Interpolatory Methods for Model Reduction of Large-Scale Dynamical Systems

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von **Dipl.-Math. techn. Tobias Breiten**

geb. am **05.11.1985** in Neuwied

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: **Prof. Dr. Peter Benner**

Prof. Dr. Athanasios Antoulas

eingereicht am: **05.11.2012**

Verteidigung am: **25.03.2013**

PUBLICATIONS

Large parts of this thesis have been published or are submitted for publication.

Chapter 3 is a revised and extended version of

> [19]: Peter Benner, Tobias Breiten: On optimality of interpolation-based low-rank approximations of large-scale matrix equations, MPIMD/11-10, Max Planck Institute Magdeburg Preprints, 2011.

The first part of Chapter 4 has been published in

> [20]: Peter Benner, Tobias Breiten: Interpolation-Based $\mathcal{H}_2$-Model Reduction of Bilinear Control Systems, SIAM Journal on Matrix Analysis and Applications, 33 (2012), pp. 859–885.

The second part of Chapter 4 is accepted for publication in Numerische Mathematik and is available as preprint

> [22]: Peter Benner, Tobias Breiten: Low rank methods for a class of generalized Lyapunov equations and related issues, MPIMD/12-03, Max Planck Institute Magdeburg Preprints, 2012.

Chapter 5 is available as preprint

> [21]: Peter Benner, Tobias Breiten: Two-sided moment matching methods for nonlinear model reduction, MPIMD/12-12, Max Planck Institute Magdeburg Preprints, 2012.

ACKNOWLEDGEMENTS

Writing a PhD thesis without having adequate support, be it of a professional or of a private nature, is hardly possible. Fortunately, I have been surrounded by several people that I want to thank for their assistance during the past three years.

I want to thank my supervisor Professor Peter Benner for his guidance and his constant support throughout the time in Chemnitz and Magdeburg. I am indebted to his friendly and uncomplicated manner, but also to his interest in my educational as well as personal development that helped to keep the number of setbacks at a minimum. I would also like to thank Professor Athanasios C. Antoulas who agreed to be one of the referees for this thesis. Since his book 'Approximation of Large-Scale Dynamical Systems' laid the groundwork for the following results, I think there could not have been a more suitable choice.

I am further grateful for lots of pleasant colleagues that I worked with during my time in the research groups 'Mathematik in Industrie und Technik' and 'Computational Methods in Systems and Control Theory'. Although nearly all of them contributed in their own way to this thesis, there are a few I want to point out explicitly. I am obliged to Jens Saak for his helpful advices and for giving me a good start in Chemnitz. I would like to thank my office colleagues Sara Grundel, Patrick Kürschner, André Schneider and Martin Stoll for creating such a friendly atmosphere. I really enjoyed working with you. A special thanks goes to Thomas Mach. Not only for solving countless LaTeX and TikZ related problems that I posed, but rather for keeping me company while running, for the most interesting and worthwhile work-related as well as private discussions.

Mein abschließender und größter Dank geht gleichermaßen an meine Familie sowie an meine Freundin Anne, denen ich die längste und schönste Zeit der Unterstützung zu verdanken habe. Meinen Eltern und meinen beiden Brüdern möchte ich für den kontinuierlichen Rückhalt über die letzten Jahre danken. Auf Eure Weise hattet Ihr den wesentlichen Einfluss auf das Entstehen dieser Arbeit. Bei Anne bedanke ich mich nicht nur für die vielen Stunden Korrekturlesen dieser Arbeit, sondern vor allem für die wundervolle gemeinsame Zeit. Durch Dich habe ich feststellen dürfen, wie leicht und unbeschwert das letzte Jahr einer Promotion doch verlaufen kann; vielen Dank dafür!

ABSTRACT

In this thesis, we study interpolation-based model order reduction techniques for large-scale linear, bilinear and quadratic-bilinear control systems. A particular focus lies on the $\mathcal{H}_2$-optimal model reduction problem. Based on existing theory for linear $\mathcal{H}_2$-optimal model reduction, we derive several new results that find application in the approximate solution of large-scale linear matrix equations. This includes a new connection between the topic of Riemannian optimization on matrix manifolds and the concept of rational interpolation. We further propose a method for locally minimizing the residual of the Lyapunov equation for a given rank $\hat{n}$. As is shown, the idea can be interpreted as a special case of the $\mathcal{H}_2$-optimal model order reduction problem for bilinear control systems. Moreover, for this special class of nonlinear control systems, we derive an abstract interpolation-based model reduction technique that aims at minimizing the bilinear $\mathcal{H}_2$-norm. New optimality conditions are computed and compared with existing ones that are based on generalized Lyapunov equations arising in the context of bilinear control theory. These matrix equations so far constituted a bottleneck within the method of balanced truncation for bilinear systems. Based on results from linear control theory, we show that under certain assumptions a fast exponential singular value decay of the solution matrix allows to approximately solve these equations via appropriate low rank methods. By means of numerical examples ranging up to dimensions $n = 562\ 500$, we demonstrate the efficiency of several new approaches. Finally, we investigate a recently introduced framework for model reduction of more general nonlinear control systems. This leads to the analysis of so-called quadratic-bilinear control systems. We show how tools and results from tensor theory can be used to improve the existing method with regard to computational efficiency as well as approximation accuracy. Again, numerical examples resulting from the spatial discretization of nonlinear partial differential equations are used to compare our method with current state-of-the-art techniques.

ZUSAMMENFASSUNG

Die vorliegende Arbeit behandelt interpolationsbasierte Modellordnungsreduktionstechniken für große lineare, bilineare sowie quadratisch-bilineare Regelungssysteme. Ein spezieller Fokus liegt hierbei auf dem Problem der $\mathcal{H}_2$-optimalen Modellreduktion. Basierend auf existierender Theorie für das lineare $\mathcal{H}_2$-Modellreduktionsproblem leiten wir diverse neue Resultate her, die Anwendung in der approximativen Lösung von großen, linearen Matrixgleichungen finden. Unter anderem beinhaltet das eine neue Beziehung zwischen dem Bereich der Riemannoptimierung auf Matrixmannigfaltigkeiten und dem Konzept der rationalen Interpolation. Weiterhin entwickeln wir eine Methode, die für einen vorgegebenen Rang $\hat{n}$ das Residuum der Lyapunovgleichung lokal minimiert. Wir zeigen, dass die dahintersteckende Idee als Spezialfall des $\mathcal{H}_2$-optimalen Modellreduktionsproblems für bilineare Regelungssysteme interpretiert werden kann. Für diese spezielle Klasse von nichtlinearen Regelungssystemen entwickeln wir eine abstrakte interpolationsbasierte Modellreduktionsmethode die zum Ziel hat, die bilineare $\mathcal{H}_2$-Norm lokal zu minimieren. Neue Optimalitätsbedingungen werden berechnet und mit existierenden Bedingungen verglichen. Letzere beruhen auf verallgemeinerten Lyapunovgleichungen, die im Bereich der bilinearen Regelungstheorie auftreten. Diese Lyapunovgleichungen wiederum wurden bislang als großer Nachteil der Methode des balancierten Abschneidens für bilineare Systeme angesehen. Mit Hilfe von Resultaten aus der linearen Regelungstheorie zeigen wir, dass die Lösungen dieser Matrixgleichungen unter gewissen Annahmen einen starken Singulärwertabfall aufweisen, der es ermöglicht die Gleichung durch Niedrigrangmethoden approximativ zu lösen. Anhand von numerischen Beispielen bis zur Größenordnung $n = 562500$ demonstrieren wir den Nutzen von verschiedenen, neuen Ansätzen. Schließlich untersuchen wir einen kürzlich eingeführten Ansatz zur Modellreduktion einer allgemeineren Klasse von nichtlinearen Regelungssystemen. Das wird uns zur Analyse von sogenannten quadratisch-bilinearen Regelungssystemen führen. Wir bedienen uns bestimmter Techniken aus der Tensortheorie, die es ermöglichen, den existierenden Ansatz im Hinblick auf numerische Effizienz sowie Approximationsgüte zu verbessern. Anhand von numerischen Beispielen, resultierend aus der räumlichen Diskretisierung von partiellen Differentialgleichungen, vergleichen wir unsere Methode mit anerkannten Methoden der aktuellen Wissenschaft.

CONTENTS

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ACRONYMS

ADI	alternating directions implicit
BIBO	bounded-input-bounded-output
BiCG	biconjugate gradient
BIRKA	bilinear iterative rational Krylov algorithm
BPIM	best points interpolation method
CG	conjugate gradient
DAE	differential-algebraic equation
(D)EIM	(discrete) empirical interpolation method
FEM	finite element method
GMRES	generalized minimal residual
IRKA	iterative rational Krylov algorithm
KPIK	Krylov-Plus-Inverted-Krylov
LPV	linear parameter-varying
LRCF	low rank Cholesky factor
LTI	linear time-invariant
MEMS	micro-electro-mechanical systems
MIMO	multiple-input and multiple-output
MinRes	minimal residual
MIRIAm	MIMO iterative rational interpolation algorithm
MATLAB	software from The MathWorks Inc.
MOR	model order reduction
MPE	missing point estimation
ODE	ordinary differential equation
PDE	partial differential equation
POD	proper orthogonal decomposition
TPWL	trajectory piecewise linear
QBDAE	quadratic-bilinear differential-algebraic equation
SISO	single-input and single-output
SLICOT	subroutine library for control theory
SVD	singular value decomposition
Xeon®	processor series from Intel®

NOTATION

$\mathbb{N}$	natural numbers $\{0, 1, 2, \dots\}$
$\mathbb{R}, \mathbb{C}$	field of real, complex numbers
$\mathbb{C}_-$	left half of the complex plane
$\mathbb{D}$	open unit disc around 0
$\mathbb{R}[s]^{p\times m}$	ring of $p \times m$ polynomial matrices in s with real coefficients
$\mathbb{R}(s)^{p\times m}$	quotient field of $\mathbb{R}^{p\times m}[s]$
$\mathrm{Re}\,(z)$	real part of a complex number z
i	imaginary unit or index, depending on context
$\mathbb{R}^{n\times m}$	vector space of real matrices with m rows and n columns
$\mathbb{R}^n$	equal to $\mathbb{R}^{n\times 1}$
$\mathcal{M}$	manifold of symmetric positive semi-definite matrices of rank $\hat{n}$
$\mathbf{x}$	vector $\in \mathbb{R}^n$
$\mathbf{A}$	matrix $\in \mathbb{R}^{n\times m}$
$\mathbf{A}^T$	transpose of a matrix $\mathbf{A}$
$\mathbf{A}^*$	complex conjugate transpose of a matrix $\mathbf{A}$
$\mathrm{diag}\,(\mathbf{d})$	diagonal matrix with diagonal $\mathbf{d} \in \mathbb{R}^n$
$\mathbf{I}_n, \mathbf{I}$	identity matrix of size $n \times n$ resp. of suitable size
$\mathbf{0}_{n\times m}, \mathbf{0}$	zero matrix of size $n \times m$ resp. of suitable size
$\mathbf{e}_i$	i-th column of the identity matrix $\mathbf{I}$
$\kappa_{\mathbf{A}}$	condition number of a matrix $\mathbf{A}$
$\lambda_i(\mathbf{A})$	i-th eigenvalue of a matrix $\mathbf{A}$
$\sigma(\mathbf{A})$	spectrum of a matrix $\mathbf{A}$
$\mathbf{A}_{ij}$	entry (i, j) of a matrix $\mathbf{A}$
$\mathbf{A}_{k:l,m:n}$	submatrix of $\mathbf{A}$ with entries $\mathbf{A}_{i,j},\ i \in \{k, \dots, l\}$ and $j \in \{m, \dots, n\}$
$\mathbf{A}_k$	k-th column of a matrix $\mathbf{A}$
$\mathbf{A} = \mathbf{A}^T \succ 0$	$\mathbf{A}$ is symmetric positive definite
$\mathrm{rank}\,(\mathbf{A})$	rank of a matrix $\mathbf{A}$
$\mathrm{span}\,(\mathbf{A})$	subspace spanned by the columns of a matrix $\mathbf{A}$

$\operatorname{orth}(\mathbf{A})$	orthonormal subspace spanned by the columns of a matrix $\mathbf{A}$
$\operatorname{tr}(\mathbf{A})$	trace of a matrix $\mathbf{A}$, $\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$
$\operatorname{vec}(\cdot)$	vectorization operator .... Definition 2.1.1
$\operatorname{vec}^{-1}(\cdot)$	inverse function of $\operatorname{vec}(\cdot)$
ξ_m	vectorized identity matrix of dimension m, i.e. $\xi_m = \operatorname{vec}(\mathbf{I}_m)$
$\otimes$	Kronecker product .... Definition 2.1.1
$\mathcal{L}$	(generalized) Lyapunov operator $\mathbf{E} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{E}$
$\mathcal{L}(\mathbf{X})$	(generalized) Lyapunov operator $\mathbf{AXE}^T + \mathbf{EXA}^T$
$\Pi(\mathbf{X})$	positive linear operator $\sum_{k=1}^{m} \mathbf{N}_k \mathbf{X} \mathbf{N}_k^T$
Π	linear operator $\sum_{j=1}^{m} \mathbf{N}_j \otimes \mathbf{N}_j$
$\mathcal{D}$	(generalized) Stein operator $\mathbf{A} \otimes \mathbf{A} - \mathbf{E} \otimes \mathbf{E}$
$\mathcal{L}_S$	(generalized) Sylvester operator $\mathbf{E} \otimes \mathbf{A} + \mathbf{H} \otimes \mathbf{M}$
$\mathcal{L}_d$	d-dimensional linear operator $\sum_{i=1}^{d} \mathbf{I} \otimes \cdots \otimes \mathbf{I} \otimes \mathbf{A}_i \otimes \mathbf{I} \otimes \cdots \otimes \mathbf{I}$
$\mathcal{A}_d$	d-dimensional linear operator $\mathcal{L}_d + \sum_{j=1}^{k} \mathbf{N}_{j_1} \otimes \cdots \otimes \mathbf{N}_{j_d}$
$\mathcal{T}$	lin. operator $\mathcal{T} = \mathbf{I} \otimes \mathbf{A}^T\mathbf{A} + \mathbf{A}^T\mathbf{A} \otimes \mathbf{I} + \mathbf{A}^T \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{A}^T$ Eq. 3.35
$\mathbb{L}$	Laplace transform $\mathbf{x}(t) \mapsto \mathbf{x}(s) = \int_0^\infty e^{-st}\mathbf{x}(t)\mathrm{d}t$ ...... Equation 2.6
$\|\|\mathbf{A}\|\|_2$	spectral norm $\max_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\|\mathbf{Ax}\|\|_2}{\|\|\mathbf{x}\|\|_2}$ of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$
$\|\|\mathbf{A}\|\|_F$	Frobenius norm $\|\|\mathbf{A}\|\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} \|\mathbf{A}_{ij}\|^2}$ of a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$
$\|\|\mathbf{P}\|\|_{\mathcal{L}}$	energy norm of a matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ induced by the $\mathcal{L}$-inner product $\langle \operatorname{vec}(\mathbf{P}), \operatorname{vec}(\mathbf{Q})\rangle_{\mathcal{L}} = \langle -\mathcal{L}\operatorname{vec}(\mathbf{P}), \operatorname{vec}(\mathbf{Q})\rangle$
$\|\|\mathbf{P}\|\|_{\mathcal{D}}$	energy norm of a matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$ induced by the $\mathcal{D}$-inner product $\langle \operatorname{vec}(\mathbf{P}), \operatorname{vec}(\mathbf{Q})\rangle_{\mathcal{D}} = \langle -\mathcal{D}\operatorname{vec}(\mathbf{P}), \operatorname{vec}(\mathbf{Q})\rangle$
$\|\|\mathbf{X}\|\|_{\mathcal{L}_S}$	energy norm of a matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ induced by the $\mathcal{L}_S$-inner product $\langle \operatorname{vec}(\mathbf{X}), \operatorname{vec}(\mathbf{Y})\rangle_{\mathcal{L}_S} = \langle -\mathcal{L}_S\operatorname{vec}(\mathbf{X}), \operatorname{vec}(\mathbf{Y})\rangle$
$\mathbf{\Sigma}$	linear control system .... Equation 2.2
$\mathbf{\Sigma}_B$	bilinear control system .... Equation 4.1
$\mathbf{\Sigma}_Q$	quadratic-bilinear control system .... Equation 5.3
$\|\|\mathbf{\Sigma}\|\|_{\mathcal{H}_2}$	$\mathcal{H}_2$-norm $\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \operatorname{tr}\left(\overline{\mathbf{H}(i\omega)}\mathbf{H}(i\omega)\right) \mathrm{d}\omega\right)^{\frac{1}{2}}$ of a linear control system $\mathbf{\Sigma}$ with transfer function $\mathbf{H}$ .... Equation 2.27
$\operatorname{res}[H(s), \mu]$	residue $\lim_{s\to\mu}(s-\mu)\cdot H(\mu)$ of a rational function $H(s)$
$\mathcal{H}_2$	Hardy space denoting the set of square integrable functions
$\|\|\mathbf{\Sigma}_B\|\|_{\mathcal{H}_2}$	$\mathcal{H}_2$-norm of a bilinear dynamical system $\mathbf{\Sigma}_B$ .... Definition 4.2.1
$\mathcal{A}^j_{\mathbf{E},\sigma}$	abbreviation for $((\sigma\mathbf{E}-\mathbf{A})^{-1}\mathbf{E})^j\,(\sigma\mathbf{E}-\mathbf{A})^{-1}$ ...... Definition 5.4.1
$\mathcal{A}^{T,j}_{\mathbf{E},\sigma}$	abbreviation for $\left((\sigma\mathbf{E}^T-\mathbf{A}^T)^{-1}\mathbf{E}^T\right)^j(\sigma\mathbf{E}^T-\mathbf{A}^T)^{-1}$ Definition 5.4.1
$\mathcal{K}_q(\mathbf{A}, \mathbf{b})$	Krylov subspace spanned by $\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{q-1}\mathbf{b}\}$
$\mathcal{K}_q(\mathbf{E}, \mathbf{A}, \mathbf{b}, \sigma)$	rational Krylov subspace $\mathcal{K}_q((\sigma\mathbf{E}-\mathbf{A})^{-1}\mathbf{E}, (\sigma\mathbf{E}-\mathbf{A})^{-1}\mathbf{b})$

CHAPTER 1

INTRODUCTION

Contents

1.1 Dynamical control systems and model order reduction

The mathematical study and analysis of dynamical processes, i.e., processes that vary with time t, certainly is one of the most important and challenging topics in the wide field of numerical analysis. In general, almost all real-life applications can be modeled as systems of partial differential equations (PDEs) and/or ordinary differential equations (ODEs). Sometimes, these are subject to additional algebraic constraints, leading to differential-algebraic equations (DAEs). The demand of having accurate models frequently leads to very complex mathematical systems that require a large amount of computational resources when studied and analyzed on a computer. At this point, the term *complex* can reflect rather different meanings. For example, the number of equations of the underlying mathematical system, i.e., the *state dimension* n, can define a complex model. In fact, it is not uncommon that one encounters systems with $n \sim 10^6$. It is clear that the simulation time of such systems, usually called *large-scale*, directly depends on n. However, there are other properties that can define a complex dynamical system. For example, a system whose dynamics is of linear nature in general is less complex than one that is nonlinear. Despite the fact that, so far, we have not explicitly defined a linear system, for readers with a mathematical background it should not be too surprising that such systems belong to a special and, in some sense, easier case and thus can be treated with exclusive methods. Finally, the already mentioned presence of algebraic constraints can easily complicate the desired analysis and, hence, also determines the complexity of a system.

Let us now proceed with a more rigorous and formal introduction to the topic of this thesis. Throughout this work, we study dynamical control systems, i.e., a set of ODEs whose dynamics can be influenced through external forces by means of a control input. In a general form, these systems are given by a *state equation*

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) + \mathbf{g}(\mathbf{x}(t), \mathbf{u}(t), t),$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ is the solution trajectory of the system, $\mathbf{f} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ and $\mathbf{g} : \mathbb{R}^{n+m+1} \to \mathbb{R}^n$ are functions with smoothness properties that are specified later on and $\mathbf{u}(t) : \mathbb{R} \to \mathbb{R}^m$ is a bounded input function. Anticipating later examples, one can think of, e.g., industrial cooling processes that are mathematically described by the heat equation. Assuming that we can specify the temperature on some region of the work piece that should be cooled allows to model the dynamics by a system of the previous form. Moreover, it also explains why the state dimension n indeed can become inconveniently large. Since in our setting we assume that $\mathbf{x}(t)$ only varies with time, processes that exhibit a spatial distribution first have to be semi-discretized in space. This can be done by, e.g., a finite difference method (FDM) or a finite element method (FEM), respectively. However, even for one-dimensionally distributed processes, often a very fine resolution of the discretization is required in order to guarantee an accurate approximation of the underlying PDE. As a consequence, the state dimension n also increases uncomfortably. Unfortunately, in this case, classical control theoretic concepts such as stability analysis, frequency response analysis or optimal control problems are stretched to their limits and can no longer be efficiently realized. On the other hand, in most applications one often is not interested in the entire system state $\mathbf{x}(t)$ anyway. Instead, a dynamical control system comes along with an *output equation*

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t), t) + \mathbf{k}(\mathbf{x}(t), \mathbf{u}(t), t),$$

where $\mathbf{h} : \mathbb{R}^{n+1} \to \mathbb{R}^p$ and $\mathbf{k} : \mathbb{R}^{n+m+1} \to \mathbb{R}^p$ again are smooth functions. Coming back to the mentioned example from industrial cooling, a typical interpretation of $\mathbf{y}(t)$ is given by the average temperature of the work piece. The key to control dynamical processes even for large n now is the fact that usually the dimensions m, p of the input and output functions $\mathbf{u}(t), \mathbf{y}(t)$, respectively, are much smaller than the actual system dimension n. Hence, if we consider the system as a black box model for $\mathbf{x}(t)$ and rather decide to analyze the mapping $\mathbf{z} : \mathbb{R}^m \to \mathbb{R}^p,\ \mathbf{u}(t) \mapsto \mathbf{y}(t)$, there might be parts of $\mathbf{x}(t)$ that are less important than others and for this reason can be neglected without influencing $\mathbf{z}(t)$ significantly. The mathematical concept of *model order reduction (MOR)* is motivated by exactly this consideration and tries to replace the black box model for $\mathbf{x}(t)$ by another one of much smaller state dimension $\hat{n} \ll n$, such that one can come up with a modified input-output mapping $\hat{\mathbf{z}}(t)$ approximating the original one. To be more precise, in this thesis we are interested in the construction of a reduced-order model

$$\begin{aligned}\dot{\hat{\mathbf{x}}}(t) &= \hat{\mathbf{f}}(\hat{\mathbf{x}}(t), t) + \hat{\mathbf{g}}(\hat{\mathbf{x}}(t), \mathbf{u}(t), t),\\ \hat{\mathbf{y}}(t) &= \hat{\mathbf{h}}(\hat{\mathbf{x}}(t), t) + \hat{\mathbf{k}}(\hat{\mathbf{x}}(t), \mathbf{u}(t), t),\end{aligned}$$

with $\hat{\mathbf{f}} : \mathbb{R}^{\hat{n}+1} \to \mathbb{R}^{\hat{n}}$, $\hat{\mathbf{g}} : \mathbb{R}^{\hat{n}+m+1} \to \mathbb{R}^{\hat{n}}$, $\hat{\mathbf{h}} : \mathbb{R}^{\hat{n}+1} \to \mathbb{R}^{p}$, and $\hat{\mathbf{k}} : \mathbb{R}^{\hat{n}+m+1} \to \mathbb{R}^{p}$. Of course, the essential goal of MOR lies in minimizing $||\mathbf{y}(t) - \hat{\mathbf{y}}(t)||$ as well as $\hat{n}$ for a large class of system inputs $\mathbf{u}(t)$. Depending on the specific nature of the functions $\mathbf{f}, \mathbf{g}, \mathbf{h}$ and $\mathbf{k}$, the development of appropriate numerical algorithms is of different complexity. As we mentioned in the beginning, a special position among all models is taken by linear time-invariant control systems

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

where $\mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{B} \in \mathbb{R}^{n \times m}, \mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{D} \in \mathbb{R}^{p \times m}$. Although theory is quite well-established for such systems, there still exist open problems that are clearly worth studying. Unfortunately, most real-life applications can hardly be described by a linear model. A remedy is given by the linearization of an actually nonlinear model around an operating point. However, a linearization often only allows for locally accurate approximations. A more sophisticated approach can be realized by so-called *bilinear control systems*, an important subclass of nonlinear control systems. Since we provide a detailed description of these models later on, at this point we refrain from a further discussion of the topic. In situations where even the latter systems fail to give a faithful representation of the true dynamics, one can transform the model into a system of *quadratic-bilinear differential algebraic equations (QBDAEs)* which than can be reduced by suitable MOR techniques.

1.2 Motivating examples

In order to get a better understanding of processes that indeed can be handled by tools from the area of MOR, we subsequently present three different motivating examples that underscore its practical use.

A model of a CD player

According to the structure of this thesis, we start with a model from linear control theory. The subject of interest is a classical CD player. From a control theoretic point of view, following the description in [38], the task is to construct a controller that ensures that the laser spot points to the track of pits on a rotating CD. The system theoretic

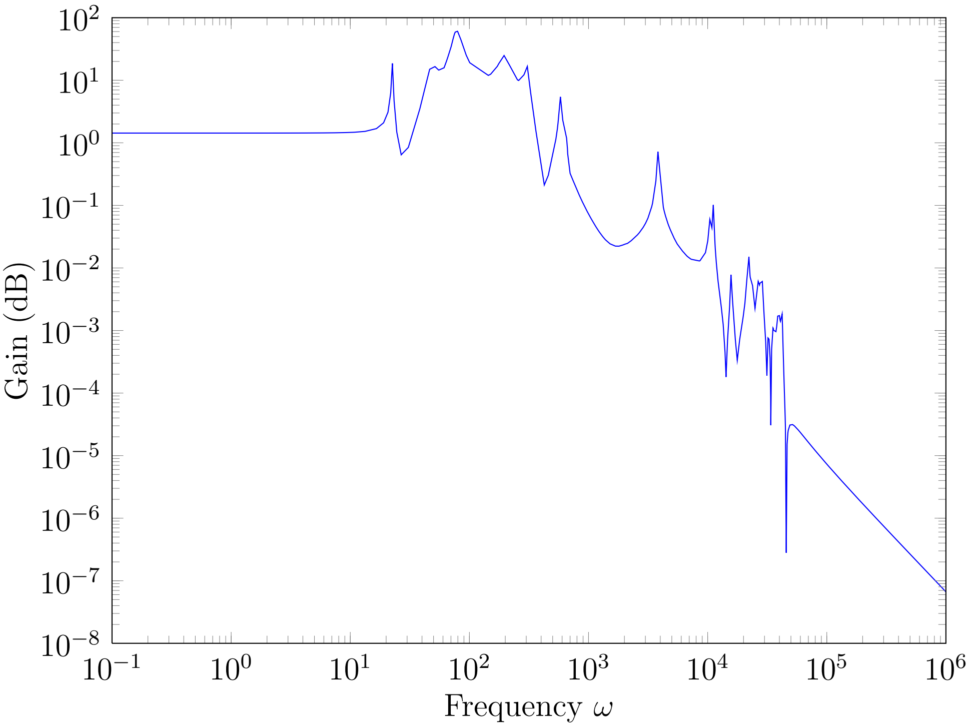


Figure 1.1: Transfer function of a CD player model.

description of the underlying process can be found as part of the SLICOT benchmark collection[1] and is one of the standard test problems for MOR of linear systems. Despite the fact that this model has been used in the literature for several years now, due to its complicated nature, it is still useful if the efficiency of linear reduction techniques should be tested.

In engineering applications, the behavior of a system often is studied in the frequency domain rather than in the time domain. In particular, a clear picture of the system dynamics can be drawn from the transfer function. Since we discuss the detailed concept in Chapter 2, at this point, we interpret a transfer function $H(s) = \mathbf{c}^T(\mathbf{sI} - \mathbf{A})^{-1}\mathbf{b}$ as a rational function in the frequency variable s, determined by the system matrices $(\mathbf{A}, \mathbf{b}, \mathbf{c})$. Basically, if this function is known, one can compute the output response of a linear system to arbitrary input signals. Consequently, it is one of the most important tools for classical linear control theory. For readers not familiar with those concepts, in Figure 1.1, we see the transfer function of the CD player over the frequency range $[10^{-1}, 10^{6}]$. For the construction of a satisfying reduced-order model it is now important that its transfer function is close to the one from Figure 1.1. We have chosen this particular example to demonstrate possible difficulties that can arise throughout the reduction process. As is shown, the original transfer function exhibits many peaks. Interpreting

[1] http://www.slicot.org/index.php?site=benchmodred

MOR as an interpolation problem, coming up with a *good* reduced interpolant is very difficult since the function is not very smooth and, hence, a large number of interpolation points is needed for a successful reproduction. Moreover, a central question for the performance of a reduced-order model is the location of the interpolation points and throughout the thesis we provide several (known) statements about optimality with respect to a specific accuracy measure.

The Fokker-Planck equation

Despite the fact that in this thesis we are basically dealing with deterministic processes, the second example actually has its origin in stochastics. However, as we see later on, there is an interesting and very close connection between linear stochastic dynamical processes and bilinear ones. For now, we simply describe the stochastic model and its application. We give a brief recapitulation of the more detailed explanation in [77]. Let us assume that we are interested in the motion of a dragged Brownian particle on the real line assuming states $x \in \mathbb{R}$. According to [77], if the particle is confined by a double-well potential

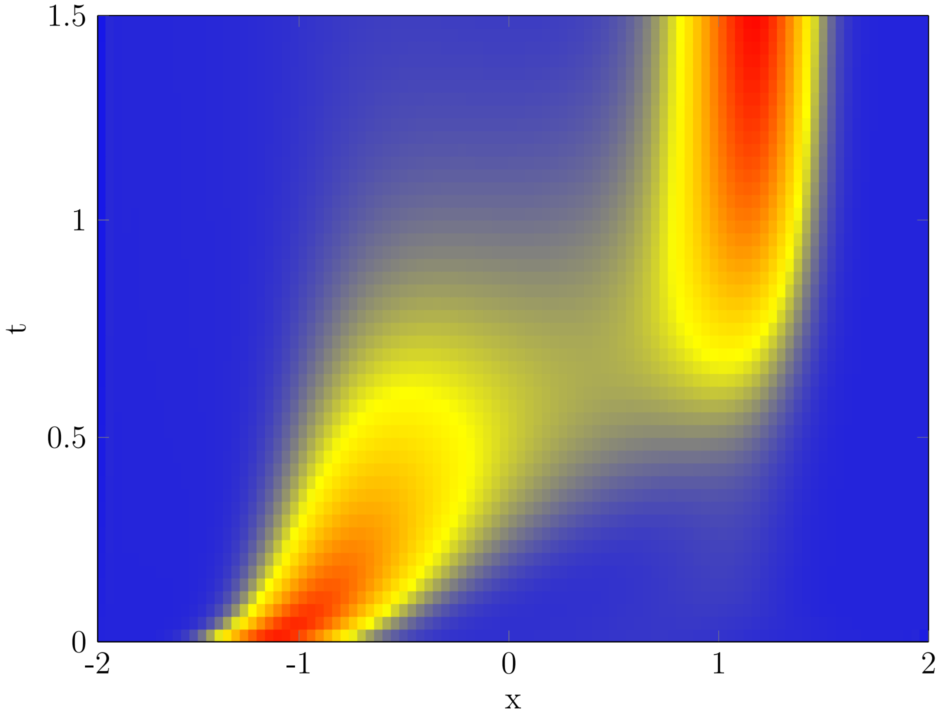


Figure 1.2: Spatio-temporal probability discribution of a Brownian particle.

$$W(x) = (x^2 - 1)^2,$$

the dynamics of the motion can be described by the stochastic partial differential equation

$$\mathrm{d}X_t = -\nabla V(X_t,t)\mathrm{d}t + \sqrt{2\sigma}\mathrm{d}W_t,$$

with $0 < \sigma < \frac{1}{2}$ and $V(x,t) = W(x) - ux$. Here, $X_t \in \mathbb{R}$ denotes the location of the particle at time point t, $-\nabla V(X_t,t)$ denotes the drift term of the process and σ is referred to as the diffusion coefficient. If we instead consider the probability distribution function

$$\rho(x,t)\mathrm{d}x = \mathbb{P}\left[X_t \in [x, x+\mathrm{d}x]\right],$$

the system can be replaced by means of the Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \sigma\Delta\rho + \nabla\cdot(\rho\nabla V),$$

that, after a spatial discretization, automatically leads to a bilinear system. In Figure 1.2, we present a possible probability distribution evolving with time t. Here, the finite difference discretization leads to a system of dimension $n = 100$. As we can see, initially the particle is located in the left potential well and is dragged to the right potential well by means of a suitable control function u. A reproduction of the entire system state, i.e., of the probability distribution unfortunately is hard to realize. Instead, we can consider the probability of the particle being located in the right potential well as an output of the system. As a consequence, the dimension of the output function is very small compared to n, representing the desired setup for MOR purposes.

The FitzHugh-Nagumo system

Finally, from the area of nonlinear control systems, we present a model for the activation and deactivation dynamics of a spiking neuron that goes back to FitzHugh and Nagumo. Formally, the model is described by the following coupled nonlinear PDEs

$$\begin{aligned}\epsilon v_t(x,t) &= \epsilon^2 v_{xx}(x,t) + f(v(x,t)) - w(x,t) + g,\\ w_t(x,t) &= hv(x,t) - \gamma w(x,t) + g,\end{aligned}$$

with $f(v) = v(v-0.1)(1-v)$ and constant terms ϵ, γ, h, g. Here, the spatio-temporal variables v and w denote a voltage and a recovery term associated with the neuron that is subject to an external excitation source. As soon as this excitation exceeds a certain threshold value, the neuron begins spiking, i.e., the voltage increases. After a while, the variables v and w return to their rest values. Mathematically, this phenomenon can be nicely described by means of a characteristic limit cycle behavior as shown in Figure 1.3, where we see a three-dimensional phase space diagram given by the system state consisting of v and w. Although discretizing these coupled PDEs does neither lead to a linear nor a bilinear control system, in Chapter 5, we discuss possible model reduction techniques appropriate for this case as well.

