\mathbf{G}_{bal,k}$, for $k = 2, 4$.

(b) Using MATLAB (or equivalent), compute the $\mathcal{H}_\infty$-norms of the error systems $\mathbf{G} - \mathbf{G}_{bal,k}$ for $k = 2, 4$. Compare them with the theoretical upper bounds. Plot the amplitude Bode plots of $\mathbf{G}$, $\mathbf{G}_{bal,k}$, $\mathbf{G} - \mathbf{G}_{bal,k}$ for $k = 2, 4$.

(c) Using MATLAB (or equivalent), compute the reachability and observability gramians of $\mathbf{H}_N$ for $N = 2, 4, 6$. Hence compute the Hankel singular values of $\mathbf{H}_N$ for $N = 2, 4, 6$. Compute a balanced truncation of order one, for $N = 2, 4, 6$. What is the $\mathcal{H}_\infty$-error in each case? How does it compare with the theoretical value? Compute a balanced truncation of order two for $N = 4, 6$. Again compare the actual error with the upper bound predicted by the theory. Finally, tabulate the poles and the zeros of $\mathbf{H}_N$, $N = 2, 4, 6$, and compare with poles and zeros of the various approximants.

Problem 28. *Finite horizon balancing.* Instead of the infinite horizon control and observation energy windows, use

$$\int_0^T \| \mathbf{u}(\tau) \|_2^2 \, d\tau, \quad \int_0^T \| \mathbf{y}(\tau) \|_2^2 \, d\tau.$$

Describe the resulting balancing and reduction algorithms.

• Consider the system with transfer function $\mathbf{H}_4$ given in the previous problem. Let $\mathbf{H}_{2,T}$ be the second-order system obtained from $\mathbf{H}$ by means of length-T finite horizon balanced truncation.

(a) Plot the impulse response of $\mathbf{H}_4$ and $\mathbf{H}_{2,T}$ for $T = \frac{1}{10}, 1, 10, \infty$, on the same plot. Can you describe what happens as $T \to 0$?

(b) Plot the amplitude Bode plots of the corresponding error systems. What is their $\mathcal{H}_\infty$-norm?

• Discuss finite-horizon balancing for $T = 1$ of the unstable system

$$\mathbf{G}(s) = \frac{1}{s^2 - \alpha^2}, \qquad \alpha \in \mathbb{R},$$

by reducing it to a first-order system.

Problem 29. *Balanced and Hankel approximants.* Consider the rational function

$$\mathbf{G}(s) = \frac{-44s^3 + 44.4s^2 + 3803s + 1449}{s^6 + 29.1s^5 + 297s^4 + 1805s^3 + 9882s^2 + 19015s + 7245}.$$

1. Compute the *Hankel* singular values of $\mathbf{G}$.
2. Compute a realization of $\mathbf{G}$ which is *balanced* (in the Lyapunov sense). Hence determine balanced truncations $\mathbf{G}_{bal,k}$ for $k = 2, 4$.
3. Compute the *optimal Hankel-norm* approximations $\mathbf{G}_{H,k}$ for $k = 2, 4$.
4. Compute the $\mathcal{H}_\infty$-norms of the error systems $\mathbf{G} - \mathbf{G}_{bal,k}$, $\mathbf{G} - \mathbf{G}_{H,k}$ for $k = 2, 4$. Compare these norms with the theoretical upper bounds. Plot the amplitude Bode diagrams and the Nyquist diagrams of $\mathbf{G}$, the various approximants, and the various error systems.
5. Compute the dyadic decomposition of the rational function

$$\mathbf{K}(s) = \frac{1}{(s+1)(s+2)(s+3)}.$$

Compare the amplitude Bode plots of the first-order approximants obtained by balanced truncation, optimal Hankel-norm approximation, and truncation of the dyadic decomposition.

Problem 30. Give a proof of Proposition 5.4.

Hint. First, by scaling, that is, substituting $\mathbf{D}$, $\mathbf{B}$, $\mathbf{C}$, for $\frac{1}{\gamma}\mathbf{D}$, $\frac{1}{\sqrt{\gamma}}\mathbf{B}$, $\frac{1}{\sqrt{\gamma}}\mathbf{C}$, respectively, the problem can be reduced to the case $\gamma = 1$.

Next show that

$$\left[\begin{array}{c|c} \tilde{\mathbf{A}} & \tilde{\mathbf{B}} \\ \hline \tilde{\mathbf{C}} & \tilde{\mathbf{D}} \end{array}\right] = \left[\begin{array}{cc|c} \mathbf{A} & \mathbf{0} & \mathbf{B}\mathbf{D}^* + \mathcal{P}\mathbf{C}^* \\ \mathbf{0} & -\mathbf{A}^* & \mathbf{C}^* \\ \hline -\mathbf{C} & \mathbf{C}\mathcal{P} + \mathbf{D}\mathbf{B}^* & \mathbf{I} - \mathbf{D}\mathbf{D}^* \end{array}\right]$$

is a realization of Φ, where $\mathcal{P}$ is the reachability gramian $\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^* = \mathbf{0}$. Using the state space transformation $\mathbf{T} = \begin{pmatrix} \mathbf{I} & \mathcal{P} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$, compute the equivalent realization $\mathbf{A}' = \mathbf{T}^{-1}\tilde{\mathbf{A}}\mathbf{T}$, $\mathbf{B}' = \mathbf{T}^{-1}\tilde{\mathbf{B}}$, $\mathbf{C}' = \tilde{\mathbf{C}}\mathbf{T}^{-1}$, $\mathbf{D}' = \tilde{\mathbf{D}}$. Finally, provided that $\mathbf{D}$ is nonsingular, use the formula

$$[\mathbf{D}+\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}]^{-1} = \mathbf{D}^{-1}[\mathbf{D}+\mathbf{C}(s\mathbf{I}-\mathbf{A}+\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}]\mathbf{D}^{-1}$$

to prove the desired formula (5.18).

Problem 31. This problem refers to Example 5.28. First derive a state variable representation of the system where the four states should be chosen as $x_1 = q_1, x_2 = \dot{q}_1 - \frac{b_1}{m_1}u_1$, $x_3 = q_2$, $x_4 = \dot{q}_2$; the inputs are $u_1 = q_0$, $u_2 = f$; and the output (observation) is $y = \dot{q}_2$. Show that the system is observable if $\frac{b_1}{k_1} \neq \frac{b2}{k2}$. If this assumption is satisfied, express the state as a function of the input, output, and derivatives thereof and consequently the supply function in terms of the input and the output. Finally, assuming that $m_1 = m_2 = 1$, $b_1 = 0$, $b_2 = 1$, $k_1 = k_2 = 1$, use Theorem 5.26 to derive an expression for the available storage and the required supply and compare each one with the storage function given in the above mentioned example.

Problem 32. *Comparison of $\mathcal{H}_\infty$-norm bounds* [149].

- Consider the systems described by

$$\mathbf{H}(s) = \sum_{i=1}^{n} \frac{a^{2i}}{s+a^{2i}}, \qquad a > 0.$$

1. Compute a minimal realization $\left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right)$ of $\mathbf{H}(s)$, where $\mathbf{A}$ is diagonal and $\mathbf{B} = \mathbf{C}^*$. Hence show that the two gramians can be chosen as follows:

$$\mathcal{P} = \mathcal{Q} = \left[\frac{a^{i+j}}{a^{2i}+a^{2j}}\right].$$

2. Deduce that the sum of the Hankel singular values of this system is $\frac{n}{2}$. Moreover, the $\mathcal{H}_\infty$-norm of $\mathbf{H}$ is n. Hence in this case the upper bound given by formula (8.20) is tight.

   The system under consideration has the property that its poles and zeros are on the negative real axis and interlace one another. It can be shown in general (see [312]) that for such systems the error bound (8.20) is tight.

3. What conclusion can you draw concerning the error (8.21) between $\mathbf{H}$ and a reduced model $\mathbf{H}_{bal,k}$ of McMillan degree $k < n$ obtained by balanced truncation?

- Consider a system in balanced canonical form given by formula (7.24), where $n = 5, \sigma_1 = 1, \sigma_i = \sigma_{i-1} - 10^{-3}, i = 2, \ldots, n$.

1. Compute the $\mathcal{H}_\infty$-norm of the system for various choices of the γ_i and s_i; compare this norm with the upper bound given by formula (8.20).

2. For a specific choice of the γ_i and s_i, repeat the above exercise, considering the reduced-order system of order 2, obtained by balanced truncation; again compare the upper bound given by formula (8.21) with the actual $\mathcal{H}_\infty$-norm of the error system.

Problem 33. *Continuity of suboptimal approximants.* Consider the linear system defined as follows:

$$\mathbf{A} = \begin{pmatrix} -\frac{\alpha^2}{2\sigma_1} & -\frac{\alpha\beta}{\sigma_1+\sigma_2} \\ -\frac{\alpha\beta}{\sigma_1+\sigma_2} & -\frac{\beta^2}{2\sigma_2} \end{pmatrix}, \quad \mathbf{B} = \mathbf{C}^* = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \qquad \alpha, \beta > 0.$$

Compute the *Hankel* singular values of this system. Hence give a general expression for the transfer function of a γ-suboptimal Hankel-norm approximation with γ between the two singular values. What is the Hankel-norm of this first-order system? Examine in detail the case where γ approaches the smallest singular value. Is the corresponding optimal solution obtained?

Problem 34. Consider the pair $\mathbf{A} = \text{diag}\,(-1, -2, \ldots, -n)$, $\mathbf{B} = \texttt{ones}(n, 1)$, $n = 10$, $n = 50$. Find the Cholesky ordering of the eigenvalues of **A**, defined by formula (9.6). Hence compute the estimate of the decay rate of the eigenvalues of the reachability gramian $\mathcal{P}$ (which satisfies $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$) according to formula (9.7) and the decay rate bound given by (9.8). By computing $\mathcal{P}$, compare the estimate and the bound with the true decay rate of the eigenvalues.

Problem 35. Repeat the previous exercise if $\mathbf{A} = \text{diag}\,(\mathbf{A}_1, \ldots, \mathbf{A}_N)$, where $\mathbf{A}_k \in \mathbb{R}^{2\times 2}$ has eigenvalues $-k \pm ik\gamma$, $k = -1, \ldots, -N$, where $\gamma \geq 0$; **B** is as before, a matrix of ones. In other words, the eigenvalues of **A** consist of complex conjugate pairs that lie in the left half of the complex plane, on two straight lines with slope $\pm\gamma$. This situation is common in structural systems with proportional damping.

Problem 36. The diffusion of heat through a perfectly insulated, heat-conducting rod shown in Figure 15.1, is described by the following linear partial differential equation:

$$\frac{\partial \mathbf{T}}{\partial t}(x, t) = \frac{\partial^2 \mathbf{T}}{\partial x^2}(x, t), \qquad t \geq 0,\ x \in [0, 1],$$

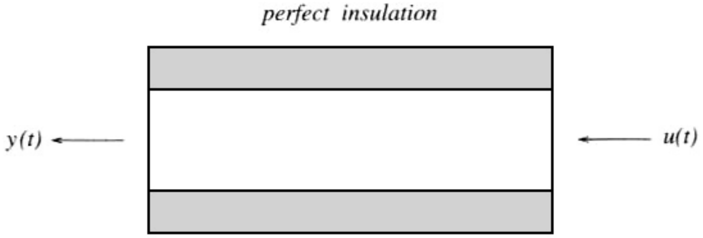


Figure 15.1. *Heat-conducting rod.*

where $\mathbf{T}(x,t)$ is the temperature at distance x from the origin and time t. The boundary conditions are

$$\frac{\partial \mathbf{T}}{\partial x}(0,t) = 0 \quad \text{and} \quad \frac{\partial \mathbf{T}}{\partial x}(1,t) = u(t),$$

where $\mathbf{u}$ is the input function (supplied heat). Finally, the output (which is measured) is $\mathbf{y}(t) = \mathbf{T}(0,t)$.

(i) Show that the transfer function $\mathbf{Z}(s) = \frac{\hat{\mathbf{y}}(s)}{\hat{\mathbf{u}}(s)}$ of the above system is given by

$$\mathbf{Z}(s) = \frac{1}{\sqrt{s}\,\sinh\sqrt{s}}.$$

Hence conclude that the poles of $\mathbf{Z}(s)$ are

$$s = -k^2\pi^2, \qquad k = 0, 1, 2, \ldots.$$

(ii) Show that $\mathbf{Z}(s)$ has the following partial fraction expansion:

$$\mathbf{Z}(s) = \sum_{k=0}^{\infty} \frac{\alpha_k}{s + k^2\pi^2}, \qquad \alpha_0 = 1, \ \ \alpha_k = (-1)^k 2, \ k > 0.$$

Conclude that the impulse response of the above system is

$$\mathbf{z}(t) \ = \ 1 \ + \ 2\mathbf{h}(t), \ \ \text{where } \mathbf{h}(t) = \sum_{k=1}^{\infty}(-1)^k e^{-k^2\pi^2 t}, \quad t \geq 0.$$

The objective of this problem is to investigate three different approximations of $\mathbf{z}$ *as a linear combination of three exponentials.*

(iii) The first approximation is the *modal approximation*, which is simply

$$\mathbf{z}_{mod}(t) = 1 + 2(-e^{-\pi^2 t} + e^{-4\pi^2 t}), \qquad t \geq 0.$$

(iv) Let $\tilde{\mathbf{h}}$ be a high-order, finite-dimensional approximation of the infinite-dimensional system defined by $\mathbf{h}$, obtained as follows:

$$\tilde{\mathbf{h}}(t) = \sum_{k=1}^{8}(-1)^k e^{-k^2\pi^2 t}.$$

The next two approximations are now obtained by means of a second-order *Lyapunov balanced truncation* $\mathbf{h}_L$ and a second-order *Hankel-norm approximation* $\mathbf{h}_H$ of $\tilde{\mathbf{h}}$. Let $\mathbf{z}_L = 1 + 2\mathbf{h}_L$ and $\mathbf{z}_H = 1 + 2\mathbf{h}_H$.

(v)a. Plot $\mathbf{z}(t)$ (50 terms should be enough) together with $\tilde{\mathbf{z}}(t)$, $\mathbf{z}_{mod}(t)$, $\mathbf{z}_L(t)$, and $\mathbf{z}_H(t)$.

(v)b. Plot the step responses of the systems corresponding to the above approximations.

(v)c. Plot the Bode amplitude diagrams of the approximants in (v)a as well as those for the error systems. Compare the $\mathcal{H}_\infty$-norms of the error systems $\mathbf{h}-\mathbf{h}_L$ and $\mathbf{h}-\mathbf{h}_H$ with the upper bounds of these differences, predicted by the theory.

Problem 37. *Heat conducting rod with convection.* Assume now that in the previous example convection is taking place as well. The equation describing the evolution of the temperature $\mathbf{T}(t,x)$ along the rod is

$$\frac{\partial \mathbf{T}}{\partial t}(x,t) = \frac{\partial^2 \mathbf{T}}{\partial x^2}(x,t) - 2\eta\frac{\partial \mathbf{T}}{\partial x}(x,t), \qquad t \geq 0,\ x \in [0,1],\ \eta \geq 0.$$

The boundary conditions, and hence the input and the output of the system, are the same as before.

Show that the transfer function $\mathbf{Z} = \frac{\mathbf{Y}}{\mathbf{U}}$ in this case is

$$\mathbf{Z}(s) = \frac{1}{s\, e^{\eta}} \cdot \frac{\sqrt{s+\eta^2}}{\sinh\sqrt{s+\eta^2}}.$$

Thus, $\mathbf{Z}$ has poles at $s=0$ and $s=-\eta^2-k^2\pi^2$, $k=1,2,\ldots$. This in turn implies that the partial fraction expansion of $\mathbf{Z}$ is

$$e^{\eta}\,\mathbf{Z}(s) = \frac{\alpha_0}{s} + \sum_{k=1}^{\infty}\frac{\alpha_k}{s+\eta^2+k^2\pi^2}, \quad \alpha_0 = \frac{\eta}{\sinh\eta},$$

$$\alpha_k = 2(-1)^k\left(\frac{k^2\eta^2}{k^2\eta^2+k^2\pi^2}\right), \qquad k>0.$$

It follows that the impulse response is

$$e^{\eta}\,\mathbf{z}(t) = \alpha_0 + 2\mathbf{h}(t), t\geq 0, \ \text{ where } \mathbf{h}(t) = \sum_{k=1}^{\infty}(-1)^k\left(1+\frac{\eta^2}{k^2\pi^2}\right)e^{-(\eta^2+k^2\pi^2)t}.$$

For the convection-diffusion case $\eta \neq 0$, answer questions (iv) and (v)a–c as in the preceding problem.

Problem 38. The preceding two problems are to be solved now by discretizing the corresponding PDEs. In particular, assume that the heat conducting rod has been divided in $N+1$ intervals of length h; the resulting variables are $\mathbf{T}(kh)$, $k=0,1,\ldots,N+1$. Given the initial conditions, namely, $(\mathbf{T}(0)-\mathbf{T}(h))/h = 0$, $\mathbf{y}=\mathbf{T}(0)$, $\mathbf{u} = (\mathbf{T}((N+1)h)-\mathbf{T}(Nh))/h$, the state $\mathbf{x}$ becomes $\mathbf{x} = (\mathbf{T}(1h),\ldots,\mathbf{T}(Nh))^*$, $\in \mathbb{R}^N$, where

$$\dot{\mathbf{T}}(kh) = \frac{1}{h^2}(\mathbf{T}((k+1)h) - 2\mathbf{T}(kh) + \mathbf{T}((k-1)h))$$
$$- \frac{2\eta}{h}(\mathbf{T}(kh)-\mathbf{T}((k-1)h)), \qquad h = \frac{1}{N+1}.$$

Thus deduce that the resulting system has the form $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$, $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)$, where $\mathbf{A}$ is tridiagonal, while $\mathbf{B}$, $\mathbf{C}$ are multiples of canonical unit vectors:

$$\mathbf{A} = \frac{1}{h^2}\begin{bmatrix} -1 & 1 & 0 & & & \\ 1+2h\eta & -2(1+h\eta) & 1 & & & \\ & & & \ddots\ \ddots\ \ddots & & \\ & & & 1+2h\eta & -2(1+h\eta) & 1 \\ & & & 0 & 1+2h\eta & -(1+2h\eta) \end{bmatrix},$$

$$\mathbf{B} = \frac{1}{h}\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

while $\mathbf{C} = [1, 0, 0, \ldots, 0]$. Compare the approximants obtained by reducing this Nth-order system with those based on the transfer function approach.

Problem 39. Consider a cascade of low-pass filters:

$$\mathbf{G}_N(s) = \frac{1}{(s+1)^N}, \qquad N = 8.$$

The purpose of this problem is to study the approximation of this cascade by means of kth-order systems, $k = 1, 3, 5, 7$, in particular, by

(1) $\mathbf{G}_k(s)$,

(2) balanced truncation,

(3) unweighted Hankel-norm approximation,

(4) weighted Hankel-norm approximation with weights,

$$\mathbf{W}_1(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(s)},\ \mathbf{W}_2(s) = \frac{\mathbf{n}(s)}{\mathbf{d}(-s)},\ \mathbf{W}_3(s) = \frac{\mathbf{n}(-s)}{\mathbf{d}(s)},\ \mathbf{W}_4(s) = \frac{\mathbf{n}(-s)}{\mathbf{d}(-s)},$$

where

$$\mathbf{n}(s) = s^2 + \omega_0 s + \omega_0^2,\ \ \mathbf{d}(s) = s^2 + 2\omega_0 s + \omega_0^2, \qquad \omega_0 = .1,\ 1,\ 10.$$

Plot the magnitude of the approximation errors on one plot for each different type of approximation. Compare with the upper bounds of the error magnitudes whenever possible.

Problem 40. Consider the following sixth-order system:

$$\mathbf{G}(s) = \frac{-s+1}{s^6+3s^5+5s^4+7s^3+5s^2+3s+1}.$$

This system is doubly resonant with poles close to the imaginary axis. Find third-order approximants using the methods below:

(1) modal approximation keeping the dominant poles,

(2) balanced truncation,

(3) unweighted Hankel-norm approximation,

(4) weighted Hankel-norm approximation with weight

$$\mathbf{W}(s) = \frac{s^2-s+1}{s^2-2s+1}.$$

Plot the magnitude of the approximation errors on one plot for each different type of approximation. Compare with the upper bounds of the error magnitudes whenever possible.

Problem 41. *A property of the Arnoldi procedure.* Given the reachable pair of matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{B} \in \mathbb{R}^n$, let $\mathcal{P}$ be the reachability gramian: $\mathbf{A}\mathcal{P}+\mathcal{P}\mathbf{A}^*+\mathbf{B}\mathbf{B}^* = \mathbf{0}$. Consider the k-step Arnoldi procedure, $\mathbf{A}\mathbf{V}_k = \mathbf{V}_k\mathbf{H}_k + h_{k+1,k}\mathbf{v}_{k+1}\mathbf{e}_k^*$, and let $\mathcal{P}_k \in \mathbb{R}^{k\times k}$ be the reachability gramian of the pair $\mathbf{H}_k$, $\beta\mathbf{e}_k$, where $\beta = \|\mathbf{B}\|$: $\mathbf{H}_k\mathcal{P}_k+\mathcal{P}_k\mathbf{H}_k^*+\beta^2\mathbf{e}_k\mathbf{e}_k^* = \mathbf{0}$. The gramian $\mathcal{P}_k$ is embedded in the bigger space as follows: $\tilde{\mathcal{P}} = \mathbf{V}_k\mathcal{P}_k\mathbf{V}_k^* \in \mathbb{R}^{n\times n}$.

(i) Show that the nonzero eigenvalues of $\tilde{\mathcal{P}}$ and $\mathbf{H}_k$ are the same.

(ii) Let $\tilde{\mathbf{A}} = \mathbf{A} - \Delta$, where $\Delta = h_{k+1,k}\mathbf{v}_{k+1}\mathbf{v}_k^*$ ($\tilde{\mathbf{A}}$ is a rank one correction of $\mathbf{A}$). Show also that $\tilde{\mathcal{P}}$ satisfies the Lyapunov equation: $\tilde{\mathbf{A}}\tilde{\mathcal{P}} + \tilde{\mathcal{P}}\tilde{\mathbf{A}}^* + \mathbf{B}\mathbf{B}^* = \mathbf{0}$,

Problem 42. Prove the fundamental lemma, Lemma 10.14, of the Arnoldi procedure.

Problem 43. **(a)** Given are the matrices $\mathbf{A}$ and $\mathbf{B}$:

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

Compute the Arnoldi (Lanczos) factorization. Notice that it stops prematurely. What can one conclude concerning the eigenvalues and corresponding eigenvectors of A?

(b) Prove that for the symmetric Lanczos (Arnoldi) process, the projected matrices $\mathbf{V}_k^*\mathbf{A}\mathbf{V}_k$ have eigenvalues that lie in the interval $[\lambda_{\min}(\mathbf{A}), \lambda_{\max}(\mathbf{A})]$. Therefore, if $\mathbf{A}$ is stable, so are the resulting reduced systems.

Problem 44. Consider elliptic continuous-time low-pass filters which can be obtained in MATLAB as follows: `[A,B,C,D]=ellip(n,1,100,1,'s')`. The order of the filter should be taken $n = 25$. The reduced order is $k = 10, 15, 20$.

Hint. The realization obtained from MATLAB is not well-conditioned. It is advisable to multiply $\mathbf{C}$ by 10^4 and divide $\mathbf{B}$ by the same number.

Problem 45. Given is the system described by

$$\mathbf{A} = -10^{-3}\mathrm{diag}\,(10^3,\ 9,\ 8,\ 7,\ 6,\ 5,\ 4,\ 3,\ 2,\ 1) \in \mathbb{R}^{10\times 10},$$
$$\mathbf{B} = (1\ \ 1\ \ \cdots\ \ 1)^T \text{ and } \mathbf{C}^* = \mathbf{B} \in \mathbb{R}^{10}.$$

1. Using the square root method and the sign function method, compute the reachability and observability gramians and hence compute and plot the Hankel singular values.
2. Compute reduced-order systems of order $k = 3, 5$ by means of balanced truncation, the Arnoldi method, the Lanczos method, and modal truncation.
3. Compute the $\mathcal{H}_\infty$-norm of the resulting error systems in each case.
4. Tabulate the poles and the zeros of the approximants.
5. Plot the frequency response of the original system and the error systems (on the same plot for each k).

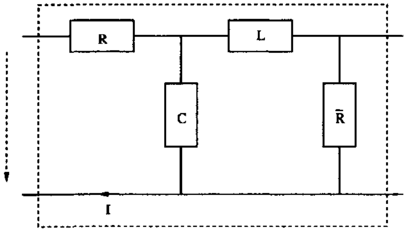


Figure 15.2. *A section of an RLC cascade circuit:* $R = .1\Omega$, $C = 1pF$, $\bar{R} = 100\Omega$, $L = 10pH$.