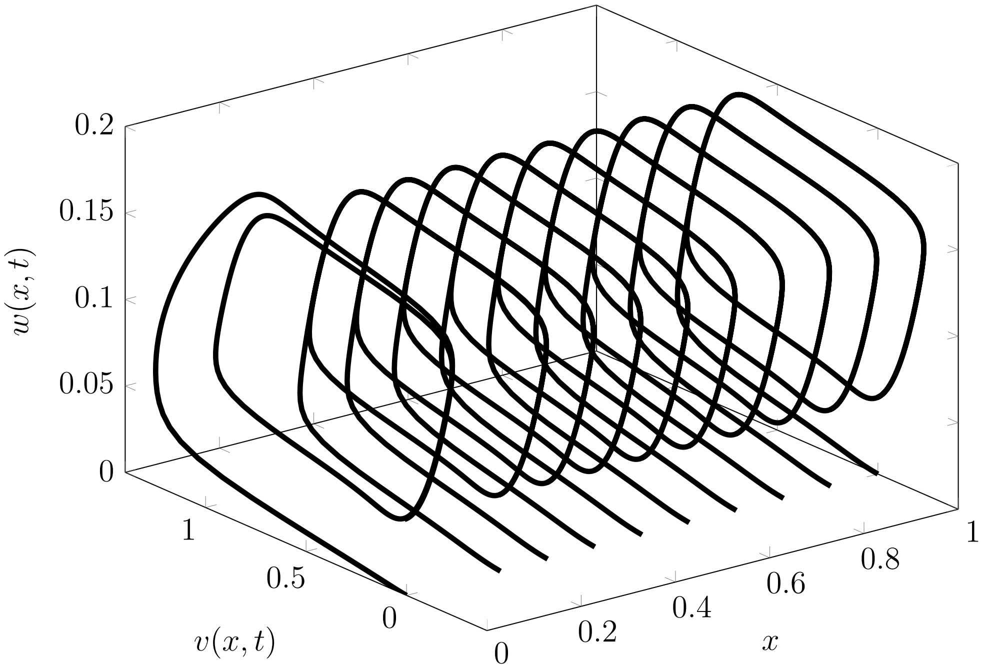


Figure 1.3: Typical limit cycle behavior for the FitzHugh-Nagumo system.

1.3 Mathematical background

The essential part of this thesis certainly is based on fundamental concepts from numerical linear algebra. However, since this field is a very wide field within numerical mathematics, we give a brief summary of the different mathematical disciplines we use throughout this work and refer to the most important references that provide the background for all following results.

Here we deal with MOR of dynamical systems and, hence, most of the ideas heavily depend on results from classical *control theory.* Concepts of stability, reachability or observability, respectively, are of particular importance for linear model reduction techniques. Lots of fundamental developments have been established by, e.g., Hinrichsen and Pritchard and can be found in one of the standard textbooks on control theory, see [78]. For nonlinear model reduction, in Chapter 4 and Chapter 5 we make use of basic results from nonlinear control theory. A crucial role plays, e.g., the Volterra series representation of a nonlinear system as well as the concept of variational analysis, allowing for a generalized input-output map which is the key tool for all our methods. A good introduction and very detailed explanation of the corresponding techniques can be found in, e.g., [82, 100, 115].

Furthermore, we build upon numerous results from *linear and multilinear algebra* such as linear matrix equations and their multidimensional counterpart, tensorized linear systems. While most of the results on algebraic operations such as the Kronecker product and the vectorization of matrices are well-known and can be studied in, e.g., [80, 94], for multidimensional purposes a rather new mathematical area has emerged over the last few years. To be more precise, although tensor theoretic ideas have been used within the chemical and physical community for over 50 years, the study of its mathematical foundation has recently experienced a lot of attention. Besides the detailed and rigorous discussions in [74, 95], we refer to a very nice overview given in [87]. Additionally, for a better understanding of the second part of Chapter 4, we also point to [68] for a detailed insight into tensor theory.

One probably cannot get around the ideas of *projection* and *iterative methods* when it comes to the construction of numerical algorithms that efficiently compute a reduced-order model. Common techniques such as the Arnoldi method, Lanczos procedure or the (Petrov-)Galerkin framework are fundamental for the remainder of this thesis and for an entire understanding of the statements, the reader should be familiar with these concepts. The author has mainly benefited from studying [117].

Finally, the central model reduction technique studied in this thesis has its origin in the classical concept of rational *interpolation*. We have already discussed the significance of transfer functions when it comes to linear control theory. Throughout all chapters, we are constantly faced with finding (optimal) interpolation points needed for the construction of appropriate rational interpolants. Although the rational interpolation framework has a long and interesting history, for our purposes, the works by [3, 71, 73, 99] have played an extraordinary role for the development of the results presented here.

1.4 Outline of the thesis

We begin with a review on tensor theory as well as linear control theory in Chapter 2. Regarding the tensor theoretic ideas, we present the most important properties of the Kronecker product and the closely related operation of vectorization of a matrix. Moreover, we define the tensor rank and matricizations of vectors which are important in Chapter 4 and Chapter 5. For linear control systems, we give an introduction into the previously mentioned concepts of reachability, observability and stability. Although we restrict ourselves to the linear continuous time-invariant case, we point out differences that occur in the discrete-time setting. We conclude the chapter with an explanation of projection-based model reduction for linear systems, including the special cases of interpolation and balancing, respectively.

In Chapter 3, we focus on model reduction of linear control systems. Due to its importance for this thesis, we discuss the problem of $\mathcal{H}_2$-optimal model reduction in detail and state different optimality conditions. Subsequently, we derive new results concerning low rank approximations of large-scale linear matrix equations. In particular, we pick up an idea from Riemannian optimization, introduced in [125], and show how to achieve the same results by means of the concept of rational interpolation. We differentiate between the symmetric and the unsymmetric case. While for the first case, the goal is to minimize the canonical energy norm induced by the Lyapunov operator, for unsymmetric matrices, we extend the ideas to a more general setting which reappears in Chapter 4.

Chapter 4 reflects the main contributions of this work. Here, we deal with the problem of MOR for bilinear control systems. After an introduction into the basic theory for this class of systems, we extend the ideas from $\mathcal{H}_2$-optimal model reduction for linear systems to the bilinear case. We derive new abstract interpolatory optimality conditions that we show to be equivalent to existing ones based on generalized linear matrix equations. We further propose two iterative algorithms that theoretically as well as numerically are proven to outperform other state-of-the-art techniques with respect to the bilinear $\mathcal{H}_2$-norm. In the second part of Chapter 4, we discuss low rank approximation methods for generalized matrix equations arising in the method of balanced truncation for bilinear control systems. Besides a theoretical explanation for the often observed fast singular value decay of the solution matrix, we investigate the generalization of several successful low rank approximation methods known for the case of linear control systems.

In Chapter 5, we discuss a recently introduced method, see [72], for more general nonlinear control systems. The fundamentals for this approach again have their origin in the idea of rational interpolation by projection. Here, the new contribution on the one hand is an efficient construction of a reduced-order model and, on the other hand, is the development of a two-sided projection method theoretically improving the existing technique. We further extensively test the method by several examples arising from the semi-discretization of nonlinear PDEs and compare the results with those obtained by the proper orthogonal decomposition (POD) method, a commonly used method in nonlinear MOR.

We conclude with a summary of the results and an overview of open questions for further research in Chapter 6.

CHAPTER 2

MATHEMATICAL FOUNDATIONS

Contents

In this chapter, we collect basic concepts and ideas that we use and assume to be known throughout the rest of this thesis. Most of the tools presented in the first section are well-known in the the context of matrix and tensor theory and can be found in, e.g., [65, 68, 80, 87]. The mathematical foundations of classical linear control theory are discussed in nearly every textbook like, e.g., [78]. For a detailed introduction into the topic of model order reduction, we refer to [3] and the references therein.

2.1 Tensors and matricizations

For what follows, one of the most important operations is the Kronecker product of matrices together with the closely related $\operatorname{vec}(\cdot)$-operator defined as follows.

Definition 2.1.1. *([65, Section 12.1]) Let* $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n\times m}$ *and* $\mathbf{Y} \in \mathbb{R}^{p\times q}$.

Then

$$\operatorname{vec}(\mathbf{X}) := \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{bmatrix} \in \mathbb{R}^{n \cdot m \times 1}, \quad \mathbf{X} \otimes \mathbf{Y} = \begin{bmatrix} x_{11}\mathbf{Y} & \dots & x_{1m}\mathbf{Y} \\ \vdots & & \vdots \\ x_{n1}\mathbf{Y} & \dots & x_{nm}\mathbf{Y} \end{bmatrix} \in \mathbb{R}^{n \cdot p \times m \cdot q}.$$

From the above definition, one can immediately show the following useful properties, see, e.g., [65, Section 12.1].

Proposition 2.1.1. *Let* $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\mathbf{B} \in \mathbb{R}^{p \times q}$, $\mathbf{C} \in \mathbb{R}^{m \times r}$, $\mathbf{D} \in \mathbb{R}^{q \times s}$ *and* $\mathbf{E} \in \mathbb{R}^{r \times \ell}$. *Then it holds*

a) $\operatorname{vec}(\mathbf{ACE}) = (\mathbf{E}^T \otimes \mathbf{A}) \operatorname{vec}(\mathbf{C})$,

b) $\operatorname{tr}(\mathbf{AC}) = \operatorname{vec}\left(\mathbf{A}^T\right)^T \operatorname{vec}(\mathbf{C})$,

c) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$.

We further need some properties of the derivative of the trace operator. If $\mathbf{f}(\mathbf{X})$ is a matrix function, let

$$\frac{\partial \operatorname{tr}(\mathbf{f}(\mathbf{X}))}{\partial \mathbf{X}} = \left[\frac{\partial \operatorname{tr}(\mathbf{f}(\mathbf{X}))}{\partial \mathbf{X}_{ij}}\right]_{ij}$$

denote its derivative with respect to $\mathbf{X}$. From [47], we cite a very useful result on its computation. Note that the first part of the following statement is due to Kleinman and Athans and can be found in [8, 85].

Theorem 2.1.1. *Let* $\mathbf{f}(\mathbf{X})$ *be some matrix function, then*

1) (by Kleinman and Athans) if
$\mathbf{f}(\mathbf{X} + \epsilon\Delta\mathbf{X}) - \mathbf{f}(\mathbf{X}) = \epsilon\mathbf{M}(\mathbf{X})\Delta\mathbf{X}, \epsilon \to 0,$ *we have*

$$\frac{\partial \operatorname{tr}(\mathbf{f}(\mathbf{X}))}{\partial \mathbf{X}} = \mathbf{M}^T(\mathbf{X});$$

2) (by Dulov and Andrianova) if
$\mathbf{f}(\mathbf{X} + \epsilon\Delta\mathbf{X}) - \mathbf{f}(\mathbf{X}) = \epsilon\mathbf{M}_1(\mathbf{X})\Delta\mathbf{X}\mathbf{M}_2(\mathbf{X}), \epsilon \to 0,$ *then*

$$\frac{\partial \operatorname{tr}(\mathbf{f}(\mathbf{X}))}{\partial \mathbf{X}} = \left[\mathbf{M}_2^T(\mathbf{X})\mathbf{M}_1(\mathbf{X})\right]^T.$$

For later purposes, we introduce a special permutation matrix which simplifies the computation of Kronecker products for certain block matrices.

Proposition 2.1.2. *([20]) Let* $\mathbf{A}$, $\mathbf{E} \in \mathbb{R}^{m \times m}$, $\mathbf{B} \in \mathbb{R}^{n \times n}$, $\mathbf{C} \in \mathbb{R}^{n \times m}$, $\mathbf{D} \in \mathbb{R}^{m \times n}$. *Assume that a permutation matrix* $\mathbf{M}$ *is given as follows*

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_m \otimes \begin{bmatrix} \mathbf{I}_n \\ \mathbf{0} \end{bmatrix} & \mathbf{I}_m \otimes \begin{bmatrix} \mathbf{0}^T \\ \mathbf{I}_m \end{bmatrix} \end{bmatrix}. \tag{2.1}$$

Then it holds

$$\mathbf{M}^T \left(\mathbf{A} \otimes \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix} \right) \mathbf{M} = \begin{bmatrix} \mathbf{A} \otimes \mathbf{B} & \mathbf{A} \otimes \mathbf{C} \\ \mathbf{A} \otimes \mathbf{D} & \mathbf{A} \otimes \mathbf{E} \end{bmatrix}.$$

Furthermore, we constantly make use of the *tensor rank* of a vectorized matrix.

Definition 2.1.2. *([68, 87]) Let* $\mathbf{x} = \operatorname{vec}(\mathbf{X}) \in \mathbb{R}^{n^2}$*. Then the minimal number* k *s.t.*

$$\mathbf{x} = \sum_{i=1}^{k} \mathbf{u}_i \otimes \mathbf{v}_i,$$

where $\mathbf{u}_i, \mathbf{v}_i \in \mathbb{R}^n$*, is called the tensor rank of the vector* $\mathbf{x}$*.*

Remark 2.1.1. *Due to the properties of the Kronecker product, it is easily seen that the tensor rank of a vectorized matrix* $\mathbf{X}$ *coincides with* $\operatorname{rank}(\mathbf{X})$*.*

In recent years, more and more attention has been paid to tensors. Formally, a tensor $\mathcal{H}$ is a vector indexed by a product index set

$$\mathcal{I} = \mathcal{I}_1 \times \cdots \times \mathcal{I}_d, \quad |\mathcal{I}_j| = n_j.$$

Besides the concept of the above mentioned tensor rank, several tensor decompositions have been discussed in detail in [68, 87, 89]. An important idea in understanding the nature of tensors is to transform them into matrices. For a given tensor $\mathcal{H}$, the corresponding tensor operation is called t-matricization $\mathcal{H}^{(t)}$ and is defined as

$$\mathcal{H}^{(t)} \in \mathbb{R}^{\mathcal{I}_t \times \mathcal{I}_{t'}}, \quad \mathcal{H}^{(t)}_{(i_\mu)_{\mu \in t},\, (i_\mu)_{\mu \in t'}} := \mathcal{H}_{(i_1,\ldots,i_d)}, \quad t' := \{1, \ldots, d\} \backslash t.$$

Since the concept is rather abstract, it might be helpful to consider a simple example. Due to its importance later on, we restrict ourselves to a 3-tensor. For example, here we can think of the Hessian matrix of a vector valued function.

Example 2.1.1. *For a given 3-tensor* $\mathcal{H}_{(i_1,i_2,i_3)}$ *with* $i_1, i_2, i_3 \in \{1, 2\}$*, we have the following matricizations:*

$$\mathcal{H}^{(1)} = \begin{bmatrix} \mathcal{H}_{(1,1,1)} & \mathcal{H}_{(1,2,1)} & \mathcal{H}_{(1,1,2)} & \mathcal{H}_{(1,2,2)} \\ \mathcal{H}_{(2,1,1)} & \mathcal{H}_{(2,2,1)} & \mathcal{H}_{(2,1,2)} & \mathcal{H}_{(2,2,2)} \end{bmatrix},$$

$$\mathcal{H}^{(2)} = \begin{bmatrix} \mathcal{H}_{(1,1,1)} & \mathcal{H}_{(2,1,1)} & \mathcal{H}_{(1,1,2)} & \mathcal{H}_{(2,1,2)} \\ \mathcal{H}_{(1,2,1)} & \mathcal{H}_{(2,2,1)} & \mathcal{H}_{(1,2,2)} & \mathcal{H}_{(2,2,2)} \end{bmatrix},$$

$$\mathcal{H}^{(3)} = \begin{bmatrix} \mathcal{H}_{(1,1,1)} & \mathcal{H}_{(2,1,1)} & \mathcal{H}_{(1,2,1)} & \mathcal{H}_{(2,2,1)} \\ \mathcal{H}_{(1,1,2)} & \mathcal{H}_{(2,1,2)} & \mathcal{H}_{(1,2,2)} & \mathcal{H}_{(2,2,2)} \end{bmatrix}.$$

Roughly speaking, for the t-matricization, the t-th index of the tensor $\mathcal{H}_{i_1,i_2,i_3}$ determines the row. The columns then are sorted according to a reverse lexicographic ordering.

2.2 Linear time-invariant systems

In the following, we state the essential ideas and concepts developed in the context of control theory of linear time-invariant (LTI) dynamical systems. All of the statements can be found in standard textbooks like, e.g., [78, 86, 113, 122].

2.2.1 The continuous-time case

Since the results discussed in this thesis are mainly dedicated to continuous-time systems, we give a more detailed background on continuous-time systems and only briefly mention the differences in the discrete-time setting.

Time-domain characterization

Let us begin with the so-called *state space representation* of linear continuous-time control systems of the form

$$\mathbf{\Sigma}: \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \end{cases} \tag{2.2}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}, \mathbf{C} \in \mathbb{R}^{p\times n}$ and $\mathbf{D} \in \mathbb{R}^{p\times m}$. Unless otherwise stated, we always assume that $\mathbf{A}$ is an invertible matrix so that the above system indeed is a system of ODEs. Here, $\mathbf{x}(t) \in \mathbb{R}^n$ is the *state* of the system, $\mathbf{x}_0$ the initial condition, $\mathbf{u}(t) \in \mathbb{R}^m$ denotes an *input* signal while $\mathbf{y}(t)$ is a measurable *output* of the system. The dimension n of the state vector $\mathbf{x}(t)$ is called the *order* of the system. Finally, $\mathbf{D}$ is the *throughput* of the system which often is assumed to be zero. In case of $m = p = 1$, i.e., $\mathbf{B} = \mathbf{b}$ and $\mathbf{C} = \mathbf{c}^T$, we speak of a single-input and single-output (SISO) system, otherwise we call it a multiple-input and multiple-output (MIMO) system. From now on, we abbreviate the state space representation (2.2) by $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$. Note that by choosing an arbitrary invertible matrix $\mathbf{T}$ and introducing a new state variable as $\mathbf{x} = \mathbf{T}^{-1}\tilde{\mathbf{x}}$, we can perform a so-called state space transformation that changes the *realization* of $\mathbf{\Sigma}$ according to

$$\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \to \tilde{\mathbf{\Sigma}} = (\mathbf{T}\mathbf{A}\mathbf{T}^{-1}, \mathbf{T}\mathbf{B}, \mathbf{C}\mathbf{T}^{-1}, \mathbf{D}).$$

Although most of the following concepts remain invariant under such transformations, they are important for the method of balanced truncation.

As can easily be verified, the solution of the *state equation* (2.2) is determined as

$$\Phi(\mathbf{u}; \mathbf{x}_0; t) = e^{\mathbf{A}t}\mathbf{x}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\, \mathrm{d}\tau. \tag{2.3}$$

Assuming a zero initial condition $\mathbf{x}_0 = \mathbf{0}$, by means of the above expression, we can easily specify the so-called *impulse response* of the system, i.e., the response to an input signal with $u_i(t) = \delta(t)$ and $\delta(t)$ denoting the Dirac delta function. The result is

$$\mathbf{H}(t) = \mathbf{C}e^{\mathbf{A}t}\,\mathbf{B} + \delta(t)\mathbf{D}. \tag{2.4}$$

One of the main assumptions for most model order reduction techniques is that the system under consideration is stable in the following sense, see [3, Section 5.8].

Definition 2.2.1. *Given a dynamical system $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$. The system (2.2) is called* stable*, if for every state trajectory $\mathbf{x}(t)$ it holds $||\mathbf{x}(t)|| < M$, $\forall t$ and some constant M. We call the system* asymptotically stable*, if additionally, $\lim\limits_{t\to\infty} ||\mathbf{x}(t)|| = 0$ for an arbitrary norm $||\cdot||$.*

As is easily shown, for $\sigma(\mathbf{A}) \subset \mathbb{C}_- \cup i\mathbb{R}$ the system $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is stable. Moreover, for $\sigma(\mathbf{A}) \subset \mathbb{C}_-$ the system $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is asymptotically stable. A further very important concept now is the reachability of an LTI system which is specified in the following definition, see [3, Section 4.2.1].

Definition 2.2.2. *Given a dynamical system $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$. A state $\bar{\mathbf{x}} \in \mathbb{R}^n$ is* reachable from the zero state *if there exists an input function $\bar{\mathbf{u}}(t)$, of finite energy, and a time $\bar{T} < \infty$, such that*

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

The system $\mathbf{\Sigma}$ is called reachable *if for the reachable subspace $\mathbb{X} \subset \mathbb{R}^n$ it holds that $\mathbb{X} = \mathbb{R}^n$. Furthermore,*

$$\mathcal{R}(\mathbf{A}, \mathbf{B}) = \left[\mathbf{B}, \mathbf{AB}, \dots, \mathbf{A}^{n-1}\mathbf{B}\right] \tag{2.5}$$

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

Having defined the concept of reachability already allows us to state one of the crucial ideas behind model order reduction by balanced truncation which we discuss in detail later on. Basically, the goal is to figure out which states are hard to reach, i.e., in terms of the definition, states that require a large amount of energy to be reached. Those states are less important and thus can be neglected without influencing the behavior of the system essentially. However, so far the question remains how to check the reachability of a system $\mathbf{\Sigma}$. This can be done by means of the reachability matrix $\mathcal{R}$, see, e.g., [3, Corollary 4.8].

Proposition 2.2.1. *A linear system $\mathbf{\Sigma}$ is reachable if and only if* $\operatorname{rank}(\mathcal{R}(\mathbf{A}, \mathbf{B})) = n$.

Recall that we are in general not interested in or even able to measure the entire state vector $\mathbf{x}$ and instead focus on the output $\mathbf{y}$ which is given by a linear combination of the states. Similarly to the concept of reachability, for control theoretic purposes it is helpful to identify the observable subspace of $\mathbf{\Sigma}$. To be more precise, we want to know which states of the underlying system can actually be *observed*. Since observability and reachability are dual concepts, one can check observability of the system by analyzing reachability of the dual pair $(\mathbf{A}^T, \mathbf{C}^T)$. Before performing model order reduction of a system, one should at least think about if this is a reasonable assumption and if the system actually can be reduced or not. For this reason, we need the definition of *minimality*, see, e.g., [3, Definition 4.36].

Definition 2.2.3. *Given a dynamical system* $\mathbf{\Sigma}$ *with* Markov parameters $\mathbf{h}_k = \mathbf{C}\mathbf{A}^{k-1}\mathbf{B} \in \mathbb{R}^{p\times m}$, $k = 1, 2, \ldots$ *The triple* $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *is then called a* realization *of the sequence* $\mathbf{h}_k$. $(\mathbf{C}, \mathbf{A}, \mathbf{B})$ *is a* minimal realization *if among all realizations of the sequence, its dimension is the smallest possible.*

In other words, it is impossible to reduce a minimal system without an approximation error. Moreover, there is a useful link between the concepts of minimality, reachability and observability, see [3, Lemma 4.42].

Proposition 2.2.2. *A linear system* $\mathbf{\Sigma}$ *is minimal if and only if* $(\mathbf{A}, \mathbf{B})$ *is reachable and* $(\mathbf{A}, \mathbf{C})$ *is observable.*

The above statement means that a system that is not completely reachable or observable is not minimal. Hence, we can replace it by a system with a smaller number of states. In context of model order reduction this means that we can construct a reduced-order system which *exactly* reproduces the dynamics of the original system. Therefore, from now on we always assume that the system under consideration is reachable and observable, hence minimal.

Frequency-domain characterization

In order to derive an explicit input-output relationship, instead of the state space representation (2.2) it is useful to analyze the system in *frequency domain*. This is done by applying the *Laplace transform*

$$\mathbb{L} : \mathbf{x}(t) \mapsto \mathbf{x}(s) = \int_0^\infty e^{-st}\mathbf{x}(t)\mathrm{d}t \quad (\Rightarrow \dot{\mathbf{x}}(t) \mapsto s\mathbf{x}(s) - \mathbf{x}(0)) \tag{2.6}$$

to the equations in (2.2). What we end up with is

$$s\mathbf{x}(s) - \mathbf{x}(0) = \mathbf{A}\mathbf{x}(s) + \mathbf{B}\mathbf{u}(s), \tag{2.7}$$
$$\mathbf{y}(s) = \mathbf{C}\mathbf{x}(s) + \mathbf{D}\mathbf{u}(s). \tag{2.8}$$

If we now solve the state equation for $\mathbf{x}(s)$ and insert the result into the output equation, we obtain an explicit expression for the output

$$\mathbf{y}(s) = \left(\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}\right)\mathbf{u}(s) + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}_0.$$

Assuming a zero initial state, the function

$$\mathbf{H}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \ \in \mathbb{R}(s)^{p\times m}, \tag{2.9}$$

representing the relation between inputs and outputs is called the *transfer function* of $\mathbf{\Sigma}$. Note that $\mathbf{H}(s)$ also results from the Laplace transform of the impulse response (2.4). For the special case of a SISO system, the transfer function is a rational function of degree n in the frequency variable s. Moreover, for $H(s) = \frac{d(s)}{n(s)}$, the zeros of $n(s)$ are called the *poles* of $\mathbf{\Sigma}$. Stable systems, i.e., systems that have poles only in the left half of the complex plane are particularly important. If the system matrix $\mathbf{A}$ now is diagonalizable, a useful representation is the *pole-residue expression* of the transfer function

$$H(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} + \mathbf{d} = \sum_{i=1}^{n} \frac{R_i}{s-\lambda_i} + \mathbf{d}, \tag{2.10}$$

where

$$R_i = \lim_{s\to\lambda_i} H(s)(s-\lambda_i),$$

and $\lambda_i,\ i = 1,\dots,n$ denote the eigenvalues of $\mathbf{A}$. Note that the assumption of a minimal system implies that there is no so-called *pole-zero cancellation.* If this was the case, we could have replaced the system by one of smaller dimension.

The Lyapunov equation

Most of the properties we have considered so far can be characterized by means of a certain type of matrix equation for which we need the *infinite reachability Gramian*

$$\mathbf{P} = \int_0^\infty e^{\mathbf{A}s}\mathbf{B}\mathbf{B}^T e^{\mathbf{A}^T s}\,\mathrm{d}s. \tag{2.11}$$

Similarly, the *infinite observability Gramian* of the system is given as

$$\mathbf{Q} = \int_0^\infty e^{\mathbf{A}^T s}\mathbf{C}^T\mathbf{C}e^{\mathbf{A}s}\,\mathrm{d}s. \tag{2.12}$$

By definition of $\mathbf{P}$ and $\mathbf{Q}$, it is clear that the Gramians are symmetric and positive semidefinite. Moreover, for minimal and asymptotically stable systems, the Gramians are positive-definite and additionally satisfy the so-called *Lyapunov equations.* The following result is from [3, Section 4.3].

Proposition 2.2.3. *Let* $\mathbf{\Sigma} = (\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})$ *denote a stable minimal dynamical system. Then* $\mathbf{P} = \mathbf{P}^T \succ 0$ *and* $\mathbf{Q} = \mathbf{Q}^T \succ 0$ *are the unique solutions of the Lyapunov equations*

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}, \tag{2.13a}$$

$$\mathbf{A}^T\mathbf{Q} + \mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0}. \tag{2.13b}$$

The special name of the matrix equation results from the fact that for a positive definite right hand side, the solution $\mathbf{P}$ can be used to prove that the system is stable in the sense of the classical Lyapunov stability. For this, one can show that $V(\mathbf{z}) = \mathbf{z}^T\mathbf{P}\mathbf{z}$ indeed is a Lyapunov function for the system $\boldsymbol{\Sigma}$.

Besides their importance in the context of stability of LTI systems, the above equations play an important role in the concept of balancing and balancing related methods as we see in the following section. In particular, they allow to measure a certain energy associated with each system state $\mathbf{x}(t)$. However, for large-scale systems, computing and storing the solution matrices $\mathbf{P}$ and $\mathbf{Q}$, respectively, is a major challenge. Note that in general $\mathbf{P}$ and $\mathbf{Q}$ are dense matrices even if the system $\boldsymbol{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ is sparse. As a trivial example one might think of $\mathbf{A} = -\mathbf{I}$ which leads to $\mathbf{P} = \frac{1}{2}\mathbf{B}\mathbf{B}^T$ and $\mathbf{Q} = \frac{1}{2}\mathbf{C}^T\mathbf{C}$. On the other hand, for a variety of real-life applications, it is well-known that if the number of inputs and outputs is small compared to the order of the system, i.e., $m, p \ll n$, the Gramians $\mathbf{P}$ and $\mathbf{Q}$ exhibit a very strong singular value decay which allows for accurate low rank approximations $\mathbf{P} \approx \mathbf{L}\mathbf{L}^T,\ \mathbf{L} \in \mathbb{R}^{n\times k}$ and $\mathbf{Q} \approx \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T,\ \tilde{\mathbf{L}} \in \mathbb{R}^{n\times \tilde{k}}$, where $k, \tilde{k} \ll n$. For example, let us assume that the system matrix $\mathbf{A}$ results from a finite-difference discretization of the two-dimensional heat equation on the unit square and that the input matrix is given as $\mathbf{B} = \left[1, \cdots, 1\right]^T$. In Figure 2.1, we see the normalized singular values of the reachability Gramian $\mathbf{P}$ for a discretization with $n = 100$ grid points. We show the data corresponding to computations done in single precision as well as computations done in double precision. The comparison underscores a very important drawback of numerical computations generated on a computer architecture. For singular values close to machine precision, we can no longer trust the results. In Figure 2.1, we see that the red circles coincide with the blue circles as long as the numerical values are large enough. A similar phenomenon would occur if we had further used quadruple precision. The point where the blue circles seem to stagnate indicates the limit of computations done in finite arithmetic. Nevertheless, we can record that for this example, a low rank approximation of order 10 already suffices to reproduce $\mathbf{P}$ up to the common machine precision used in MATLAB.

Besides the practical occurrence itself, the theoretical explanation for this phenomenon, of course, is of interest as well. Indeed, there exist several different approaches that show why one can expect an exponential singular value decay in certain situations. Probably the earliest results are discussed in [109], where the author focuses on the case of a symmetric system matrix $\mathbf{A} = \mathbf{A}^T$. Making use of an error expression based on the so-called ADI iteration for Lyapunov equations, one can show that

$$\frac{\lambda_{mk+1}(\mathbf{P})}{\lambda_1(\mathbf{P})} \leq \left(\prod_{j=0}^{k-1} \frac{\kappa_{\mathbf{A}}^{(2j+1)/(2k)} - 1}{\kappa_{\mathbf{A}}^{(2j+1)/(2k)} + 1} \right)^2,$$

where $\kappa_{\mathbf{A}} = ||\mathbf{A}|| \cdot ||\mathbf{A}^{-1}||$ denotes the condition number of $\mathbf{A}$. The unsymmetric case is studied in [6]. Here, the authors derive an approximation result based on properties of Cauchy matrices. For the SISO case, the final error bound for a rank-k approximation

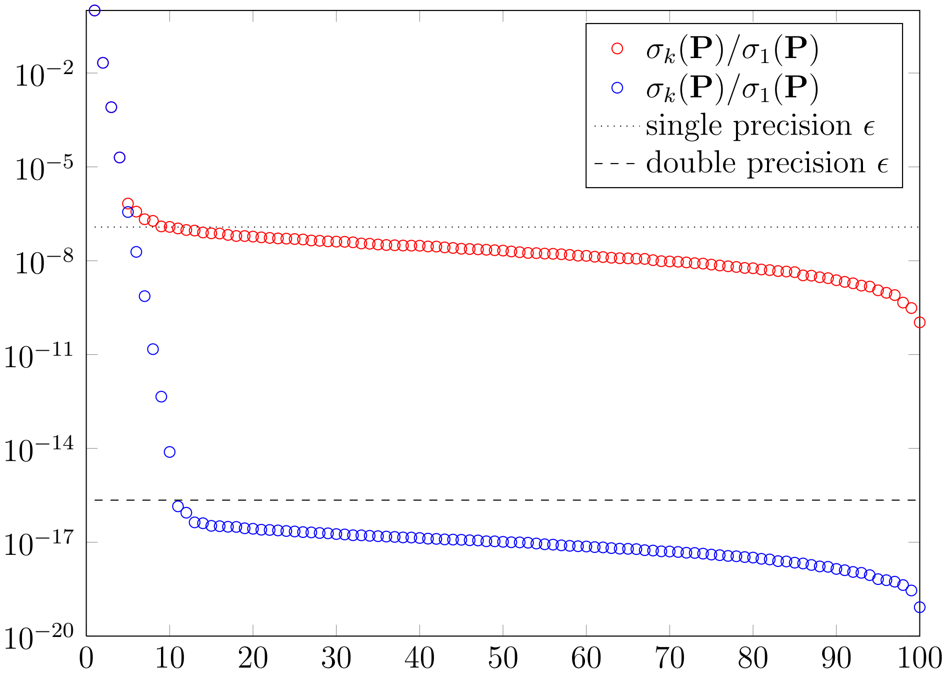


Figure 2.1: Singular value decay of the controllability Gramian for the 2D heat equation. Computations in single precision vs computations in double precision.

$\mathbf{P}_k$ then is as follows

$$||\mathbf{P} - \mathbf{P}_k||_2 \leq (n-k)^2 \delta_{k+1} (\kappa_2(\mathbf{X})||\mathbf{b}||_2)^2,$$

where $\mathbf{X}$ is the matrix containing the eigenvectors of $\mathbf{A}$ and

$$\delta_{k+1} = \frac{-1}{2\,\mathrm{Re}\,(\lambda_{k+1})} \prod_{j=1}^{k} \left| \frac{\lambda_{k+1} - \lambda_j}{\lambda_{k+1}^* + \lambda_j} \right|^2 .$$

Finally, a similar bound for the unsymmetric case is given in [123].

However, for our purposes a rather different approach is more suitable. For convenience, let us for a moment consider the SISO case, i.e., $\mathbf{B} = \mathbf{b}$. Using the tools from Section 2.1, instead of equation (2.13a), equivalently we can consider its vectorization. Together with the properties of the Kronecker product, this leads to a linear system of n^2 equations of the form

$$\underbrace{(\mathbf{I} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{I})}_{\mathcal{L}} \underbrace{\mathrm{vec}\,(\mathbf{P})}_{\mathbf{p}} = \underbrace{-\mathbf{b} \otimes \mathbf{b}}_{\mathcal{B}}. \tag{2.14}$$

As has been shown in [66, 89], the main advantage is that most of the low rank approaches dealing with the above structure can be even generalized to the d-dimensional case with additional mass matrices appearing within the tensor structure

$$\underbrace{\left(\sum_{i=1}^{d} \mathbf{E}_1 \otimes \cdots \otimes \mathbf{E}_{i-1} \otimes \mathbf{A}_i \otimes \mathbf{E}_{i+1} \otimes \cdots \otimes \mathbf{E}_d\right)}_{\mathcal{L}_d} \mathcal{X} = \bigotimes_{i=1}^{d} \mathbf{b}_i. \tag{2.15}$$

The important observation is that the special structure of equation (2.15) allows to diagonalize the left-hand side by a matrix of tensor rank 1, meaning that the approximation procedure basically amounts to an approximation problem for the function

$$f(x_1, \ldots, x_d) = \frac{1}{x_1 + \cdots + x_d}.$$

In [66], for linear systems of the form (2.15), it is shown that there exists a vector $\tilde{\mathcal{X}}$ of tensor rank k that fulfills a profitable error bound.