Problem 46. Consider a circuit consisting of 25 sections interconnected in cascade; each section is as shown in Figure 15.2. The input to this system is the voltage $\mathbf{V}$ applied to the first section; the output is the current $\mathbf{I}$ of the first section, as shown.

- Derive a state space representation of this circuit.
- Approximate this system by means of $k = 10, 20, 30$ states.

Problem 47. Consider Chebyshev I continuous-time low-pass filters which can be obtained in MATLAB as follows: `[A,B,C,D]=cheby1(n,r,1,'s')`, where r denotes the admissible ripple in the passband and should be taken as $r = 1db$ and corresponds approximately to a 10% ripple. The order of the filter should be taken $n = 1 : 25$. The first goal of this problem is to compare the three balancing transformations $\mathbf{T}_i$, $i = 1, 2, 3$, and the corresponding balanced triples:

$$\left.\begin{array}{l}\mathbf{E}_i^{(1)} = \text{norm}\,(\mathbf{I}_n - (\mathbf{T}_i)(\mathbf{T}_i^{-1}))\\ \mathbf{E}_i^{(2)} = \text{norm}\,(\mathbf{A}_i\Sigma + \Sigma\mathbf{A}_i^* + \mathbf{B}_i\mathbf{B}_i^*)\\ \mathbf{E}_i^{(3)} = \text{norm}\,(\mathbf{A}_i^*\Sigma + \Sigma\mathbf{A}_i + \mathbf{C}_i^*\mathbf{C}_i)\end{array}\right\} i = 1, 2, 3.$$

Plot the above errors as a function of the order n of the filter. Plot also the Hankel singular values for $n = 5, 10, 15, 20, 25$ on the same plot. The y-axis should be logarithmic (log10). Notice that one should transform the original triple $\mathbf{C}, \mathbf{A}, \mathbf{B}$ to the Schur basis first.

The transformations are $\mathbf{T}_1$, the usual balancing transformation; $\mathbf{T}_2$, the square-root balancing transformation; $\mathbf{T}_3$, the square root algorithm, where the balanced matrices are obtained up to diagonal scaling with $\Sigma^{-1/2}$.

Balance and truncate. Each filter above should be reduced approximately to one-quarter of its original dimension; more precisely, the reduced order should be $k = \text{round}\,(n/4) + 1$. Let $\mathbf{T}_{ik} = [I_k \;\; 0]T_i$ and $\hat{\mathbf{T}}_{ik} = \mathbf{T}_i^{-1}[\mathbf{I}_k; \;\; \mathbf{0}]$ for $i = 1, 2, 3$. Compute and plot the corresponding errors $\mathbf{E}_{ik}^{(1)}$, $\mathbf{E}_{ik}^{(2)}$, $\mathbf{E}_{ik}^{(3)}$. Plot also the condition number of the three transformations $\mathbf{T}_{ik}$, $i = 1, 2, 3$, for the above k.

Finally, given the Hankel singular values of the 25th-order filter, what are the possible dimensions of a reduced system obtained by balanced truncation so as to avoid (almost) marginally stable poles (i.e., poles on the imaginary axis)?

Problem 48. *Application of the POD method to a linear system.* Consider the following system:

$$\Sigma = \left(\begin{array}{c|c}\mathbf{A} & \mathbf{B}\\ \hline \mathbf{C} & \mathbf{D}\end{array}\right)$$

$$= \left[\begin{array}{cccccc|c} -9 & -33.0101 & -63.0705 & -66.180901 & -36.200703 & -8.080202 & 10\\ 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ \hline 0 & -20 & 0 & 0 & -10 & 10 & 0\end{array}\right].$$

The poles of this system are $-1 \pm \frac{i}{10}$, $-2 \pm \frac{i}{100}$, -1, -2. Obtain the system representation of Σ in MATLAB using the command `sys = ss(A,B,C,D)`.

Part (a)

1. Using the MATLAB command `[y,t,X] = impulse(sys,T)`, obtain the impulse response of the system, where T is the time interval, y is the output, X are the states. Choose $T = 0 : 0.1 : 10$. Notice that the matrix of snapshots $X' = [x(t_1) \;\; \cdots \;\; x(t_N)]$ in the POD method is simply given by X^*.

2. Compute the SVD of $X' = U\Sigma V^*$ and plot the *normalized* singular values on a logarithmic scale. Choose $k = 2, 3, 4, 5$ and obtain the reduced-order system for $\hat{\mathbf{x}} = \mathbf{U}_k^*\mathbf{x}$, where $\mathbf{U}_k$ is a matrix composed by the leading k columns of $\mathbf{U}$.

3. Apply the impulse response to the reduced-order system for the same time interval T as above and obtain $\mathcal{X}_k$.

4. Plot the states for both the full-order and the reduced-order model on the same plot (one state per plot). You can have more than one figure on the same page using the `subplot` command. Compute both the error norm $\|X - X_k\|_2$ and the relative error norm $\frac{\|X - X_k\|_2}{\|X\|_2}$. How do these numbers relate to the neglected singular values?

Part (b)

1. Obtain the impulse response of the full-order model for the time interval $T1 = 10$. Construct $\mathcal{X}$, compute the SVD, and again obtain the reduced-order system of order 2, 3, 4, 5 for the reduced states $\hat{\mathbf{x}} = \mathbf{U}_k^*\mathbf{x}$.

2. Continue the simulation for time interval $T2 = 1 : 5$ using the reduced system. Notice that you should use the command `initial` rather than `impulse`.

3. After obtaining data X_k, transform the states back into original space and append the data to what you had for $T1 = 0 : 0.02 : 10$. Compare the whole data you obtained with the data obtained for the full-order model for $T = 0 : 0.02 : 20$ in step 1 of Part (a) above.

4. Again plot the data and compute the error norms.

Problem 49. If we apply the sign function method to the Sylvester equation (6.1) $\mathbf{AX} + \mathbf{XB} = \mathbf{C}$, where the eigenvalues of $\mathbf{A}$ are in the (open) left half of the complex plane, the uniqueness condition is satisfied, and the eigenvalues of $\mathbf{B}$ are both in the (open) left half and (open) right half of the complex plane, we will obtain a projected form of the $\mathbf{X}$. Derive an expression for this projected $\mathbf{X}$.

Problem 50. Consider the unstable system given by

$$\Sigma = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right) = \left[\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & -2 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -\frac{1}{10} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -100 & \frac{1}{20} & 1 \\ \hline 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{array}\right].$$

Reduce this system following the steps outlined in Examples 7.24 and 7.25.

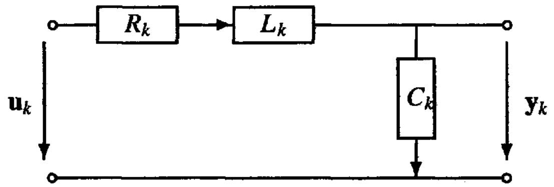


Figure 15.3. *The kth-section of an RLC circuit.*

Problem 51. This problem is concerned with an RLC circuit which consists of several sections, as shown in Figure 15.3; $\mathbf{u}_k$, $\mathbf{y}_k$ are the input, output voltage of the kth section. N sections are cascaded by letting $\mathbf{y}_k = \mathbf{u}_{k+1}$, $k = 1, \ldots, N-1$; the input of the system is $\mathbf{u}_1$ and the output is $\mathbf{y}_N$.

Write state equations by taking as state variables for the kth section the current through the inductor L_k and the voltage across the capacitor C_k.

Reduce the cascade system composed of $N = 50$ sections to a system of order 10 by means of the following methods: (a) balanced truncation, (b) Arnoldi method, (c) Lanczos method. In the last two cases, choose as interpolation (projection) points the mirror image of the poles of the system with the largest real part.

The values of the components should be chosen as follows: $R_i = 1\Omega$, $L_i = 1H$, $C_i = 10^{-6}F$.

Problem 52. Use the same set-up as for Problem 51. In this case, weighted approximations are to be compared, in particular, by taking a low-pass Butterworth filter of order 4 as weight and applying the method without weights as described in section 7.6.2. The cutoff frequency of the low-pass filter and the upper limit of integration in the second method are to be chosen to lie (a) between the first and the second resonances and (b) between the fourth and fifth resonances.

Problem 53. Use the same set-up as for Problem 51. In this case, compare the approximations by means of the three passive approximation methods described in this book. In particular, two methods have to do with positive real balanced truncation and are described in section 7.5.4, and the third is described in section 11.3.3.

Problem 54. Consider the SISO system Σ of order n, where $\mathbf{A}$ is diagonal with $-\ell$, as ℓth entry, while $\mathbf{B}^* = \mathbf{C}$ with ℓth entry equal to 1. For $n = 10$ write the solution of the Lyapunov equations explicitly. Then use the sign iteration method; how many iterations are needed for the 2-error between the two expressions to drop below machine precision?

Then, reduce the system to order $k = 2$ and compare balanced truncation and Lanczos approximation by taking as projection points the mirror image of the largest poles.

Estimate the decay rate of the Hankel singular values for this example, using the decay estimate (9.7) and the decay bounds (9.8) and (9.11).

Problem 55. Given is the pair $(\mathbf{A}, \mathbf{b})$ as in (4.71), with $\mathbf{A}$ stable (eigenvalues in the left half plane). Furthermore, let

$$\mathbf{V}(s) = \frac{1}{\chi_{\mathbf{A}}(s)} \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{\nu-1} \end{bmatrix}, \text{ where } \chi_{\mathbf{A}}(s) = \det(s\mathbf{I} - \mathbf{A}).$$

Show that the associated reachability gramian $\mathcal{P} \in \mathbb{R}^{\nu\times\nu}$, that is, the solution of the Lyapunov equation $\mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^* + \mathbf{b}\mathbf{b}^* = \mathbf{0}$, can be expressed as follows:

$$\mathcal{P} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{V}(i\omega)\mathbf{V}^*(-i\omega)\, d\omega.$$

With the notation as in (4.71), verify using the above formula that for $\nu = 2$, the gramian is

$$\mathcal{P} = \frac{1}{2q_0q_1} \begin{pmatrix} 1 & 0 \\ 0 & q_0 \end{pmatrix},$$

while for $\nu = 3$, the gramian is

$$\mathcal{P} = \frac{1}{2q_0(q_2q_1 - q_0)} \begin{pmatrix} q_2 & 0 & -q_0 \\ 0 & q_0 & 0 \\ -q_0 & 0 & q_0q_1 \end{pmatrix}.$$

Notice that the conditions for positive definiteness of the gramian are equivalent with the Routh conditions for stability of the polynomial $\chi_{\mathbf{A}}$.

Problem 56. (See [354].) Let the triple $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ be minimal with $\mathbf{A}$ stable. Let $\mathcal{X}_\epsilon$ satisfy the Riccati equation

$$\mathbf{A}^*\mathcal{X}_\epsilon + \mathcal{X}_\epsilon\mathbf{A} + \mathbf{C}^*\mathbf{C} - \epsilon^2\mathcal{X}_\epsilon\mathbf{B}\mathbf{B}^*\mathcal{X}_\epsilon = \mathbf{0},$$

where ϵ is a positive number. Denote by $\mathcal{X}_\epsilon^+ > 0$, $\mathcal{X}_\epsilon^- < 0$, the two extremal solutions of this equation. Show that

$$\lim_{\epsilon\to 0} \mathcal{X}_\epsilon^+ = \mathcal{Q}, \quad \lim_{\epsilon\to 0} \epsilon^2\mathcal{X}_\epsilon^- = -\mathcal{P}^{-1},$$

where $\mathcal{P}$, $\mathcal{Q}$ are the reachability, observability, gramians of the give triple, respectively.

Problem 57. This problem has to do with condition (11.12). Show that $\hat{\mathbf{H}}(s) = \frac{1}{s}$ can be obtained by the rational Krylov procedure from $\mathbf{H}(s) = \frac{1}{s^n}$ for $n = 3$ but not for $n = 2$. Furthermore, prove (11.12) in the general case.

Problem 58. Consider a closed curve in the plane described parametrically as $x = f(t)$, $y = g(t)$, where t is a real parameter belonging to the interval $[a, b]$. It can be shown that the area $\mathcal{A}$ enclosed by this curve is given by

$$\mathcal{A} = \frac{1}{2}\int_a^b \left[g(t)\frac{d}{dt}f(t) - f(t)\frac{d}{dt}g(t) \right] dt.$$

First show that the area $\mathcal{A}$ enclosed by the Nyquist plot of a SISO system with transfer function $\mathbf{H}(s)$ is given by half the integral of the imaginary part of the expression $\mathbf{H}(i\omega)\frac{d}{d\omega}\mathbf{H}(-i\omega)$:

$$\mathcal{A} = \frac{1}{2}\int_{-\infty}^{\infty} \mathcal{I}m \left[\mathbf{H}(i\omega)\frac{d}{d\omega}\mathbf{H}(-i\omega)\right] d\omega.$$

Consequently, verify Proposition 5.10 for the Butterworth filter given in Example 5.5.

Problem 59. Show that the square root algorithm for solving the Lyapunov equation described in section 6.3.3 holds in the case when the solution $\mathcal{P}$ is *semidefinite*, i.e., it has zero eigenvalues.

Problem 60. Show that if the pair of matrices $(\mathbf{A}, \mathbf{B})$ with $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, is reachable, the geometric multiplicity of the eigenvalues of $\mathbf{A}$ is at most m. Therefore, if this pair is reachable, $m = 1$, and $\mathbf{A}$ has a multiple eigenvalue, it corresponds to a Jordan block. Another way of saying this is that the smallest number of input channels m needed to ensure reachability given $\mathbf{A}$ is equal to the largest geometric multiplicity of the eigenvalues of $\mathbf{A}$.

This page intentionally left blank

Bibliography

[1] V.M. Adamjan, D.Z. Arov, and M.G. Krein, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR Sbornik, **15**: 31–73 (1971).

[2] V.M. Adamjan, D.Z. Arov, and M.G. Krein, *Infinite block Hankel matrices and related extension problems*, Amer. Math. Soc. Trans., **111**: 133–156 (1978).

[3] K. Afanasiev and M. Hinze, *Adaptive control of a wake flow using proper orthogonal decomposition*, Reprint 648, Department of Mathematics, Technical University, Berlin (1999).

[4] H. Aling, S. Banerjee, A.K. Bangia, V. Cole, J.L. Ebert, A. Emami-Naeini, K.F. Jensen, I.G. Kevrekidis, and S. Shvartsman, *Nonlinear model reduction for simulation and control of rapid thermal processing*, Proceedings of the 1997 American Control Conference, 2233–2238 (1997).

[5] T. Amdeberhan and D. Zeilberger, *Determinants through the looking glass*, Adv. Appl. Math., **27**: 225–230 (2001).

[6] B.D.O. Anderson and S. Vongpanitlerd, *Network analysis and synthesis*, Prentice–Hall, Englewood Cliffs, NJ (1973).

[7] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, 3rd ed., *LAPACK users' guide*, SIAM, Philadelphia, (1999). http://www.netlib.org/lapack/lug/lapack_lug.html

[8] A.C. Antoulas, *On canonical forms for linear constant systems*, Internat. J. Control, **33**: 95–122 (1981).

[9] A.C. Antoulas, *On recursiveness and related topics in linear systems*, IEEE Trans. Automat. Control, **31**: 1121–1135 (1986).

[10] A.C. Antoulas and R.H. Bishop, *Continued fraction decomposition of linear systems in the space state*, Systems Control Lett., **9**: 43–53 (1987).

[11] A.C. Antoulas, *Recursive modeling of discrete-time series*, in IMA Vol. Linear Algebra Control, P.M. Van Dooren and B.F. Wyman, eds., 1–22 (1993).

[12] A.C. Antoulas and J.C. Willems, A behavioral approach to linear exact modeling, *IEEE Trans. Automat. Control*, **38**: 1776–1802 (1993).

[13] A.C. Antoulas, *On the approximation of Hankel matrices*, in Operators, systems, and linear algebra, U. Helmke, D. Prätzel-Wolters, and E. Zerz, eds., Teubner Verlag, Stuttgart, 17–22 (1997).

[14] A.C. Antoulas, *Approximation of linear operators in the* 2-*norm*, Linear Algebra Appl., Special Issue on Challenges in Matrix Theory (1998).

[15] A.C. Antoulas and J.C. Willems, *Minimal rational interpolation and Prony's method*, Lecture Notes in Control and Inform. Sci., **144**: 297–306 (1990).

[16] A.C. Antoulas, E.D. Sontag, and Y. Yamamoto, *Controllability and observability*, in Wiley Encyclopedia of Electrical and Electronics Engineering, vol. 4, J.G. Webster, ed., 264–281 (1999).

[17] A.C. Antoulas, *Approximation of linear dynamical systems*, in Wiley Encyclopedia of Electrical and Electronics Engineering, vol. 11, J.G. Webster, ed., 403–422 (1999).

[18] A.C. Antoulas, *On eigenvalues and singular values of linear dynamical systems*, Proceedings of the Householder XIV Symposium, Chateau Whistler, BC, 8–11 (1999).

[19] A.C. Antoulas and D.C. Sorensen, *Lyapunov, Lanczos and inertia*, Linear Algebra Appl., **326**: 137–150 (2001).

[20] A.C. Antoulas, D.C. Sorensen, and S. Gugercin, *A survey of model reduction methods for large-scale systems*, Contemp. Math., **280**: 193–219 (2001).

[21] A.C. Antoulas, D.C. Sorensen, and Y. Zhou, *On the decay rates of the Hankel singular values*, Systems Control Lett., **46**: 323–342 (2002).

[22] A.C. Antoulas, J.A. Ball, J. Kang, and J.C. Willems, *On the solution of the minimal rational interpolation problem*, Linear Algebra Appl., **137/138**: 511–573 (1990).

[23] A.C. Antoulas and B.D.O. Anderson, *State-space and polynomial approaches to rational interpolation*, in Realization and modelling in system theory, Proceedings of MTNS-89, vol. I, M.A. Kaashoek, J.H. van Schuppen, and A.C.M. Ran, eds., 73–81 (1990).

[24] B.D.O. Anderson and A.C. Antoulas, *Rational interpolation and state-variable realizations*, Linear Algebra Appl., **137/138**: 479–509 (1990).

[25] A.C. Antoulas and B.D.O. Anderson, *On the scalar rational interpolation problem*, IMA J. Math. Control Inform., **3**: 61–88 (1986).

[26] A.C. Antoulas and B.D.O. Anderson, *On the problem of stable rational interpolation*, Linear Algebra Appl., **122-124**: 301–329 (1989).

[27] A.C. Antoulas, *A new result on passivity preserving model reduction*, Systems Control Lett. **54**: 361–374 (2005).

[28] W.E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenproblem*, Quart. Appl. Math., **9**: 17–29 (1951).

[29] J.A. Atwell and B.B. King, *Proper orthogonal decomposition for reduced basis feedback controllers for parabolic equations*, Math. Comput. Modelling, **33**: 1–19 (2001).

[30] J.A. Atwell and B.B. King, *Reduced order controllers for spatially distributed systems via proper orthogonal decomposition*, SIAM J. Sci. Comput., **26**: 128–151 (2004).

[31] Z. Bai, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Appl. Numer. Math., **43**: 9–44 (2002).

[32] Z. Bai and Q. Ye, *Error estimation of the Padé approximation of transfer functions via the Lanczos process*, Electron. Trans. Numer. Anal., **7**: 1–17 (1998).

[33] Z. Bai, P. Feldmann, and R.W. Freund, *Stable and passive reduced-order models based on partial Padé approximation via the Lanczos process*, Bell Laboratories, Murray Hill, NJ (1997).

[34] Z. Bai and R. Freund, *A partial Padé-via-Lanczos method for reduced order modeling*, Linear Algebra Appl., **332-334**: 139–164 (2001).

[35] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the solution of algebraic eigenvalue problems: A practical guide*, SIAM, Philadelphia (2000).

[36] Z. Bai and Y. Su, *Dimension reduction of second-order dynamical systems via a second-order Arnoldi method*, Linear Algebra Appl., to appear.

[37] Z. Bai, P.M. Dewilde, and R.W. Freund, *Reduced-order modeling*, Bell Laboratories, Murray Hill, NJ (2002).

[38] J.A. Ball, I. Gohberg, and L. Rodman, *Interpolation of rational matrix functions*, Birkhäuser Verlag, Basel (1990).

[39] B. Bamieh and J. B. Pearson, *A general framework for linear periodic systems with applications to H_∞ sampled-data control*, IEEE Trans. Automat. Control, **37**: 418–435 (1992).

[40] B. Bamieh and M. Dahleh, *Energy amplification of channel flows with stochastic excitation*, Phys. Fluids, **13**: 3258–3269 (2001).

[41] B. Bamieh, *Distributed systems and distributed control*, 40th IEEE Conference on Decision and Control (2001).

[42] H.T. Banks, *Control and estimation in distributed parameter systems*, SIAM, Philadelphia (1992).

[43] R.H. Bartels and G.W. Stewart, *Solution of the matrix equation $AX+XB=C$*, Commun. ACM, **15**: 820–826 (1972).

[44] J. Bastian and J. Haase, *Order reduction for second order systems*, Proceedings of the 4th MATHMOD Conference, Vienna, I. Troch and F. Breitenbecker, eds., 418–424 (2003).

[45] F.S.V. Bazán and Ph.L. Toint, *Conditioning of infinite Hankel matrices of finite rank*, Systems Control Lett., **41**: 347–359 (2000).

[46] C. Beattie, M. Embree, and J. Rossi, *Convergence of restarted Krylov subspaces to invariant subspaces*, SIAM J. Matrix Anal. Appl., **25**: 1074–1109 (2004).

[47] C. Beck, *Coprime factors reduction methods for linear parameter varying and uncertain systems*, Systems Control Lett. (2005).

[48] B. Beckermann, *The condition number of real Vandermonde, Krylov and positive definite Hankel matrices*, Numer. Math., **85**: 553–577 (2000).

[49] S.C. Beeler, G.M. Kepler, H.T. Tran, and H.T. Banks, *Reduced order modeling and control of thin film growth in an HPCVD reactor*, Technical Report CRSC-00-33, Center for Research in Scientific Computation, North Carolina State University (1999).

[50] V. Belevitch, *Interpolation matrices*, Philips Research Reports, **25**: 337–369 (1970).

[51] P. Benner, E.S. Quintana-Ortí, and G. Quintana-Ortì, *Singular perturbation approximation for large, dense linear systems*, IEEE Internat. Symposium on Computer-Aided Control System Design, Anchorage, 255–260 (2000).

[52] P. Benner and E.S. Quintana-Ortí, *Solving Stable Generalized Lyapunov Equations with the Matrix Sign Function*, Numer. Algorithms, **20**: 75–100 (1999).

[53] P. Benner, E.S. Quintana-Ortì, and G. Quintana-Ortì, *PSLICOT routines for model reduction of stable large-scale systems*, Proceedings of the 3rd NICONET Workshop on Numerical Software in Control Engineering, Louvain-la-Neuve, Belgium, 39–44 (2001).

[54] P. Benner, E.S. Quintana-Ortì, and G. Quintana-Ortì, *Efficient numerical algorithms for balanced stochastic truncation*, Internat. J. Appl. Math. Comput. Sci., **11**: 1123–1150 (2001).

[55] P. Benner, E.S. Quintana-Ortì, and G. Quintana-Ortì, *Balanced truncation model reduction of large-scale dense systems on parallel computers*, Math. Comput. Modelling Dynam. Syst., **6**: 383–405 (2000).

[56] P. Benner, E.S. Quintana-Ortì, and G. Quintana-Ortì, *Numerical solution of discrete stable linear matrix equations on multicomputers*, Parallel Algorithms Appl., **17**: 127–146 (2002).

[57] P. Benner, E.S. Quintana-Ortì, and G. Quintana-Ortì, *Parallel algorithms for model reduction of discrete-time systems*, Internat. J. System Sci., **34**: 319–333 (2003).

[58] P. Benner, E.S. Quintana-Ortì, and G Quintana-Ortì, *State-space truncation methods for parallel model reduction of large-scale systems*, Parallel Comput., **29**: 1701–1722 (2003).

[59] P. Benner, V. Mehrmann, V. Sima, S. Van Huffel, and A. Varga, *SLICOT-A subroutine library in systems and control theory*, Appl. Comput. Control Signals Circuits, **1**: 499–532 (1999).

[60] P. Benner, *Solving large-scale control problems*, Control Syst. Magazine, **24**: 44–59 (2004).

[61] P. Benner and H. Fassbender, *SLICOT drives tractors!*, NICONET Report 1999-2, NICONET Newsletter, **2**: 17–22 (1999).

[62] P. Benner and H. Fassbender, *An implicitly restarted symplectic Lanczos method for the symplectic eigenvalue problem*, SIAM J. Matrix Anal. Appl., **22**: 682–713 (2000).

[63] G. Berkooz, P. Holmes, and J.L. Lumley, *Coherent structures, dynamical systems and symmetry*, Cambridge Monogr. Mechanics, Cambridge University Press, Cambridge, UK (1996).

[64] R. Byers, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., **85**: 267–279 (1987).

[65] R. Bhatia, *Matrix analysis*, Graduate Texts in Math. **169**, Springer-Verlag, New York (1997).