Theorem 2.2.1. *([66]) Let $\mathcal{L}_d$ denote a matrix of tensor product structure as in (2.15) with tensor right-hand side $\mathcal{B}$. Assume that the sum of the spectra of the $\mathbf{E}_i^{-1}\mathbf{A}_i$ is contained in the strip $\Omega := -[\lambda_{\min}, \lambda_{\max}] \oplus i[-\mu, \mu] \subseteq \mathbb{C}_-$ and let Γ denote the boundary of $-[1, 2\lambda_{\max}/\lambda_{\min} + 1] \oplus i[-2\mu/\lambda_{\min} - 1, 2\mu/\lambda_{\min} + 1]$. Let $k \in \mathbb{N}$ and for $j = -k, \ldots, k$, define the following quadrature weights and points*

$$h_{st} := \frac{\pi}{\sqrt{k}}, \tag{2.16}$$

$$t_j := \log\left(\exp(jh_{st}) + \sqrt{1 + \exp(2jh_{st})}\right), \tag{2.17}$$

$$w_j := \frac{h_{st}}{\sqrt{1 + \exp(-2jh_{st})}}. \tag{2.18}$$

Then there exists C_{st} s.t. the solution $\mathcal{X}$ to $\mathcal{L}_d\mathcal{X} = \mathcal{B}$ can be approximated by

$$\tilde{\mathcal{X}} := -\sum_{j=-k}^{k} \frac{2w_j}{\lambda_{\min}} \bigotimes_{i=1}^{d} \exp\left(\frac{2t_j}{\lambda_{\min}} \mathbf{E}_i^{-1}\mathbf{A}_i\right) \mathbf{E}_i^{-1}\mathbf{b}_i, \tag{2.19}$$

with approximation error

$$||\mathcal{X} - \tilde{\mathcal{X}}||_2 \leq \frac{C_{st}}{\pi\lambda_{\min}} \exp\left(\frac{2\mu\lambda_{\min}^{-1} + 1}{\pi} - \pi\sqrt{k}\right) \oint_{\Gamma} \left|\left|\lambda\mathbf{I} - 2\frac{\mathcal{L}_d}{\lambda_{\min}}\right|\right|_2 \mathrm{d}_{\Gamma}\lambda \quad \left|\left|\bigotimes_{i=1}^{d} \mathbf{E}_i^{-1}\mathbf{b}_i\right|\right|_2.$$

Obviously, in the special case $d = 2$, $\mathbf{E}_1 = \mathbf{E}_2 = \mathbf{E}$, $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{A}$, the above statement immediately reveals that the solution to the (generalized) Lyapunov equation

$$\mathbf{APE}^T + \mathbf{EPA}^T + \mathbf{BB}^T = \mathbf{0},$$

can be approximated by a low rank matrix $\tilde{\mathbf{P}} = \mathbf{LL}^T$, $\mathbf{L} \in \mathbb{R}^{n \times k}$, with almost exponentially decreasing approximation error $||\mathbf{P} - \tilde{\mathbf{P}}||_2$. The basic ideas for proving the assertion are, on the one hand, the exponential character of the solution matrix $\mathcal{L}_d^{-1}$ corresponding to a system of linear equations $\mathcal{L}_d\mathcal{X} = \mathcal{B}$ as well as the Dunford-Cauchy representation of the underlying matrix exponential. On the other hand, one can exploit the special tensor structure which allows to decompose the approximant $\tilde{\mathcal{X}}$ and thus leads to the above tensor structure. However, for a more detailed analysis, we refer to [66].

Remark 2.2.1. *The quadrature weights and points from Theorem 2.2.1 go back to the quadrature formula of Stenger, see, e.g., [124]. Note that the constant C_{st} is independent of the individual problem and has been experimentally determined as $C_{st} \approx 2.75$, see [89].*

Remark 2.2.2. *As has been shown in [89], at least for the symmetric and supersymmetric cases, respectively, one can construct even better approximations $\tilde{\mathcal{X}}$ that, although depending on the condition number of $\mathcal{L}_d$, exhibit a true exponentially decreasing approximation error, i.e., the bound depends on $\exp(-k)$ rather than on $\exp(-\sqrt{k})$ as in Theorem 2.2.1. Moreover, for the unsymmetric case the bound is only of theoretical interest since the spectrum and its bound usually is not known and one thus might largely overestimate the true approximation error, see also the numerical study in [67]. Nevertheless, Theorem 2.2.1 theoretically provides an explanation for the often observed low numerical rank of the reachability Gramian $\mathbf{P}$.*

In summary, we conclude that it is often possible to approximate the solutions to the Lyapunov equations (2.13) by low rank factors appropriately. In Chapter 3, we briefly review some important methods used in the context of model order reduction of LTI systems that solely operate on these low rank factors making an efficient computation possible for system dimensions up to $n \sim 10^6$.

System norms

Let us come back to linear control systems of the form (2.2). For now, we assume that the feedthrough term $\mathbf{D}$ vanishes. Especially for model reduction purposes, the latter assumption is no restriction as we show within the next section. On the other hand, if $d = 0$, the transfer function $H(s)$ of a stable SISO dynamical system belongs to the Hardy space $\mathcal{H}_2$. Recall that the latter space denotes the set of square integrable functions that are analytic in the open right half of the complex plane, i.e., functions $f(x+iy)$ with

$$\sup_{x>0} \int_{-\infty}^{\infty} |f(x+iy)|^2 \,\mathrm{d}y < \infty.$$

Throughout the thesis, $\mathcal{H}_2$ is of particular interest. According to, e.g., [73], with two stable dynamical systems described by their transfer functions G and H, we can associate the $\mathcal{H}_2$-inner product via

$$\langle G, H\rangle_{\mathcal{H}_2} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{G(i\omega)}H(i\omega)\,\mathrm{d}\omega \tag{2.20}$$

and the corresponding norm via

$$||H||_{\mathcal{H}_2} = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |H(i\omega)|^2 \, \mathrm{d}\omega \right)^{\frac{1}{2}}. \tag{2.21}$$

Since the above definition is rather of theoretical interest, in [73] the authors give two alternatives to compute the $\mathcal{H}_2$-norm which we summarize below.

Lemma 2.2.1. *Suppose* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m\times m}$ *are stable. Given* $\mathbf{b},\ \mathbf{c} \in \mathbb{R}^n$ *and* $\tilde{\mathbf{b}},\ \tilde{\mathbf{c}} \in \mathbb{R}^m$*, define associated transfer functions,*

$$G(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \quad \textit{and} \quad H(s) = \tilde{\mathbf{c}}^T(s\mathbf{I} - \mathbf{B})^{-1}\tilde{\mathbf{b}}.$$

Then the inner product $\langle G, H\rangle_{\mathcal{H}_2}$ *is associated with solutions to Sylvester equations as:*

$$\textit{If } \ \mathbf{P} \ \textit{ solves } \ \mathbf{AP} + \mathbf{PB}^T + \mathbf{b}\tilde{\mathbf{b}}^T = \mathbf{0} \ \textit{ then } \ \langle G, H\rangle_{\mathcal{H}_2} = \mathbf{c}^T\mathbf{P}\tilde{\mathbf{c}}. \tag{2.22}$$

$$\textit{If } \ \mathbf{Q} \ \textit{ solves } \ \mathbf{QA} + \mathbf{B}^T\mathbf{Q} + \tilde{\mathbf{c}}\mathbf{c}^T = \mathbf{0} \ \textit{ then } \ \langle G, H\rangle_{\mathcal{H}_2} = \tilde{\mathbf{b}}^T\mathbf{Q}\mathbf{b}. \tag{2.23}$$

$$\textit{If } \ \mathbf{R} \ \textit{ solves } \ \mathbf{AR} + \mathbf{RB} + \mathbf{b}\tilde{\mathbf{c}}^T = \mathbf{0} \ \textit{ then } \ \langle G, H\rangle_{\mathcal{H}_2} = \mathbf{c}^T\mathbf{R}\tilde{\mathbf{b}}. \tag{2.24}$$

Note that if $\mathbf{A} = \mathbf{B},\ \mathbf{b} = \tilde{\mathbf{b}}$*, and* $\mathbf{c} = \tilde{\mathbf{c}}$ *then* $\mathbf{P}$ *is the reachability Gramian of* $G(s)$*,* $\mathbf{Q}$ *is the observability Gramian of* $G(s)$*, and* $\mathbf{R}$ *is the so-called cross Gramian of* $G(s)$*; and*

$$||G||^2_{\mathcal{H}_2} = \mathbf{c}^T\mathbf{P}\mathbf{c} = \mathbf{b}^T\mathbf{Q}\mathbf{b} = \mathbf{c}^T\mathbf{R}\mathbf{b}. \tag{2.25}$$

Another formula can be provided by the pole-residue expression of the transfer function (2.10). Again, we refer to [73] for a proof of the following useful statement.

Lemma 2.2.2. *Suppose that* $G(s)$ *has simple poles at* $\lambda_1, \ldots, \lambda_n$ *and* $H(s)$ *has simple poles at* $\mu_1, \ldots, \mu_m$*, where both sets are contained in the open left half of the complex plane. Then*

$$\langle G, H\rangle_{\mathcal{H}_2} = \sum_{k=1}^{m} G(-\mu_k)\, \mathrm{res}[H(s), \mu_k]. \tag{2.26}$$

So far, we only considered the SISO case. However, the extension to MIMO systems is rather straightforward. For example, the $\mathcal{H}_2$-norm is given as

$$||\mathbf{\Sigma}||_{\mathcal{H}_2} = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{tr}\left(\overline{\mathbf{H}(i\omega)}\mathbf{H}^T(i\omega) \right) \, \mathrm{d}\omega \right)^{\frac{1}{2}}. \tag{2.27}$$

Moreover, all the formulas from Lemma 2.2.1 still hold true when the right hand sides are changed appropriately. Also, the pole-residue expression can be generalized to the MIMO setting as well. Since the corresponding computation formula is rather technical and does not yield additional insight, we refrain from a more detailed discussion here.

Before we proceed with the discrete-time case, we also introduce the $\mathcal{H}_\infty$-norm for a linear dynamical system $\mathbf{\Sigma}$.

Definition 2.2.4. *Let* $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ *denote a stable dynamical linear control sytem. Then, the* $\mathcal{H}_\infty$*-norm of* $\mathbf{\Sigma}$ *is defined as*

$$||\mathbf{\Sigma}||_{\mathcal{H}_\infty} := \sup_{\omega \in \mathbb{R}} \sigma_{max}(\mathbf{H}(i\omega)), \tag{2.28}$$

where σ_{max} *denotes the maximal singular value of the matrix valued transfer function* $\mathbf{H}(s) \in \mathbb{R}(s)^{p \times m}$.

In conclusion, there are two system norms that obviously open up the possibility to measure the approximation of a reduced-order model in different ways. While the $\mathcal{H}_\infty$-norm is of greater importance for balancing-type model reduction methods, in this thesis, we mainly focus on interpolation-based methods that try to minimize the $\mathcal{H}_2$-norm.

2.2.2 The discrete-time case

Though the thesis is mainly dedicated to the continuous-time case, for the sake of completeness we give some details and differences that arise in the discrete-time setting. Hence, let us have a look at systems of the form

$$\mathbf{\Sigma}_d : \begin{cases} \mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k), \\ \quad\;\; \mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k), \quad \mathbf{x}(0) = \mathbf{x}_0, \end{cases} \tag{2.29}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}, \mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{D} \in \mathbb{R}^{p \times m}$. Analogously to the previous subsection, an explicit solution of the state equation can be derived as

$$\Phi(\mathbf{u}; \mathbf{x}_0; k) = \mathbf{A}^k \mathbf{x}_0 + \sum_{j=0}^{k-1} \mathbf{A}^{k-1-j} \mathbf{B}\mathbf{u}(j). \tag{2.30}$$

The concepts of reachability and observability now remain the same as in the continuous case. Moreover, the transfer function exhibits the same structure, i.e.,

$$\mathbf{H}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}. \tag{2.31}$$

However, the frequency variable usually is denoted by z, obtained from a discrete-time Laplace or $\mathcal{Z}$-transform of the system. For the infinite reachability and observability Gramians

$$\mathbf{P} = \sum_{k=0}^{\infty} \mathbf{A}^k \mathbf{B}\mathbf{B}^T (\mathbf{A}^T)^k, \tag{2.32}$$

$$\mathbf{Q} = \sum_{k=0}^{\infty} (\mathbf{A}^T)^k \mathbf{C}^T \mathbf{C} \mathbf{A}^k, \tag{2.33}$$

one can show the following counterpart to Proposition 2.2.3, see [3, Section 4.3].

Proposition 2.2.4. *Let* $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ *denote a stable minimal dynamical system. Then* $\mathbf{P} = \mathbf{P}^T \succ 0$ *and* $\mathbf{Q} = \mathbf{Q}^T \succ 0$ *are the solutions of the Stein equations*

$$\mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{P}, \tag{2.34a}$$
$$\mathbf{A}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{Q}. \tag{2.34b}$$

Finally, we want to state the definitions of the discrete-time system norms that require the evaluation of the transfer function on the unit circle instead of the imaginary axis. As a result, one can define the h_2-norm as

$$||\mathbf{\Sigma}_d||_{h_2} = \left(\frac{1}{2}\int_0^{2\pi} \operatorname{tr}\left(\overline{\mathbf{H}(e^{i\theta})}\mathbf{H}^T(e^{i\theta})\right)\,\mathrm{d}\theta\right)^{\frac{1}{2}} \tag{2.35}$$

and the h_∞-norm as

$$||\mathbf{\Sigma}_d||_{h_\infty} := \sup_{\theta\in[0,2\pi]} \sigma_{max}(\mathbf{H}(e^{i\theta})). \tag{2.36}$$

Unsurprisingly, for the h_2-norm, computation formulas similar to the continuous case can be shown based on the solution of the Stein equations (see [3]) and the pole-residue expression (see [36]). However, here we do not give the exact statements but instead refer the interested reader to the given references.

2.3 Model reduction by projection

As we have seen in the preceding section, the analysis of important control theoretic concepts like stability, minimality and frequency domain behavior require the solution of linear matrix equations or the evaluation of complex integrals of the transfer function matrix $\mathbf{H}(s)$ of the system. Moreover, we have already discussed the presence of high-dimensional systems, i.e., state dimensions n reaching up to 10^6, in the context of typical real-life applications. In what follows, we give a compact review of the most important MOR techniques that can be found in, e.g., [3, 5, 28, 61, 106, 118]. Mathematically speaking, we want to replace the system (2.2) by the following one

$$\hat{\mathbf{\Sigma}} : \begin{cases} \dot{\hat{\mathbf{x}}}(t) = \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \hat{\mathbf{B}}\mathbf{u}(t), \\ \hat{\mathbf{y}}(t) = \hat{\mathbf{C}}\hat{\mathbf{x}}(t) + \hat{\mathbf{D}}\mathbf{u}(t), \quad \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0, \end{cases} \tag{2.37}$$

where $\hat{\mathbf{A}} \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $\hat{\mathbf{B}} \in \mathbb{R}^{\hat{n}\times m}$, $\hat{\mathbf{C}} \in \mathbb{R}^{p\times\hat{n}}$ and $\hat{\mathbf{D}} \in \mathbb{R}^{p\times m}$. Obviously, for $\hat{\mathbf{\Sigma}}$ we require $\hat{n} \ll n$ and the error $||\mathbf{y} - \hat{\mathbf{y}}||$ to be small. Depending on the specific norm we choose for the minimization problem, there are different techniques that have been proven to be very successful. On the one hand, there are interpolation-based model reduction techniques that try to minimize the error in the $\mathcal{H}_2$-norm and, on the other hand, methods like balanced truncation focus on a small $\mathcal{H}_\infty$-error of the reduced-order system.

In order to measure the quality of a reduced-order approximation, we define the so-called *error system* $\mathbf{\Sigma}_{err} = (\mathbf{A}_{err}, \mathbf{B}_{err}, \mathbf{C}_{err}, \mathbf{D}_{err})$, where the system matrices are given as

$$\mathbf{A}_{err} = \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \hat{\mathbf{A}} \end{bmatrix}, \quad \mathbf{B}_{err} = \begin{bmatrix} \mathbf{B} \\ \hat{\mathbf{B}} \end{bmatrix}, \quad \mathbf{C}_{err} \begin{bmatrix} \mathbf{C} & -\hat{\mathbf{C}} \end{bmatrix}, \quad \mathbf{D}_{err} = \mathbf{D} - \hat{\mathbf{D}}. \tag{2.38}$$

The special block structure of the error system is motivated by the fact that for the transfer function it holds $\mathbf{H}_{err} = \mathbf{H} - \hat{\mathbf{H}}$. We now also see why it is reasonable to assume that the feedthrough term $\mathbf{D}$ of the original system is the zero matrix. If this is not the case, the feedthrough term of the reduced-order system obviously can be set to $\hat{\mathbf{D}} = \mathbf{D}$ so that the feedthrough term of the error system vanishes. Some approaches such as, e.g., Hankel norm approximation and a recently proposed method from [57], use the $\hat{\mathbf{D}}$-term of the reduced-order model to improve the approximation quality even in the case $\mathbf{D} = \mathbf{0}$. However, this is beyond the scope of this thesis and we therefore always assume that $\hat{\mathbf{D}} = \mathbf{D} = \mathbf{0}$.

The question that immediately arises is how to construct $\hat{\mathbf{\Sigma}}$, given an original system $\mathbf{\Sigma}$. As it turns out, a reduced-order system can be obtained by a projection-type framework. For this, let us briefly state the most important properties of projection matrices, see, e.g., [117, Section 1.12] and [17].

Definition 2.3.1. *a) A matrix* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *is a* projector (onto a subspace $\mathcal{V} \subset \mathbb{R}^n$) *if* range $(\mathbf{P}) = \mathcal{V}$ *and* $\mathbf{P}^2 = \mathbf{P}$.

b) Let $\mathbf{Z} \in \mathbb{R}^{n \times n}$ *with spectrum* $\Lambda(\mathbf{Z}) = \Lambda_1 \cup \Lambda_2$, $\Lambda_1 \cap \Lambda_2 = \emptyset$, *and let* $\mathcal{V}_1$ *be the (right)* $\mathbf{Z}$*-invariant subspace corresponding to* Λ_1*. Then a projector onto* $\mathcal{V}_1$ *is called a* spectral projector.

c) If $\mathbf{P} = \mathbf{P}^T$*, then* $\mathbf{P}$ *is an* orthogonal projector (Galerkin projection)*, otherwise an* oblique projector (Petrov-Galerkin projection).

The following useful properties of (spectral) projectors can be shown, see, e.g., [117, Section 1.12] and [17].

Proposition 2.3.1. *Let* $\mathbf{Z} \in \mathbb{R}^{n \times n}$ *be as in Definition 2.3.1, and let* $\mathbf{P} \in \mathbb{R}^{n \times n}$ *be a (spectral) projector onto* $\mathcal{V}$ *(in the spectral case, onto the right* $\mathbf{Z}$*-invariant subspace corresponding to* Λ_1*). Then the following assertions hold true*

a) $\operatorname{rank}(\mathbf{P}) = \dim \mathcal{V} = |\Lambda_1| := r$,

b) $\mathbf{P}$ *is the identity operator on* $\mathcal{V}$*, i.e.,* $\mathbf{P}\mathbf{v} = \mathbf{v}, \ \forall \mathbf{v} \in \mathcal{V}$,

c) $\operatorname{range}(\mathbf{P}) = \operatorname{range}(\mathbf{Z}\mathbf{P})$,

d) $\ker(\mathbf{P}) = \operatorname{range}(\mathbf{I} - \mathbf{P}), \ \operatorname{range}(\mathbf{P}) = \ker(\mathbf{I} - \mathbf{P})$,

e) $\mathbf{I} - \mathbf{P}$ *is a spectral projector onto the right* $\mathbf{Z}$*-invariant subspace corresponding to* Λ_2.

f) Let $\mathbf{V} = \begin{bmatrix}\mathbf{v}_1, \ldots, \mathbf{v}_r\end{bmatrix}$ *be an orthonormal basis matrix for* $\mathcal{V}$*, then* $\mathbf{P} = \mathbf{V}\mathbf{V}^T$ *is an orthogonal projector onto* $\mathcal{V}$*.*

g) Let $\mathcal{W} \subset \mathbb{R}^n$ *be another subspace of the same dimension as* $\mathcal{V}$ *with basis matrix* $\mathbf{W} = \begin{bmatrix}\mathbf{w}_1, \ldots, \mathbf{w}_r\end{bmatrix}$ *, then* $\mathbf{P} = \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T$ *is an oblique projector onto* $\mathcal{V}$ *along* $\mathcal{W}$*.*

In terms of the above properties, let us assume that for an original system $\mathbf{\Sigma}$ of the form (2.2), the state vector $\mathbf{x}$ is approximated by an oblique projection $\mathbf{P} = \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T$, i.e., let $\mathbf{x} \approx \mathbf{P}\mathbf{x}$. Inserting our approximation into the state equation of (2.2) and imposing a common Petrov-Galerkin condition on the residual, we obtain that

$$\mathbf{P}\dot{\mathbf{x}} - \mathbf{A}\mathbf{P}\mathbf{x} + \mathbf{B}\mathbf{u} \perp \mathcal{W}$$

which implies that

$$(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T(\mathbf{P}\dot{\mathbf{x}} - \mathbf{A}\mathbf{P}\mathbf{x} + \mathbf{B}\mathbf{u}) = \mathbf{0}. \tag{2.39}$$

Introducing a new state variable as $\hat{\mathbf{x}} = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{x}$, the last equation can be interpreted as

$$\dot{\hat{\mathbf{x}}} - \underbrace{(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V}}_{\hat{\mathbf{A}}}\,\hat{\mathbf{x}} + \underbrace{(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{B}}_{\hat{\mathbf{B}}}\,\mathbf{u} = \mathbf{0}. \tag{2.40}$$

Moreover, with this notation, for the output equation we thus get

$$\mathbf{y} \approx \hat{\mathbf{y}} = \mathbf{C}\mathbf{V}\hat{\mathbf{x}}. \tag{2.41}$$

Hence, a reduced-order system as in (2.37) can be obtained by a Petrov-Galerkin type projection $\mathbf{P} = \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T$. It still remains open how to choose the projection subspaces $\mathbf{V}$ and $\mathbf{W}$ such that $||\mathbf{y} - \hat{\mathbf{y}}||$ is minimized. Below, we present two well-known choices for $\mathbf{V}$ and $\mathbf{W}$ that are based on different ideas.

2.3.1 Interpolation-based model reduction

Recall that the transfer function $H(s)$ of a SISO dynamical system is a rational function in s of degree equal to the state dimension of the system under consideration. Hence, following classical approximation theory, one might think of constructing a reduced-order system described by a rational function of lower degree that interpolates the original function at certain prescribed points within the complex plane. Similarly, in the MIMO case, if we can ensure that for a complex number $\sigma \in \mathbb{C}$ we have

$$\mathbf{H}(\sigma) = \hat{\mathbf{H}}(\sigma), \tag{2.42}$$

then for frequencies $s \approx \sigma$, the reduced-order system can be expected to faithfully reproduce the original system dynamics. Using the properties of projections, one can show that this can be achieved by solving a certain linear system, see [71, Theorem 3.1] and [127].

Theorem 2.3.1. *Let an original system* $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C})$ *be given. Assume that a reduced-order model is given as*

$$\hat{\mathbf{A}} = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V}, \quad \hat{\mathbf{B}} = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{B}, \quad \hat{\mathbf{C}} = \mathbf{C}\mathbf{V},$$

with $\sigma \in \mathbb{C}\backslash(\Lambda(\mathbf{A}) \cup \Lambda(\hat{\mathbf{A}}))$ *and either*

- $(\sigma\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \in \operatorname{range}(\mathbf{V})$, *or*
- $(\sigma\mathbf{I} - \mathbf{A})^{-*}\mathbf{C} \in \operatorname{range}(\mathbf{W})$.

Then it holds

$$\mathbf{H}(\sigma) = \hat{\mathbf{H}}(\sigma),$$

i.e., the reduced transfer function $\hat{\mathbf{H}}$ *is a rational matrix-valued interpolant of* $\mathbf{H}$ *in* $s = \sigma$.

Although the above interpolation can easily be ensured, for MIMO systems it is often desirable to keep the projection subspace as small as possible, leading to the concept of so-called tangential interpolation, see, e.g., [62]. Here, for a given set of interpolation points $S = \{\sigma_1, \ldots, \sigma_r\}$, together with left and right tangential directions $\tilde{\mathbf{b}}_i \in \mathbb{R}^m$, $i = 1, \ldots, r$ and $\tilde{\mathbf{c}}_i \in \mathbb{R}^p$, $i = 1, \ldots, r$, respectively, the goal is to come up with a reduced-order system such that

$$\mathbf{H}(\sigma_i)\tilde{\mathbf{b}}_i = \hat{\mathbf{H}}(\sigma_i)\tilde{\mathbf{b}}_i, \qquad i = 1, \ldots, r, \tag{2.43a}$$

$$\tilde{\mathbf{c}}_i^T\mathbf{H}(\sigma_i) = \tilde{\mathbf{c}}_i^T\hat{\mathbf{H}}(\sigma_i), \qquad i = 1, \ldots, r. \tag{2.43b}$$

It is clear that the choice of *good* or even *optimal* interpolation points is of particular interest in order to guarantee a satisfactory performance of the reduced-order system compared to the original behavior. However, if the frequency range of interest is not known, this step within the model reduction process is far from being trivial. On the contrary, many authors have discussed reasonable interpolation points that also try to take care of preserving system properties such as stability or passivity, see e.g. [3, 4, 36, 44, 46, 52, 60, 62, 71, 73, 127].

2.3.2 Balancing-based model reduction

A more system theoretic model reduction approach is given by the method of balanced truncation where the solutions of the Lyapunov equations (2.13) play an important role. The idea originated within the design of digital filters (see [104]) and, in context of model reduction of linear control systems, is further discussed in [103]. The main motivation of this approach is to transform the system into a realization from which one can easily read off system states that are *important* and *less important* for the input-output behavior, respectively. For this, it is noteworthy that there is a direct connection between the system Gramians $\mathbf{P}$ and $\mathbf{Q}$ and the energy associated with an arbitrary system state $\mathbf{x}_*$.

To be more precise, due to the assumption of a reachable system, the smallest amount of energy needed to reach $\mathbf{x}_*$ from the zero initial state $\mathbf{x}_0 = \mathbf{0}$ is given by

$$\mathbf{J}_r = \mathbf{x}_*^T \mathbf{P}^{-1} \mathbf{x}_*. \tag{2.44}$$

Similarly, the energy obtained by observing the output of an uncontrolled system $\boldsymbol{\Sigma}$ with zero initial condition can be derived to satisfy

$$\mathbf{J}_o = \mathbf{x}_*^T \mathbf{Q} \mathbf{x}_*. \tag{2.45}$$

Hence, if a system satisfies $\mathbf{P} = \mathbf{Q} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ with decreasing eigenvalues σ_i, then states that need a large amount of energy to be reached yield only a small amount of energy if they are observed. This fact immediately suggest that these states do not contribute much to the system behavior and thus may be neglected to obtain a reduced-order system. In fact, systems with equal and diagonal Gramians are called *balanced realizations*. Transforming a system into a balanced realization determines the first step in the balanced truncation model reduction procedure. Moreover, the matrix $\operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ consists of the *Hankel singular values* of $\boldsymbol{\Sigma}$ which basically are the nonzero singular values of the *Hankel operator*. Since it is of little importance for this thesis, we refrain from a detailed discussion on the latter operator and instead just refer to [3, Chapter 8]. Coming back to the balanced truncation method, the goal is to transform a given minimal and stable system $\boldsymbol{\Sigma}$ by means of a state space transformation $\mathbf{T}_b$ such that a balanced realization is obtained. This can be achieved by, e.g., the *square root balancing* method proposed in [64]. According to Proposition 2.2.3, for a minimal and stable system $\boldsymbol{\Sigma}$, the Gramians $\mathbf{P}$ and $\mathbf{Q}$ are symmetric positive-definite. Hence, assume that $\mathbf{P} = \mathbf{S}^T\mathbf{S}$ and $\mathbf{Q} = \mathbf{R}^T\mathbf{R}$. By simple algebraic multiplications it can be shown that a state space transformation with $\mathbf{T}_b = \mathbf{D}^{-\frac{1}{2}}\mathbf{Z}^T\mathbf{R}$ and $\mathbf{T}_b^{-1} = \mathbf{S}^T\mathbf{U}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{S}\mathbf{R}^T = \mathbf{U}\mathbf{D}\mathbf{Z}^T$ is the singular value decomposition, yields a balanced realization which in partitioned form looks like

$$\dot{\tilde{\mathbf{x}}}(t) = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \tilde{\mathbf{x}}(t) + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{u}(t) \qquad \tilde{\mathbf{y}}(t) = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix} \tilde{\mathbf{x}}(t), \tag{2.46}$$

with $\mathbf{A}_{11} \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $\mathbf{B}_1 \in \mathbb{R}^{\hat{n}\times m}$ and $\mathbf{C}_1 \in \mathbb{R}^{p\times\hat{n}}$. As indicated by its name, the next step consists of truncating the system states in order to obtain a reduced-order model. If we set $\hat{\boldsymbol{\Sigma}} = (\mathbf{A}_{11}, \mathbf{B}_1, \mathbf{C}_1)$, following [103] and [3, Chapter 7], we conclude that $\hat{\boldsymbol{\Sigma}}$ is a balanced realization, i.e., $\hat{\mathbf{P}} = \hat{\mathbf{Q}} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\hat{n}})$. Moreover, the Hankel singular values of the reduced-order system coincide with those of the original system and one can show (see [49]) the following error bound

$$||\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}||_{\mathcal{H}_\infty} \leq 2(\sigma_{\hat{n}+1} + \cdots + \sigma_n). \tag{2.47}$$

In view of the theory of projections, the same model $\hat{\boldsymbol{\Sigma}}$ is obtained if we construct a Petrov-Galerkin type projection $\mathbf{P} = \mathbf{V}\mathbf{W}^T$ with

$$\mathbf{V} = \mathbf{S}^T\mathbf{U}_{\hat{n}}\mathbf{D}_{\hat{n}}^{-\frac{1}{2}}, \quad \mathbf{W} = \mathbf{D}_{\hat{n}}^{-\frac{1}{2}}\mathbf{Z}_{\hat{n}}^T\mathbf{R},$$

and

$$\mathbf{U}_{\hat{n}} = \left[\mathbf{u}_1, \ldots, \mathbf{u}_{\hat{n}}\right], \ \mathbf{D} = \operatorname{diag}\left(\sigma_1, \ldots, \sigma_{\hat{n}}\right) \quad \text{and} \quad \mathbf{Z}_{\hat{n}} = \left[\mathbf{z}_1, \ldots, \mathbf{z}_{\hat{n}}\right].$$

CHAPTER 3

LINEAR SYSTEMS

Contents

3.1 Introduction

In this chapter, we focus on linear dynamical systems. Though the intrinsic nature of most complex physical processes indeed is nonlinear, linear models often help to give a first approximation of the underlying dynamics. Moreover, in several real-life applications, linearizing a system around a known operating point already yields a sufficient reflection of the original model allowing to design, e.g., controllers based on the analysis of the linear system. The resulting trade-off of giving up accuracy is then compensated by the fact that, as we have seen in the previous chapter, theory for linear control systems can be considered as well-understood. The goal is to point out some new aspects of model order reduction of linear control systems that provide a unifying framework for different reduction approaches that so far have been considered unrelated. In particular, we study the solution of large-scale matrix equations of the form

$$\mathbf{AXE} + \mathbf{MXH} + \mathbf{BC} = \mathbf{0},$$

where $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{n\times n}$, $\mathbf{E}, \mathbf{H} \in \mathbb{R}^{q\times q}, \mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times q}$. We now proceed as follows. We start with a brief review of $\mathcal{H}_2$-optimal model reduction of linear systems, including different necessary optimality conditions as well as algorithms converging to corresponding local optima. Subsequently we show that for symmetric state space systems constructing locally $\mathcal{H}_2$-optimal models is equivalent to minimizing a certain energy norm that naturally arises for the Lyapunov equation of the system. For unsymmetric systems, we give an interpretation of the $\mathcal{H}_2$-norm as an error measure for the approximation of the controllability and the observability Gramian. Furthermore, we extend some of the ideas to a more general setting which in Chapter 4 is shown to be interpretable as a generalized interpolation-based framework as well. By means of some numerical examples, we compare new and existing methods and underscore our theoretical results. We emphasize that the results of this chapter are primarily of theoretical interest. In particular, the discussed and proposed algorithms will not be competitive in terms of computational efficiency when compared to state-of-the-art low rank solution methods. However, they are optimal with respect to certain energy norms and should help to improve the current understanding of low rank approximations to linear matrix equations.

3.2 $\mathcal{H}_2$-optimal model reduction

Let us consider a continuous LTI control system of the form (2.2) with zero feedthrough term and zero initial condition, i.e., $\mathbf{x}_0 = \mathbf{0}$. For systems with non-zero initial condition, a transformation via introducing an artificial constant input signal as well as a reference trajectory allows to embed all of the following concepts into the above case. Let us now study the optimal $\mathcal{H}_2$-model reduction problem, where the goal is to find the best stable $\hat{n}$-dimensional model with transfer function $\hat{\mathbf{H}}$ s.t.

$$||\mathbf{H} - \hat{\mathbf{H}}||_{\mathcal{H}_2} = \min_{\substack{\dim(\tilde{\mathbf{H}}) \,=\, \hat{n} \\ \tilde{\mathbf{H}} \text{ stable}}} ||\mathbf{H} - \tilde{\mathbf{H}}||_{\mathcal{H}_2}. \tag{3.1}$$

3.2.1 Necessary optimality conditions

Since the set of all $\hat{n}$-dimensional systems whose transfer functions are in $\mathcal{H}_2$ is nonconvex, finding a global minimum is infeasible and thus one usually aims at constructing locally optimal models that fulfill first order necessary optimality conditions. Over the last decades, several conditions have been derived and they have been shown to be equivalent. Here, we give a brief review of the conditions that are of particular interest for this thesis.

The Wilson conditions

Recall from Chapter 2 that the $\mathcal{H}_2$-error of a reduced-order model $\hat{\mathbf{\Sigma}}$ can be measured by means of the solution of the Lyapunov equations (2.13). According to Lemma 2.2.1 and the corresponding extension to the MIMO case, we obtain

$$||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2}^2 = \operatorname{tr}\left(\mathbf{C}_{err}\mathbf{P}_{err}\mathbf{C}_{err}^T\right) = \operatorname{tr}\left(\mathbf{B}_{err}^T\mathbf{Q}_{err}\mathbf{B}_{err}\right), \tag{3.2}$$

where $\mathbf{A}_{err}$, $\mathbf{B}_{err}$ and $\mathbf{C}_{err}$ are defined as in (2.38) and $\mathbf{P}_{err}$ and $\mathbf{Q}_{err}$ are the solutions of the Lyapunov equations of the error system, i.e.,

$$\mathbf{A}_{err}\mathbf{P}_{err} + \mathbf{P}_{err}\mathbf{A}_{err}^T + \mathbf{B}_{err}\mathbf{B}_{err}^T = \mathbf{0}, \tag{3.3a}$$

$$\mathbf{A}_{err}^T\mathbf{Q}_{err} + \mathbf{Q}_{err}\mathbf{A}_{err} + \mathbf{C}_{err}^T\mathbf{C}_{err} = \mathbf{0}. \tag{3.3b}$$

In order to derive necessary conditions for $\mathcal{H}_2$-optimality, we can now interpret $J := ||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2}^2$ as a cost functional depending on the reduced-order system matrices $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$. Hence, by computing $\nabla J_{\hat{\mathbf{A}}}$, $\nabla J_{\hat{\mathbf{B}}}$ and $\nabla J_{\hat{\mathbf{C}}}$ and setting these expressions equal to zero, we obtain

$$\mathbf{P}_{12}^T\mathbf{Q}_{12} + \mathbf{P}_{22}\mathbf{P}_{22} = \mathbf{0}, \tag{3.4a}$$

$$\mathbf{Q}_{12}^T\mathbf{B} + \mathbf{Q}_{22}\hat{\mathbf{B}} = \mathbf{0}, \tag{3.4b}$$

$$\hat{\mathbf{C}}\mathbf{P}_{22} - \mathbf{C}\mathbf{P}_{12} = \mathbf{0}, \tag{3.4c}$$

where we assume that $\mathbf{P}_{err}$ and $\mathbf{Q}_{err}$ are partitioned as follows:

$$\mathbf{P}^{err} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{bmatrix}, \quad \mathbf{Q}^{err} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}^T & \mathbf{Q}_{22} \end{bmatrix}.$$

The above conditions, usually referred to as the Wilson conditions for $\mathcal{H}_2$-optimality, were first discussed in [131] and, in [73], were shown to be equivalent to the subsequently following interpolation-based optimality conditions.