[66] V.D. Blondel and J.N. Tsitsiklis, *A survey of computational complexity results in systems and control theory*, Automatica, **36**: 1249–1274 (2000).

[67] B. Bohnhorst, A. Bunse-Gerstner, and H. Fassbender, *On the perturbation theory for the unitary eigenvalue problem*, SIAM J. Matrix Anal. Appl., **21**: 809–824 (2000).

[68] D.L. Boley, *Krylov space methods on state-space control models*, Circuits Systems Signal Process., **13**: 733–758 (1994).

[69] R. Bott and R.J. Duffin, *Impedance synthesis without transformers*, J. Appl. Phys., **20**: 816 (1949).

[70] A. Böttcher and B. Silbermann, *Introduction to large truncated Toeplitz matrices*, Springer, New York (1999).

[71] S.P. Boyd and C.H. Barratt, *Linear controller design: Limits of performance*, Prentice–Hall, Englewood Cliffs, NJ (1991).

[72] J. Brandts, *Projection methods for oversized linear algebra problems*, Nieuw Arch. Wiskd., **5**: 264–277 (2000).

[73] J. Brands, *A comparison of subspace methods for Sylvester equations*, Preprint 1183, Mathematics Institute, Universiteit Utrecht (2001).

[74] K.E. Brenan, S.L. Campbell, and L.R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, SIAM, Philadelphia (1995).

[75] N.A. Bruinsma and M. Steinbuch, *A fast algorithm to compute the $\mathcal{H}_\infty$-norm of a transfer function*, Systems Control Lett., **14**: 287–293 (1990).

[76] W.L. Brogan, *Modern control theory*, Prentice–Hall, Englewood Cliffs, NJ (1991).

[77] O. Brune, *Synthesis of a finite two-terminal network whose driving point impedance is a prescribed function of frequency*, J. Math. Phys., **10**: 191–236 (1931).

[78] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N.K. Nichols, *Feedback design for regularizing descriptor systems*, Linear Algebra Appl., **299**: 119–151 (1999).

[79] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *Numerical methods for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl., **14**: 927–949 (1993).

[80] A. Bunse-Gerstner, *Large linear systems in electromagnetic field simulation*, AMS/IMS/SIAM Summer Research Conference on Fast Algorithms in Mathematics, Engineering and Computer Science, South Hadley, MA, August 5-9 (2001).

[81] J.A. Burns, *Proper orthogonal decomposition and model reduction for control*, Workshop on POD and Its Applications, Graz (2000).

[82] J.A. Burns and B.B. King, *A reduced basis approach to the design of low order compensators for nonlinear partial differential systems*, J. Vibrations Control, **4**: 297–323 (1998).

[83] Y. Chahlaoui, K. Gallivan, and P.M. Van Dooren, *Recursive calculation of dominant singular subspaces*, SIAM J. Matrix Anal. Appl., **25**: 445–463 (2003).

[84] Y. Chahlaoui, D. Lemonnier, A. Vandendorpe, and P.M. Van Dooren, *Second order balanced truncation*, Linear Algebra Appl., to appear.

[85] V.-S. Chellaboina and W.M. Haddad, *Is the Frobenius norm induced?*, IEEE Trans. Automatic Control, **40**: 2137–2139 (1995).

[86] V.-S. Chellaboina, W.M. Haddad, D.S. Bernstein, and D.A. Wilson, *Induced convolution norms of linear dynamical systems*, Mathematics of Control Signals and Systems, **13**: 216–239 (2000).

[87] E. Chiprout and M.S. Nakhla, *Asymptotic waveform evaluation*, Kluwer, Norwell, MA (1994).

[88] M.T. Chu, R.E. Funderlic, and G.H. Golub, *A rank-one reduction formula and its applications to matrix factorizations*, SIAM Review, **37**: 512–530 (1995).

[89] J.V. Clark, D. Bindel, W. Kao, E. Zhu, A. Kuo, N. Zhou, J. Nie, J. Demmel, Z. Bai, S. Govindjee, K.S.J. Pister, M. Gu, and A. Agogino, *Addressing the needs of complex MEMS design*, Proceedings of MEMS 2002, Las Vegas (2002).

[90] J. Cullum, A. Ruehli, and T. Zhang, *A method for reduced-order modeling and simulation of large interconnect circuits and its application to PEEC models with retardation*, IEEE Trans. Circuits Systems II: Analog Digital Signal Processing, **47**: 261–273 (2000).

[91] B.N. Datta, *Numerical methods for linear control systems design and analysis*, Elsevier, New York (2003).

[92] B.N. Datta and Y. Saad, Arnoldi methods for large Sylvester-like observer matrix equations and an associated algorithm for partial spectrum assignment, *Linear Algebra and Its Applications*, **154-156**, 225–244 (1991).

[93] C. de Villemagne and R.E. Skelton, *Model reduction using a projection formulation*, International Journal of Control, **46**: 2141–2169 (1987).

[94] J.W. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numerische Mathematik, **51**: 251–289 (1987).

[95] U.B. Desai and D. Pal, *A transformation approach to stochastic model reduction*, IEEE Trans. Automatic Control, **29**: 1097–1110 (1984).

[96] C.A. Desoer and M. Vidyasagar, *Feedback systems: Input-output properties*, Academic Press, New York (1975).

[97] E. de Souza and S.P. Bhattacharyya, *Controllability, observability and the solution of* $AX - XB = C$, Linear Algebra and its Applications, **39**: 167–188 (1981).

[98] A.E. Deane, I.G. Kevrekidis, G.E. Karniadakis, and S.A. Orszag, *Low-dimensional models for complex geometry flows: Application to grooved channels and circular cylinders*, Phys. Fluids A, **3**: 2337–2354 (1991).

[99] S.M. Djouadi, *Comments on "Is the Frobenius norm induced?"*, IEEE Trans. Automat. Control, **48**: 518–519 (2003). V.-S. Chellaboina and W.M. Haddad, *Authors' reply*, IEEE Trans. Automat. Control, **48**: 519–520 (2003).

[100] D.L. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, Applied and Computational Harmonic Analysis, **1**: 100–115 (1993).

[101] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, **1**: 211–218 (1936).

[102] N. Eichler, *Mathematische Modellierung der Mechanischen Eigenschaften einer Bienenwabe*, Universität Würzburg (1996).

[103] H. Elbern and H. Schmidt, *Ozone episode analysis by four-dimensional variational chemistry data assimilation scheme*, J. Geophysical Research, **106**: 3569–3590 (2001).

[104] N. Ellner and E. Wachspress, *Alternating direction implicit iteration for systems with complex spectra*, SIAM Journal on Numerical Analysis, **28**: 859–870 (1991).

[105] M. Embree, *Private communication*, 29 May 2001.

[106] M. Embree and L.N. Trefethen, *Generalizing eigenvalue theorems to pseudospectra theorems*, SIAM Journal on Scientific Computing, **23**: 583–590 (2001).

[107] D.F. Enns, *Model reduction with balanced realizations: An error bound and frequency weighted generalization*, Proceedings of the IEEE Conference on Decision and Control, 127–132 (1984).

[108] Y. Fang, K.A. Loparo, and X. Feng, *New estimates for solutions of Lyapunov equations*, IEEE Trans. Automatic Control, **42**: 408–411 (1997).

[109] B.F. Farrell and P.J. Ioannou, *Perturbation growth and structure in time dependent flows*, Journal Atmospheric Sciences, **56**: 3622–3639 (1999).

[110] B.F. Farrell and P.J. Ioannou, *State estimation using a reduced order Kalman Filter*, Journal Atmospheric Sciences, **58**: 3666–3680 (2001).

[111] H. Fassbender, *Error analysis of the symplectic Lanczos method for the symplectic eigenvalue problem*, BIT, **40**: 471–496 (2000).

[112] H. Fassbender, *Structured linear algebra problems in control*, 8th SIAM Conference on Applied Linear Algebra, Williamsburg (2003).

[113] K.V. Fernando and H. Nicholson, *On the structure of balanced and other principal representations of SISO systems*, IEEE Trans. Automat. Control, **28**: 228–231 (1983).

[114] P. Feldmann and R.W. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Trans. Comput. Aided Design, **14**: 639–649 (1995).

[115] L. Fortuna, G. Nunnari, and A. Gallo, *Model order reduction techniques with applications in electrical engineering*, Springer-Verlag, London, New York (1992).

[116] B.A. Francis, *A course in $\mathcal{H}_\infty$ control theory*, Lecture Notes in Control and Inform. Sci., **88** (1987).

[117] R.W. Freund, *Reduced-order modeling techniques based on Krylov subspaces and their use in circuit simulation*, in Applied and computational control, signals, and circuits, vol. 1, Birkhäuser, Berlin, 435–498 (1999).

[118] R.W. Freund, *Passive reduced-order modeling via Krylov subspace methods*, Bell Laboratories, Murray Hill, NJ (2000).

[119] P.A. Fuhrmann, *Linear spaces and operators in Hilbert space*, McGraw–Hill, New York (1981).

[120] P.A. Fuhrmann, *Duality in polynomial models with applications to geometric control theory*, IEEE Trans. Automat. Control, **26**: 284–295 (1981).

[121] P.A. Fuhrmann, *A polynomial approach to Hankel norm and balanced approximations*, Linear Algebra Appl., **146**: 133–220 (1991).

[122] P.A. Fuhrmann, *On the Hamiltonian structure in the computation of singular values for a class of Hankel operators*, in H_∞-Control Theory, Lecture Notes in Math., E. Mosca and L. Pandolfi, eds., **1496**: 250–276 (1991).

[123] P.A. Fuhrmann, *A polynomial approach to linear algebra*, Springer-Verlag, New York (1996).

[124] K. Fujimoto and J.M.A. Scherpen, *Nonlinear balanced realization based on the differential eigenstructure of Hankel operators*, Proceedings of the IEEE Conference on Decision and Control, Maui (2003).

[125] K. Gallivan, E.J. Grimme, and P.M. Van Dooren, *Asymptotic waveform evaluation via a restarted Lanczos method*, Appl. Math. Lett., **7**: 75–80 (1994).

[126] K. Gallivan, E. Grimme and P.M. Van Dooren, *Padé approximation of large-scale dynamic systems with Lanczos methods*, Proceedings of the 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL (1994).

[127] K. Gallivan, E. Grimme, and P.M. Van Dooren, *A rational Lanczos algorithm for model reduction*, CSRD Report 1411, University of Illinois, Urbana-Champaign, IL (1995).

[128] K. Gallivan, A. Vandendorpe, and P.M. Van Dooren, *On the generality of multi-point Padé approximations*, J. Comp. Appl. Math., **162**: 213–229 (2004).

[129] K. Gallivan, A. Vandendorpe, and P.M. Van Dooren, *Sylvester equations and projection-based model reduction*, Technical Report, CESAME, University of Louvain-la-Neuve (2001).

[130] K. Gallivan, A. Vandendorpe, and P.M. Van Dooren, *Model reduction via truncation: An interpolation point of view*, Linear Algebra and Its Applications, **375**: 115–134 (2003).

[131] K. Gallivan, A. Vandendorpe, and P.M. Van Dooren, *Model reduction of MIMO systems via tangential interpolation*, SIAM Journal on Matrix Analysis and Applications, **26**: 328–349 (2004).

[132] K. Gallivan, E. Grimme, and P.M. Van Dooren, *Model reduction of large-scale systems: Rational Krylov versus balancing techniques*, in Error control and adaptivity in scientific computing, H. Bulgak and C. Zenger, eds., **536**: 177–190 (1999).

[133] K. Gallivan and P.M. Van Dooren, *Rational approximations of pre-filtered transfer functions via the Lanczos algorithm*, Numer. Algorithms, **20**: 331–342 (1999).

[134] F.R. Gantmacher, *The theory of matrices*, vols. I, II, Chelsea Publishing Company, New York (1959).

[135] W. Gawronski and J.-N. Juang, *Model reduction in limited time and frequency intervals*, Internat. J. Systems Sci., **21**: 349–376 (1990).

[136] W. Gawronski, *Balanced control of flexible structures*, Springer-Verlag, New York (1996).

[137] Y.V. Genin and S.Y. Kung, *A two-variable approach to the model reduction problem with Hankel norm criterion*, IEEE Trans. Circuits Systems, **28**: 912–924 (1981).

[138] Y. Genin, A. Vandendorpe, and P.M. Van Dooren, *Projection and embedding of rational transfer functions*, 8th SIAM Conference on Applied Linear Algebra, Williamsburg, VA (2003).

[139] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, **39**: 1115–1193 (1984).

[140] K. Glover, *Multiplicative approximation of linear multivariable systems with $\mathcal{L}_\infty$ error bounds*, Proceedings of the American Control Conference, pp. 1705–1709 (1986).