Interpolation-based conditions

Alternatively, one can characterize $\mathcal{H}_2$-optimality in terms of the transfer functions of the original and the reduced-order system. For this, we cite a rather recently introduced expression for the $\mathcal{H}_2$-norm of the error system based on the pole-residue expression (2.26) and Lemma 2.2.2. For convenience we begin with the SISO case and briefly state the extension for MIMO systems from [73].

Corollary 3.2.1. *Given the original system $\mathbf{\Sigma}$ and a reduced-order system $\hat{\mathbf{\Sigma}}$, let λ_i and $\hat{\lambda}_j$ be the simple poles of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$, respectively, and suppose that the poles of $\hat{\mathbf{\Sigma}}$ are distinct. Let Φ_i and $\hat{\Phi}_j$ denote the residues of the transfer functions $H(s)$ and $\hat{H}(s)$ at their poles λ_i and $\hat{\lambda}_j$, respectively: $\Phi_i = \operatorname{res}[H(s), \lambda_i] := \lim_{s\to\lambda_i} H(s)(s-\lambda_i)$, $i = 1, \ldots, n$ and $\hat{\Phi}_j = \operatorname{res}[\hat{H}(s), \lambda_j]$ for $j = 1, \ldots, \hat{n}$. The $\mathcal{H}_2$-norm of the error system is given by*

$$||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2}^2 = \sum_{i=1}^{n} \Phi_i \left(H(-\lambda_i) - \hat{H}(-\lambda_i) \right) - \sum_{j=1}^{\hat{n}} \hat{\Phi}_j \left(H(-\hat{\lambda}_j) - \hat{H}(-\hat{\lambda}_j) \right). \tag{3.5}$$

Initially investigated in [99] in terms of orthogonality conditions for the transfer function, a locally optimal SISO model now satisfies certain Hermite interpolation conditions. To be more precise, the transfer function of a locally $\mathcal{H}_2$-optimal model interpolates the transfer function and its derivative of the original system at its own system poles reflected at the imaginary axis. One way to derive these conditions is to use the above formula together with the reduced-order system poles $\hat{\lambda}_j$ and the residues $\hat{\Phi}_j$ as optimization parameters for $\hat{H}$. This leads to the optimality conditions

$$H(-\hat{\lambda}_j) = \hat{H}(-\hat{\lambda}_j), \quad H'(-\hat{\lambda}_j) = \hat{H}'(-\hat{\lambda}_j), \quad j = 1, \ldots, \hat{n}. \tag{3.6}$$

Although from Chapter 2 we know how to ensure that $\hat{H}(s)$ interpolates $H(s)$ at arbitrary prescribed interpolation points $\sigma \in \mathbb{C}$, here the problem we are faced with is that these points are not known a priori. Nevertheless, in [73], the authors have proposed the iterative rational Krylov algorithm (IRKA), see Algorithm 3.2.1, a very reliable iterative scheme that, upon convergence, yields a reduced-order model fulfilling (3.6).

We already have seen that in the case of MIMO systems, the transfer function is a matrix-valued rational function, i.e., $\mathbf{H}(s) \in \mathbb{R}(s)^{p\times n}$ and the concept of interpolation is usually understood as tangential interpolation as described in (2.43). Making use of this framework, a locally $\mathcal{H}_2$-optimal reduced-order model has to satisfy

$$\tilde{\mathbf{c}}_j^T \mathbf{H}(-\hat{\lambda}_j) = \tilde{\mathbf{c}}_j^T \hat{\mathbf{H}}(-\hat{\lambda}_j), \qquad j = 1, \dots, \hat{n}, \tag{3.7a}$$

$$\mathbf{H}(-\hat{\lambda}_j)\tilde{\mathbf{b}}_j = \hat{\mathbf{H}}(-\hat{\lambda}_j)\tilde{\mathbf{b}}_j, \qquad j = 1, \dots, \hat{n}, \tag{3.7b}$$

$$\tilde{\mathbf{c}}_j^T \mathbf{H}'(-\hat{\lambda}_j)\tilde{\mathbf{b}}_j = \tilde{\mathbf{c}}_j^T \hat{\mathbf{H}}'(-\hat{\lambda}_j)\tilde{\mathbf{b}}_j, \qquad j = 1, \dots, \hat{n}, \tag{3.7c}$$

where $\mathbf{R}\hat{\mathbf{\Lambda}}\mathbf{R}^{-1} = \hat{\mathbf{A}}$ and $\hat{\mathbf{\Lambda}} = \operatorname{diag}\left(\hat{\lambda}_1, \dots, \hat{\lambda}_{\hat{n}}\right)$, $\tilde{\mathbf{B}} = \hat{\mathbf{B}}^T\mathbf{R}^{-T}$ and $\tilde{\mathbf{C}} = \hat{\mathbf{C}}\mathbf{R}$ is the spectral decomposition of the system.

Algorithm 3.2.1 Iterative rational Krylov algorithm (IRKA)

```
Input: Initial selection of interpolation points $\sigma_i$, $i = 1, \dots, \hat{n}$ that is closed under
    conjugation and a convergence tolerance $tol$.
Output: $\hat{\mathbf{A}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{c}}$
Choose $\mathbf{V}$ and $\mathbf{W}$ s.t. $\mathcal{V} = \operatorname{span}\{(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \dots, (\sigma_{\hat{n}}\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\}$ and $\mathcal{W} =$
    $\operatorname{span}\{(\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \dots, (\sigma_{\hat{n}}\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\}$ and $\mathbf{W}^T\mathbf{V} = \mathbf{I}$.
while relative change in $\{\sigma_i\} > tol$ do
    $\hat{\mathbf{A}} = \mathbf{W}^T\mathbf{A}\mathbf{V}$,
    assign $\sigma_i \leftarrow -\lambda_i(\hat{\mathbf{A}})$ for $i = 1, \dots, \hat{n}$,
    update $\mathbf{V}$ and $\mathbf{W}$ s.t. $\mathcal{V} = \operatorname{span}\{(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \dots, (\sigma_{\hat{n}}\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\}$ and $\mathcal{W} =$
        $\operatorname{span}\{(\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \dots, (\sigma_{\hat{n}}\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\}$ and $\mathbf{W}^T\mathbf{V} = \mathbf{I}$.
end while
$\hat{\mathbf{A}} = \mathbf{W}^T\mathbf{A}\mathbf{V}$, $\hat{\mathbf{b}} = \mathbf{W}^T\mathbf{b}$, $\hat{\mathbf{c}} = \mathbf{c}^T\mathbf{V}$
```

Remark 3.2.1. *Note that the demand of having a set S of interpolation points that are closed under conjugation means that if $\sigma_i \in S$ is complex, then we also have $\bar{\sigma}_i \in S$.*

3.2.2 The discrete-time case

For the sake of completeness, we briefly state the corresponding optimality conditions for linear discrete-time systems. The Lyapunov-based approach then is replaced by the solution of the Stein equations of the error system

$$\mathbf{A}_{err}\mathbf{P}_{err}\mathbf{A}_{err}^T + \mathbf{B}_{err}\mathbf{B}_{err}^T = \mathbf{P}_{err}, \tag{3.8a}$$

$$\mathbf{A}_{err}^T\mathbf{Q}_{err}\mathbf{A}_{err} + \mathbf{C}_{err}^T\mathbf{C}_{err} = \mathbf{Q}_{err}. \tag{3.8b}$$

Using the same partitioning as for the continuous-time case, if a reduced-order model $\hat{\mathbf{\Sigma}}$ is locally h_2-optimal, it satisfies

$$\mathbf{Q}_{12}^T\mathbf{A}\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{A}}\mathbf{P}_{22} = \mathbf{0}, \tag{3.9a}$$

$$\mathbf{Q}_{12}^T\mathbf{B} + \mathbf{Q}_{22}\hat{\mathbf{B}} = \mathbf{0}, \tag{3.9b}$$

$$\hat{\mathbf{C}}\mathbf{P}_{22} - \mathbf{C}\mathbf{P}_{12} = \mathbf{0}. \tag{3.9c}$$

Similar to [73], in [36], the authors derive an interpolatory framework which states the optimality conditions in terms of tangential interpolation of the transfer function. The result from [36] is given in the following theorem.

Theorem 3.2.1. *Given a large-scale linear discrete-time MIMO control system with transfer function* $\mathbf{H}(s)$*. Let* $\hat{\mathbf{H}}(s)$ *be the transfer function of the reduced-order system given in an eigenvector basis* $\hat{\mathbf{A}} = \operatorname{diag}\left(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{n}}\right)$, $\hat{\mathbf{B}} = \left[\hat{\mathbf{b}}_1^*, \ldots, \hat{\mathbf{b}}_{\hat{n}}^*\right]^*$ *and* $\hat{\mathbf{C}} = \left[\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_{\hat{n}}\right]$. *If* $\hat{\mathbf{H}}(s)$ *solves the* h_2*-optimal problem, then the following conditions are satisfied*

$$\mathbf{c}_j^*\mathbf{H}\left(\frac{1}{\hat{\lambda}_j^*}\right) = \mathbf{c}_j^*\hat{\mathbf{H}}\left(\frac{1}{\hat{\lambda}_j^*}\right), \qquad j = 1, \ldots, \hat{n}, \tag{3.10a}$$

$$\mathbf{H}\left(\frac{1}{\hat{\lambda}_j^*}\right)\mathbf{b}_j^* = \hat{\mathbf{H}}\left(\frac{1}{\hat{\lambda}_j^*}\right)\mathbf{b}_j^*, \qquad j = 1, \ldots, \hat{n}, \tag{3.10b}$$

$$\mathbf{c}_j^*\mathbf{H}'\left(\frac{1}{\hat{\lambda}_j^*}\right)\mathbf{b}_j^* = \mathbf{c}_j^*\hat{\mathbf{H}}'\left(\frac{1}{\hat{\lambda}_j^*}\right)\mathbf{b}_j^*, \qquad j = 1, \ldots, \hat{n}, \tag{3.10c}$$

where $\frac{1}{\hat{\lambda}_j^*}$ *are the mirror images with respect to the unit circle of the poles of* $\hat{\mathbf{H}}(s)$.

Analog to the continuous-time case, an iterative algorithm (MIRIAm) which fulfills the above conditions upon convergence is discussed in [36].

3.3 Interpolatory methods for large-scale matrix equations

In this section, we show that the problem of $\mathcal{H}_2$-model reduction is closely related to the approximation of solutions of large-scale linear matrix equations. In particular, for symmetric state space systems, we prove that a locally $\mathcal{H}_2$-optimal reduced-order system automatically leads to a low rank approximation of the solution of the Lyapunov equation that minimizes the naturally induced energy norm of the underlying linear operator. For unsymmetric systems, we investigate the use of the $\mathcal{H}_2$-norm as an adequate error measure and provide a method to minimize the residual norm for the Lyapunov equation by means of a generalized $\mathcal{H}_2$-optimality framework.

3.3.1 Existing low rank approaches

As we have already seen, in typical examples such as, e.g., the heat equation, the Gramian $\mathbf{P}$ often can be well approximated by a low order subspace. This fact has been exploited by several authors that proposed different methods. Here, we give a brief overview on existing low rank techniques that have been proven successful over the last years.

Projection-based methods

Due to their particular importance for this chapter, we begin with projection-based methods that were first studied in [116]. The idea is to construct a low rank approximation $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$ according to the following two steps. At first, one has to specify a suitable projection subspace $\mathbf{V} \in \mathbb{R}^{n \times \hat{n}}$ such that $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Then, one solves the reduced-order Lyapunov equation

$$\underbrace{\mathbf{V}^T\mathbf{A}\mathbf{V}}_{\hat{\mathbf{A}}}\hat{\mathbf{P}} + \hat{\mathbf{P}}\underbrace{\mathbf{V}^T\mathbf{A}^T\mathbf{V}}_{\hat{\mathbf{A}}^T} + \underbrace{\mathbf{V}^T\mathbf{B}}_{\hat{\mathbf{B}}}\underbrace{\mathbf{B}^T\mathbf{V}^T}_{\hat{\mathbf{B}}^T} = \mathbf{0},$$

and prolongates back to the original space $\mathbb{R}^{n \times n}$ by setting $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$. Since the reduced solution $\hat{\mathbf{P}}$ is only of dimension $\hat{n} \ll n$, it may be obtained by direct solution techniques such as the Bartels-Stewart algorithm, see [12], or Hammarling's method, see [75]. However, in general one has to be careful when projecting the equation since uniqueness and positive-definiteness of $\hat{\mathbf{P}}$ may not be guaranteed. At least for dissipative $\mathbf{A}$, i.e., $\sigma(\mathbf{A} + \mathbf{A}^T) < 0$, this always holds true. Of course, the quality of the approximation $\mathbf{P}_{\hat{n}}$ heavily depends on the choice of the projection subspace $\mathbf{V}$ for which several possibilities have been proposed. In [116], the integral representation (2.11) has motivated using $\mathbf{V} = \mathcal{K}_{\hat{n}}(\mathbf{A}, \mathbf{B})$, i.e., the (block) Krylov subspace generated by the matrix $\mathbf{A}$ and the input matrix $\mathbf{B}$. Moreover, the author has shown that this approach actually is equivalent to the approximation of the integral (2.11) by means of a numerical quadrature technique. Later on, in [120], the so-called *Krylov-Plus-Inverted-Krylov* (KPIK) method was introduced. Here, the projection subspace is constructed by the union of the Krylov subspaces $\mathcal{K}_{\hat{n}}(\mathbf{A}, \mathbf{B})$ and $\mathcal{K}_{\hat{n}}(\mathbf{A}^{-1}, \mathbf{A}^{-1}\mathbf{B})$. This slight modification has lead to a significant improvement in the approximation quality and often turns out to be a fast and easily implementable low rank method. We also refer to the works from [83, 84] which point into a similar direction. Rather recently, more and more effort has been put into the use of rational Krylov subspaces that may further improve the approximations, see [45, 55, 59]. In the next section, we give a theoretical explanation for the fact that the rational Krylov subspaces corresponding to an $\mathcal{H}_2$-optimal ROM often lead to very accurate low approximations.

The LRCF-ADI iteration

A quite different technique has its origin in solving elliptic and parabolic difference equations, see [107]. Interestingly, due to the properties of the Lyapunov operator, the alternating directions implicit (ADI) iteration can also be used to construct approximate solutions of the Lyapunov equation, see [129]. In [96, 108], it is shown how to implement the to-be-expected low rank structure of the solution within the ADI iteration. As a consequence, for an appropriate set of complex shift parameters $\{p_1, \ldots, p_q\} \in \mathbb{C}$, the following iteration converges to the true solution $\mathbf{P}$.

$$\mathbf{Z}_1 = \sqrt{2p_1}(\mathbf{A} - p_1\mathbf{I})^{-1}\mathbf{B}, \tag{3.11a}$$

$$\mathbf{Z}_j = \left[\sqrt{2p_j}(\mathbf{A} - p_j\mathbf{I})^{-1}\mathbf{B}, (\mathbf{A} - p_j I)^{-1}(\mathbf{A} + p_j\mathbf{I})\mathbf{Z}_{j-1}\right]. \tag{3.11b}$$

Besides the study of good or optimal shift parameters, over the last few years there have evolved several works on an efficient implementation of the method, its extension to Sylvester and Riccati equations and also on the use of the ADI subspaces within a projection-based approach. Since a complete overview of all details is beyond the scope of this thesis, we instead refer to some of the most standard references in this area of research, e.g., [5, 26, 27, 28, 30, 31, 96, 97, 108, 110, 135] .

Iterative solvers for the linear system

Finally, instead of considering the matrix equation itself, one might construct low rank approximations by means of the system of linear equations (2.14). The solution of the system then can be obtained by means of an iterative Krylov-based solver like, e.g., CG, BiCG or MinRes (see [50, 51, 90]). The crucial step is to observe that each intermediate iterate also exhibits a low rank structure that allows to improve the efficiency of the methods significantly. There exist further low rank techniques such as, e.g., the sign function iteration, see [29], and approaches based on hierarchical matrices and multigrid techniques, see, e.g., [14, 15, 67, 69]. However, in this thesis we mainly focus on (generalizations) of the previously mentioned methods and the subsequently following new technique from [125].

3.3.2 Riemannian optimization and the energy norm

A very recent and, at a first glance, completely different approach is proposed in [125, Chapter 4]. However, this method actually is closely connected to the problem of $\mathcal{H}_2$-optimal model reduction. Hence, let us make use of the setting of [125, Chapter 4] and focus on symmetric dynamical systems of the form

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \tag{3.12}$$
$$\mathbf{y}(t) = \mathbf{B}^T\mathbf{x}(t), \tag{3.13}$$

with $\mathbf{E} = \mathbf{E}^T \succ 0$ and $\mathbf{A} = \mathbf{A}^T \prec 0$. From now on, whenever we write $\mathbf{\Sigma} = (\mathbf{E}; \mathbf{A}, \mathbf{B}, \mathbf{B})$, we refer to a system of this form. Although in Chapter 2, we mainly discussed standard state space systems with $\mathbf{E} = \mathbf{I}$, we already indicated that for a generalized state space system (3.12), the Gramians now satisfy

$$\mathbf{A}\mathbf{P}\mathbf{E}^T + \mathbf{E}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}, \tag{3.14}$$
$$\mathbf{A}^T\mathbf{Q}\mathbf{E} + \mathbf{E}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0}. \tag{3.15}$$

Moreover, as we have seen in Theorem 2.2.1, the singular values of the solutions $\mathbf{P}$ and $\mathbf{Q}$ tend to decay exponentially fast also in the generalized state space setting, motivating the search for accurate low rank approximations. Due to our symmetry assumptions, the Gramians coincide and we are faced with solving

$$\mathbf{A}\mathbf{P}\mathbf{E} + \mathbf{E}\mathbf{P}\mathbf{A} + \mathbf{B}\mathbf{B}^T = \mathbf{0}. \tag{3.16}$$

According to Chapter 2, equivalently we might consider the system of linear equations

$$\mathcal{L}\,\mathrm{vec}\,(\mathbf{P}) = -\,\mathrm{vec}\left(\mathbf{B}\mathbf{B}^T\right), \tag{3.17}$$

with $\mathcal{L} = \mathbf{E}\otimes\mathbf{A} + \mathbf{A}\otimes\mathbf{E}$. Since by assumption we have $\mathbf{E} = \mathbf{E}^T \succ 0$ and $\mathbf{A} = \mathbf{A}^T \prec 0$, it follows that $\mathcal{L} = \mathcal{L}^T \prec 0$ and, consequently, we can define an energy norm via

$$||\cdot||_{\mathcal{L}} = \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{L}}} \quad \text{with } \langle\mathbf{U},\mathbf{V}\rangle_{\mathcal{L}} = \langle -\mathcal{L}\,\mathrm{vec}\,(\mathbf{U})\,,\mathrm{vec}\,(\mathbf{V})\rangle, \tag{3.18}$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times n}$. In [125, 126], the authors construct a method based on Riemannian optimization that computes a low rank approximation $\mathbf{P}_{\hat{n}}$ by minimizing the objective function

$$f : \mathcal{M} \to \mathbb{R},\ \mathbf{P} \mapsto -\operatorname{tr}(\mathbf{PAPE}) - \operatorname{tr}\left(\mathbf{PBB}^T\right)$$

on the manifold $\mathcal{M}$ of symmetric positive semi-definite matrices of rank $\hat{n}$, i.e.,

$$\mathcal{M} = \{\mathbf{P} : \mathbf{P} \in S_n^{sym}, \mathbf{P} \succeq 0, \operatorname{rank}(\mathbf{P}) = \hat{n}\}. \tag{3.19}$$

The specific function f is motivated by the fact that it holds

$$\begin{aligned} ||\mathbf{P} - \mathbf{P}_{\hat{n}}||_{\mathcal{L}}^2 &= -2\operatorname{tr}(\mathbf{P}_{\hat{n}}\mathbf{E}\mathbf{P}_{\hat{n}}\mathbf{A}) - 2\operatorname{tr}\left(\mathbf{P}_{\hat{n}}\mathbf{BB}^T\right) - 2\operatorname{tr}(\mathbf{PEPA}) \\ &= 2f(\mathbf{P}_{\hat{n}}) - 2\operatorname{tr}(\mathbf{PEPA}). \end{aligned}$$

Since the second term depends only on the true solution $\mathbf{P}$, it is constant and it thus suffices to minimize f. The first step now is to realize that there is a close relationship between the elements in $\mathcal{M}$ and approximations constructed by a projection-based approach. This is seen as follows. Let $\mathbf{P}_{\hat{n}} \in \mathcal{M}$. Hence it can be written as $\mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$, where $\mathbf{V} \in \mathbb{R}^{n\times\hat{n}}$ is an orthogonal matrix and $\hat{\mathbf{P}} = \hat{\mathbf{P}}^T \in \mathbb{R}^{\hat{n}\times\hat{n}}$. In order to minimize the objective function f, we compute the derivative of f with respect to $\hat{\mathbf{P}}$ and, according to Theorem 2.1.1, obtain:

$$\begin{aligned} \frac{\partial f}{\partial \hat{\mathbf{P}}} &= \frac{\partial}{\partial \hat{\mathbf{P}}}\left(\operatorname{tr}\left(\mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T\mathbf{E}\mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T\mathbf{A}\right) + \operatorname{tr}\left(\mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T\mathbf{BB}^T\right)\right) \\ &= \hat{\mathbf{A}}\hat{\mathbf{P}}\hat{\mathbf{E}} + \hat{\mathbf{E}}\hat{\mathbf{P}}\hat{\mathbf{A}} + \hat{\mathbf{B}}\hat{\mathbf{B}}^T. \end{aligned} \tag{3.20}$$

Consequently, as a necessary optimality condition we obtain that $\hat{\mathbf{P}}$ has to be the solution of the Lyapunov equation associated with the projected system matrices $\hat{\mathbf{E}} = \mathbf{V}^T\mathbf{EV}, \hat{\mathbf{A}} = \mathbf{V}^T\mathbf{AV}$ and $\hat{\mathbf{B}} = \mathbf{V}^T\mathbf{B}$. In the context of matrix equations, this is also known as the typical Galerkin condition, see [116]. For this reason, instead of using the Riemannian optimization approach, we want to construct an approximation by projecting onto a suitable subspace $\mathbf{V}$.

3.3.3 A lower bound property of the $\mathcal{H}_2$-Norm

In the following, we discuss a link between IRKA and the Riemannian optimization method. The most important observation is that the energy norm of every low rank approximant is bounded below by the $\mathcal{H}_2$-norm of the associated error system. For this, we need the following result.

Lemma 3.3.1. *Let* $\mathbf{\Sigma} = (\mathbf{E}; \mathbf{A}, \mathbf{B}, \mathbf{B}^T)$ *denote a symmetric dynamical system with* $\mathbf{E} \succ 0$ *and* $\mathbf{A} \prec 0$. *Further assume that* $\hat{\mathbf{\Sigma}} = (\hat{\mathbf{E}}; \hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{B}}^T)$ *is a reduced-order system obtained by a Galerkin-type projection* $\mathbf{P} = \mathbf{V}\mathbf{V}^T$. *Then it holds*

$$||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2}^2 \leq ||\mathbf{\Sigma}||_{\mathcal{H}_2}^2 - ||\hat{\mathbf{\Sigma}}||_{\mathcal{H}_2}^2,$$

with equality in case of $\hat{\mathbf{\Sigma}}$ *being a locally* $\mathcal{H}_2$*-optimal reduced-order system.*

Proof. By the definition of the $\mathcal{H}_2$-inner product, we know that it holds:

$$\begin{aligned}\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \mathbf{\Sigma} - \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} &= \langle \mathbf{\Sigma}, \mathbf{\Sigma} \rangle_{\mathcal{H}_2} - 2\langle \mathbf{\Sigma}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} + \langle \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} \\ &= \langle \mathbf{\Sigma}, \mathbf{\Sigma} \rangle_{\mathcal{H}_2} - 2\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} - \langle \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2}.\end{aligned}$$

By the computation formulas from Lemma 2.2.1, we have that

$$\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} = \operatorname{tr}\left(\mathbf{C}_{err}\tilde{\mathbf{P}}\hat{\mathbf{B}}\right) = \operatorname{vec}\left(\mathbf{C}_{err}^T\hat{\mathbf{B}}^T\right)^T \operatorname{vec}\left(\tilde{\mathbf{P}}\right),$$

where $\mathbf{C}_{err} = \begin{bmatrix}\mathbf{B}^T & -\hat{\mathbf{B}}^T\end{bmatrix}$ and $\tilde{\mathbf{P}}$ is the solution of the generalized Sylvester equation

$$\underbrace{\begin{bmatrix}\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{A}}\end{bmatrix}}_{\mathbf{A}_{err}}\tilde{\mathbf{P}}\hat{\mathbf{E}} + \underbrace{\begin{bmatrix}\mathbf{E} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{E}}\end{bmatrix}}_{\mathbf{E}_{err}}\tilde{\mathbf{P}}\hat{\mathbf{A}} + \mathbf{B}_{err}\hat{\mathbf{B}}^T = \mathbf{0}.$$

Multiplying from the left with $\begin{bmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}\end{bmatrix}$, we get

$$\begin{bmatrix}\mathbf{A} & \mathbf{0} \\ \mathbf{0} & -\hat{\mathbf{A}}\end{bmatrix}\tilde{\mathbf{P}}\hat{\mathbf{E}} + \begin{bmatrix}\mathbf{E} & \mathbf{0} \\ \mathbf{0} & -\hat{\mathbf{E}}\end{bmatrix}\tilde{\mathbf{P}}\hat{\mathbf{A}} + \mathbf{C}_{err}^T\hat{\mathbf{B}}^T = \mathbf{0}.$$

Next, note that by Proposition 2.1.2, we have

$$
\begin{aligned}
&\operatorname{vec}\left(\mathbf{C}_{err}^T\hat{\mathbf{B}}^T\right)^T\left(\hat{\mathbf{E}}\otimes\begin{bmatrix}-\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{A}}\end{bmatrix}+\hat{\mathbf{A}}\otimes\begin{bmatrix}-\mathbf{E} & \mathbf{0}\\ \mathbf{0} & \hat{\mathbf{E}}\end{bmatrix}\right)^{-1}\operatorname{vec}\left(\mathbf{C}_{err}^T\hat{\mathbf{B}}^T\right)\\
&=\underbrace{\operatorname{vec}\left(\mathbf{B}\hat{\mathbf{B}}^T\right)^T}_{\mathbf{x}^T}\underbrace{\left(-\hat{\mathbf{E}}\otimes\mathbf{A}-\hat{\mathbf{A}}\otimes\mathbf{E}\right)^{-1}}_{\mathbf{M}^{-1}}\underbrace{\operatorname{vec}\left(\mathbf{B}\hat{\mathbf{B}}^T\right)}_{\mathbf{x}}\\
&\quad-\operatorname{vec}\left(\hat{\mathbf{B}}\hat{\mathbf{B}}^T\right)^T\left(-\hat{\mathbf{E}}\otimes\hat{\mathbf{A}}-\hat{\mathbf{A}}\otimes\hat{\mathbf{E}}\right)^{-1}\operatorname{vec}\left(\hat{\mathbf{B}}\hat{\mathbf{B}}^T\right).
\end{aligned}
$$

Hence, if we set $\mathbf{Z}=(\mathbf{I}\otimes\mathbf{V})$, it holds that

$$
\langle\mathbf{\Sigma}-\hat{\mathbf{\Sigma}},\hat{\mathbf{\Sigma}}\rangle_{\mathcal{H}_2}=\mathbf{x}^T\underbrace{\left(\mathbf{M}^{-1}-\mathbf{Z}\left(\mathbf{Z}^T\mathbf{M}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\right)}_{\mathcal{S}}\mathbf{x}.
$$

However, $\mathcal{S}$ is the Schur complement of $\mathbf{S}=\begin{bmatrix}\mathbf{Z}^T\mathbf{M}\mathbf{Z} & \mathbf{Z}^T\\ \mathbf{Z} & \mathbf{M}^{-1}\end{bmatrix}$ in $\mathbf{M}^{-1}$. Let $\mathbf{s}=\begin{bmatrix}\mathbf{y}\\ \mathbf{z}\end{bmatrix}$ now be an arbitrary vector. Then, it holds that

$$
\mathbf{s}^T\mathbf{S}\mathbf{s}=\mathbf{y}^T\mathbf{Z}^T\mathbf{M}\mathbf{Z}\mathbf{y}+\mathbf{y}^T\mathbf{Z}^T\mathbf{z}+\mathbf{z}^T\mathbf{Z}\mathbf{y}+\mathbf{z}^T\mathbf{M}^{-1}\mathbf{z}.
$$

Defining $\mathbf{q}:=\mathbf{M}\mathbf{Z}\mathbf{y}$, it follows that

$$
\mathbf{s}^T\mathbf{S}\mathbf{s}=\left(\mathbf{q}^T+\mathbf{z}^T\right)\mathbf{M}^{-1}(\mathbf{q}+\mathbf{z})\geq 0.
$$

This means that $\mathbf{S}$ as well as its Schur complement $\mathcal{S}$ are positive semi-definite. Hence, this shows that $\langle\mathbf{\Sigma}-\hat{\mathbf{\Sigma}},\hat{\mathbf{\Sigma}}\rangle_{\mathcal{H}_2}\geq 0$. Assume now that $\tilde{\mathbf{P}}:=\begin{bmatrix}\tilde{\mathbf{P}}_1\\ \tilde{\mathbf{P}}_2\end{bmatrix}$. Then, it follows that

$$
\mathbf{A}\tilde{\mathbf{P}}_1\hat{\mathbf{E}}+\mathbf{E}\tilde{\mathbf{P}}_1\hat{\mathbf{A}}+\mathbf{B}\hat{\mathbf{B}}^T=\mathbf{0},\quad \hat{\mathbf{A}}\tilde{\mathbf{P}}_2\hat{\mathbf{E}}+\hat{\mathbf{E}}\tilde{\mathbf{P}}_2\hat{\mathbf{A}}+\hat{\mathbf{B}}\hat{\mathbf{B}}^T=\mathbf{0},
$$

and

$$
\langle\mathbf{\Sigma}-\hat{\mathbf{\Sigma}},\hat{\mathbf{\Sigma}}\rangle_{\mathcal{H}_2}=\operatorname{tr}\left(\mathbf{B}^T\tilde{\mathbf{P}}_1\hat{\mathbf{B}}-\hat{\mathbf{B}}^T\tilde{\mathbf{P}}_2\hat{\mathbf{B}}\right).
$$

Finally, due to the Wilson conditions for $\mathcal{H}_2$-optimality, a locally optimal reduced-order model satisfies $\mathbf{B}^T\tilde{\mathbf{P}}_1-\hat{\mathbf{B}}^T\tilde{\mathbf{P}}_2=\mathbf{0}$ which proves the statement. □

Remark 3.3.1. *An alternative way of proving the above statement is given by the pole-residue expression for the $\mathcal{H}_2$-error and results on the residues of symmetric state space systems as well as on the difference of the transfer functions which have been shown in [58].*

However, it now easily follows that for symmetric state space systems, IRKA yields low rank approximations $\mathbf{P}_{\hat{n}}$ that minimize the distance to the true solution $\mathbf{P}$ with respect to the $\mathcal{L}$-norm.

Theorem 3.3.1. *Let $\mathbf{\Sigma} = (\mathbf{E}; \mathbf{A}, \mathbf{B}, \mathbf{B}^T)$ denote a symmetric dynamical system $\mathbf{\Sigma}$ with $\mathbf{E} \succ 0$ and $\mathbf{A} \prec 0$ and let $\mathbf{V}$ denote a projection matrix corresponding to a reduced-order model $\hat{\mathbf{\Sigma}} = (\hat{\mathbf{E}}; \hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{B}}^T)$. Let further $\mathbf{P}$ and $\hat{\mathbf{P}}$ denote the solutions of the associated Lyapunov equations and set $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$. Then*

$$||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2} \leq ||\mathbf{P} - \mathbf{P}_{\hat{n}}||_{\mathcal{L}}, \tag{3.21}$$

with equality in case of $\hat{\mathbf{\Sigma}}$ being a locally $\mathcal{H}_2$-optimal reduced-order system.