[141] K. Glover, R.F. Curtain, and J.R. Partington, *Realisation and approximation of linear infinite-dimesional systems with error bounds*, SIAM Journal on Control and Optimization, **26**: 863–898 (1988).

[142] S.K. Godunov, *Modern aspects of linear algebra*, Trans. Math. Monogr., **175** (1998).

[143] I. Gohberg and I. Koltracht, *Triangular factors of Cauchy and Vandermonde matrices*, Integral Equations Oper. Theory, **46**: 46–59 (1996).

[144] G.H. Golub and C.F. Van Loan, *Matrix computations*, 3rd ed., The Johns Hopkins University Press, Baltimore (1996).

[145] G.H. Golub and J.M. Varah, *On the characterization of the best l_2-scaling of a matrix*, SIAM Journal on Numerical Analysis, **11**: 472–479 (1974).

[146] W.B. Gragg and A. Lindquist, *On the partial realization problem*, Linear Algebra Appl., **50**: 277–319 (1983).

[147] M. Green, *Balanced stochastic realizations*, Linear Algebra Appl., **98**: 211–247 (1988).

[148] M. Green, *A relative-error bound for balanced stochastic truncation*, IEEE Trans. Automat. Control, **33**: 961–965 (1988).

[149] M. Green and D.J.N. Limebeer, *Linear robust control*, Prentice–Hall, Englewood Cliffs, NJ (1995).

[150] A. Greenbaum, *Iterative methods for solving linear systems*, SIAM, Philadelphia (1997).

[151] A. Greenbaum, *Using the Cauchy integral formula and partial fractions decomposition of the resolvent to estimate* $\|f(A)\|$, Technical Report, Department of Mathematics, University of Washington, Seattle (2000).

[152] E.J. Grimme, *Krylov projection methods for model reduction*, Ph.D. Thesis, ECE Department, University of Illinois, Urbana-Champaign (1997).

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

[154] E. Grimme, K. Gallivan, and P.M. Van Dooren, *On some recent developments in projection-based model reduction*, ENUMATH-97, 98–113 (1998).

[155] T. Gudmundsson, A. J. Laub, *Approximate solution of large sparse Lyapunov equations*, IEEE Trans. Automat. Control, **39**: 1110–1114 (1994).

[156] S. Gugercin, *Projection methods for model reduction of large-scale systems*, Ph.D. Thesis, ECE Department, Rice University, Houston, TX (2003).

[157] S. Gugercin, D.C. Sorensen, and A.C. Antoulas, *A modified low-rank Smith method for large-scale Lyapunov equations*, Numer. Algorithms, **32**: 27–55 (2003).

[158] S. Gugercin and A.C. Antoulas, *Model reduction of large scale systems by least squares*, Linear Algebra Appl., to appear.

[159] S. Gugercin and A.C. Antoulas, *A survey of model reduction by balanced truncation and some new results*, International Journal of Control, **77**: 748–766 (2004).

[160] M.D. Gunzburger and J.S. Petersen, *The reduced basis method in control problems*, in Computation and Control III, Birkhäuser, Berlin, 211–218 (1993).

[161] M.H. Gutknecht, *The Lanczos process and Padé approximation*, in Proceedings of the Cornelius Lanczos International Centenary Conference, J.D. Brown et al., eds., SIAM, Philadelphia, 61–75 (1994).

[162] J. Hahn and T.F. Edgar, *An improved method for nonlinear model reduction using balancing of empirical gramians*, Comput. Chemical Engrg., **26**: 1379–1397 (2002).

[163] J. Hahn, T.F. Edgar, and W. Marquardt, *Controllability and observability covariance matrices for the analysis and order reduction of stable nonlinear systems*, J. Process Control, **13**: 115–127 (2003).

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

[165] B. Hanzon, *The area enclosed by the (oriented) Nyquist diagram and the Hilbert-Schmidt-Hankel norm of a linear system*, IEEE Trans. Automat. Control, **37**: 835–839 (1992).

[166] B. Hanzon and R.L.M. Peeters, *A Fadeev sequence method for solving Lyapunov and Sylvester equations*, Linear Algebra Appl., **241-243**: 401–430 (1996).

[167] M.T. Heath, A.J. Laub, C.H. Paige, and R.C. Ward, *Computing the singular value decomposition of a product of two matrices*, SIAM Journal on Scientific and Statistical Computing, **7**: 1147–1159 (1986).

[168] A.W. Heemink, M. Verlaan, and A.J. Segers, *Variance reduced ensemble Kalman filter*, Report 00-03, Department of Applied Mathematical Analysis, Delft University of Technology, Delft,The Netherlands (2000).

[169] N.J. Higham, *Matrix nearness problems and applications*, in Applications of matrix theory, M.J.C. Gover and S. Barnett, eds., Oxford University Press, Oxford, UK (1989).

[170] N.J. Higham, *Perturbation theory and backward error for $AX - XB = C$*, BIT, **33**: 124–136 (1993).

[171] N.J. Higham, *Accuracy and stability of numerical algorithms*, 2nd ed., SIAM, Philadelphia (2002).

[172] N.J. Higham, M. Konstantinov, V. Mehrmann, and P. Petkov, *Sensitivity of computational control problems*, Control Syst. Magazine, **24**: 28–43 (2004).

[173] D. Hinrichsen and A.J. Pritchard, *An improved error estimate for reduced-order models of discrete-time systems*, IEEE Trans. Automat. Control, **35**: 317–320 (1990).

[174] M. Hinze and K. Kunisch, *Three control methods for time-dependent fluid flow*, Flow Turbulence Combustion, **65**: 273–298 (2000).

[175] M. Hochbruck and G. Starke, *Preconditioned Krylov subspace methods for Lyapunov matrix equations*, SIAM Journal on Matrix Analysis and Applications, **16**: 156–171 (1995).

[176] A. Hodel, K. Poolla, B. Tenison, *Numerical solution of the Lyapunov equation by approximate power iteration*, Linear Algebra Appl., **236**: 205–230 (1996)

[177] K. Hoffman, *Banach spaces of analytic functions*, Prentice–Hall, Englewood Cliffs, NJ (1962).

[178] K. Horiguchi, T. Nishimura, and A. Nagata, *Minimal realizations interpolating first and second order information*, Internat. J. Control, **52**: 389–704 (1990).

[179] L.W. Horowitz, S. Walters, D.L. Mauzerall, L.K. Emmons, P.J. Rasch, C. Granier, X. Tie, J.-F. Lamarque, M.G. Schultz, G.S. Tyndall, J.J. Orlando, and G.P. Brasseur, *A global simulation of tropospheric ozone and related tracers: Description and evaluation of MOZART, version* 2, J. Geophys. Res., **108(D24)**: 4784 (2003).

[180] A. Horn, *On the eigenvalues of a matrix with prescribed singular values*, Proc. Amer. Math. Soc., **5**: 4–7 (1954).

[181] R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, UK (1985).

[182] R.A. Horn and C.R. Johnson, *Topics in matrix analysis*, Cambridge University Press, Cambridge, UK (1991).

[183] J. Hu, *A study of discrete-time, linear, periodic, time-varying systems*, MS Thesis, Department of Electrical and Computer Engineering, Rice University, Houston, TX (2003).

[184] D. Hu and L. Reichel, *Krylov-subspace methods for the Sylvester equation*, Linear Algebra and Its Applications, **172**, 283–313, (1992).

[185] L. Hubert, J. Meulman, and W. Heiser, *Two purposes for matrix factorization: A historical appraisal*, SIAM Review, **42**: 68–82 (2000).

[186] *IEEE standard for floating point arithmetic*, ANSI/IEEE Standard 745-1985, IEEE, New York (1985).

[187] I. Jaimoukha and E. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM Journal on Numerical Analysis, **31**: 227–251 (1994).

[188] I.M. Jaimoukha and E.M. Kasenally, *Implicitly restarted Krylov subspace methods for stable partial realizations*, SIAM Journal on Matrix Analysis and Its Applications, **18**: 633–652 (1997).

[189] I. Jonsson and B. Kågström, *Recursive blocked algorithms for solving triangular systems, Part* I: *One-sided and coupled Sylvester-type matrix equations. Part* II: *Two-sided and generalized Sylvester and Lyapunov equations*, ACM Trans. Math. Software, **28**: 392–415, 416–435 (2002).

[190] B. Kågström, P. Johansson, and E. Elmroth, *Bounds on the distance between nearby Jordan and Kronecker structures in a closure hierarchy*, J. Math. Sci. (N.Y.), **114**: 1765–1779 (2003).

[191] T. Kailath, *Linear systems*, Prentice–Hall, Englewood Cliffs, NJ (1980).

[192] R.E. Kalman, P.L. Falb, and M.A. Arbib, *Topics in mathematical system theory*, McGraw–Hill, New York (1969).

[193] R.E. Kalman, *On partial realizations, transfer functions, and canonical forms*, Acta Polytech. Scand. Math., **31**: 9–32 (1979).

[194] M. Kamon, F. Wang, and J. White, *Generating nearly optimally compact models from Krylov-subspace based reduced-order models*, IEEE Trans. Circuits Systems II: Analog Digital Signal Process., **47**: 239–248 (2000).

[195] O. Kaneko and P. Rapisarda, *Recursive exact $\mathcal{H}_\infty$ identification from impulse-response measurements*, Systems and Control Letters, **49**: 323–334 (2003).

[196] M. Karow, *Private communication*, Feb. 2003.

[197] A. Katavolos, *Introduction to operator theory*, Department of Mathematics, University of Athens (2000) (in Greek).

[198] C. Kenney and A.J. Laub, *The matrix sign function*, IEEE Trans. Automat. Control, **40**: 1330–1348 (1995).

[199] G.M. Kepler, H.T. Tran, and H.T. Banks, *Reduced order model compensator control of species transport in a CVD reactor*, Technical Report CRSC-99-15, Center for Research in Scientific Computation, North Carolina State University (1999).

[200] G.M. Kepler, H.T. Tran, and H.T. Banks, *Compensator control for chemical vapor deposition film growth using reduced order design models*, Technical Report CRSC-99-41, Center for Research in Scientific Computation, North Carolina State University (1999).

[201] S.W. Kim, B.D.O. Anderson, and A.G. Madievski, *Error bound for transfer function order reduction using frequency weighted balanced truncation*, Systems Control Lett., **24**: 183–192 (1995).

[202] S.-Y. Kim, N. Gopal, and L.T. Pillage, *Time-domain macromodels for VLSI interconnect analysis*, IEEE Trans. Comput.-Aided Design Integrated Circuits Syst., **13**: 1257–1270 (1994).

[203] B.B. King and E.W. Sachs, *Optimization techniques for stable reduced order controllers for partial differential equations*, in Optimal control: Theory, algorithms, and applications, W.W. Hager and P. Pardalos, eds., Kluwer, Dordrecht, The Netherlands, 278–297 (1998).

[204] B.B. King and E.W. Sachs, *Semidefinite programming techniques for reduced order systems with guaranteed stability margins*, Comput. Optim. Appl., **17**: 37–59 (2000).

[205] L. Knockaert and D. De Zutter, *Passive reduced order multiport modeling: The Padé-Laguerre, Krylov-Arnoldi-SVD connection*, Internat. J. Electron. Commun., **53**: 254–260 (1999).

[206] L. Knockaert, B. De Backer, and D. De Zutter, *SVD compression, unitary transforms, and computational complexity*, Technical Report, INTEC, University of Gent (2002).

[207] L. Knockaert and D. De Zutter, *Stable Laguerre-SVD reduced-order modeling*, IEEE Trans. Circuits Systems I: Fundamental Theory Appl., **50**: 576–579 (2003).

[208] P. Kokotovic, R. O'Malley, and P. Sannuti, *Singular perturbations and order reduction in control theory: An overview*, Automatica, **12**: 123–132 (1976).

[209] L. Komzsik, *The Lanczos method: Evolution and application*, SIAM, Philadelphia (2003).

[210] M.M. Konstantinov, D.W. Gu, V. Mehrmann, and P.Hr. Petkov, *Perturbation theory for matrix equations*, Studies in Computational Math. 9, North–Holland, Amsterdam (2003).

[211] D. Kressner, *Large periodic Lyapunov equations: Algorithms and applications*, Technical report, Institut für Mathematik, Technische Universität, Berlin (2003).

[212] K. Kunisch and S. Volkwein, *Control of Burgers' equation by a reduced order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., **102**: 345–371 (1999).

[213] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math., **90**: 117–148 (2001).

[214] W. Kwon, Y. Moon, S. Ahn, *Bounds in algebraic Riccati and Lyapunov equations: A survey and some new results*, Internat. J. Control, **64**: 377–389 (1996).