Proof. First, note that the vectorized solutions of the original and reduced Lyapunov equations are obtained as follows:

$$\begin{aligned} \operatorname{vec}(\mathbf{P}) &= -\underbrace{(\mathbf{E} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{E})^{-1}}_{\mathcal{L}^{-1}} \operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right), \\ \operatorname{vec}\left(\hat{\mathbf{P}}\right) &= -\underbrace{(\hat{\mathbf{E}} \otimes \hat{\mathbf{A}} + \hat{\mathbf{A}} \otimes \hat{\mathbf{E}})^{-1}}_{\hat{\mathcal{L}}^{-1}} \operatorname{vec}\left(\hat{\mathbf{B}}\hat{\mathbf{B}}^T\right). \end{aligned}$$

Hence, we subsequently derive

$$\begin{aligned} &||\mathbf{P} - \mathbf{P}_{\hat{n}}||_{\mathcal{L}}^2 = \operatorname{vec}(\mathbf{P} - \mathbf{P}_{\hat{n}})^T (-\mathcal{L}) \operatorname{vec}(\mathbf{P} - \mathbf{P}_{\hat{n}}) \\ &= ||\mathbf{\Sigma}||_{\mathcal{H}_2}^2 - 2 \ \operatorname{vec}(\mathbf{P}_{\hat{n}})^T (-\mathcal{L}) \operatorname{vec}(\mathbf{P}) + \operatorname{vec}(\mathbf{P}_{\hat{n}})^T (-\mathcal{L}) \operatorname{vec}(\mathbf{P}_{\hat{n}}) \\ &= ||\mathbf{\Sigma}||_{\mathcal{H}_2}^2 - 2 \ \operatorname{vec}\left(\hat{\mathbf{P}}\right)^T \left(\mathbf{V}^T \otimes \mathbf{V}^T\right) (-\mathcal{L}) \operatorname{vec}(\mathbf{P}) + \operatorname{vec}\left(\hat{\mathbf{P}}\right)^T (-\hat{\mathcal{L}}) \operatorname{vec}\left(\hat{\mathbf{P}}\right) \\ &= ||\mathbf{\Sigma}||_{\mathcal{H}_2}^2 - 2 \ \operatorname{vec}\left(\hat{\mathbf{P}}\right)^T \left(\mathbf{V}^T \otimes \mathbf{V}^T\right) \operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right) + \operatorname{vec}\left(\hat{\mathbf{P}}\right)^T (-\hat{\mathcal{L}}) \operatorname{vec}\left(\hat{\mathbf{P}}\right) \\ &= ||\mathbf{\Sigma}||_{\mathcal{H}_2}^2 - 2 \ \operatorname{vec}\left(\hat{\mathbf{P}}\right)^T (-\hat{\mathcal{L}}) \operatorname{vec}\left(\hat{\mathbf{P}}\right) + \operatorname{vec}\left(\hat{\mathbf{P}}\right)^T (-\hat{\mathcal{L}}) \operatorname{vec}\left(\hat{\mathbf{P}}\right) \\ &= ||\mathbf{\Sigma}||_{\mathcal{H}_2}^2 - ||\hat{\mathbf{\Sigma}}||_{\mathcal{H}_2}^2 \end{aligned}$$

The assertion follows with the previous Lemma. □

Remark 3.3.2. *In summary, we conclude that the error with respect to the energy norm of each low rank approximation obtained by orthogonally prolongating the solution* *of a reduced Lyapunov equation is bounded below by the $\mathcal{H}_2$-norm of its associated error system. Since for IRKA this lower bound is not only minimized but at the same time equality is attained, we thus know that the right-hand side of (3.21) is locally minimized as well. However, this means that the corresponding low rank approximation to the solution of the Lyapunov equation is locally optimal with respect to the energy norm.*

3.3.4 Embedding the discrete-time case

We already presented the h_2-optimal MOR framework for the discrete-time case before. One might wonder if similar optimality properties for approximations to the solutions of the Stein equations can be shown here as well. As it turns out, this can easily be answered by means of the previous results. Let a stable symmetric discrete-time control system be given, i.e., consider

$$\mathbf{\Sigma}_d : \begin{cases} \mathbf{E}_d\, \mathbf{x}(k+1) = \mathbf{A}_d\, \mathbf{x}(k) + \mathbf{B}_d\, \mathbf{u}(k), \\ \qquad\quad \mathbf{y}(k) = \mathbf{B}_d^T\, \mathbf{x}(k), \end{cases} \tag{3.22}$$

with $\mathbf{E} = \mathbf{E}_d^T \succ 0$, $\mathbf{A}_d = \mathbf{A}_d^T \in \mathbb{R}^{n\times n}$ and $\mathbf{B}_d \in \mathbb{R}^{n\times m}$. Along the lines of this chapter, assume that we are interested in a low rank approximation $\mathbf{P}_{d,\hat{n}} = \mathbf{V}\hat{\mathbf{P}}_d\mathbf{V}^T$ to the true solution $\mathbf{P}_d$ of the Stein equation

$$\mathbf{A}_d\mathbf{P}_d\mathbf{A}_d + \mathbf{B}_d\mathbf{B}_d^T = \mathbf{E}_d\mathbf{P}_d\mathbf{E}_d. \tag{3.23}$$

Analog to the continuous case, the approximation $\mathbf{P}_{d,\hat{n}}$ should be determined via the solution of a reduced Stein equation. Hence, given a projection matrix $\mathbf{V}$, we first solve

$$\hat{\mathbf{A}}_d\hat{\mathbf{P}}_d\hat{\mathbf{A}}_d + \hat{\mathbf{B}}_d\hat{\mathbf{B}}_d^T = \hat{\mathbf{E}}_d\hat{\mathbf{P}}_d\hat{\mathbf{E}}_d, \tag{3.24}$$

where $\hat{\mathbf{A}}_d = \mathbf{V}^T\mathbf{A}_d\mathbf{V}$, $\hat{\mathbf{E}} = \mathbf{V}^T\mathbf{E}_d\mathbf{V}$ and $\hat{\mathbf{B}} = \mathbf{V}^T\mathbf{B}$ before we prolongate to get the approximation $\mathbf{P}_{d,\hat{n}}$. As a measure of accuracy for the approximation $\mathbf{P}_{d,\hat{n}}$, we use the Stein operator

$$\mathcal{D} = \mathbf{A}_d \otimes \mathbf{A}_d - \mathbf{E}_d \otimes \mathbf{E}_d \tag{3.25}$$

in order to introduce an energy norm via

$$||\cdot||_{\mathcal{D}} = \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{D}}} \qquad \text{with } \langle\mathbf{U},\mathbf{V}\rangle_{\mathcal{D}} = \langle -\mathcal{D}\,\mathrm{vec}\,(\mathbf{U})\,, \mathrm{vec}\,(\mathbf{V})\rangle, \tag{3.26}$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times n}$. Obviously, symmetry of $\mathbf{E}_d$ and $\mathbf{A}_d$ implies symmetry of $\mathcal{D}$. Moreover, $\mathcal{D}$ is negative definite as is easily seen as follows. Assume that $\mathbf{E}_d = \mathbf{L}\mathbf{L}^T, \mathbf{L} \in \mathbb{R}^{n \times n}$ is the Cholesky decomposition of $\mathbf{E}_d$. Hence, we have

$$\sigma(\mathbf{E}_d^{-1}\mathbf{A}_d) = \sigma(\mathbf{L}^{-T}\mathbf{L}^{-1}\mathbf{A}_d^{-1}) = \sigma(\mathbf{L}^T(\mathbf{L}^{-T}\mathbf{L}^{-1}\mathbf{A}_d)\mathbf{L}^{-T}) = \sigma(\mathbf{L}^{-1}\mathbf{A}_d\mathbf{L}^{-T}).$$

However, for an arbitrary vector $\mathbf{x} \in \mathbb{C}^{n^2}$, we now obtain

$$\mathbf{x}^*\mathcal{D}\mathbf{x} = \mathbf{x}^*(\mathbf{L} \otimes \mathbf{L}) \left((\mathbf{L}^{-1} \otimes \mathbf{L}^{-1})(\mathbf{A}_d \otimes \mathbf{A}_d)(\mathbf{L}^{-T} \otimes \mathbf{L}^{-T}) - \mathbf{I} \otimes \mathbf{I}\right) (\mathbf{L}^T \otimes \mathbf{L}^T)\mathbf{x} < 0.$$

The last step is due to the fact that the eigenvalues of $\mathbf{E}_d^{-1}\mathbf{A}_d$ lie within the unit disc. Accordingly, so do the eigenvalues of $\mathbf{L}^{-1}\mathbf{A}_d\mathbf{L}^{-T}$ and we have that $\mathcal{D} = \mathcal{D}^T \prec 0$. Let us now have a closer look at the Stein equation (3.23). Equivalently, we might simply solve a special Lyapunov equation. For this, note that it holds

$$\mathbf{A}_d\mathbf{P}_d\mathbf{A}_d - \mathbf{E}_d\mathbf{P}_d\mathbf{E}_d = \frac{1}{2}(\mathbf{A}_d - \mathbf{E}_d)\mathbf{P}_d(\mathbf{A}_d + \mathbf{E}_d) + \frac{1}{2}(\mathbf{A}_d + \mathbf{E}_d)\mathbf{P}_d(\mathbf{A}_d - \mathbf{E}_d).$$

In view of the previous subsection, this means we have to solve the Lyapunov equation

$$\mathbf{A}\mathbf{P}_d\mathbf{E} + \mathbf{E}\mathbf{P}_d\mathbf{A} + \mathbf{B}_d\mathbf{B}_d^T = \mathbf{0}, \tag{3.27}$$

where $\mathbf{A} = \mathbf{A}_d - \mathbf{E}_d$ and $\mathbf{E} = \frac{1}{2}(\mathbf{A}_d + \mathbf{E}_d)$. Due to symmetry of $\mathbf{A}_d$ and $\mathbf{E}_d$, it trivially follows that $\mathbf{A} = \mathbf{A}^T$ and $\mathbf{E} = \mathbf{E}^T$. With the same arguments as before, for a stable pencil $(\mathbf{A}_d, \mathbf{E}_d)$, we additionally have $\mathbf{A} \prec 0$ and $\mathbf{E} \succ 0$. Similarly, this holds true for the reduced system and we conclude that it holds that

$$||\mathbf{P}_d - \mathbf{P}_{d,\hat{n}}||_{\mathcal{D}} = ||\mathbf{P}_d - \mathbf{P}_{d,\hat{n}}||_{\mathcal{L}}.$$

A similar analysis for the continuous-time error system yields $\mathbf{P}_{err} = \mathbf{P}_{d,err}$ and it therefore follows that

$$||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2} = ||\mathbf{\Sigma}_d - \hat{\mathbf{\Sigma}}_d||_{h_2}.$$

Finally, let us discuss what happens if we construct an $\mathcal{H}_2$-optimal continuous-time reduced-order system. At the beginning of the chapter we stated the Wilson conditions only for standard state space systems with $\mathbf{E} = \mathbf{I}$. However, it is easy to show that in the generalized setting, we have

$$\mathbf{Q}_{12}^T\mathbf{A}\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{A}}\mathbf{P}_{22} = \mathbf{0}, \tag{3.28a}$$

$$\mathbf{Q}_{12}^T\mathbf{E}\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{E}}\mathbf{P}_{22} = \mathbf{0}, \tag{3.28b}$$

$$\mathbf{Q}_{12}^T\mathbf{B} + \mathbf{Q}_{22}\hat{\mathbf{B}} = \mathbf{0}, \tag{3.28c}$$

$$\hat{\mathbf{C}}\mathbf{P}_{22} - \mathbf{C}\mathbf{P}_{12} = \mathbf{0}. \tag{3.28d}$$

Comparing these conditions with the h_2-optimality conditions (3.9), we see that the last two coincide with those from the continuous-time case. Moreover, due to the special structure of $\mathbf{A}$ and $\mathbf{E}$, the first two equations from (3.28) are equivalent to

$$\mathbf{Q}_{12}^T(\mathbf{A}_d - \mathbf{E}_d)\mathbf{P}_{12} + \mathbf{Q}_{22}(\hat{\mathbf{A}}_d - \hat{\mathbf{E}}_d)\mathbf{P}_{22} = \mathbf{0}, \tag{3.29a}$$

$$\mathbf{Q}_{12}^T\frac{1}{2}(\mathbf{A}_d + \mathbf{E}_d)\mathbf{P}_{12} + \mathbf{Q}_{22}\frac{1}{2}(\hat{\mathbf{A}}_d + \mathbf{E}_d)\mathbf{P}_{22} = \mathbf{0}. \tag{3.29b}$$

The simple calculation $\frac{1}{2}(3.29a) + (3.29b)$ leads to

$$\mathbf{Q}_{12}^T\mathbf{A}_d\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{A}}_d\mathbf{P}_{22} = \mathbf{0}. \tag{3.30}$$

Analogously, from $(3.29b) - \frac{1}{2}(3.29a)$, we can conclude that

$$\mathbf{Q}_{12}^T\mathbf{E}_d\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{E}}_d\mathbf{P}_{22} = \mathbf{0}. \tag{3.31}$$

In other words, constructing a continuous-time $\mathcal{H}_2$-optimal ROM is equivalent to constructing a discrete-time h_2-optimal ROM. Altogether, the previous analysis shows the following discrete-time version of Theorem 3.3.1.

Theorem 3.3.2. *Let $\mathbf{\Sigma}_d = (\mathbf{E}_d; \mathbf{A}_d, \mathbf{B}_d, \mathbf{B}_d^T)$ denote a stable symmetric dynamical system $\mathbf{\Sigma}_d$ with $\mathbf{E}_d = \mathbf{E}_d^T \succ 0,\ \mathbf{A}_d = \mathbf{A}_d^T$ and let $\mathbf{V}$ denote a projection matrix corresponding to a reduced-order model $\hat{\mathbf{\Sigma}}_d = (\hat{\mathbf{E}}_d; \hat{\mathbf{A}}_d, \hat{\mathbf{B}}_d, \hat{\mathbf{B}}_d^T)$. Let further $\mathbf{P}_d$ and $\hat{\mathbf{P}}_d$ denote the solutions of the associated Stein equations and set $\mathbf{P}_{d,\hat{n}} = \mathbf{V}\hat{\mathbf{P}}_d\mathbf{V}^T$. Then*

$$||\mathbf{\Sigma}_d - \hat{\mathbf{\Sigma}}_d||_{h_2} \leq ||\mathbf{P}_d - \mathbf{P}_{d,\hat{n}}||_{\mathcal{D}}, \tag{3.32}$$

with equality in case of $\hat{\mathbf{\Sigma}}_d$ being a locally h_2-optimal reduced-order system.

Recapitulating what we have just shown, the MIRIAm algorithm from [36] automatically yields projection matrices that allow to construct locally optimal low rank approximations to the solution of the underlying Stein equation of the system. Due to the structural similarity between continuous-time and discrete-time systems, this observation is not too surprising. However, for later purposes we keep in mind that the results immediately follow as special cases from each other.

3.3.5 Problems in the unsymmetric case

So far, we have assumed that the system under consideration is symmetric. However, these assumptions characterize a rather limited class of dynamical systems. For example, let us consider the following simple two-dimensional system

$$\mathbf{E} = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \mathbf{c}.$$

Although the above system is stable, dissipative and has only real eigenvalues, it can be trivially shown that we can never transform it into a symmetric state space system which would allow defining an energy norm. This is due to the fact that the spectra of $\mathbf{E}$ and $\mathbf{A}$ lie on both sides of the imaginary axis and thus $\mathcal{L} = -\mathbf{E} \otimes \mathbf{A} - \mathbf{A} \otimes \mathbf{E}$ will be indefinite. Otherwise, if it is transformed into a definite matrix, the inputs and outputs will no longer be equal which is necessary for applying the techniques from the proof of Theorem 3.3.1.

Moreover, all systems with complex poles automatically exclude the possibility of an induced energy norm of the form $-\mathbf{E} \otimes \mathbf{A} - \mathbf{A} \otimes \mathbf{E}$. This is seen as follows. Assume that an unsymmetric dynamical system $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{b}, \mathbf{c}^T)$ is given, with $\mathbf{A}$ having complex eigenvalues. Assume now that the system can be transformed into a generalized symmetric state space system of the form $(\tilde{\mathbf{E}}; \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{\mathbf{b}}^T)$ and that the operator

$$\tilde{\mathcal{L}} = -\tilde{\mathbf{E}} \otimes \tilde{\mathbf{A}} - \tilde{\mathbf{A}} \otimes \tilde{\mathbf{E}}$$

is positive definite. Due to the Theorem of Stephanos, see, e.g., [94, Section 12.2], the eigenvalues of $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{A}}$ must all have equal or opposite sign since otherwise $\tilde{\mathcal{L}}$ would be indefinite. W.l.o.g. we assume that $\sigma(\tilde{\mathbf{E}}) \subset \mathbb{C}_+$. This means that $\tilde{\mathbf{E}}$ is symmetric positive definite and the eigenvalue problem for the pencil $(\tilde{\mathbf{A}}, \tilde{\mathbf{E}})$ can be transformed into a symmetric one. However, this would imply that all eigenvalues of $(\tilde{\mathbf{A}}, \tilde{\mathbf{E}})$ are real. Since the poles of a dynamical system are invariant under state space transformations, this would mean that all eigenvalues of $\mathbf{A}$ are real which is a contradiction to our assumption. Thus we cannot define the desired energy norm in a straightforward way.

Nevertheless, it remains the question if low rank Lyapunov approximations obtained by an IRKA reduced-order model still can be expected to be accurate even if the underlying dynamical system is unsymmetric and exhibits complex poles. For this, it is an interesting observation that the $\mathcal{H}_2$-norm of the error system vanishes if and only if the corresponding Lyapunov approximations that are generated by the reduced system are exact.

Theorem 3.3.3. *Let* $\mathbf{\Sigma} = (\mathbf{A}, \mathbf{B}, \mathbf{C})$ *denote a minimal stable dynamical system with* $\mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{B} \in \mathbb{R}^{n \times m}$ *and* $\mathbf{C} \in \mathbb{R}^{p \times n}$*. Assume that a stable reduced-order model* $\hat{\mathbf{\Sigma}} = (\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ *is constructed by a Petrov-Galerkin projection* $\mathbf{P} = \mathbf{V}\mathbf{W}^T$ *with* $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times \hat{n}}, \mathbf{V}^T\mathbf{V} = \mathbf{I}$ *and* $\mathbf{W}^T\mathbf{V} = \mathbf{I}$*. Let further* $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$ *and* $\mathbf{Q}_{\hat{n}} = \mathbf{W}\hat{\mathbf{Q}}\mathbf{W}^T$ *be obtained by solving the reduced Lyapunov equations*

$$\hat{\mathbf{A}}\hat{\mathbf{P}} + \hat{\mathbf{P}}\hat{\mathbf{A}}^T + \hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{0}, \quad \hat{\mathbf{A}}^T\hat{\mathbf{Q}} + \hat{\mathbf{Q}}\hat{\mathbf{A}} + \hat{\mathbf{C}}^T\hat{\mathbf{C}} = \mathbf{0}.$$

Then, the $\mathcal{H}_2$*-norm of the error system is zero if and only if* $\mathbf{P}_{\hat{n}} = \mathbf{P}$ *and* $\mathbf{Q}_{\hat{n}} = \mathbf{Q}$*, where* $\mathbf{P}$ *and* $\mathbf{Q}$ *are the exact solutions of the original Lyapunov equations*

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}, \quad \mathbf{A}^T\mathbf{Q} + \mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0}.$$

Proof. Let us assume that $||\mathbf{\Sigma}_{err}|| = ||\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}||_{\mathcal{H}_2} = 0$. By the definition of the $\mathcal{H}_2$-norm this means that

$$\int_0^\infty ||\mathbf{C}_{err}\ e^{\mathbf{A}_{err}t}\mathbf{B}_{err}||_F \ \mathrm{d}t = 0.$$

Hence, since $\mathbf{C}_{err}\ e^{\mathbf{A}_{err}t}\mathbf{B}_{err}$ is continuous it has to be the constant zero function and thus its derivatives evaluated at zero have to be zero as well, i.e.,

$$\mathbf{C}_{err}\ \mathbf{A}_{err}^i\mathbf{B}_{err} = \mathbf{0}, \quad i \geq 0.$$

Due to the structure of the error system this means that

$$\mathbf{C}\ \mathbf{A}^i\ \mathbf{B} = \hat{\mathbf{C}}\ \hat{\mathbf{A}}^i\ \hat{\mathbf{B}}, \quad i \geq 0.$$

Thus, the Markov parameters of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$ coincide. Since we assumed $\mathbf{\Sigma}$ to be a minimal realization, from Definition 2.2.3, it follows that $\hat{n} = n$. Consequently, the projection matrices $\mathbf{V}$ and $\mathbf{W}$ are (bi-)orthogonal. Let us now have a look at the transformed Lyapunov equation

$$\hat{\mathbf{A}}\hat{\mathbf{P}} + \hat{\mathbf{P}}\hat{\mathbf{A}}^T + \hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{0}.$$

Inserting the definition of $\hat{\mathbf{\Sigma}}$, we have

$$\mathbf{W}^T\mathbf{A}\mathbf{V}\hat{\mathbf{P}} + \hat{\mathbf{P}}\mathbf{V}^T\mathbf{A}^T\mathbf{W} + \mathbf{W}^T\mathbf{B}\mathbf{B}^T\mathbf{W} = \mathbf{0}.$$

Multiplying from the left with $\mathbf{V}$ and from the right with $\mathbf{V}^T$, we see that $\mathbf{P}_{\hat{n}}$ solves the original Lyapunov equation. Similarly, one can show that $\mathbf{Q}_{\hat{n}} = \mathbf{Q}$.

Conversely, let us assume that the approximation is exact, i.e., $\mathbf{P}_{\hat{n}} = \mathbf{P}$. As we have seen in the proof of Lemma 3.3.1, for the $\mathcal{H}_2$-norm of the error system, it holds that

$$\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \mathbf{\Sigma} - \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} = \langle \mathbf{\Sigma}, \mathbf{\Sigma} \rangle_{\mathcal{H}_2} - 2\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} - \langle \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2}.$$

Since $\mathbf{P}_{\hat{n}} = \mathbf{P}$, it follows that

$$\langle \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} = \hat{\mathbf{C}}\hat{\mathbf{P}}\hat{\mathbf{C}}^T = \mathbf{C}\mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T\mathbf{C}^T = \mathbf{C}\mathbf{P}_{\hat{n}}\mathbf{C}^T = \mathbf{C}\mathbf{P}\mathbf{C}^T = \langle \mathbf{\Sigma}, \mathbf{\Sigma} \rangle_{\mathcal{H}_2}.$$

Hence, in order to prove the assertion, it remains to show that it holds

$$\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} = 0.$$

Once again, analog to the proof of Lemma 3.3.1, we know that

$$\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} = \operatorname{tr}\left(\mathbf{C}\mathbf{M}\hat{\mathbf{C}}^T - \hat{\mathbf{C}}\hat{\mathbf{P}}\hat{\mathbf{C}}^T\right),$$

where $\mathbf{M}$ is the solution of

$$\mathbf{A}\mathbf{M} + \mathbf{M}\hat{\mathbf{A}}^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0}.$$

Since $\mathbf{A}$ and $\hat{\mathbf{A}}$ are assumed to be stable, the solution $\mathbf{M}$ is unique. However, since $\mathbf{P}_{\hat{n}}$ is the exact solution of the Lyapunov equation, we have

$$\mathbf{A}\mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T + \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}.$$

Multiplying from the right with $\mathbf{W}$, it follows that

$$\mathbf{A}\mathbf{V}\hat{\mathbf{P}} + \mathbf{V}\hat{\mathbf{P}}\hat{\mathbf{A}}^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0}.$$

Thus, it holds $\mathbf{V}\hat{\mathbf{P}} = \mathbf{M}$ and also $\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2} = 0$. □

Remark 3.3.3. *Theorem 3.3.3 shows that the $\mathcal{H}_2$-norm of the error system is an objective function which is zero if and only if the low rank Lyapunov approximations are the exact solutions. Hence, it seems reasonable to minimize this objective function in order to obtain approximations which are close to the exact solutions. However, this is exactly what the iterative rational Krylov algorithm aims at.*

3.3.6 Minimizing the residual norm

An alternative way for measuring the quality of a low rank approximation $\mathbf{P}_{\hat{n}}$ for unsymmetric systems clearly is given by the residual. In view of the formulation of the Lyapunov equation as a system of linear equations

$$\underbrace{(\mathbf{I} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{I})}_{\mathcal{L}} \underbrace{\operatorname{vec}(\mathbf{P})}_{\mathbf{p}} = \underbrace{\operatorname{vec}\left(-\mathbf{B}\mathbf{B}^T\right)}_{\mathcal{B}}, \tag{3.33}$$

for a given rank $\hat{n}$, we can thus try to construct an approximation $\mathbf{p}_{\hat{n}}$ that minimizes the norm of

$$\mathbf{r} = \mathcal{B} - \mathcal{L}\mathbf{p}_{\hat{n}}.$$

Again, this has been discussed within the framework of Riemannian optimization in [126, Chapter 4], where it was further shown that it holds

$$||\mathbf{R}||_F := ||\mathbf{r}||_2 = (\mathbf{p} - \mathbf{p}_{\hat{n}})^T \mathcal{L}^T \mathcal{L} (\mathbf{p} - \mathbf{p}_{\hat{n}}). \tag{3.34}$$

Despite the fact that the residual often is a less accurate estimator for the true error $||\mathbf{P} - \hat{\mathbf{P}}||_F$, it exhibits the obvious advantage that we do not have to assume that the system is symmetric, see also the discussion in [126]. Furthermore, note that there exist several other approaches that try to minimize the residual of the Lyapunov equation, see, e.g. [83]. However, the big difference to the setting here is that these methods construct an approximation that is optimal for a fixed projection matrix $\mathbf{V}$, not for a fixed rank $\hat{n}$.

Since in this thesis we are especially interested in the topic of $\mathcal{H}_2$-model reduction, let us have a closer look at the term $\mathcal{L}^T\mathcal{L}$ for which we obtain

$$\mathcal{T} := \mathcal{L}^T\mathcal{L} = \mathbf{I} \otimes \mathbf{A}^T\mathbf{A} + \mathbf{A}^T\mathbf{A} \otimes \mathbf{I} + \mathbf{A}^T \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{A}^T. \tag{3.35}$$

Interestingly, this operator can be associated with the generalized Lyapunov-type equation

$$\mathbf{A}^T\mathbf{A}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{A}^T\mathbf{A} + \mathbf{A}\tilde{\mathbf{P}}\mathbf{A} + \mathbf{A}^T\tilde{\mathbf{P}}\mathbf{A}^T + \mathbf{A}^T\mathbf{B}\mathbf{B}^T + \mathbf{B}\mathbf{B}^T\mathbf{A} = \mathbf{0}. \tag{3.36}$$

If we factor out $\mathbf{A}$ and $\mathbf{A}^T$ from both sides, the previous equation is equivalent to

$$\mathbf{A}^T \underbrace{(\mathbf{A}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)}_{\mathbf{T}} + \underbrace{(\mathbf{A}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T)}_{\mathbf{T}} \mathbf{A} = \mathbf{0}. \tag{3.37}$$

Making use of the Kronecker product notation, the above can be transformed into

$$(\mathbf{I} \otimes \mathbf{A}^T + \mathbf{A}^T \otimes \mathbf{I}) \operatorname{vec}(\mathbf{T}) = \mathbf{0}. \tag{3.38}$$

Assuming that $\mathbf{A}$ is a stable matrix, we can conclude that $\operatorname{vec}(\mathbf{T}) = \mathbf{0}$ and thus $\tilde{\mathbf{P}} = \mathbf{P}$. If we now denote $\mathbf{M} = \mathbf{A}^T\mathbf{A},\ \mathbf{F} = \begin{bmatrix}\mathbf{A}^T\mathbf{B} & \mathbf{B}\end{bmatrix}$ and $\mathbf{G} = \begin{bmatrix}\mathbf{B}^T \\ \mathbf{B}^T\mathbf{A}\end{bmatrix}$, we can rewrite equation (3.36) as

$$\mathbf{M}\tilde{\mathbf{P}} + \tilde{\mathbf{P}}\mathbf{M} + \mathbf{A}\tilde{\mathbf{P}}\mathbf{A} + \mathbf{A}^T\tilde{\mathbf{P}}\mathbf{A}^T + \mathbf{F}\mathbf{G} = \mathbf{0}. \tag{3.39}$$

Analog to the derivation in (3.20), for constructing an optimal rank-$\hat{n}$ approximation $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$ to $\tilde{\mathbf{P}}$, as a necessary optimality condition it follows that $\hat{\mathbf{P}}$ has to fulfill the reduced matrix equation

$$\hat{\mathbf{M}}\hat{\mathbf{P}} + \hat{\mathbf{P}}\hat{\mathbf{M}} + \hat{\mathbf{A}}\hat{\mathbf{P}}\hat{\mathbf{A}} + \hat{\mathbf{A}}^T\hat{\mathbf{P}}\hat{\mathbf{A}}^T + \hat{\mathbf{F}}\hat{\mathbf{G}} = \mathbf{0}, \tag{3.40}$$

where $\hat{\mathbf{M}} = \mathbf{V}^T\mathbf{M}\mathbf{V},\ \hat{\mathbf{A}} = \mathbf{V}^T\mathbf{A}\mathbf{V}, \hat{\mathbf{F}} = \mathbf{V}^T\mathbf{F}$ and $\hat{\mathbf{G}} = \mathbf{G}\mathbf{V}$. For a set of matrices $\mathbf{\Psi} = (\mathbf{M}, \mathbf{A}, \mathbf{F}, \mathbf{G})$, we introduce the objective function $f(\mathbf{\Psi}) = \operatorname{tr}\left(-\mathbf{G}\tilde{\mathbf{P}}\mathbf{F}\right)$, where $\tilde{\mathbf{P}}$ is determined via equation (3.39). Hence, for the residual (3.34), we can easily show that it holds

$$\begin{aligned}
||\mathbf{R}||_F &= \operatorname{vec}\left(\tilde{\mathbf{P}} - \mathbf{P}_{\hat{n}}\right)^T \mathcal{T} \operatorname{vec}\left(\tilde{\mathbf{P}} - \mathbf{P}_{\hat{n}}\right) \\
&= \operatorname{vec}\left(\tilde{\mathbf{P}}\right)^T \mathcal{T} \operatorname{vec}\left(\tilde{\mathbf{P}}\right) - 2 \operatorname{vec}(\mathbf{P}_{\hat{n}})^T \mathcal{T} \operatorname{vec}\left(\tilde{\mathbf{P}}\right) + \operatorname{vec}(\mathbf{P}_{\hat{n}})^T \mathcal{T} \operatorname{vec}(\mathbf{P}_{\hat{n}}) \\
&= f(\mathbf{\Psi}) - f(\hat{\mathbf{\Psi}}),
\end{aligned}$$

where $\hat{\mathbf{\Psi}} = (\hat{\mathbf{M}}, \hat{\mathbf{A}}, \hat{\mathbf{F}}, \hat{\mathbf{G}})$ and $\hat{\mathbf{P}}$ is given by equation (3.40). Again, we can try to find a lower bound for the difference $f(\mathbf{\Psi}) - f(\hat{\mathbf{\Psi}})$ whose minimization automatically minimizes the residual itself as well. For this, let us consider the set of matrices $\mathbf{\Psi}_{err}$ given as

$$\mathbf{M}_{err} = \begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{M}} \end{bmatrix}, \quad \mathbf{A}_{err} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{A}} \end{bmatrix}, \quad \mathbf{F}_{err} = \begin{bmatrix} \mathbf{F} \\ \hat{\mathbf{F}} \end{bmatrix}, \quad \mathbf{G}_{err} = \begin{bmatrix} \mathbf{G} & -\hat{\mathbf{G}} \end{bmatrix}.$$

Due to the special structure of $\mathbf{F}$ and $\mathbf{G}$, we have the following helpful relations

$$\mathbf{F}\mathbf{G} = \mathbf{G}^T\mathbf{F}^T, \quad \hat{\mathbf{F}}\hat{\mathbf{G}} = \hat{\mathbf{G}}^T\hat{\mathbf{F}}^T, \quad \hat{\mathbf{G}}^T\mathbf{F}_{err}^T = \hat{\mathbf{F}}\mathbf{G}_{err}\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix}. \tag{3.41}$$

Assume now that $\mathbf{P}_{err} = \begin{bmatrix} \tilde{\mathbf{P}} & \mathbf{X} \\ \mathbf{Y} & -\hat{\mathbf{P}} \end{bmatrix}$ is the solution of the matrix equation of the form (3.39) associated with $\mathbf{\Psi}_{err}$. Hence, it follows that

$$\mathbf{M}\mathbf{X} + \mathbf{X}\hat{\mathbf{M}} + \mathbf{A}\mathbf{X}\hat{\mathbf{A}} + \mathbf{A}^T\mathbf{X}\hat{\mathbf{A}}^T - \mathbf{F}\hat{\mathbf{G}} = \mathbf{0}, \tag{3.42}$$

$$\hat{\mathbf{M}}\mathbf{Y} + \mathbf{Y}\mathbf{M} + \hat{\mathbf{A}}\mathbf{Y}\mathbf{A} + \hat{\mathbf{A}}^T\mathbf{Y}\mathbf{A}^T + \hat{\mathbf{F}}\mathbf{G} = \mathbf{0}. \tag{3.43}$$

In particular, it holds that $\mathbf{X} = -\mathbf{Y}^T$. Making use of the fact that $\tilde{\mathbf{P}}$ and $\hat{\mathbf{P}}$ are the solutions of equations (3.39) and (3.40) and the relation between $\mathbf{F}$ and $\mathbf{G}$, we subsequently obtain

$$\begin{aligned} f(\mathbf{\Psi}_{err}) &= \operatorname{tr}\left(-\mathbf{G}_{err}\mathbf{P}_{err}\mathbf{F}_{err}\right) \\ &= \operatorname{tr}\left(-\mathbf{G}_{err}\left(\begin{bmatrix} \tilde{\mathbf{P}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{0} & -\hat{\mathbf{P}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{Y} & -\hat{\mathbf{P}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{P}} \end{bmatrix}\right)\mathbf{F}_{err}\right) \\ &= f(\mathbf{\Psi}) - f(\hat{\mathbf{\Psi}}) + \operatorname{tr}\left(-\mathbf{G}_{err}\left(\begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{0} & -\hat{\mathbf{P}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{Y} & -\hat{\mathbf{P}} \end{bmatrix}\right)\mathbf{F}_{err}\right) \\ &= f(\mathbf{\Psi}) - f(\hat{\mathbf{\Psi}}) - 2\operatorname{tr}\left(\mathbf{G}_{err}\begin{bmatrix} \mathbf{X} \\ -\hat{\mathbf{P}} \end{bmatrix}\hat{\mathbf{F}}\right). \end{aligned}$$

From here, one can proceed completely analog to what we have seen before in order to show that it holds

$$f(\mathbf{\Psi}_{err}) \le f(\mathbf{\Psi}) - f(\hat{\mathbf{\Psi}}). \tag{3.44}$$

Basically, the important point lies in realizing the analogy to the computation formula for $\langle \mathbf{\Sigma} - \hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}} \rangle_{\mathcal{H}_2}$ which we used in the previous section. Then by the exact same arguments, one can show that a special Schur complement is positive semi-definite which immediately leads to the previous bound.