[215] S. Lall, J.E. Marsden, and S. Glavaski, *Empirical model reduction of controlled nonlinear systems*, Technical Report CIT-CDS-98-008, California Institute of Technology, (1998).

[216] S. Lall, J.E. Marsden, and S. Glavaski, *A subspace approach to balanced truncation for model reduction of nonlinear control systems*, Internat. J. Robust Nonlinear Control, **12**: 519–535 (2002).

[217] S. Lall and C. Beck, *Error-bounds for balanced model-reduction of linear time-varying systems*, IEEE Trans. Automat. Control, **48**: 946–956 (2003).

[218] J. Lam and B.D.O. Anderson, *$\mathcal{L}_1$ impulse response error bound for balanced truncation*, Systems Control Lett., **18**: 129–138 (1992).

[219] P. Lancaster, *Explicit solutions of linear matrix equations*, SIAM Rev., **12**: 544–566 (1970).

[220] P. Lancaster and M. Tismenetsky, *The theory of matrices*, Academic Press, New York (1985).

[221] C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. National Bureau Standards, **45**: 255–282 (1950).

[222] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. National Bureau Standards, **49**: 33–53 (1952).

[223] V. Larin and F. Aliev, *Construction of square root factor for solution of the Lyapunov matrix equation*, Systems Control Lett., **20**: 109–112 (1993).

[224] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, *Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms*, IEEE Trans. Automat. Control, **32**: 115–121 (1987).

[225] P.D. Lax, *Linear algebra*, in Pure and applied mathematics, John Wiley, New York (1997).

[226] R.B. Lehoucq and A.G. Salinger, *Large-scale eigenvalue calculations for stability analysis of steady flows on massively parallel computers*, Internat. J. Numer. Methods Fluids, **36**: 309–327 (2001).

[227] R.B. Lehoucq, D.C. Sorensen, and C. Yang, *ARPACK users' guide: Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, SIAM, Philadelphia (1998). http://www.caam.rice.edu/software/ARPACK.

[228] J. Li, F. Wang, and J. White, *An efficient Lyapunov equation-based approach for generating reduced-order models of interconnect*, Proceedings of the 36th IEEE/ACM Design Automation Conference, New Orleans (1999).

[229] C.-A. Lin and T.-Y. Chiu, *Model reduction via frequency weighted balanced realization*, Control Theory Adv. Technol., **8**: 341–351 (1992).

[230] Y. Liu and B.D.O. Anderson, *Singular perturbation approximation of balanced systems*, Internat. J. Control, **50**: 1379–1405 (1989).

[231] E.N. Lorenz, *Empirical orthogonal functions and statistical weather prediction*, Scientific Report 1, Statistical Forecasting Project, MIT, Cambridge, MA (1956).

[232] T.A. Lypchuk, M.C. Smith, and A. Tannenbaum, *Weighted sensitivity minimization: General plants in H^∞ and rational weights*, Linear Algebra Appl., **109**: 71–90 (1988).

[233] A.W. Marshall and I. Olkin, *Inequalities: Theory of majorization and its applications*, in Mathematics in Science and Engineering, vol. 143, Academic Press, New York (1979).

[234] A.J. Mayo and A.C. Antoulas, *A behavioral approach to positive real interpolation*, J. Math. Comput. Modelling Dynam. Syst., **8**: 445–455 (2002).

[235] K. Meerbergen, *Model Reduction techniques for parametrized linear systems arising from vibro-acoustics*, SIAM Conference on Linear Algebra in Signal Systems and Control, Boston (2001).

[236] V. Mehrmann and H. Xu, *Numerical methods in control*, J. Comput. Appl. Math., **123**: 371–394 (2000).

[237] J. Meinguet, *On the Glover concretization of the Adamjan-Arov-Krein approximation theory*, in Modelling, identification, and robust control, C.I. Byrnes and A. Lindquist, eds., 325–334 (1986).

[238] C.D. Meyer, *Matrix analysis and applied linear algebra*, SIAM, Philadelphia (2000).

[239] D.G. Meyer and S. Srinivasan, *Balancing and model reduction for second-order form linear systems*, IEEE Trans. Automat. Control, **41**: 1632–1644 (1996).

[240] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math., **11**: 50–59 (1960).

[241] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., **20**: 801–836 (1978).

[242] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Review, **45**: 3–49 (2003).

[243] B.C. Moore, *Principal component analysis in linear systems: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, **26**: 17–32 (1981).

[244] J.B. Moore, R.E. Mahoney, and U. Helmke, *Numerical gradient algorithms for eigenvalue and singular value calculations*, SIAM J. Matrix Anal. Appl., **15**: 881–902 (1994).

[245] C.T. Mullis and R.A. Roberts, *Synthesis of minimum roundoff noise fixed point digital filters*, Trans. Circuits Syst., **23**: 551–562 (1976).

[246] I.M. Navon, *Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography*, Dyn. Atmospheres Oceans, **27**: 55–79 (1998).

[247] B.R. Noack, K. Afanasiev, M. Morzynski, G. Tadmore, and F. Thiele, *A hierarchy of low dimensional models for the transient and post-transient cylinder wake*, J. Fluid Mech., **497**: 335–363 (2003).

[248] R.J. Ober and D. McFarlane, *Balanced canonical forms for minimal systems: A normalized coprime factorization approach*, Linear Algebra Appl., **122–124**: 23–64 (1989).

[249] R. Ober, *Balanced parametrization of classes of linear systems*, SIAM J. Control Optim., **29**: 1251–1287 (1991).

[250] R.J. Ober and A. Gheondea, *A trace formula for Hankel operators*, Proc. Amer. Math. Soc., **127**: 2007–2012 (1999).

[251] P.C. Opdenacker and E.A. Jonckheere, *A contraction mapping preserving balanced reduction scheme and its infinity norm error bounds*, IEEE Trans. Circuits Systems, **35**: 184–189, 1988.

[252] G. Obinata and B.D.O. Anderson, *Model reduction for control system design*, Springer-Verlag, New York (2000).

[253] Y. Ohta, *Hankel singular values and vectors of a class of infinite dimensional systems: Exact Hamiltonian formulas for control and approximation problems*, Math. Control, Signals Syst., **12**: 361–375 (1999).

[254] Y. Ohta, *Private communication*, 18 Oct. 2000.

[255] D.P. O'Leary and S. Peleg, *Digital image compression by outer product expansion*, IEEE Trans. Commun., **31**: 441–444 (1983).

[256] M. Overton, *Numerical computing with IEEE floating point arithmetic*, SIAM, Philadelphia (2001).

[257] A.M. Ostrowski and H. Schneider, *Some theorems on the inertia of matrices*, J. Math. Anal. Appl., **4**: 72–84 (1962).

[258] V. Papakos, *Restarted Lanczos algorithms for model reduction*, Ph.D. Thesis, Department of Electrical and Electronic Engineering, Imperial College, London (2003).

[259] B.N. Parlett, *The symmetric eigenvalue problem*, Prentice–Hall, Englewood Cliffs, NJ (1980).

[260] B.N. Parlett, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., **13**: 567–593 (1992).

[261] J.R. Partington, *An introduction to Hankel operators*, London Math. Soc., Stud. Texts, **13** (1988).

[262] R.V. Patel, A. Laub, and P.M. Van Dooren, eds., *Numerical linear algebra techniques for systems and control*, IEEE Press, Piscataway, NJ (1993).

[263] D.W. Peaceman and H.H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. SIAM, **3**: 28–41 (1955).

[264] R.L.M. Peeters and P. Rapisarda, *A two-variable approach to solve the polynomial Lyapunov equation*, Syst. Control Lett., **42**: 117–126 (2001).

[265] T. Penzl, *Algorithms for model reduction of large dynamical systems*, Technical Report SFB393/99-40, Technische Universität Chemnitz (1999).

[266] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: The symmetric case*, Syst. Control Lett., **40**: 139–144 (2000).

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

[268] J.E. Perkins, U. Helmke, and J.B. Moore, *Balanced realizations via gradient flows*, Syst. Control Lett., **14**: 369–380 (1990).

[269] L.T. Pillage and R.A. Rohrer, *Asymptotic waveform evaluation for timing analysis*, IEEE Trans. Comput. Aided Design, **9**: 352–366 (1990).

[270] J.W. Polderman and J.C. Willems, *Introduction to mathematical systems and control: A behavioral approach*, Text Appl. Math., **26** (1998).

[271] P. Rapisarda and J.C. Willems, *The subspace Nevanlinna interpolation problem and the most powerful unfalsified model*, Syst. Control Lett., **32**: 291–300 (1997).

[272] P. Rapisarda and J.C. Willems, *State maps for linear systems*, SIAM J. Control Optim., **35**: 1053–1091 (1997).

[273] M. Rathinam and L. Petzold, *A new look at proper orthogonal decomposition*, Technical Report, Computational Sciences and Engineering, University of California, Santa Barbara (2001).

[274] J.D. Roberts, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, **32**: 677–687 (1980).

[275] T.D. Romo, J.B. Clarage, D.C Sorensen, and G.N. Phillips, Jr., *Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time-averaged crystallographic refinements*, Proteins Structure Function Genetics, **22**: 311–321 (1995).

[276] W.E. Roth, *The equation $AX - YB = C$ and $AX - XB = C$ in matrices*, Proc. Amer. Math. Soc., **3**: 392–396 (1952).

[277] C.W. Rowley, T. Colonius, and R.M. Murray, *Model reduction of compressible flows using POD and Galerkin projection*, Physica D, **189**: 115–129 (2003).

[278] E.B. Rudnyi and J.G. Korvink, *Review: Automatic model reduction for transient simulation of MEMS-based devises*, Sensors Update, **11**: 3–33 (2002).

[279] A.E. Ruehli and A. Cangellaris, *Progress in the methodologies for the electrical modeling of interconnects and electronic packages*, Proc. IEEE, **89**: 740–771 (2001).

[280] W.J. Rugh, *Linear system theory*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ (1996).

[281] A. Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, Linear Algebra Appl., **58**: 391–405 (1984).

[282] S.M. Rump, *Ill-conditioned matrices are componentwise near to singularity*, SIAM Rev., **41**: 102–112 (1999).

[283] M.G. Safonov and R.Y. Chiang, *A Schur method for balanced truncation model reduction*, IEEE Trans. Automat. Control, **34**: 729–733 (1989).

[284] Y. Saad, *Numerical solution of large Lyapunov equations*, in Signal processing, scattering and operator theory, and numerical methods, Proc. MTNS-89, **3**: 503–511 (1990).

[285] Y. Saad, *Iterative methods for sparse linear systems*, 2nd ed., SIAM, Philadelphia (2003).

[286] H. Sandberg and A. Rantzer, *Balanced truncation of linear time-varying systems*, IEEE Trans. Automatic Control, **49**: 217–229 (2004).

[287] A.J. van der Schaft, *Duality for linear systems: External and state space characterization of the adjoint system*, in Analysis of controlled dynamical systems, B. Bonnard et al., eds., Birkhäuser, Boston, 393–403 (1991).

[288] C. Scherer and S. Weiland, *Linear matrix inequalities in control*, DISC lecture notes, version 3.0, 2000.

[289] E. Schmidt, *Zur Theorie der linearen und nichtlinearen Intergralgleichungen, Teil* I: *Entwicklung willkürlicher Funktionen nach System vorgeschriebener*, Math. Ann., **63**: 433–476 (1907).

[290] P. Schwarz, *Physically oriented modeling of heterogeneous systems*, in Proceedings of the 3rd IMACS Symposium on Mathematical Modeling, Wien, vol. 1, 309–318 (2000).

[291] P. Schwarz and P. Schneider, *Model library and tool support for MEMS simulation*, in MEMS Design, Fabrication, Characterization, and Packaging, Y.F. Behringer and D.G Uttamchandani, eds., SPIE Proceedings Series, Volume 4407, 10–23 (2001).

[292] N.L. Seaman, *Meteorological modeling for air quality assessments*, Atmos. Environ., **34**: 2231–2259 (2000).

[293] J.H. Seinfeld and S.N. Pandis, *Atmospheric chemistry and physics*, John Wiley, New York (1998).

[294] B. Shapiro, *Frequency weighted sub-system model truncation to minimize systems level model reduction errors*, Technical Report, Department of Aeronautics, University of Maryland, College Park (2001).

[295] S. Sherman and C.J. Thompson, *Equivalences on eigenvalues*, Indiana U. Math. J., **21**: 807–814 (1972).

[296] S. Shvartsman and I.G. Kevrekidis, *Nonlinear model reduction for control of distributed parameter systems*, J. AIChE, **44**: 1579 (1998).