Hence, if we can construct a local minimizer of $f(\mathbf{\Psi}_{err})$ that equals the previous bound, this would automatically lead to a local minimizer of the Lyapunov residual. Since the derivation of necessary optimality conditions for $f(\mathbf{\Psi}_{err})$ is very similar to the procedure of finding optimality conditions for the $\mathcal{H}_2$-model reduction for bilinear systems from [133], at some points we shorten our discussion. Let us now have a look at the objective function $J := f(\mathbf{\Psi}_{err})$ that we want to minimize. For the derivative with respect to an arbitrary parameter γ, such as, e.g., $\gamma = (\mathbf{P}_{err})_{ij}$, we know that it holds

$$\frac{\partial J}{\partial \gamma} = -\operatorname{tr}\left(\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{F}_{err}\mathbf{G}_{err}\right) - \operatorname{tr}\left(\mathbf{P}_{err}\frac{\partial(\mathbf{F}_{err}\mathbf{G}_{err})}{\partial \gamma}\right).$$

Recall that $\mathbf{P}_{err}$ is the solution of the associated matrix equation of the form (3.39), i.e.,

$$\mathbf{M}_{err}\mathbf{P}_{err} + \mathbf{P}_{err}\mathbf{M}_{err} + \mathbf{A}_{err}\mathbf{P}_{err}\mathbf{A}_{err} + \mathbf{A}_{err}^T\mathbf{P}_{err}\mathbf{A}_{err}^T + \mathbf{F}_{err}\mathbf{G}_{err} = \mathbf{0}. \tag{3.45}$$

Hence, we obtain that

$$\begin{aligned}\frac{\partial J}{\partial \gamma} = {} & \operatorname{tr}\left(\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{M}_{err}\mathbf{P}_{err}\right) + \operatorname{tr}\left(\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{P}_{err}\mathbf{M}_{err}\right) + \operatorname{tr}\left(\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{A}_{err}\mathbf{P}_{err}\mathbf{A}_{err}\right) \\ & + \operatorname{tr}\left(\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{A}_{err}^T\mathbf{P}_{err}\mathbf{A}_{err}^T\right) - \operatorname{tr}\left(\mathbf{P}_{err}\frac{\partial(\mathbf{F}_{err}\mathbf{G}_{err})}{\partial \gamma}\right).\end{aligned}$$

Next, we compute the derivate of (3.45) with respect to γ, leading to

$$\begin{aligned}-\frac{\partial(\mathbf{F}_{err}\mathbf{G}_{err})}{\partial \gamma} = {} & \frac{\partial \mathbf{M}_{err}}{\partial \gamma}\mathbf{P}_{err} + \mathbf{M}_{err}\frac{\partial \mathbf{P}_{err}}{\partial \gamma} + \frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{M}_{err} + \mathbf{P}_{err}\frac{\partial \mathbf{M}_{err}}{\partial \gamma} \\ & + \frac{\partial \mathbf{A}_{err}}{\partial \gamma}\mathbf{P}_{err}\mathbf{A}_{err} + \mathbf{A}_{err}\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{A}_{err} + \mathbf{A}_{err}\mathbf{P}_{err}\frac{\partial \mathbf{A}_{err}}{\partial \gamma} \\ & + \frac{\partial \mathbf{A}_{err}^T}{\partial \gamma}\mathbf{P}_{err}\mathbf{A}_{err}^T + \mathbf{A}_{err}^T\frac{\partial \mathbf{P}_{err}}{\partial \gamma}\mathbf{A}_{err}^T + \mathbf{A}_{err}^T\mathbf{P}_{err}\frac{\partial \mathbf{A}_{err}^T}{\partial \gamma}.\end{aligned}$$

Making use of a trick suggested in [133], we can multiply the last equation by $\mathbf{P}_{err}$ from the left and take the trace. If we insert the result into the expression for $\frac{\partial J}{\partial \gamma}$, we arrive at

$$\begin{aligned}\frac{\partial J}{\partial \gamma} = {} & -2\operatorname{tr}\left(\mathbf{P}_{err}\frac{\partial \mathbf{M}_{err}}{\partial \gamma}\mathbf{P}_{err}\right) - 2\operatorname{tr}\left(\frac{\partial \mathbf{A}_{err}}{\partial \gamma}\mathbf{P}_{err}\mathbf{A}_{err}\mathbf{P}_{err}\right) \\ & - 2\operatorname{tr}\left(\frac{\partial \mathbf{A}_{err}^T}{\partial \gamma}\mathbf{P}_{err}\mathbf{A}_{err}^T\mathbf{P}_{err}\right) - 2\operatorname{tr}\left(\mathbf{P}_{err}\frac{\partial(\mathbf{F}_{err}\mathbf{G}_{err})}{\partial \gamma}\right).\end{aligned}$$

From here, it is straightforward (cf. the derivation in [133]) to show that a locally optimal reduced set of matrices $\hat{\mathbf{\Psi}} = (\hat{\mathbf{M}}, \hat{\mathbf{A}}, \hat{\mathbf{F}}, \hat{\mathbf{G}})$ has to fulfill the following optimality conditions:

$$\mathbf{YX} + \hat{\mathbf{P}}\hat{\mathbf{P}} = \mathbf{0}, \quad \mathbf{YAX} + \hat{\mathbf{P}}\hat{\mathbf{A}}\hat{\mathbf{P}} = \mathbf{0}, \tag{3.46a}$$

$$\mathbf{GX} + \hat{\mathbf{G}}\hat{\mathbf{P}} = \mathbf{0}, \quad \mathbf{YF} - \hat{\mathbf{P}}\hat{\mathbf{F}} = \mathbf{0}. \tag{3.46b}$$

In particular, the latter conditions imply that $f(\mathbf{\Psi}_{err}) = f(\mathbf{\Psi}) - f(\hat{\mathbf{\Psi}})$. Hence, we propose Algorithm 3.3.1 in order to iteratively construct an approximation $\mathbf{P}_{\hat{n}}$ which aims at locally minimizing the residual for a given rank $\hat{n}$. Although at this part of the thesis, it might not be completely obvious why this method upon convergence indeed fulfills the optimaliy conditions, in the next chapter, we show that the foregoing generalized matrix equations play a crucial role in the context of $\mathcal{H}_2$-optimal model reduction for bilinear control systems. Since there the goal is to construct an iterative algorithm minimizing an objective function which basically coincides with $f(\mathbf{\Psi}_{err})$, we just refer to the next chapter for further details.

```
Algorithm 3.3.1 Minimization of the Lyapunov residual
Input: $\mathbf{A}, \mathbf{B}, \hat{\mathbf{M}}, \hat{\mathbf{A}}, \hat{\mathbf{F}}, \hat{\mathbf{G}}$
Output: $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$ fulfilling (3.46)
Set $\mathbf{M} = \mathbf{A}^T\mathbf{A},\ \mathbf{F} = \begin{bmatrix}\mathbf{A}^T\mathbf{B} & \mathbf{B}\end{bmatrix},\ \mathbf{G}^T = \begin{bmatrix}\mathbf{B} & \mathbf{A}^T\mathbf{B}\end{bmatrix}$.
while (not converged) do
    Solve $\mathbf{MX} + \mathbf{X}\hat{\mathbf{M}} + \mathbf{AX}\hat{\mathbf{A}} + \mathbf{A}^T\mathbf{X}\hat{\mathbf{A}}^T - \mathbf{F}\hat{\mathbf{G}} = \mathbf{0}$.
    $\mathbf{V} = \operatorname{orth}(\mathbf{X})$
    $\hat{\mathbf{M}} = \mathbf{V}^T\mathbf{MV},\ \hat{\mathbf{A}} = \mathbf{V}^T\mathbf{AV},\ \hat{\mathbf{F}} = \mathbf{V}^T\mathbf{F},\ \hat{\mathbf{G}} = \mathbf{GV}$
end while
Solve $\hat{\mathbf{M}}\hat{\mathbf{P}} + \hat{\mathbf{P}}\hat{\mathbf{M}} + \hat{\mathbf{A}}\hat{\mathbf{P}}\hat{\mathbf{A}} + \hat{\mathbf{A}}^T\hat{\mathbf{P}}\hat{\mathbf{A}}^T + \hat{\mathbf{F}}\hat{\mathbf{G}} = \mathbf{0}$.
Set $\mathbf{P}_{\hat{n}} = \mathbf{V}\hat{\mathbf{P}}\mathbf{V}^T$.
```

Remark 3.3.4. *Recapitulating the essential steps of this section, it is worthwhile to note the analogy to a system of linear equations* $\mathbf{Ax} = \mathbf{b}$. *For a symmetric positive definite* $\mathbf{A}$, *the Conjugate Gradient (CG) minimizes the error with respect to the energy norm induced by the matrix* $\mathbf{A}$. *On the other hand, for an unsymmetric* $\mathbf{A}$, *one can instead solve the normal equations* $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$ *and apply CG to the transformed system* $\mathbf{A}^T\mathbf{A}$ *which again is symmetric and positive-definite. As a result one obtains the so-called CGNR method, see [117, Section 8.3.1]. Since we also obtain a linear system when we use the Kronecker product notation of the Lyapunov equation, we can interpret the foregoing theory as an abstract extension of this method which in our case aims at minimizing the residual for a given rank* $\hat{n}$.

Remark 3.3.5. *Algorithm 3.3.1 should be understood more as a theoretical tool than as a compatible algorithm for constructing low rank approximations. We already mentioned the conceptual similarity between Algorithm 3.3.1 and the CGNR method. As is* *well-known, see [117, Section 8.3.1], the latter approach often results in a very slow convergence rate for common PDEs. Since the efficiency of Algorithm 3.3.1 also depends on the convergence rate, we cannot expect it to outperform state-of-the-art low rank techniques when it comes to computational efficiency. On the other hand, it obviously has the advantage of locally minimizing the residual for a given rank.*

3.3.7 Sylvester equations

Finally, let us briefly discuss the extension to more general matrix equations of the form

$$\mathbf{AXE} + \mathbf{MXH} + \mathbf{BC} = \mathbf{0}, \tag{3.47}$$

where $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{n\times n}$, $\mathbf{E}, \mathbf{H} \in \mathbb{R}^{q\times q}, \mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times q}$. Once again, we want to assume that all involved square matrices are symmetric and have eigenvalues either in $\mathbb{C}_-$ or in $\mathbb{C}_+$. To be more precise, we require $\mathbf{A} = \mathbf{A}^T \prec 0, \mathbf{H} = \mathbf{H}^T \prec 0, \mathbf{E} = \mathbf{E}^T \succ 0$ and $\mathbf{M} = \mathbf{M}^T \succ 0$. This allows us to define an energy norm based on the following symmetric negative definite matrix

$$\mathcal{L}_S = \mathbf{E} \otimes \mathbf{A} + \mathbf{H} \otimes \mathbf{M}.$$

We now seek for approximations of the form $\mathbf{X}_{\hat{n}} = \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T$ of rank $\hat{n}$ which minimize the $\mathcal{L}_S$-norm between the original solution $\mathbf{X}$ and $\mathbf{X}_{\hat{n}}$, i.e.

$$||\mathbf{X} - \mathbf{X}_{\hat{n}}||_{\mathcal{L}_S}^2 = \operatorname{vec}\left(\mathbf{X} - \mathbf{X}_{\hat{n}}\right)^T \left(-\mathcal{L}_S\right) \operatorname{vec}\left(\mathbf{X} - \mathbf{X}_{\hat{n}}\right).$$

Here, $\hat{\mathbf{X}}$ again is determined by solving a reduced Sylvester equation

$$\hat{\mathbf{A}}\hat{\mathbf{X}}\hat{\mathbf{E}} + \hat{\mathbf{M}}\hat{\mathbf{X}}\hat{\mathbf{H}} + \hat{\mathbf{B}}\hat{\mathbf{C}} = \mathbf{0},$$

while $\mathbf{V}$ and $\mathbf{W}$ denote projection matrices with $\mathbf{V}^T\mathbf{V} = \mathbf{W}^T\mathbf{W} = \mathbf{I}$. From now on, let

$$\mathbf{\Theta} = (\mathbf{A}, \mathbf{E}, \mathbf{M}, \mathbf{H}, \mathbf{B}, \mathbf{C})$$

denote an associated Sylvester equation of the form (3.47). Furthermore, let us consider the following objective function

$$f(\mathbf{\Theta}) = \operatorname{tr}\left(\mathbf{B}^T\mathbf{X}\mathbf{C}^T\right), \tag{3.48}$$

with $\mathbf{X}$ fulfilling (3.47). As it is easily seen, this function results from a slight modification of the $\mathcal{H}_2$-norm of a dynamical system and thus can be computed as

$$f(\mathbf{\Theta}) = \operatorname{vec}(\mathbf{BC})^T(-\mathbf{E}\otimes\mathbf{A} - \mathbf{H}\otimes\mathbf{M})^{-1}\operatorname{vec}(\mathbf{BC}). \tag{3.49}$$

For later purposes, it is helpful to note that f is invariant under orthonormal transformations.

Lemma 3.3.2. *Let $\mathbf{\Theta} = (\mathbf{A},\mathbf{E},\mathbf{M},\mathbf{H},\mathbf{B},\mathbf{C})$ denote a set of matrices and let $\mathbf{X}$ be the solution of the associated Sylvester equation (3.47). Assume that $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of the matrix pencil $(\mathbf{H},\mathbf{E})$ and that $\mathbf{Q}$ is a matrix containing an orthogonal set of eigenvectors. Then it holds*

$$f(\mathbf{\Theta}) = \operatorname{tr}\left(\mathbf{B}^T\mathbf{X}\mathbf{C}^T\right) = \operatorname{tr}\left(\mathbf{B}^T\mathbf{Y}\tilde{\mathbf{C}}^T\right),$$

where $\tilde{\mathbf{C}} = \mathbf{CQ}$ and $\mathbf{Y}$ is the solution of $\mathbf{AY} + \mathbf{MY\Lambda} + \mathbf{B}\tilde{\mathbf{C}} = \mathbf{0}$.

Proof. Let $\mathbf{Q}$ be the matrix of eigenvectors for the matrix pencil $(\mathbf{H},\mathbf{E})$, i.e., assume that it holds $\mathbf{Q}^T\mathbf{EQ} = \mathbf{I}$ and $\mathbf{Q}^T\mathbf{HQ} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix consisting of the eigenvalues. Since $\mathbf{H} = \mathbf{H}^T \prec 0$ and $\mathbf{E} = \mathbf{E}^T \succ 0$ this is always possible. If we now postmultiply equation (3.47) with $\mathbf{Q}$, we get

$$\mathbf{AXEQ} + \mathbf{MXHQ} + \mathbf{BCQ} = \mathbf{0}.$$

Due to the orthonormality of $\mathbf{Q}$, this can be transformed into

$$\mathbf{AXQQ}^T\mathbf{EQ} + \mathbf{MXQQ}^T\mathbf{HQ} + \mathbf{BCQ} = \mathbf{0}.$$

If we denote $\mathbf{Y} = \mathbf{XQ}$ and $\tilde{\mathbf{C}} = \mathbf{CQ}$, it follows that

$$\mathbf{AY} + \mathbf{MY\Lambda} + \mathbf{B}\tilde{\mathbf{C}} = \mathbf{0},$$

which implies that

$$\operatorname{tr}\left(\mathbf{B}^T\mathbf{Y}\tilde{\mathbf{C}}^T\right) = \operatorname{tr}\left(\mathbf{B}^T\mathbf{XQQ}^T\mathbf{C}\right) = \operatorname{tr}\left(\mathbf{B}^T\mathbf{X}\mathbf{C}^T\right).$$

□

Assume now that we have constructed a reduced set of matrices

$$\hat{\mathbf{\Theta}} = (\hat{\mathbf{A}}, \hat{\mathbf{E}}, \hat{\mathbf{M}}, \hat{\mathbf{H}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$$

by the following projection:

$$\begin{aligned} &\hat{\mathbf{A}} = \mathbf{V}^T\mathbf{A}\mathbf{V}, \quad \hat{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{W}, \quad \hat{\mathbf{M}} = \mathbf{V}^T\mathbf{M}\mathbf{V}, \\ &\hat{\mathbf{H}} = \mathbf{W}^T\mathbf{H}\mathbf{W}, \quad \hat{\mathbf{B}} = \mathbf{V}^T\mathbf{B}, \quad \hat{\mathbf{C}} = \mathbf{C}\mathbf{W}. \end{aligned} \tag{3.50}$$

Next, for $\mathbf{\Theta}$ and $\hat{\mathbf{\Theta}}$ we define the corresponding error set

$$\mathbf{\Theta}_{err} = (\mathbf{A}_{err}, \mathbf{E}_{err}, \mathbf{M}_{err}, \mathbf{H}_{err}, \mathbf{B}_{err}, \mathbf{C}_{err}),$$

with

$$\begin{aligned} \mathbf{A}_{err} &= \begin{bmatrix} -\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{A}} \end{bmatrix}, \quad \mathbf{E}_{err} = \begin{bmatrix} -\mathbf{E} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{E}} \end{bmatrix}, \quad \mathbf{M}_{err} = \begin{bmatrix} -\mathbf{M} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{M}} \end{bmatrix} \\ \mathbf{H}_{err} &= \begin{bmatrix} -\mathbf{H} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{H}} \end{bmatrix}, \quad \mathbf{B}_{err} = \begin{bmatrix} \mathbf{B} \\ \hat{\mathbf{B}} \end{bmatrix}, \quad \mathbf{C}_{err} = \begin{bmatrix} \mathbf{C} & \hat{\mathbf{C}} \end{bmatrix}. \end{aligned}$$

Similar to the previous cases, it is easy to show that a crucial lower bound is given by the objective function f evaluated in the error set $\mathbf{\Theta}_{err}$.

Corollary 3.3.1. *Let $\mathbf{\Theta}$ and $\hat{\mathbf{\Theta}}$ denote two sets of matrices associated with large and reduced generalized Sylvester equations of the form (3.47), respectively. Then, for the associated error set $\mathbf{\Theta}_{err}$, it holds that*

$$f(\mathbf{\Theta}_{err}) \leq f(\mathbf{\Theta}) - f(\hat{\mathbf{\Theta}}).$$

Hence, analog to the topic of $\mathcal{H}_2$-optimal model order reduction, we want to find a local minimizer of $f(\mathbf{\Theta}_{err})$. This can be done by deriving first order necessary conditions based on the computation formula of $f(\mathbf{\Theta}_{err})$. Due to the structural similarity to the previous sections, we only briefly mention how to proceed. For convenience, let us start with the case of $\mathbf{B} = \mathbf{b}$ and $\mathbf{C} = \mathbf{c}^T$. First of all, according to Lemma 3.3.2, we may w.l.o.g. assume that $\mathbf{H}_{err} = \mathbf{\Lambda}_{err}$ and $\mathbf{E}_{err} = \mathbf{I}$. Consequently, the objective function simplifies according to

$$\begin{aligned} f(\mathbf{\Theta}_{err}) &= \mathbf{b}_{err}^T \mathbf{X}_{err} \mathbf{c}_{err} \\ &= \left(\mathbf{c}_{err}^T \otimes \mathbf{b}_{err}^T\right) \left(-\mathbf{\Lambda}_{err} \otimes \mathbf{M}_{err} - \mathbf{I} \otimes \mathbf{A}_{err}\right)^{-1} \left(\mathbf{c}_{err} \otimes \mathbf{b}_{err}\right) \\ &= \sum_{i=1}^{n+\hat{n}} \mathbf{b}_{err}^T (-\lambda_i \mathbf{M}_{err} - \mathbf{A}_{err})^{-1} \mathbf{b}_{err} \; (\mathbf{c}_{err}^{(i)})^2, \end{aligned}$$

where $\mathbf{c}_{err}^{(i)}$ denotes the i-th component of $\mathbf{c}_{err}$. Setting the derivative of $f(\mathbf{\Theta}_{err})$ with respect to $\hat{\mathbf{c}}^{(j)}$ equal to zero yields

$$2 \; \hat{\mathbf{c}}^{(j)} \; \mathbf{b}_{err}^T (-\hat{\lambda}_j \mathbf{M}_{err} - \mathbf{A}_{err})^{-1} \mathbf{b}_{err} = 0,$$

with $\hat{\lambda}_j$ being the j-th eigenvalue of $(\hat{\mathbf{H}}, \hat{\mathbf{E}})$. However, in terms of interpolation, the above means that

$$\mathbf{b}^T (-\hat{\lambda}_j \mathbf{M} - \mathbf{A})^{-1} \mathbf{b} = \hat{\mathbf{b}}^T (-\hat{\lambda}_j \hat{\mathbf{M}} - \hat{\mathbf{A}})^{-1} \hat{\mathbf{b}}. \tag{3.51}$$

Similarly, for the derivative with respect to $\hat{\lambda}_j$, we obtain

$$\mathbf{b}^T (-\hat{\lambda}_j \mathbf{M} - \mathbf{A})^{-1} \mathbf{M} (-\hat{\lambda}_j \mathbf{M} - \mathbf{A})^{-1} \mathbf{b} = \hat{\mathbf{b}}^T (-\hat{\lambda}_j \hat{\mathbf{M}} - \hat{\mathbf{A}})^{-1} \hat{\mathbf{M}} (-\hat{\lambda}_j \hat{\mathbf{M}} - \hat{\mathbf{A}})^{-1} \hat{\mathbf{b}}. \tag{3.52}$$

Hence, these conditions are obviously an extension of the Hermite interpolation conditions for $\mathcal{H}_2$-optimality. On the other hand, we have

$$f(\mathbf{\Theta}_{err}) = \mathbf{b}_{err}^T \mathbf{X}_{err} \mathbf{c}_{err} = \mathbf{c}_{err}^T \mathbf{X}_{err}^T \mathbf{b}_{err}$$

and the same argumentation leads to

$$G(-\hat{\mu}_j) = \hat{G}(-\hat{\mu}_j), \quad G'(-\hat{\mu}_j) = \hat{G}'(-\hat{\mu}_j), \tag{3.53}$$

with

$$G(s) = \mathbf{c}^T (s\mathbf{E} - \mathbf{H})^{-1} \mathbf{c}, \quad \hat{G}(s) = \hat{\mathbf{c}}^T (s\hat{\mathbf{E}} - \hat{\mathbf{H}})^{-1} \hat{\mathbf{c}}$$

and μ_j being the eigenvalues of the matrix pencil $(\hat{\mathbf{A}}, \hat{\mathbf{M}})$. Hence, we propose Algorithm 3.3.2 for iteratively constructing a reduced set of matrices fulfilling these conditions.

Remark 3.3.6. *Due to the connection to optimal $\mathcal{H}_2$-model reduction, it should be mentioned that instead of Step 5 of Algorithm 3.3.2, one can alternatively solve two*

Algorithm 3.3.2 IRKA for symmetric Sylvester equations ((Sy)^{2}IRKA)

```
Input: Initial selection of real interpolation points $\sigma_i$ and $\mu_i$ for $i = 1, \ldots, \hat{n}$ and a convergence tolerance $tol$.
Output: $\mathbf{X}_{\hat{n}} = \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T$ fulfilling first order necessary conditions
Choose $\mathbf{V}$ and $\mathbf{W}$ s.t. $\mathcal{V} = \operatorname{span}\{(\sigma_1\mathbf{M}-\mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_{\hat{n}}\mathbf{M}-\mathbf{A})^{-1}\mathbf{b}\}$ and $\mathcal{W} = \operatorname{span}\{(\mu_1\mathbf{E}-\mathbf{H})^{-1}\mathbf{c}, \ldots, (\mu_{\hat{n}}\mathbf{E}-\mathbf{H})^{-1}\mathbf{c}\}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{W}^T\mathbf{W} = \mathbf{I}$.
while relative change in $\{\sigma_i, \mu_i\} > tol$ do
    $\hat{\mathbf{A}} = \mathbf{V}^T\mathbf{A}\mathbf{V}, \hat{\mathbf{M}} = \mathbf{V}^T\mathbf{M}\mathbf{V}, \hat{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{W}, \hat{\mathbf{H}} = \mathbf{W}^T\mathbf{H}\mathbf{W}$
    assign $\sigma_i \leftarrow -\lambda_i(\hat{\mathbf{H}}, \hat{\mathbf{E}})$ and $\mu_i \leftarrow -\lambda_i(\hat{\mathbf{A}}, \hat{\mathbf{M}})$ for $i = 1, \ldots, \hat{n}$,
    update $\mathbf{V}$ and $\mathbf{W}$ s.t. $\mathcal{V} = \operatorname{span}\{(\sigma_1\mathbf{M}-\mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_{\hat{n}}\mathbf{M}-\mathbf{A})^{-1}\mathbf{b}\}$ and $\mathcal{W} = \operatorname{span}\{(\mu_1\mathbf{E}-\mathbf{H})^{-1}\mathbf{c}, \ldots, (\mu_{\hat{n}}\mathbf{E}-\mathbf{H})^{-1}\mathbf{c}\}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{W}^T\mathbf{W} = \mathbf{I}$.
end while
$\hat{\mathbf{b}} = \mathbf{V}^T\mathbf{b}, \hat{\mathbf{c}} = \mathbf{W}^T\mathbf{c}$
Solve $\hat{\mathbf{A}}\hat{\mathbf{X}}\hat{\mathbf{E}} + \hat{\mathbf{M}}\hat{\mathbf{X}}\hat{\mathbf{H}} + \hat{\mathbf{b}}\hat{\mathbf{c}}^T$.
Set $\mathbf{X}_{\hat{n}} = \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T$.
```

reduced Sylvester equations of the form

$$\mathbf{A}\mathbf{V}\hat{\mathbf{E}} + \mathbf{M}\mathbf{V}\hat{\mathbf{H}} + \mathbf{b}\hat{\mathbf{c}}^T = \mathbf{0},$$
$$\mathbf{E}\mathbf{W}\hat{\mathbf{A}} + \mathbf{H}\mathbf{W}\hat{\mathbf{M}} + \mathbf{c}\hat{\mathbf{b}}^T = \mathbf{0}.$$

For a robust solver for these types of equations, we refer to, e.g., [25].

It remains to show that in the case of convergence of Algorithm 3.3.2, the lower bound of Corollary 3.3.1 is actually attained. For this, we assume the following splitting of the solution of (3.47) for the error system

$$\mathbf{X}_{err} = \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \hat{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{Y} \\ \mathbf{0} & \hat{\mathbf{X}} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{Z} & \hat{\mathbf{X}} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{X}} \end{bmatrix}.$$

Hence, we get

$$f(\boldsymbol{\Theta}_{err}) = f(\boldsymbol{\Theta}) - f(\hat{\boldsymbol{\Theta}}) + \mathbf{b}_{err}^T \begin{bmatrix} \mathbf{Y} \\ \hat{\mathbf{X}} \end{bmatrix} \hat{\mathbf{c}} + \hat{\mathbf{b}}^T \begin{bmatrix} \mathbf{Z} & \hat{\mathbf{X}} \end{bmatrix} \mathbf{c}_{err}.$$

A closer look at the right hand side of the previous equation reveals that

$$\mathbf{b}_{err}^T \begin{bmatrix} \mathbf{Y} \\ \hat{\mathbf{X}} \end{bmatrix} \hat{\mathbf{c}} = \mathbf{b}^T\mathbf{Y}\hat{\mathbf{c}} + \hat{\mathbf{b}}^T\hat{\mathbf{X}}\hat{\mathbf{c}},$$

where $\mathbf{Y}$ is the solution of

$$-\mathbf{A}\mathbf{Y} - \mathbf{M}\mathbf{Y}\boldsymbol{\Xi} + \mathbf{b}\hat{\mathbf{c}}^T = \mathbf{0}.$$

Here, we again assumed that the reduced matrix pencil $(\hat{\mathbf{H}}, \hat{\mathbf{E}})$ is given in its eigenvalue decomposition and that the eigenvalues are contained in the diagonal matrix $\boldsymbol{\Xi}$. As a consequence, it holds that

$$\left(\hat{\mathbf{c}}^T \otimes \mathbf{b}^T\right) \operatorname{vec}\left(\mathbf{Y}\right) = -\left(\hat{\mathbf{c}}^T \otimes \mathbf{b}^T\right) \left(-\boldsymbol{\Xi} \otimes \mathbf{M} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\hat{\mathbf{c}} \otimes \mathbf{b}\right).$$

On the other hand, we know that

$$\left(\hat{\mathbf{c}}^T \otimes \hat{\mathbf{b}}^T\right) \operatorname{vec}\left(\hat{\mathbf{X}}\right) = \left(\hat{\mathbf{c}}^T \otimes \hat{\mathbf{b}}^T\right) \left(-\boldsymbol{\Xi} \otimes \hat{\mathbf{M}} - \mathbf{I} \otimes \hat{\mathbf{A}}\right)^{-1} \left(\hat{\mathbf{c}} \otimes \hat{\mathbf{b}}\right),$$

which, together with the interpolation conditions, yields $\mathbf{b}_{err}^T \begin{bmatrix} \mathbf{Y} \\ \hat{\mathbf{X}} \end{bmatrix} \hat{\mathbf{c}} = 0$. Similarly, we can show that $\hat{\mathbf{b}}^T \begin{bmatrix} \mathbf{Z} & \hat{\mathbf{X}} \end{bmatrix} \mathbf{c}_{err} = 0$.

Analog to the proof of Theorem 3.3.1, one can eventually show that

$$\operatorname{vec}\left(\mathbf{X} - \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T\right)^T (-\mathcal{L}_S) \operatorname{vec}\left(\mathbf{X} - \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T\right) = f(\boldsymbol{\Theta}) - f(\hat{\boldsymbol{\Theta}}).$$

Altogether, we have thus proven our main result.

Theorem 3.3.4. *Let $\boldsymbol{\Theta} = (\mathbf{A}, \mathbf{E}, \mathbf{M}, \mathbf{H}, \mathbf{b}, \mathbf{c}^T)$ denote a set of matrices determining a Sylvester equation as in (3.47) with solution $\mathbf{X}$. Let further $\mathbf{X}_{\hat{n}}$ be computed by Algorithm 3.3.2 with convergence tolerance 0. Then $\mathbf{X}_{\hat{n}}$ is a local minimizer of*

$$\min_{\mathbf{X}_k \in \mathcal{M}} \{\operatorname{vec}\left(\mathbf{X} - \mathbf{X}_k\right)^T (-\mathcal{L}_S) \operatorname{vec}\left(\mathbf{X} - \mathbf{X}_k\right)\}.$$

Extension to the MIMO case

So far, we have proved the result for a right hand side of rank 1, i.e., $\mathbf{b}$ and $\mathbf{c}$ being vectors. As the extension to the 'MIMO' case is straightforward, we only sketch the necessary steps in the following. What remains to be clarified are suitable optimality conditions for the MIMO case in terms of either matrix equations or tangential interpolation conditions.

For this, let us have a look at the objective function f evaluated in the error set

$$f(\boldsymbol{\Theta}_{err}) = \operatorname{tr}\left(\mathbf{B}_{err}^T \mathbf{X}_{err} \mathbf{C}_{err}^T\right),$$

where $\mathbf{X}_{err} = \begin{bmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{Z} & \hat{\mathbf{X}} \end{bmatrix}$ is partitioned as before. As we have done for the case of the Lyapunov residual, the first step is to compute the derivate of f with respect to an arbitrary parameter γ that might be one of the entries of $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{E}, \mathbf{H}$ or $\mathbf{M}$, respectively. Accordingly, we obtain

$$\frac{\partial f}{\partial \gamma} = \operatorname{tr}\left(\frac{\partial \mathbf{X}_{err}}{\partial \gamma} \mathbf{C}_{err}^T \mathbf{B}_{err}^T\right) + \operatorname{tr}\left(\mathbf{X}_{err} \frac{\partial (\mathbf{C}_{err}^T \mathbf{B}_{err}^T)}{\partial \gamma}\right).$$

Taking into account that $\mathbf{X}_{err}$ is the solution of the generalized Sylvester equation

$$\mathbf{A}_{err} \mathbf{X}_{err} \mathbf{E}_{err} + \mathbf{M}_{err} \mathbf{X}_{err} \mathbf{H}_{err} + \mathbf{B}_{err} \mathbf{C}_{err} = \mathbf{0},$$

a careful analysis leads to

$$\begin{aligned} \frac{\partial f}{\partial \gamma} &= \operatorname{tr}\left(\mathbf{X}_{err} \frac{\partial \mathbf{E}_{err}}{\partial \gamma} \mathbf{X}_{err}^T \mathbf{A}_{err}\right) + \operatorname{tr}\left(\mathbf{X}_{err} \mathbf{E}_{err} \mathbf{X}_{err}^T \frac{\partial \mathbf{A}_{err}}{\partial \gamma}\right) \\ &+ \operatorname{tr}\left(\mathbf{X}_{err} \frac{\partial \mathbf{H}_{err}}{\partial \gamma} \mathbf{X}_{err}^T \mathbf{M}_{err}\right) + \operatorname{tr}\left(\mathbf{X}_{err} \mathbf{H}_{err} \mathbf{X}_{err}^T \frac{\partial \mathbf{M}_{err}}{\partial \gamma}\right) \\ &+ 2 \operatorname{tr}\left(\mathbf{X}_{err} \frac{\partial (\mathbf{C}_{err}^T \mathbf{B}_{err}^T)}{\partial \gamma}\right). \end{aligned}$$

Depending on the specific choice of γ, we can derive different optimality conditions. For example, setting $\gamma = \hat{\mathbf{E}}_{i,j}$, leads to the condition

$$-\mathbf{Y}^T \mathbf{A} \mathbf{Y} + \hat{\mathbf{X}}^T \hat{\mathbf{A}} \hat{\mathbf{X}} = \mathbf{0}. \tag{3.54a}$$

Similarly, for the derivatives with respect to $\hat{\mathbf{A}}_{i,j}, \hat{\mathbf{H}}_{i,j}, \hat{\mathbf{M}}_{i,j}, \hat{\mathbf{B}}_{i,j}$ and $\hat{\mathbf{C}}_{i,j}$, we get

$$-\mathbf{Z}\mathbf{E}\mathbf{Z}^T + \hat{\mathbf{X}}\hat{\mathbf{E}}\hat{\mathbf{X}}^T = \mathbf{0}, \tag{3.54b}$$

$$-\mathbf{Y}^T\mathbf{M}\mathbf{Y} + \hat{\mathbf{X}}^T\hat{\mathbf{M}}\hat{\mathbf{X}} = \mathbf{0}, \tag{3.54c}$$

$$-\mathbf{Z}\mathbf{H}\mathbf{Z}^T + \hat{\mathbf{X}}\hat{\mathbf{H}}\hat{\mathbf{X}}^T = \mathbf{0}, \tag{3.54d}$$

$$\mathbf{Y}^T\mathbf{B} + \hat{\mathbf{X}}^T\hat{\mathbf{B}} = \mathbf{0}, \tag{3.54e}$$

$$\mathbf{Z}\mathbf{C}^T + \hat{\mathbf{X}}\hat{\mathbf{C}}^T = \mathbf{0}. \tag{3.54f}$$

We already know that there is an equivalence between Sylvester equations and the concept of tangential interpolation, see again [62]. Hence, it is not surprising that the above matrix equation based conditions can alternatively be replaced by demanding that a reduced-order transfer function matrix tangentially interpolates the original transfer function matrix at given interpolations points. To be precise, let

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{H})^{-1}\mathbf{C}^T \in \mathbb{R}(s)^{m\times m} \quad \text{and} \quad \mathbf{F}(s) = \mathbf{B}^T(s\mathbf{M} - \mathbf{A})^{-1}\mathbf{B} \in \mathbb{R}(s)^{m\times m}.$$

Moreover, assume that $(\mathbf{Q}, \mathbf{\Lambda})$ and $(\mathbf{R}, \mathbf{\Xi})$ are the eigenvalue decompositions of the matrix pencils $(\hat{\mathbf{H}}, \hat{\mathbf{E}})$ and $(\hat{\mathbf{A}}, \hat{\mathbf{M}})$, respectively. Here, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_{\hat{n}})$ and $\mathbf{\Xi} = \operatorname{diag}(\mu_1, \dots, \mu_{\hat{n}})$ contain the eigenvalues while $\mathbf{Q}$ and $\mathbf{R}$ consist of a set of $\hat{\mathbf{E}}$ and $\hat{\mathbf{M}}$-orthogonal eigenvectors. A locally optimal reduced set $\hat{\mathbf{\Theta}}$ of matrices now has to fulfill

$$\mathbf{G}(-\mu_j)\tilde{\mathbf{b}}_j = \hat{\mathbf{G}}(-\mu_j)\tilde{\mathbf{b}}_j, \tag{3.55a}$$

$$\tilde{\mathbf{b}}_j^T\mathbf{G}(-\mu_j) = \tilde{\mathbf{b}}_j^T\hat{\mathbf{G}}(-\mu_j), \tag{3.55b}$$

$$\tilde{\mathbf{b}}_j^T\mathbf{G}'(-\mu_j)\tilde{\mathbf{b}}_j = \tilde{\mathbf{b}}_j^T\hat{\mathbf{G}}'(-\mu_j)\tilde{\mathbf{b}}_j, \tag{3.55c}$$

$$\mathbf{F}(-\lambda_j)\tilde{\mathbf{c}}_j = \hat{\mathbf{F}}(-\lambda_j)\tilde{\mathbf{c}}_j, \tag{3.55d}$$

$$\tilde{\mathbf{c}}_j^T\mathbf{F}(-\lambda_j) = \tilde{\mathbf{c}}_j^T\hat{\mathbf{F}}(-\lambda_j), \tag{3.55e}$$

$$\tilde{\mathbf{c}}_j^T\mathbf{F}'(-\lambda_j)\tilde{\mathbf{c}}_j = \tilde{\mathbf{c}}_j^T\hat{\mathbf{F}}'(-\lambda_j)\tilde{\mathbf{c}}_j, \tag{3.55f}$$

with $\tilde{\mathbf{B}} = \hat{\mathbf{B}}^T\mathbf{R}$ and $\tilde{\mathbf{C}} = \hat{\mathbf{C}}\mathbf{Q}$ denoting tangential directions. For the sake of completeness, in Algorithm 3.3.3 we now see a matrix version that upon convergence yields a local minimizer for the MIMO case. Consequently, we have the following result extending Theorem 3.3.4.