[297] V. Simoncini, *On the numerical solution of* $AX - XB = C$, BIT, **36**: 814–830 (1996).

[298] L. Sirovich, *Turbulence and the dynamics of coherent structures, Parts* I–III, Quart. Appl. Math., **45**: 561–590 (1987).

[299] R.E. Skelton, T. Iwasaki, and K. Grigoriadis, *A unified algebraic approach to linear control system design*, Taylor & Francis, Philadelphia (1998).

[300] G.L.G. Sleijpen and H.A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., **17**: 401–425 (1996).

[301] S. Smale, *Complexity theory and numerical analysis*, Technical Report, Department of Mathematics, City University of Hong Kong (2000).

[302] M.C. Smith, *Singular values and vectors of a class of Hankel operators*, Systems Control Lett., **12**: 301–308 (1989).

[303] R.A. Smith, *Matrix Equation* $XA + BX = C$, SIAM J. Appl. Math, **16**: 198–201 (1968).

[304] E.D. Sontag, *Mathematical control theory*, Springer-Verlag, Berlin (1990).

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

[306] D.C. Sorensen, *Lecture Notes on Numerical Methods for Large Scale Eigenvalue Problems*, Draft of Lectures Delivered at the Royal Institute of Technology KTH, Stockholm, Spring 1999.

[307] D.C. Sorensen, *Numerical methods for large eigenvalue problems*, Acta Numer., **11**: 519–584 (2002).

[308] D.C. Sorensen, *Numerical linear algebra*, CAAM Department, Rice University, Houston, TX (2002).

[309] D.C. Sorensen, *Passivity preserving model reduction via interpolation of spectral zeros*, Systems Control Lett., **54**: 347–360 (2005).

[310] D.C. Sorensen and A.C. Antoulas, *The Sylvester equation and approximate balanced reduction*, Linear Algebra Appl., **351-352**: 671–700 (2002).

[311] D.C. Sorensen and Y. Zhou, *Direct methods for matrix Sylvester and Lyapunov equations*, J. Appl. Math., **6**: 277–304 (2003).

[312] B. Srinivasan and P. Myszkorowski, *Model reduction of systems with zeros interlacing the poles*, Systems Control Lett., **30**: 19–24 (1997).

[313] G. Starke, *Optimal alternating direction implicit parameters for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., **28**: 1431–1445 (1991).

[314] G.W. Stewart and J. Sun, *Matrix perturbation theory*, Academic Press, Boston (1990).

[315] G.W. Stewart, *The decompositional approach to matrix computation*, in Top 10 algorithms of the 20th century, Comput. Sci. Engrg., **2**: 50–59 (2000).

[316] D. Stroobandt, *Recent advances in system-level interconnect prediction*, IEEE Circuits Systems Newsletter, **11**: 4–20 (2000).

[317] T. Stykel, *Model reduction of descriptor systems*, Technical Report 720-2001, Institut für Mathematik, Technische Universität Berlin (2001).

[318] T. Stykel, *Analysis and numerical solution of generalized Lyapunov equations*, Ph.D. Thesis, Department of Mathematics, Technische Universität Berlin (2002).

[319] T. Stykel, *Balanced truncation model reduction for semidiscretized Stokes equation*, Technical Report 04-2003, Institut für Mathematik, Technische Universität Berlin (2003).

[320] T. Stykel, *Input-output invariants for descriptor systems*, Preprint PIMS-03-1, Pacific Institute of Mathematics (2003).

[321] J. Tautz and K. Rohrseitz, *What attracts honeybees to a waggle danser?*, J. Compar. Physiol. A, **183**: 661–667 (1998).

[322] D. Teegarden, G. Lorenz, and R. Neuf, *How to model and simulate microgyroscope systems*, IEEE Spectrum, **35**: 66–75 (1998).

[323] R. Tempo and F. Dabbene, *Randomized algorithms for analysis and control of uncertain systems: An overview*, in Perspectives in robust control, Springer-Verlag, London (2001).

[324] R.C. Thompson, *Matrix Spectral Theory*, Research Conference and Lecture Series, Notes of 10 Lectures Delivered at The Johns Hopkins University, Baltimore, MD, June 20–24, 1988.

[325] F. Tisseur and K. Meerbergen, *The quadratic eigenvalue problem*, SIAM Rev., **43**: 235–286 (2001).

[326] L.N. Trefethen, *Pseudospectra of matrices*, in Numerical analysis 1991, D.F. Griffiths and G.A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 234–266 (1992).

[327] L.N. Trefethen, *Computation of pseudospectra*, Acta Numer., **8**: 247–295 (1999).

[328] L.N. Trefethen and D. Bau III, *Numerical linear algebra*, SIAM, Philadelphia (1997).

[329] M. van Barel and A. Bultheel, *A canonical matrix continued fraction solution of the minimal realization problem*, Linear Algebra Appl., **122-124**: 973–1002 (1990).

[330] N.P. van der Meijs, *Model reduction for VLSI physical verification*, Technical Report, Department of ITS/EE, Delft University of Technology, Delft, The Netherlands (2000).

[331] A.-J. van der Veen, *A Schur method for low-rank matrix approximation*, SIAM J. Matrix Anal. Appl., **17**: 139–160 (1996).

[332] H.A. van der Vorst, *Krylov subspace iteration*, in Top 10 algorithms of the 20th century, Comput. Sci. Engrg., **2**: 32–37 (2000).

[333] P.M. Van Dooren, *The Lanczos algorithm and Padé approximations*, Benelux Meeting on Systems and Control, Houthalen, Belgium (1995).

[334] P.M. Van Dooren, *Software for control systems analysis and design, singular value decomposition*, in Wiley encyclopedia of electrical and electronics engineering, vol. 11, J.G. Webster, ed., 464–473 (1999).

[335] P.M. Van Dooren, *Orthogonal matrix decompositions in systems and control*, in Error Control and Adaptivity in Scientific Computing, H. Bulgak and C. Zenger, eds., Kluwer, Dordrecht, The Netherlands, 159–175 (1999).

[336] P.M. Van Dooren, *Gramian based model reduction of large-scale dynamical systems*, in Numerical Analysis 2000, 231–247 (2000).

[337] P.M. Van Dooren, *Numerical linear algebra for signals systems and control*, Graduate School in Systems and Control, Université Catholique de Louvain, Louvain-la-Neuve, Belgium (2003).

[338] P.M. Van Dooren, *The basics of developing numerical algorithms*, Control Syst. Magazine, **24**: 18–27 (2004).

[339] S. Van Huffel, V. Sima, A. Varga, S. Hammarling, and F. Delebeque, *High performance numerical software for control*, Control Syst. Magazine, **24**: 60–76 (2004).

[340] A. Varga, *Balancing-free square-root algorithm for computing singular perturbation approximations*, Proceedings of the 30th IEEE CDC, Brighton, 1062–1065 (1991).

[341] A. Varga, *Enhanced modal approach for model reduction*, Math. Model. Syst., **1**: 91–105 (1995).

[342] A. Varga, *On stochastic balancing related model reduction*, Proceedings of the IEEE Conference on Decision and Control, Sydney, 2385–2390 (2000).

[343] A. Varga, *A descriptor systems toolbox for MATLAB*, Proceedings of the IEEE International Symposium on Computer Aided Control System Design, Anchorage, 150–155 (2000).

[344] A. Varga and P.M. Van Dooren, *Computational methods for periodic systems: An overview*, IFAC Workshop on Periodic Control Systems, Como, 171–176 (2001).

[345] A. Varga, *Model reduction software in the SLICOT library*, in Applied and computational control, signals and circuits, B.N. Datta, ed., Kluwer Academic Publishers, Boston, 239–282 (2001).

[346] A. Varga and B.D.O. Anderson, *Accuracy enhancing methods for the frequency-weighted balancing related model reduction*, IEEE CDC, 3659–3664 (2001).

[347] A. Varga, ed., *Special issue on numerical awareness in control*, IEEE Control Syst. Magazine, **24** (2004).

[348] M. Verlaan, *Efficient Kalman filtering algorithms for hydrodynamic models*, Ph.D. Thesis, Technische Universiteit, Delft, The Netherlands (1998).

[349] S. Volkwein, *Proper orthogonal decomposition and singular value decomposition*, Technical Report SFB-153, Institut für Mathematik, Universität Graz (1999).

[350] E. Wachspress, *The ADI model problem*, NA Digest, **96** (1996).

[351] G. Wang, V. Sreeram, and W.Q. Liu, *A new frequency-weighted balanced truncation method and an error bound*, IEEE Trans. Automat. Control, **44**: 1734–1737 (1999).

[352] J.M. Wang, C.-C. Chu, Q. Yu, and E.S. Kuh, *On projection-based algorithms for model-order reduction of interconnects*, IEEE Trans. Circuits Systems I: Fundamental Theory Appl., **49**: 1563–1585 (2002).

[353] J.H.M. Wedderburn, *Lectures on matrices*, Colloquium publications, vol. XVII, AMS, New York (1934), and Dover, New York (1964).

[354] S. Weiland, *A behavioral approach to balanced representations of dynamical systems*, Linear Algebra Appl., **205-206**: 1227–1253 (1994).

[355] J.H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press, Oxford, UK (1965).

[356] K. Willcox and J. Peraire, *Balanced model reduction via the proper orthogonal decomposition*, AIAA J., **40**: 2323–2330 (2002).

[357] J.C. Willems, *Private communication*, 9 Feb. 2000.

[358] J.C. Willems, *Private communication*, 1 June 2001.

[359] J.C. Willems, *Mathematical models of systems*, IUAP Graduate Course, Catholic University of Leuven (2002).

[360] J.C. Willems, *Dissipative dynamical systems, Part* I: *General theory*, Arch. Ration. Mech. Anal., **45**: 321–351 (1972).

[361] J.C. Willems, *Dissipative dynamical systems, Part* II: *Linear systems with quadratic supply rates*, Arch. Ration. Mech. Anal., **45**: 352–393 (1972).

[362] P.M. Wortelboer, M. Steinbuch, and O. Bosgra, *Iterative model and controller reduction using closed-loop balancing, with application to a compact disc mechanism*, Internat. J. Robust Nonlinear Control, **9**: 123–142 (1999).

[363] Y. Yamamoto, K. Hirata, and A. Tannenbaum, *Some remarks on Hamiltonians and the infinite dimensional one block $\mathcal{H}_\infty$ problem*, Systems Control Lett., **29**: 111–118 (1996).

[364] C. Yang, D.W. Noid, B.G. Sumpter, D.C. Sorensen, and R.E. Tuzun, *An efficient algorithm for calculating the heat capacity of a large-scale molecular system*, Macromolecular Theory Simulations, **10**: 756–761 (2001)

[365] N.J. Young, *An introduction to Hilbert space*, Cambridge University Press, Cambridge, UK (1988).

[366] A. Yousuff, D.A. Wagie, and R.E. Skelton, *Linear system approximation via covariance equivalent realizations*, J. Math. Anal. Appl., **106**: 91–115 (1985).

[367] A. Yousouff and R.E. Skelton, *Covariance equivalent realizations with application to model reduction of large-scale systems*, in Control and Dynamic Systems, C.T. Leondes, ed., Academic Press, Orlando, FL, **22**: 273–348 (1985).

[368] K. Zhou and P.P. Khargonekar, *On the weighted sensitivity minimization problem for delay systems*, Systems Control Lett., **8**: 307–312 (1987).

[369] K. Zhou, *Frequency-weighted $\mathcal{L}_\infty$ norm and optimal Hankel norm model reduction*, IEEE Trans. Automat. Control, **40**: 1687–1699 (1995).

[370] K. Zhou, J.C. Doyle, and K. Glover, *Robust and optimal control*, Prentice–Hall, Englewood Cliffs, NJ (1996).

[371] K. Zhou and J.C. Doyle, *Essentials of robust control*, Prentice–Hall, Englewood Cliffs, NJ (1997).

[372] K. Zhou, G. Salomon, and E. Wu, *Balanced realization and model reduction for unstable systems*, Internat. J. Robust Nonlinear Control, **9**: 183–198 (1999).

[373] Y. Zhou, *Numerical methods for large scale matrix equations with applications in LTI system model reduction*, Ph.D. Thesis, CAAM Department, Rice University, Houston, TX (2002).

This page intentionally left blank

Index

This page intentionally left blank

Πάντ᾿ ἀνοιχτά πάντ᾿ ἄγρυπνα τά μάτια τῆς ψυχῆς μου

Οἱ Ἐλεύϑεροι Πολιορϰημένοι
Σχεδίασμα Β′
Διονύσιος Σολωμός

Always open, always vigilant, the eyes of my soul

The Free Besieged
Second Draft
Dionysios Solomos