Corollary 3.3.2. *Let $\mathbf{\Theta} = (\mathbf{A}, \mathbf{E}, \mathbf{M}, \mathbf{H}, \mathbf{B}, \mathbf{C})$ denote a set of matrices determining a Sylvester equation as in (3.47) with solution $\mathbf{X}$. Let further $\mathbf{X}_{\hat{n}}$ be computed by Algorithm*

3.3.3 with convergence tolerance 0. Then $\mathbf{X}_{\hat{n}}$ is a local minimizer of

$$\min_{\mathbf{X}_k \in \mathcal{M}} \{(\mathrm{vec}\,(\mathbf{X}-\mathbf{X}_k))^T(-\mathcal{L}_S)\,\mathrm{vec}\,(\mathbf{X}-\mathbf{X}_k)\}.$$

Algorithm 3.3.3 IRKA for MIMO symmetric Sylvester equations

```
Input: $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{E},\mathbf{H},\mathbf{M},\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}},\hat{\mathbf{E}},\hat{\mathbf{H}},\hat{\mathbf{M}}$ as in (3.50)
Output: $\mathbf{X}_{\hat{n}} = \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T$ fulfilling first order necessary conditions
while (not converged) do
    Solve $\mathbf{A}\mathbf{V}\hat{\mathbf{E}} + \mathbf{M}\mathbf{V}\hat{\mathbf{H}} + \mathbf{B}\hat{\mathbf{C}} = \mathbf{0}$.
    Solve $\mathbf{E}\mathbf{W}\hat{\mathbf{A}} + \mathbf{H}\mathbf{W}\hat{\mathbf{M}} + \mathbf{C}^T\hat{\mathbf{B}}^T = \mathbf{0}$.
    $\mathbf{V} = \mathrm{orth}\,(\mathbf{X})\,, \mathbf{W} = \mathrm{orth}\,(\mathbf{W})$
    $\hat{\mathbf{A}} = \mathbf{V}^T\mathbf{A}\mathbf{V}, \hat{\mathbf{M}} = \mathbf{V}^T\mathbf{M}\mathbf{V}, \hat{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{W}, \hat{\mathbf{H}} = \mathbf{W}^T\mathbf{H}\mathbf{W}, \hat{\mathbf{B}} = \mathbf{V}^T\mathbf{B}, \hat{\mathbf{C}} = \mathbf{C}\mathbf{W}$
end while
Solve $\hat{\mathbf{A}}\hat{\mathbf{X}}\hat{\mathbf{E}} + \hat{\mathbf{M}}\hat{\mathbf{X}}\hat{\mathbf{H}} + \hat{\mathbf{B}}\hat{\mathbf{C}} = \mathbf{0}$.
Set $\mathbf{X}_{\hat{n}} = \mathbf{V}\hat{\mathbf{X}}\mathbf{W}^T$.
```

Concluding remarks

In summary, we have theoretically shown several relations between the concept of rational interpolation and the approximate solution of large-scale matrix equations. In particular, we have seen that for a symmetric dynamical system, constructing a locally $\mathcal{H}_2$-optimal reduced-order model is equivalent to the minimization of the error of the associated system Gramians with respect to to the energy norm naturally induced by the underlying Lyapunov operator. Moreover, for the unsymmetric case, we discussed an abstract and more general matrix equation approach which is further investigated in the following chapter. However, we do not claim that the previous algorithms can compete with existing Lyapunov equation solvers like, e.g., the ADI iteration and KPIK when it comes to computational efficiency. The major drawback of our methods clearly is that they heavily depend on the speed of convergence. Still, for a comprehensive understanding of the theory, it is interesting to note the close relation between the different concepts.

3.4 Numerical examples

In this section, we study the performance of the proposed algorithms by means of some standard numerical test examples. We stop the algorithms whenever the relative change of the eigenvalues of the reduced system matrix fall below 10^{-5}. All simulations were generated on an Intel® Dual-Core CPU E5400, 2 MB cache, 3 GB RAM, Ubuntu Linux 10.04 (i686), MATLAB Version 7.11.0 (R2012a) 32-bit (glnxa86). Since in the following

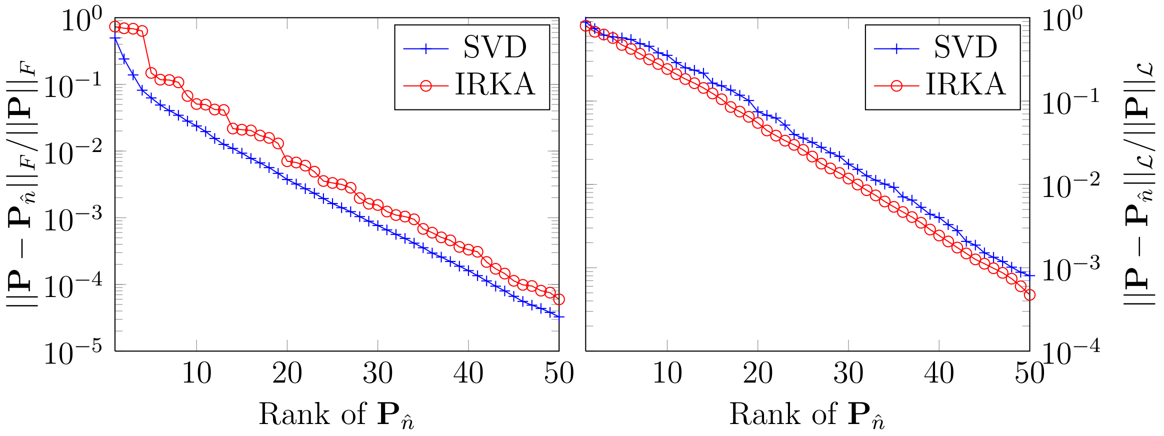


Figure 3.1: Steel profile with $n = 1357$.

we always compare the constructed low rank approximations with the singular value decomposition of the true solution, the dimensions of the considered matrix equations are only medium-sized such that a fast and reliable computation of the exact solution can actually be obtained by the *lyap* function from the MATLAB Control System Toolbox. At this point, keep in mind that this comparison allows to point out the actual difference between our approximations and the best rank-$\hat{n}$ approximation which is given by the SVD.

Energy norm for the Lyapunov equation

The first example is a semi-discretized heat transfer problem from the Oberwolfach benchmark collection[2]. Here, we use the coarsest discretization leading to symmetric matrices $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{1357 \times 1357}$ together with the input matrix $\mathbf{B} \in \mathbb{R}^{1357 \times 7}$, see [30]. In Figure 3.1, we present a comparison between the SVD-based approximation of the exact solution with the approximation given by the rational Krylov subspace obtained by IRKA. As expected from Theorem 3.3.1, the relative error in the Frobenius norm is better when the SVD approximation is used. However, the slope of the IRKA approximation is almost parallel to that and for an approximation of rank 50, the relative error ($\approx 8 \cdot 10^{-5}$) is almost as good as the best approximation ($\approx 5 \cdot 10^{-5}$) given by the SVD. On the other hand, we see that IRKA outperforms the SVD for every rank $\hat{n}$ when the relative error is measured in terms of the $\mathcal{L}$-norm. Although we have discussed a possible way of minimizing the residual, for the symmetric case we do not include a comparison here. As it has already been obtained in [125], the residual norm is a worse estimator of the true error than the energy norm. Moreover, the corresponding minimizers may be suboptimal and, thus, for symmetric state space systems it does not seem to make sense to minimize the residual at all. Note that the phenomenon of IRKA producing nearly optimal low

[2] http://portal.uni-freiburg.de/imteksimulation/downloads/benchmark

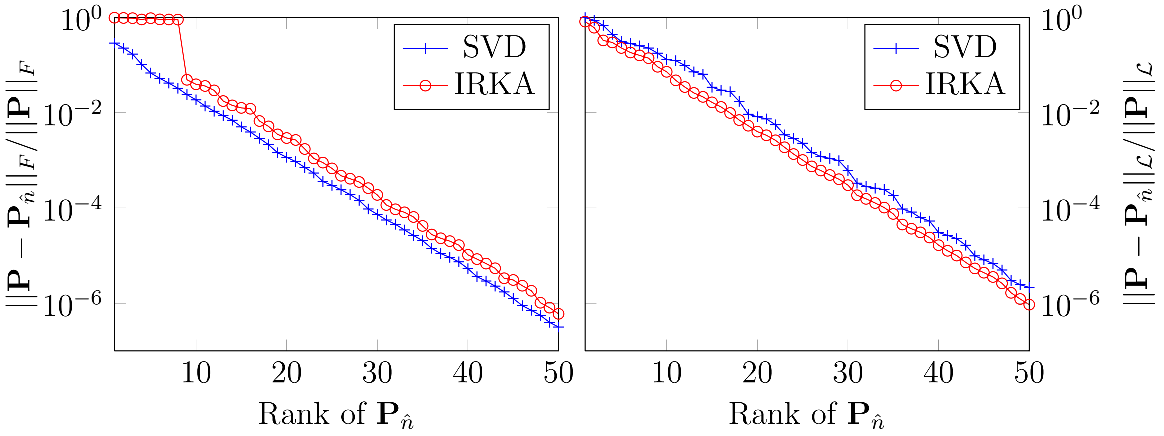


Figure 3.2: Tunable optical filter with $n = 1668$.

rank approximations has already been numerically observed and investigated in [45, 59].

The second example is also quite common in the context of model order reduction. The symmetric system matrices $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{1668\times1668}$ and $\mathbf{B} \in \mathbb{R}^{1668\times5}$ stem from the finite element discretization of a thermal model of a filter device and thus are sparse, see [79]. Similar to the previous example, from Figure 3.2 we can again conclude that the SVD approximation dominates the performance with respect to the Frobenius norm while the IRKA approach performs better when the $\mathcal{L}$-norm is taken as a basis for judgment. Moreover, starting from values $\hat{n} = 10$, the error resulting from IRKA approximations follows the slope of the error of the SVD-based approximation showing that the subspaces seem to be very close to the optimal ones. Interestingly enough, for dimensions $\hat{n} < 10$, the energy norm does not seem to be a reasonable error estimator. In particular, although the IRKA approximations outperform the SVD with respect to the energy norm, the error almost stagnates with respect to the Frobenius norm. However, since the norms are not equivalent, we cannot expect an exact one-to-one correspondence between the errors.

Residual norm for the Lyapunov equation

Next, we consider examples that result in unsymmetric dynamical systems with complex system poles that do not allow for a definition of an energy norm. Nevertheless, as we have seen, we can still try to locally minimize the residual $\mathbf{R} = \mathbf{A}\mathbf{P}_{\hat{n}} + \mathbf{P}_{\hat{n}}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T$ by the iteration specified in Algorithm 3.3.1. The first example was introduced in [110] and is one of the SLICOT benchmarks[2]. The transfer function exhibits three peaks corresponding to six complex system poles while the rest of the poles is completely real. In Figure 3.3, we compare the results given by the SVD of the true solution with the

[2] http://www.slicot.org/index.php?site=benchmodred

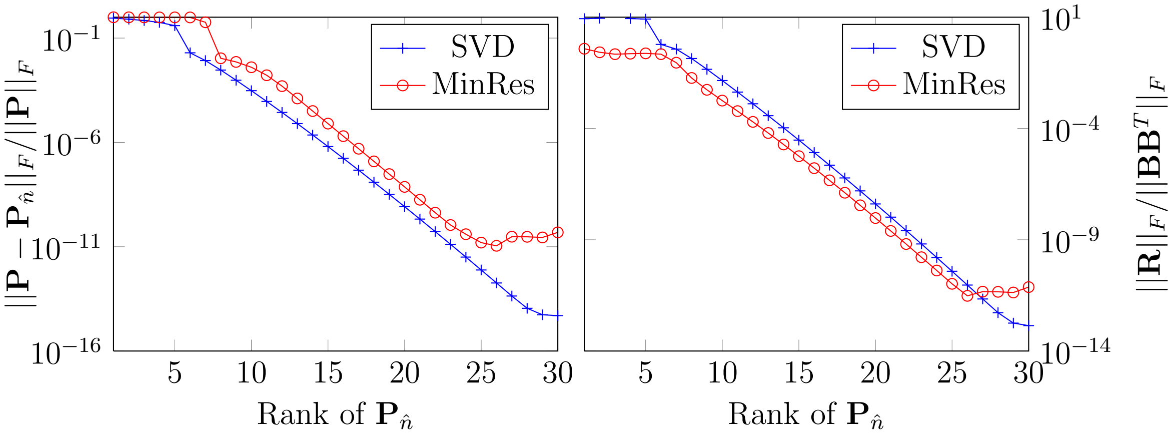


Figure 3.3: Fom model with $n = 1006$.

low rank approximations obtained by Algorithm 3.3.1 and abbreviated with *MinRes*. Similar to the symmetric case, we see that the latter approach follows the error slope resulting from the SVD. However, there seems to occur some stagnation for larger ranks $\hat{n}$. On the other hand, we see that our method indeed yields residuals that are smaller than those we get from the SVD. Again, for larger ranks, the approximations seem to be suboptimal and are outperformed by the SVD. Recall Remark 3.3.4 and Remark 3.3.5. Since we are essentially working with the normal equations, the condition number of the problem is squared and we thus cannot expect the results to be as accurate as in the symmetric case.

The second example is the CD player model we have discussed in Chapter 1. The system matrices $\mathbf{A}$ and $\mathbf{B}$ are part of the SLICOT benchmark collection and are of dimension $n = 120$ with $m = 2$ inputs. Again, the matrix $\mathbf{A}$ is unsymmetric and exhibits a complex spectrum. In Figure 3.4, we compute the results for low rank approximations varying from $\hat{n} = 1, \ldots, 20$. While for ranks up to $\hat{n} = 15$, we do not observe problems with convergence of Algorithm 3.3.1, the convergence criterion is not fulfilled for larger approximations $\mathbf{P}_{\hat{n}}$. However, we still obtain the same results as in the previous case. While the new method performs worse than the SVD when the Frobenius norm is used, we obtain smaller residuals, even if the algorithm does not converge at all. Nevertheless, note that for $\hat{n} = 8$ and $\hat{n} = 10$, the residuals are larger than those from the SVD. This might indicate that minimizing the residual is not as robust as a minimization of the energy norm as has already been pointed out in [125].

Energy norm for the Sylvester equation

Let us now briefly draw our attention to the more general case specified in (3.47). Analog to the Lyapunov case, we consider an example given by the process of optimal cooling

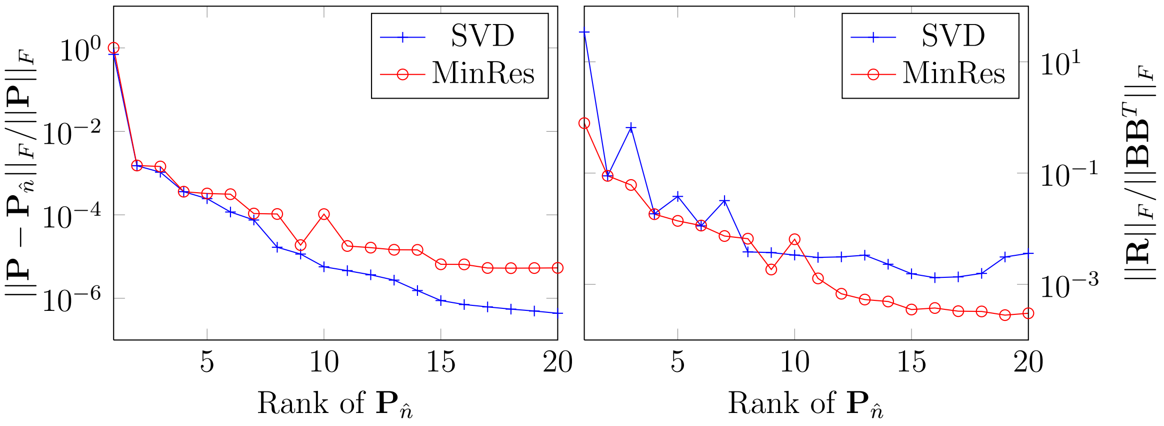


Figure 3.4: CD player with $n = 120$.

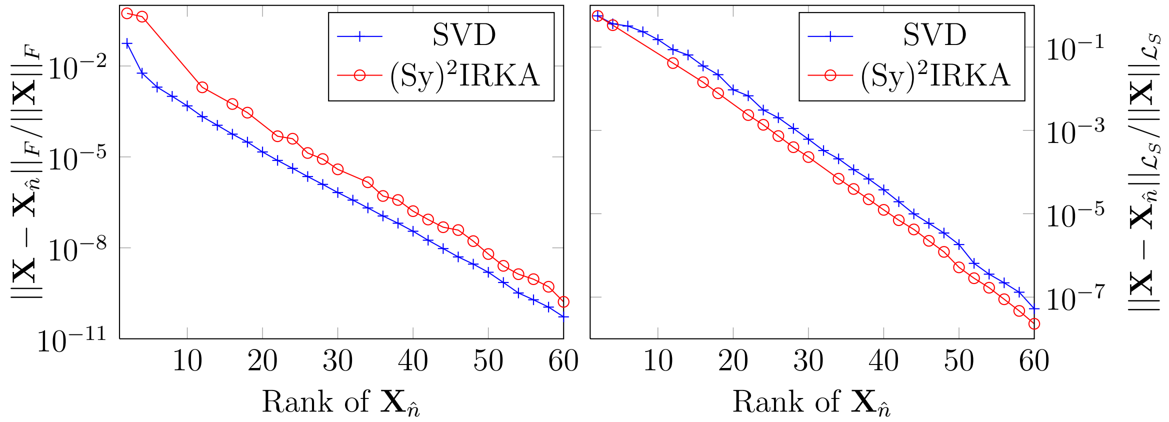


Figure 3.5: Steel profiles with discretizations $n = 5177$ and $n = 1357$.

of steel profiles. In order to end up with a generalized Sylvester equation including different matrix dimensions, we use a matrix set $(\mathbf{A}, \mathbf{E}, \mathbf{M}, \mathbf{H}, \mathbf{B}, \mathbf{C})$, where $\mathbf{A}, \mathbf{M}, \mathbf{B}$ are as specified for the Lyapunov case, while $\mathbf{E}, \mathbf{H}, \mathbf{C}$ are obtained by a finer resolution with mesh size $m = 5177$. In Figure 3.5, we see a comparison between the rational Krylov subspace approximation computed by Algorithm 3.3.2 which is abbreviated with $(\text{Sy})^2$IRKA and the SVD-based approximation. Due to the lack of an exact solver for the original Sylvester equation, we use our new method with an approximation of rank 250 for reference values. It should be mentioned that the relative residual for this approach is smaller than 10^{-13} and thus should be sufficient for comparison. Again, we see that $(\text{Sy})^2$IRKA is dominated by the SVD approximation if the quality is measured in terms of the Frobenius norm while it performs constantly better if we use the $\mathcal{L}_S$-norm.

3.5 Conclusions

In this chapter, we investigated the approximate solution of large-scale matrix equations. For a given prespecified rank $\hat{n}$, the goal was to construct approximations that are optimal with respect to different norms. For symmetric state space systems, we showed that IRKA yields subspaces that can be used to construct low rank approximations that are optimal with respect to the energy norm induced by the symmetric positive definite Lyapunov operator. So far, this could only be done by means of a Riemannian optimization method proposed in [125]. Although the latter method is guaranteed to globally converge to a local minimizer, IRKA has the advantage of being easily implementable without a deeper knowledge of interpolation-based MOR theory. We further established a first connection between the Frobenius norm of low rank approximations and the $\mathcal{H}_2$-norm of the associated error system. Moreover, for unsymmetric systems, we derived optimality conditions that allow to minimize the Lyapunov residual for a given rank $\hat{n}$. Again, there is, on the one hand, a connection to the concept of Riemannian optimization, and, on the other hand, a relation to the bilinear $\mathcal{H}_2$-optimal model reduction problem. Consequently, we came up with an iterative algorithm that, upon convergence, fulfills the necessary conditions, or equivalently, minimizes the Lyapunov residual. Finally, we extended the ideas to the more general case of Sylvester equations and slightly modified IRKA in order to ensure interpolation-based optimality conditions similar to the ones obtained in [73, 99]. By means of different standard numerical test examples, we underscored our theoretical results and demonstrated that the iterative algorithms indeed yield very accurate low rank approximations that are optimal with respect to different norms. Nevertheless, since all algorithms depend on the speed of convergence, for typical real-life applications, they cannot compete with standard low rank solvers such as, e.g., KPIK and the ADI iteration. Still, they provide a useful insight into the topic of low rank approximations and might allow for constructing a hybrid method, combining the best features of all of them.

CHAPTER 4

BILINEAR SYSTEMS

Contents

4.1 Introduction

In this chapter, we consider a more general class of dynamical systems. So far, we have always assumed the system to be jointly linear in the state and in the control, leading to LTI systems of the form (2.2). However, in many typical real-life applications this assumption does not hold true, resulting in a need for analyzing nonlinear dynamical systems. As a first step into this direction, the class of *bilinear systems* has been pointed out to be an interesting interface between fully nonlinear and linear control systems, see [34, 100, 101, 102, 115]. More precisely, from now on we consider systems of the form

$$\mathbf{\Sigma}_B : \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \sum_{k=1}^{m} \mathbf{N}_k \mathbf{x}(t) u_k(t) + \mathbf{B}\mathbf{u}(t), \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \end{cases} \tag{4.1}$$

with $\mathbf{A}, \mathbf{N}_k \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{p\times n}$, $\mathbf{D} \in \mathbb{R}^{p\times m}$, $\mathbf{u}(t) \in \mathbb{R}^m$ and $\mathbf{y}(t) \in \mathbb{R}^p$. Similar to the discussion on linear systems, for the remainder of this thesis, we assume that we have a zero initial condition $\mathbf{x}_0 = \mathbf{0}$ and a vanishing feedthrough term $\mathbf{D}$. Again, if this does not hold true, one can easily embed all our results into the above setting by incorporating $\mathbf{x}_0$ in an enlarged input vector of the form $\begin{bmatrix} \mathbf{B} & \mathbf{x}_0 \end{bmatrix}$.

As it is discussed in [34, 100, 101, 102], a variety of biological, physical and economical phenomena naturally result in bilinear models. Here, models for nuclear fusion, mechanical brakes or biological species can be mentioned as typical examples. As it might be obvious, the above systems are called *bilinear* due to the fact that for a fixed state vector $\mathbf{x}$ one obtains linearity in the control $\mathbf{u}$ and vice versa. Although bilinear systems formally belong to the class of nonlinear control systems, they clearly are a special case of those. In the following, this turns out to be very useful in order to generalize several successful linear model reduction concepts.

Following [24, 42, 43], a completely similar structure is obtained for *Itô-type linear stochastic differential equations* of the form

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= \mathbf{A}\mathbf{x}(t)\mathrm{d}t + \sum_{i=1}^{N} \mathbf{A}_0^i \mathbf{x}(t)\mathrm{d}\omega_i(t) + \mathbf{B}\mathbf{u}(t)\mathrm{d}t, \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \end{aligned} \tag{4.2}$$

where $\mathbf{A}, \mathbf{A}_i \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{p\times n}$ and $\mathrm{d}\omega_i(t)$ are *white noise* processes associated with Wiener processes $\omega_i(t)$. Formally, the above is to be understood as a notation for

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{A}\mathbf{x}(\tau)\mathrm{d}\tau + \sum_{i=1}^{N} \int_0^t \mathbf{A}_0^i \mathbf{x}(\tau)\mathrm{d}\omega_i + \int_0^t \mathbf{B}\mathbf{u}(\tau)\mathrm{d}\tau,$$

with $\mathrm{d}\omega_i$ denoting the Itô integral. Later on, we use some interesting applications like, e.g., the Fokker-Planck equation from Chapter 1, as test examples for our model reduction techniques.

Coming back to the actual MOR problem, we want to construct another bilinear system

$$\hat{\mathbf{\Sigma}}_B : \begin{cases} \dot{\hat{\mathbf{x}}}(t) = \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \sum_{k=1}^{m} \hat{\mathbf{N}}_k \hat{\mathbf{x}}(t) u_k(t) + \hat{\mathbf{B}}\mathbf{u}(t), \\ \hat{\mathbf{y}}(t) = \hat{\mathbf{C}}\hat{\mathbf{x}}(t), \quad \hat{\mathbf{x}}(0) = \mathbf{0}, \end{cases} \tag{4.3}$$

with $\hat{\mathbf{A}}, \hat{\mathbf{N}}_k \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $\hat{\mathbf{B}} \in \mathbb{R}^{\hat{n}\times m}$, $\hat{\mathbf{C}} \in \mathbb{R}^{p\times\hat{n}}$. Analog to model order reduction of linear systems, we can basically distinguish between SVD-based approaches leading to a reasonable generalization of the method of balanced truncation, see [24, 134], and interpolation-based ideas that approximate generalized transfer functions by projecting the original model onto appropriate Krylov subspaces, see [9, 10, 33, 40, 53, 111, 112].

The structure of the chapter is as follows. In the subsequent section, we provide a detailed review on concepts and theory of bilinear control systems in general. This includes an explicit solution formula, possible stability criteria, frequency domain characterizations and an extension of the already discussed system Gramians to bilinear systems. We then derive a generalized interpolation-based approach towards $\mathcal{H}_2$-optimality. We discuss two equivalent algorithms and study their performance by means of several numerical examples. Subsequently, we turn our attention to the method of balanced truncation for bilinear systems. Here, we investigate possible approximations of the required bilinear system Gramians. Besides a theoretical explanation of the fast singular value decay for those Gramians, we propose different low rank solvers that generalize well-known ideas from the linear case. Again, we show the performance of the proposed methods by means of several large-scale bilinear test examples.

4.2 Control theoretic concepts

Here, we collect some important results and ideas used in the area of bilinear control theory. Most of the statements can be found in any standard text book on bilinear systems like, e.g., [48, 82, 100, 115].

Probably the first and most important question in the study of a dynamical system is concerned with existence and uniqueness of a solution $\mathbf{x}(t)$. For bilinear control systems such as (4.1), one basically needs the same assumptions as for the linear case. Hence, let us consider a finite time interval $I = [0, T]$ and a bounded and continuous input signal $u(t)$ on I. For a SISO bilinear control system, one can show, see [115], that $\mathbf{x}(t)$ exists on I and can be computed as

$$\mathbf{x}(t) = \sum_{i=1}^{\infty} \int_0^t \int_0^{\sigma_1} \dots \int_0^{\sigma_{i-1}} \mathbf{g}_i(t, \sigma_1, \dots, \sigma_{i-1})\ u(\sigma_i) \cdots u(\sigma_1)\ \mathrm{d}\sigma_i \cdots \mathrm{d}\sigma_1, \tag{4.4}$$

where $\mathbf{g}_i(t, \sigma_1, \dots, \sigma_{i-1})$ is given as

$$\mathbf{g}_i(t, \sigma_1, \dots, \sigma_{i-1}) = \underbrace{e^{\mathbf{A}(t-\sigma_1)}\mathbf{N} \cdots e^{\mathbf{A}(\sigma_{i-2}-\sigma_{i-1})}\mathbf{N}}_{i-1 \text{ times}} e^{\mathbf{A}(\sigma_{i-1}-\sigma_i)}\mathbf{b}. \tag{4.5}$$

As it is common for nonlinear systems, this can be shown by successive approximations, each of them based on Picard-Lindelöf's theorem, see [115, Section 3.1]. As a result, instead of the bilinear system (4.1), we can now consider the following infinite series of coupled linear systems

$$\begin{aligned} \dot{\mathbf{x}}_1(t) &= \mathbf{A}\mathbf{x}_1(t) + \mathbf{b}u(t), \\ \dot{\mathbf{x}}_i(t) &= \mathbf{A}\mathbf{x}_i(t) + \mathbf{N}\mathbf{x}_{i-1}(t)u(t) + \mathbf{b}u(t), \quad i = 2, 3, \dots, \end{aligned} \tag{4.6}$$

which satisfies $\mathbf{x}_\infty(t) := \lim_{i\to\infty} \mathbf{x}_i(t) = \mathbf{x}(t)$. This interpretation of a nonlinear system as a series of coupled linear systems is important later on when we focus on more general nonlinear control systems. The expression (4.4) is a so-called *Volterra series* and is well-known in the context of nonlinear systems, see [48, 82, 115].

For our purposes, an explicit input-output representation is of particular interest. If we perform a change of variables, it is easily seen that for a linear output equation as in (4.1), we obtain that

$$\mathbf{y}(t) = \sum_{i=1}^{\infty} \int_0^t \int_0^{t_1} \dots \int_0^{t_{i-1}} \mathbf{h}_i(t_1, \dots, t_i) \prod_{j=1}^{i} u\left(t - \sum_{\ell=1}^{i-j+1} t_\ell\right)\ \mathrm{d}t_i \cdots \mathrm{d}t_1, \tag{4.7}$$

with *regular degree-i kernels* $\mathbf{h}_i(t_1, \dots, t_i)$ of the form

$$\mathbf{h}_i(t_1, \dots, t_i) = \mathbf{c}^T \underbrace{e^{\mathbf{A}t_i}\mathbf{N} \cdots e^{\mathbf{A}t_2}\mathbf{N}}_{i-1 \text{ times}} e^{\mathbf{A}t_1}\mathbf{b}. \tag{4.8}$$

In particular, note that the degree-1 kernel coincides with the impulse response of a linear system $\mathbf{h}_1(t) = \mathbf{c}^T e^{\mathbf{A}t}\mathbf{b}$. This is not too surprising after the given interpretation (4.6) of a bilinear system as a series of infinitely many coupled linear systems. Recall that in the linear case we have a one-to-one correspondence between the impulse response and the transfer function of a linear system. In a similar fashion, we can make use of a multivariate Laplace transform, see [115, Section 2.1], for each of the degree-i kernels in order to arrive at the i-th transfer function of a bilinear system. We thus obtain

$$H_i(s_1,\ldots,s_i) = \mathbf{c}^T \underbrace{(s_i\mathbf{I}-\mathbf{A})^{-1}\mathbf{N}\cdots(s_2\mathbf{I}-\mathbf{A})^{-1}\mathbf{N}}_{i-1 \text{ times}}(s_1\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}. \tag{4.9}$$

Although a clear interpretation of the frequency variables $s_1,\ldots,s_i$ and the meaning of a multivariate transfer function is very ambiguous, it is important to note that there is an abstract way of characterizing the output of a bilinear system in the frequency domain by means of a mapping that extends the transfer function of linear systems.

Let us come back to the MIMO case. Though the formulas and concepts become rather technical, everything we have seen so far still holds true. Due to the significance for this thesis, we only state the corresponding frequency-domain generalization. For a more detailed background on this topic, besides the textbooks mentioned in the beginning, we also refer to [32]. The i-th transfer function of a bilinear system of the form (4.1) is determined by

$$\mathbf{H}_i(s_1,\ldots,s_i) = \mathbf{C}\left(\prod_{j=0}^{i-2}\mathbf{I}_{m^j}\otimes(s_{i-j}\mathbf{I}-\mathbf{A})^{-1}\mathcal{N}\right)\left(\mathbf{I}_{m^{i-1}}\otimes(s_1\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\right)$$

with $\mathcal{N} = \left[\mathbf{N}_1,\ldots,\mathbf{N}_m\right]$.

In contrast to the asymptotic stability of a linear system, here we consider bounded-input-bounded-output (BIBO) stability. The following important statement can be found in [121].

Theorem 4.2.1. *Let a bilinear system $\mathbf{\Sigma}_B$ be given and assume that $\mathbf{A}$ is asymptotically stable, i.e., there exist real scalars $\beta > 0$ and $0 < \alpha \leq -\max_i(\mathrm{Re}\,(\lambda_i(\mathbf{A})))$, such that*

$$||e^{\mathbf{A}t}|| \leq \beta e^{-\alpha t}, \quad t \geq 0.$$

Further assume that $||\mathbf{u}(t)|| = \sqrt{\sum_{k=1}^m |u_k(t)|^2} \leq M$ holds uniformly on $[0,\infty[$ with $M > 0$, and set $\Gamma = \sum_{k=1}^m ||\mathbf{N}_k||$. Then, $\mathbf{\Sigma}_B$ is BIBO stable, i.e., the corresponding Volterra series of the solution $\mathbf{x}(t)$ uniformly converges on the interval $[0,\infty[$, if $\Gamma < \frac{\alpha}{M\beta}$.

Next, we have to introduce certain system Gramians that arise for bilinear control systems. First discussed in [41], the concepts of reachability and observability can be generalized as follows. According to, e.g., [1], let

$$\begin{aligned}\mathbf{P}_1(t_1) &= e^{\mathbf{A}t_1}\mathbf{B},\\ \mathbf{P}_i(t_1,\ldots,t_i) &= e^{\mathbf{A}t_i}\left[\mathbf{N}_1\mathbf{P}_{i-1},\ldots,\mathbf{N}_m\mathbf{P}_{i-1}\right], \quad i=2,3,\ldots\end{aligned}$$

and define the reachability Gramian $\mathbf{P}$ as

$$\mathbf{P} = \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \mathbf{P}_i \mathbf{P}_i^T \mathrm{d}t_1 \cdots \mathrm{d}t_i. \tag{4.10}$$

Analogously, we set

$$\begin{aligned} \mathbf{Q}_1(t_1) &= e^{\mathbf{A}^T t_1} \mathbf{C}^T, \\ \mathbf{Q}_i(t_1, \ldots, t_i) &= e^{\mathbf{A}^T t_i} \left[\mathbf{N}_1^T \mathbf{Q}_{i-1}^T, \ldots, \mathbf{N}_m^T \mathbf{Q}_{i-1}^T \right], \quad i = 2, 3, \ldots \end{aligned}$$

and define the observability Gramian $\mathbf{Q}$ as

$$\mathbf{Q} = \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \mathbf{Q}_i \mathbf{Q}_i^T \mathrm{d}t_1 \cdots \mathrm{d}t_i. \tag{4.11}$$

If $(\mathbf{A}, \mathbf{B})$ is reachable and $(\mathbf{A}, \mathbf{C})$ is controllable, all the $\mathbf{P}_i$ and $\mathbf{Q}_i$ are symmetric and positive semi-definite. Still, the series may diverge and $\mathbf{P}$ and $\mathbf{Q}$ will not exist. However, under certain assumptions this is the case and $\mathbf{P}$ and $\mathbf{Q}$ additionally satisfy linear matrix equations that extend the known Lyapunov equations for linear systems in a suitable way. For this, we give the following result from [1, 133].

Theorem 4.2.2. *The reachability Gramian* $\mathbf{P}$ *and the observability Gramian* $\mathbf{Q}$ *as given in (4.10) and (4.11) exist if*

i) $\mathbf{A}$ *is stable,*

ii) $\Gamma_1 < \frac{\sqrt{2\alpha}}{\beta}$, *where* $\Gamma_1 = \sqrt{|| \sum_{k=1}^m \mathbf{N}_k \mathbf{N}_k^T ||}$ *and* α, β *are as in Theorem 4.2.1.*

From now on, unless stated otherwise, we always assume that the system $\mathbf{\Sigma}_B$ under consideration is BIBO stable according to the meaning of Theorem 4.2.1. The above Theorem further yields the following result.

Theorem 4.2.3. *([1, 133]) Suppose that* $\mathbf{A}$ *is stable and the reachability Gramian* $\mathbf{P}$ *and the observability Gramian* $\mathbf{Q}$ *exist. Then* $\mathbf{P}$ *and* $\mathbf{Q}$ *satisfy the generalized Lyapunov equations*

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \sum_{k=1}^{m} \mathbf{N}_k \mathbf{P} \mathbf{N}_k^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}, \tag{4.12a}$$

$$\mathbf{A}^T\mathbf{Q} + \mathbf{Q}\mathbf{A} + \sum_{k=1}^{m} \mathbf{N}_k^T \mathbf{Q} \mathbf{N}_k + \mathbf{C}^T\mathbf{C} = \mathbf{0}. \tag{4.12b}$$

There exist different characterizations for $\mathbf{P}$ and $\mathbf{Q}$ to be unique and positive definite. However, at this point we do not deepen this topic and just refer to [24, 42, 43] where a more detailed discussion is presented.

Finally, in order to judge the quality of a reduced-order model, we need a system norm that allows to measure the distance from the reduced output to the original output, i.e., $\mathbf{y}(t) - \hat{\mathbf{y}}(t)$. Since in the subsequent sections our goal is to adapt and extend the interpolation-based results from [73], let us review a generalization of the $\mathcal{H}_2$-norm that first was mentioned in [133].

Definition 4.2.1. *Let $\mathbf{\Sigma}_B$ be a BIBO stable bilinear systems. Then we define the $\mathcal{H}_2$-norm as*

$$||\mathbf{\Sigma}_B||_{\mathcal{H}_2} = \sqrt{\operatorname{tr}\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}\sum_{\ell_1,\ldots,\ell_i=1}^{m}\mathbf{g}_i^{(\ell_1,\ldots,\ell_i)}(\mathbf{g}_i^{(\ell_1,\ldots,\ell_i)})^T\mathrm{d}s_1\cdots\mathrm{d}s_k\right)},$$

with $\mathbf{g}_i^{(\ell_1,\ldots,\ell_i)}(s_1,\ldots,s_i) = \mathbf{C}e^{\mathbf{A}s_k}\mathbf{N}_{\ell_1}\cdots e^{\mathbf{A}s_2}\mathbf{N}_{\ell_{i-1}}e^{\mathbf{A}s_1}\mathbf{b}_{\ell_i}$.

The above definition is reasonable in the case that the generalized reachability and observability Gramians exist. In particular, following [133], we can derive the $\mathcal{H}_2$-norm via

$$||\mathbf{\Sigma}_B||_{\mathcal{H}_2} = \sqrt{\operatorname{tr}\left(\mathbf{CPC}^T\right)} = \sqrt{\operatorname{tr}\left(\mathbf{B}^T\mathbf{QB}\right)}. \tag{4.13}$$

Hence, the derivation of the $\mathcal{H}_2$-norm amounts to a computation similar to the one for linear systems, cf. Lemma 2.2.1.

Finally, let us come to another important point that arises in the context of bilinear control systems. Although there exist several real-life phenomena that indeed exhibit a bilinear structure, often an accurate and detailed description of a process can only be given by a nonlinear model. Of course, linearizing the system around a known operating point might help to simplify the situation and opens up the way for known linear model order reduction techniques. However, often the approximation given by a linear model is not sufficient. Here, the so-called *Carleman linearization* is an interesting alternative that at least theoretically allows to approximate a nonlinear system with any desired accuracy. The first reference for this approach was given in [88]. We follow a similar discussion from [115]. Assume that a SISO nonlinear system is given by

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) + \mathbf{g}(\mathbf{x}(t), t)\, u(t), \tag{4.14}$$

$$y(t) = \mathbf{c}^T\mathbf{x}(t), \tag{4.15}$$

where $\mathbf{f}$ and $\mathbf{g}$ are functions that are analytic in $\mathbf{x}$ and continuous in t. Hence, the above equation is usually denoted as *linear-analytic*, essentially indicating that the dynamics exhibit a certain smoothness that allows to approximate the system by one of a simpler structure. As it is shown in [115], w.l.o.g. we can assume that $\mathbf{x}(0) = \mathbf{0}$ and $\mathbf{f}(\mathbf{x}(0)) = \mathbf{f}(0) = \mathbf{0}$. Consequently, $\mathbf{f}$ and $\mathbf{g}$ can be expanded into a Taylor series about $\mathbf{0}$, leading to

$$\begin{aligned}\mathbf{f}(\mathbf{x}) &= \mathbf{A}_1\mathbf{x} + \mathbf{A}_2\,\mathbf{x}\otimes\mathbf{x} + \cdots + \mathbf{A}_k\,\underbrace{\mathbf{x}\otimes\cdots\otimes\mathbf{x}}_{k} + \ldots,\\ \mathbf{g}(\mathbf{x}) &= \mathbf{B}_0 + \mathbf{B}_1\,\mathbf{x} + \cdots + \mathbf{B}_{k-1}\,\underbrace{\mathbf{x}\otimes\cdots\otimes\mathbf{x}}_{k-1} + \ldots,\end{aligned}$$

where $\mathbf{A}_i, \mathbf{B}_i \in \mathbb{R}^{n\times n^i}$ denote the matrices corresponding to the i-th derivative of $\mathbf{f}$ and $\mathbf{g}$. If we consider the enlarged state vector

$$\mathbf{x}^{\otimes} = \begin{bmatrix}\mathbf{x}\\ \mathbf{x}\otimes\mathbf{x}\\ \vdots\\ \underbrace{\mathbf{x}\otimes\cdots\otimes\mathbf{x}}_{k}\end{bmatrix},$$

we can approximately describe the dynamics of $\mathbf{x}^{\otimes}$ by a bilinear control system of the form

$$\begin{aligned}\frac{d}{dt}\mathbf{x}^{\otimes} &\approx \begin{bmatrix}\mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_k\\ \mathbf{0} & \mathbf{A}_{2,1} & \cdots & \mathbf{A}_{2,k-1}\\ \vdots & \ddots & \ddots & \vdots\\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{A}_{k,1}\end{bmatrix}\mathbf{x}^{\otimes} + \begin{bmatrix}\mathbf{B}_1 & \cdots & \mathbf{B}_{k-1} & \mathbf{0}\\ \mathbf{B}_{2,0} & \ddots & \vdots & \mathbf{0}\\ \mathbf{0} & \ddots & \mathbf{B}_{k-1,1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{k,0} & \mathbf{0}\end{bmatrix}\mathbf{x}^{\otimes}u + \begin{bmatrix}\mathbf{B}_0\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}u,\\ y &= \begin{bmatrix}\mathbf{c}^T & \mathbf{0} & \cdots & \mathbf{0}\end{bmatrix}\mathbf{x}^{\otimes}, \quad \mathbf{x}^{\otimes}(0) = \mathbf{0},\end{aligned}$$

where

$$\begin{aligned}\mathbf{A}_{i,j} &= \mathbf{A}_j\otimes\mathbf{I}\otimes\cdots\otimes\mathbf{I} + \cdots + \mathbf{I}\otimes\cdots\otimes\mathbf{I}\otimes\mathbf{A}_j,\\ \mathbf{B}_{i,j} &= \mathbf{B}_j\otimes\mathbf{I}\otimes\cdots\otimes\mathbf{I} + \cdots + \mathbf{I}\otimes\cdots\otimes\mathbf{I}\otimes\mathbf{B}_j,\end{aligned}$$

with $i-1$ Kronecker products in each of the i terms.

By increasing the number of terms in the Taylor series of $\mathbf{f}$ and $\mathbf{g}$, the quality of the approximation can be increased. However, the clear disadvantage is that the dimension of the system rapidly increases. Note that for the above system, we have that $\mathbf{x}^{\otimes} \in \mathbb{R}^{n\times n^k}$.

Nevertheless, in some situations the approach can be useful and we therefore discuss the previous technique by means of a simple test example arising in the context of

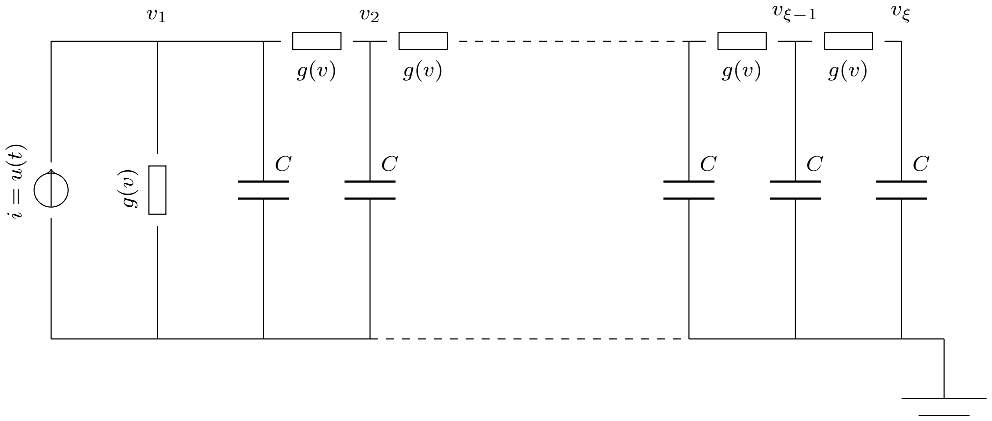


Figure 4.1: A scalable nonlinear RC circuit.

circuit theory, see also [32, 40, 111, 112]. The application of interest is an electrical RC ladder network as shown in Figure 4.1. The network consists of nonlinear resistors g and capacitors which, for simplicity, are assumed to be given by $C = 1$. Further, let $u(t)$ be the input signal to the independent current source and $v = \left[v_1, v_2, \ldots, v_N\right] \in \mathbb{R}^N$ denote the state vector consisting of the voltages between each node and the ground. Finally, we assume the voltage between node 1 and ground to be the measurable system output $y(t)$. Applying Kirchoff's current law allows to set up the state equations describing the corresponding nonlinear control system. Accordingly, for the nodes we obtain

$$\begin{aligned} C\dot{v}_1 + g(v_1) + g(v_1 - v_2) &= u, \\ C\dot{v}_k + g(v_k - v_{k+1}) &= g(v_{k-1} - v_k), \\ C\dot{v}_N &= g(v_{N-1} - v_N). \end{aligned}$$

Assuming that the current-voltage dependence of each resistor g is given as $g(\mathbf{v}) = e^{40\mathbf{v}} + \mathbf{v} - 1$, we end up with a nonlinear control system

$$\begin{aligned} \dot{\mathbf{v}}(t) &= \mathbf{f}(\mathbf{v}(t)) + \mathbf{b}u(t), \\ \mathbf{y}(t) &= \mathbf{c}^T\mathbf{v}(t), \end{aligned}$$

where

$$f(v) = f\left(\begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_k \\ \vdots \\ \mathbf{v}_N \end{bmatrix}\right) = \begin{bmatrix} -g(\mathbf{v}_1) - g(\mathbf{v}_1 - \mathbf{v}_2) \\ g(\mathbf{v}_1 - \mathbf{v}_2) - g(\mathbf{v}_2 - \mathbf{v}_3) \\ \vdots \\ g(\mathbf{v}_{k-1} - \mathbf{v}_k) - g(\mathbf{v}_k - \mathbf{v}_{k+1}) \\ \vdots \\ g(\mathbf{v}_{N-1} - \mathbf{v}_N) \end{bmatrix}, \quad \mathbf{b} = \mathbf{c} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We now replace this nonlinear system by an appropriate bilinear system by applying a second order Carleman linearization. Consequently, the resulting $(N + N^2)$-dimensional bilinear system is given by

$$\mathbf{\Sigma}_B : \quad \begin{cases} \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{N}\mathbf{x}u + \tilde{\mathbf{b}}u, \\ \mathbf{y} = \tilde{\mathbf{c}}^T\mathbf{x}, \end{cases}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \frac{1}{2}\mathbf{A}_2 \\ \mathbf{0} & \mathbf{A}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A}_1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \mathbf{v} \\ \mathbf{v} \otimes \mathbf{v} \end{bmatrix},$$
$$\mathbf{N} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{b} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{b} & \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{c}} = \begin{bmatrix} \mathbf{c} \\ \mathbf{0} \end{bmatrix}.$$

Since the computation of the matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ is straightforward, we immediately give their final structures. For $\mathbf{A}_1$ we get

$$\mathbf{A}_1 = \begin{bmatrix} -82 & 41 & & & \\ 41 & -82 & 41 & & \\ & \ddots & \ddots & \ddots & \\ & & 41 & -82 & 41 \\ & & & 41 & -41 \end{bmatrix}.$$

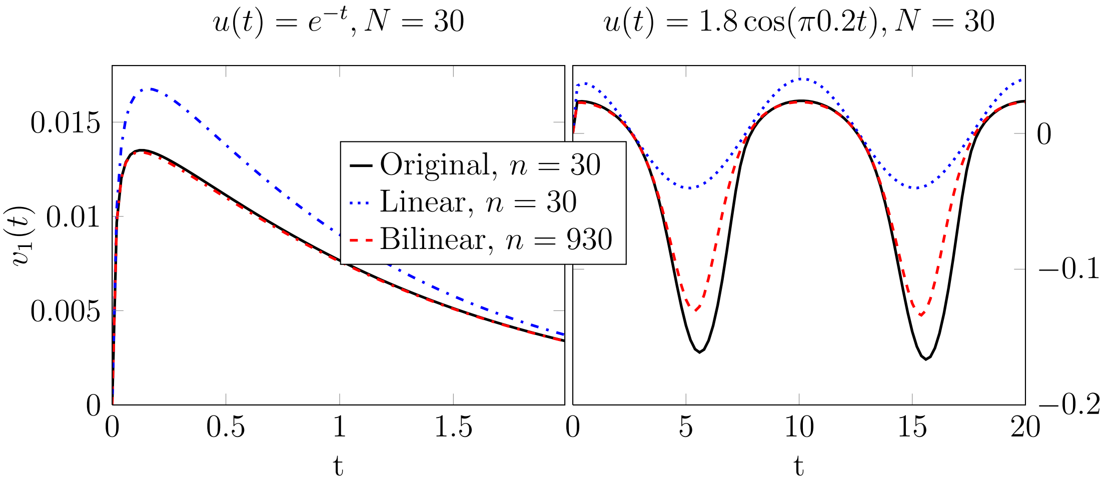


Figure 4.2: RC circuit. Transient responses for original, linearized and second order Carleman bilinearized model.

For the nonzero entries of $\mathbf{A}_2$ we arrive at

$$\begin{aligned}
\mathbf{A}_2(j,1) &= -3200, & \mathbf{A}_2(j,2) &= 1600, && \text{for } j=1,\\
\mathbf{A}_2(j,N+1) &= 1600, & \mathbf{A}_2(j,N+2) &= -1600, && \text{for } j=1,\\
\mathbf{A}_2(j,(j-2)N+j-1) &= 1600, & \mathbf{A}_2(j,(j-2)N+j) &= -1600, && \text{for } 2\le j\le N-1,\\
\mathbf{A}_2(j,(j-1)N+j-1) &= -1600, & \mathbf{A}_2(j,(j-1)N+j+1) &= 1600, && \text{for } 2\le j\le N-1,\\
\mathbf{A}_2(j,jN+j) &= 1600, & \mathbf{A}_2(j,jN+j+1) &= -1600, && \text{for } 2\le j\le N-1,\\
\mathbf{A}_2(j,(N-2)N+N-1) &= 1600, & \mathbf{A}_2(j,(N-2)N+N) &= -1600, && \text{for } j=N,\\
\mathbf{A}_2(j,(N-1)N+N-1) &= -1600, & \mathbf{A}_2(j,(N-1)N+N) &= 1600, && \text{for } j=N.
\end{aligned}$$

As it is shown in Figure 4.2, for this specific example, the previous second order Carleman linearization often yields accurate approximations that outperform a conventional linearization about an operating point by far. However, as we see for the second input, there exist excitations where a second order Carleman linearization also yields only a moderate approximation of the original dynamics.

4.3 $\mathcal{H}_2$-optimal model reduction

In this section, we study the $\mathcal{H}_2$-optimal model reduction problem for bilinear systems of the form (4.1). Similar to the linear case, we want to find iterative algorithms that construct a reduced-order system $\hat{\mathbf{\Sigma}}_B$ that solves

$$||\mathbf{\Sigma}_B - \hat{\mathbf{\Sigma}}_B||_{\mathcal{H}_2} = \min_{\substack{\dim(\tilde{\mathbf{\Sigma}}_B)\,=\,\hat{n}\\ \tilde{\mathbf{\Sigma}}_B \text{ stable}}} ||\mathbf{\Sigma}_B - \tilde{\mathbf{\Sigma}}_B||_{\mathcal{H}_2}. \tag{4.16}$$

4.3.1 Preliminaries

Recall from Chapter 3 that a locally $\mathcal{H}_2$-optimal reduced-order system has to fulfill the tangential interpolation conditions (3.7). For the generalization of these conditions to the bilinear case, it makes sense to consider the corresponding vectorized conditions. For this, let us have a look at the i-th column of the left hand side of (3.7a) which can be rewritten as follows:

$$\begin{aligned}
\tilde{\mathbf{c}}_j^T \mathbf{H}(-\hat{\lambda}_j)_i &= \tilde{\mathbf{c}}_j^T \mathbf{C}(-\hat{\lambda}_j \mathbf{I} - \mathbf{A})^{-1}\mathbf{b}_i \\
&= \left[\tilde{\mathbf{c}}_1^T \mathbf{C}, \dots, \tilde{\mathbf{c}}_{\hat{n}}^T \mathbf{C}\right] \begin{bmatrix} -\hat{\lambda}_1 \mathbf{I} - \mathbf{A} & & \\ & \ddots & \\ & & -\hat{\lambda}_{\hat{n}} \mathbf{I} - \mathbf{A} \end{bmatrix}^{-1} (\mathbf{e}_j \otimes \mathbf{b}_i) \\
&= \operatorname{vec}\left(\mathbf{C}^T \tilde{\mathbf{C}}\right)^T \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\mathbf{e}_j \mathbf{e}_i^T \otimes \mathbf{B}\right) \xi_m \\
&= \xi_p^T (\tilde{\mathbf{C}} \otimes \mathbf{C}) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\mathbf{e}_j \mathbf{e}_i^T \otimes \mathbf{B}\right) \xi_m,
\end{aligned}$$

with $\xi_m = \operatorname{vec}(\mathbf{I}_m)$ and $\hat{\mathbf{\Lambda}} = \operatorname{diag}\left(\hat{\lambda}_1, \dots, \hat{\lambda}_{\hat{n}}\right)$.

Hence, condition (3.7a) is the same as requiring that

$$\begin{aligned}
&\xi_p^T \left(\tilde{\mathbf{C}} \otimes \hat{\mathbf{C}}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}}\right)^{-1} \left(\mathbf{e}_j \mathbf{e}_i^T \otimes \hat{\mathbf{B}}\right) \xi_m \\
&\quad = \xi_p^T \left(\tilde{\mathbf{C}} \otimes \mathbf{C}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\mathbf{e}_j \mathbf{e}_i^T \otimes \mathbf{B}\right) \xi_m,
\end{aligned} \tag{4.17}$$

holds for $j = 1, \dots, \hat{n}$ and $i = 1, \dots, m$. Similarly, instead of (3.7b) we can write

$$\begin{aligned}
&\xi_p^T \left(\mathbf{e}_i \mathbf{e}_j^T \otimes \hat{\mathbf{C}}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}}\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}}\right) \xi_m \\
&\quad = \xi_p^T \left(\mathbf{e}_i \mathbf{e}_j^T \otimes \mathbf{C}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m.
\end{aligned} \tag{4.18}$$

Finally, condition (3.7c) is the same as

$$\begin{aligned}
&\xi_p^T \left(\tilde{\mathbf{C}} \otimes \hat{\mathbf{C}}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}}\right)^{-1} \left(\mathbf{e}_j \mathbf{e}_j^T \otimes \mathbf{I}\right) \times \\
&\quad \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}}\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}}\right) \xi_m \\
&= \xi_p^T \left(\tilde{\mathbf{C}} \otimes \mathbf{C}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\mathbf{e}_j \mathbf{e}_j^T \otimes \mathbf{I}\right) \times \\
&\quad \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m .
\end{aligned} \tag{4.19}$$

Furthermore, the construction of the projection matrices in each step of iterative algorithms like IRKA or MIRIAm can also be interpreted in terms of the Kronecker product notation. More precisely, in order to guarantee Hermite-type tangential interpolation, we need to compute

$$\mathbf{V}_j = \left(\sigma_j \mathbf{I} - \mathbf{A}\right)^{-1} \mathbf{B} \tilde{\mathbf{b}}_j, \tag{4.20}$$

$$\mathbf{W}_j = \left(\sigma_j \mathbf{I} - \mathbf{A}^T\right)^{-1} \mathbf{C}^T \tilde{\mathbf{c}}_j. \tag{4.21}$$

In vectorized notation, this means that

$$\operatorname{vec}(\mathbf{V}) = \left(\operatorname{diag}(\sigma_1, \ldots, \sigma_{\hat{n}}) \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m, \tag{4.22}$$

$$\operatorname{vec}(\mathbf{W}) = \left(\operatorname{diag}(\sigma_1, \ldots, \sigma_{\hat{n}}) \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}^T\right)^{-1} \left(\tilde{\mathbf{C}}^T \otimes \mathbf{C}^T\right) \xi_p. \tag{4.23}$$

Finally, from (4.13) we know that there exists a close relation between the $\mathcal{H}_2$-norm and the solutions of the generalized Lyapunov equations (4.12). In particular, in terms of the Kronecker notation we easily obtain the following useful result.

Proposition 4.3.1. *Let $\mathbf{\Sigma}_B$ be a stable bilinear system. Then it holds that*

$$||\mathbf{\Sigma}_B||_{\mathcal{H}_2}^2 = \xi_p^T \left(\mathbf{C} \otimes \mathbf{C}\right) \left(-\mathbf{A} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \mathbf{N}_k \otimes \mathbf{N}_k\right)^{-1} \left(\mathbf{B} \otimes \mathbf{B}\right) \xi_m.$$

Proof. Making use of the vectorization of (4.12), we can express $\mathbf{P}$ as

$$\operatorname{vec}(\mathbf{P}) = \left(-\mathbf{A} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \mathbf{N}_k \otimes \mathbf{N}_k\right)^{-1} \operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right).$$

Since $\operatorname{tr}\left(\mathbf{C}\mathbf{P}\mathbf{C}^T\right) = \operatorname{tr}\left(\mathbf{C}^T\mathbf{C}\mathbf{P}\right)$, the statement easily follows from the properties of the Kronecker product given in Chapter 2. □

4.3.2 $\mathcal{H}_2$-optimality conditions for bilinear systems

Obviously, since the linear $\mathcal{H}_2$-model reduction problem already is a non-convex optimization problem, for the bilinear case we cannot expect to find the global optimum of (4.16). Hence, in this section the aim is to derive first order necessary conditions. As in the linear case, for this we have to consider the norm of the error system $\mathbf{\Sigma}_{B,err} := \mathbf{\Sigma}_B - \hat{\mathbf{\Sigma}}_B$, which is defined via

$$\mathbf{A}_{err} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{A}} \end{bmatrix}, \quad \mathbf{N}_{err,k} = \begin{bmatrix} \mathbf{N}_k & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{N}}_k \end{bmatrix}, \quad \mathbf{B}_{err} = \begin{bmatrix} \mathbf{B} \\ \hat{\mathbf{B}} \end{bmatrix}, \quad \mathbf{C}_{err} = \begin{bmatrix} \mathbf{C} & -\hat{\mathbf{C}} \end{bmatrix}.$$

Based on the computation formula for the $\mathcal{H}_2$-norm, it is shown in [133] that the reduced system matrices of a locally $\mathcal{H}_2$-optimal model have to fulfill conditions that extend the Wilson conditions to the bilinear case. These are

$$\begin{aligned} \mathbf{Q}_{12}^T\mathbf{P}_{12} + \mathbf{Q}_{22}\mathbf{P}_{22} = \mathbf{0}, \quad & \mathbf{Q}_{22}\hat{\mathbf{N}}_k\mathbf{P}_{22} + \mathbf{Q}_{12}^T\mathbf{N}_k\mathbf{P}_{12} = \mathbf{0}, \quad \text{for } k = 1, \ldots, m, \\ \mathbf{Q}_{12}^T\mathbf{B} + \mathbf{Q}_{22}\hat{\mathbf{B}} = \mathbf{0}, \quad & \hat{\mathbf{C}}\mathbf{P}_{22} - \mathbf{C}\mathbf{P}_{12} = \mathbf{0}, \end{aligned} \tag{4.24}$$

where

$$\mathbf{P}_{err} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{bmatrix}, \quad \mathbf{Q}_{err} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}^T & \mathbf{Q}_{22} \end{bmatrix}, \tag{4.25}$$

are the solutions of the generalized Lyapunov equations

$$\mathbf{A}_{err}\mathbf{P}_{err} + \mathbf{P}_{err}\mathbf{A}_{err}^T + \sum_{k=1}^m \mathbf{N}_{err,k}\mathbf{P}_{err}\mathbf{N}_{err,k}^T + \mathbf{B}_{err}\mathbf{B}_{err}^T = \mathbf{0}, \tag{4.26}$$

$$\mathbf{A}_{err}^T\mathbf{Q}_{err} + \mathbf{Q}_{err}\mathbf{A}_{err} + \sum_{k=1}^m \mathbf{N}_{err,k}^T\mathbf{Q}_{err}\mathbf{N}_{err,k} + \mathbf{C}_{err}^T\mathbf{C}_{err} = \mathbf{0}. \tag{4.27}$$

The authors of [133] further proposed an algorithm that constructs a system $\hat{\mathbf{\Sigma}}_B$ fulfilling the above conditions. However, their procedure relies on a gradient flow optimization technique and, thus, in each step the solution of a large-scale system of ODEs is required, making the technique infeasible in typical real-life applications. For this reason, we want to investigate the possibility of an interpolatory approach that uses the idea of IRKA. As a first step, we have to generalize the interpolation-based optimality conditions from

Chapter 3. Here, the computation formula from Proposition 4.3.1 together with a simple analysis of the structure of the error system leads to the following expression for the error functional $J = ||\mathbf{\Sigma}_{B,err}||^2_{\mathcal{H}_2}$.

Corollary 4.3.1. *Let $\mathbf{\Sigma}_B$ denote a bilinear system. Further assume that a diagonalizable reduced-order system $\hat{\mathbf{\Sigma}}_B$ is given. Then*

$$J = \xi_p^T \left(\begin{bmatrix} \mathbf{C} & -\tilde{\mathbf{C}} \end{bmatrix} \otimes \mathbf{C}_{err} \right) \times$$
$$\left(- \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Lambda}} \end{bmatrix} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}_{err} - \sum_{k=1}^m \begin{bmatrix} \mathbf{N}_k & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{N}}_k^T \end{bmatrix} \otimes \mathbf{N}_{err} \right)^{-1} \left(\begin{bmatrix} \mathbf{B} \\ \tilde{\mathbf{B}}^T \end{bmatrix} \otimes \mathbf{B}_{err} \right) \xi_m,$$

where $\mathbf{R}\hat{\mathbf{\Lambda}}\mathbf{R}^{-1} = \hat{\mathbf{A}}$, $\tilde{\mathbf{B}} = \hat{\mathbf{B}}^T\mathbf{R}^{-T}$, $\tilde{\mathbf{C}} = \hat{\mathbf{C}}\mathbf{R}$, and $\tilde{\mathbf{N}}_k = \mathbf{R}^T\hat{\mathbf{N}}_k^T\mathbf{R}^{-T}$ is the spectral decomposition of $\hat{\mathbf{\Sigma}}_B$.

The above representation is motivated by the demand of having optimization parameters $\hat{\mathbf{\Lambda}}$, $\tilde{\mathbf{N}}_k$, $\tilde{\mathbf{B}}$, and $\tilde{\mathbf{C}}$ that can be chosen to locally minimize $||\mathbf{\Sigma}_B - \hat{\mathbf{\Sigma}}_B||^2_{\mathcal{H}_2}$. For the differentiation with respect to those optimization parameters, we need an extension of the product rule to Kronecker products.

Lemma 4.3.1. *Let $\mathbf{C}(x) \in \mathbb{R}^{p \times n}$, $\mathbf{A}(y), \mathbf{N}_k \in \mathbb{R}^{n \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$, with $x, y \in \mathbb{R}$. Let $\mathcal{L}(y) = -\mathbf{A}(y) \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}(y) - \sum_{k=1}^m \mathbf{N}_k \otimes \mathbf{N}_k$ and assume that $\mathbf{C}$ and $\mathbf{A}$ are differentiable with respect to x and y. Then,*

$$\frac{\partial}{\partial x} \left[\xi_p^T (\mathbf{C}(x) \otimes \mathbf{C}(x)) \mathcal{L}(y)^{-1} (\mathbf{B} \otimes \mathbf{B}) \xi_m \right] = 2 \cdot \xi_p^T \left(\frac{\partial}{\partial x} \mathbf{C}(x) \otimes \mathbf{C}(x) \right) \mathcal{L}(y)^{-1} (\mathbf{B} \otimes \mathbf{B}) \, \xi_m$$

and

$$\frac{\partial}{\partial y} \left[\xi_p^T (\mathbf{C}(x) \otimes \mathbf{C}(x)) \mathcal{L}(y)^{-1} (\mathbf{B} \otimes \mathbf{B}) \, \xi_m \right]$$
$$= 2 \cdot \xi_p^T \left(\mathbf{C}(x) \otimes \mathbf{C}(x) \right) \mathcal{L}(y)^{-1} \left(\frac{\partial}{\partial y} \mathbf{A}(y) \otimes \mathbf{I} \right) \mathcal{L}(y)^{-1} (\mathbf{B} \otimes \mathbf{B}) \, \xi_m.$$

Proof. For the first part, note that

$$\operatorname{vec}(\mathbf{P}(y)) := \left(-\mathbf{A}(y) \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}(y) - \sum_{k=1}^m \mathbf{N}_k \otimes \mathbf{N}_k \right)^{-1} (\mathbf{B} \otimes \mathbf{B}) \, \xi_m$$

is the solution of the parameter-dependent Lyapunov equation

$$\mathbf{A}(y)\mathbf{P}(y)+\mathbf{P}(y)\mathbf{A}(y)^T+\sum_{k=1}^{m}\mathbf{N}_k\mathbf{P}(y)\mathbf{N}_k^T+\mathbf{B}\mathbf{B}^T=\mathbf{0}.$$

Hence, we can conclude that $\mathbf{P}(y)=\mathbf{P}(y)^T$. Next, we observe that

$$\begin{aligned}&\xi_p^T\left(\mathbf{C}(x)\otimes\left(\frac{\partial}{\partial x}\mathbf{C}(x)\right)\right)\operatorname{vec}\left(\mathbf{P}(y)\right)=\operatorname{tr}\left(\mathbf{C}(x)^T\left(\frac{\partial}{\partial x}\mathbf{C}(x)\right)\mathbf{P}(y)\right)\\&=\operatorname{tr}\left(\mathbf{C}(x)\mathbf{P}(y)^T\left(\frac{\partial}{\partial x}\mathbf{C}(x)^T\right)\right)=\operatorname{tr}\left(\left(\frac{\partial}{\partial x}\mathbf{C}(x)^T\right)\mathbf{C}(x)\mathbf{P}(y)\right)\\&=\xi_p^T\left(\left(\frac{\partial}{\partial x}\mathbf{C}(x)\right)\otimes\mathbf{C}(x)\right)\operatorname{vec}\left(\mathbf{P}(y)\right).\end{aligned}$$

The last equation implies that we can interchange the derivatives with respect to x. The assertion now trivially follows. For the second part, recall that we have $\frac{\partial}{\partial y}\left(\mathbf{A}(y)^{-1}\right)=-\mathbf{A}(y)^{-1}\frac{\partial\mathbf{A}(y)}{\partial y}\mathbf{A}(y)^{-1}$. Furthermore, by $\mathbf{Q}(x,y)$ we denote the solution of the dual Lyapunov equation

$$\mathbf{A}(y)^T\mathbf{Q}(x,y)+\mathbf{Q}(x,y)\mathbf{A}(y)+\sum_{k=1}^{m}\mathbf{N}_k^T\mathbf{Q}(x,y)\mathbf{N}_k+\mathbf{C}(x)^T\mathbf{C}(x)=\mathbf{0}.$$

Hence, we end up with

$$\begin{aligned}&\operatorname{vec}\left(\mathbf{Q}(x,y)\right)^T\left(\mathbf{I}\otimes\frac{\partial}{\partial y}\mathbf{A}(y)\right)\operatorname{vec}\left(\mathbf{P}(y)\right)=\operatorname{tr}\left(\mathbf{Q}(x,y)^T\left(\frac{\partial}{\partial y}\mathbf{A}(y)\right)\mathbf{P}(y)\right)\\&=\operatorname{tr}\left(\mathbf{P}(y)^T\left(\frac{\partial}{\partial y}\mathbf{A}(y)^T\right)\mathbf{Q}(x,y)\right)=\operatorname{tr}\left(\left(\frac{\partial}{\partial y}\mathbf{A}(y)^T\right)\mathbf{Q}(x,y)^T\mathbf{P}(y)\right)\\&=\operatorname{vec}\left(\mathbf{Q}(x,y)\frac{\partial}{\partial y}\mathbf{A}(y)\right)^T\operatorname{vec}\left(\mathbf{P}(y)\right)\\&=\operatorname{vec}\left(\mathbf{Q}(x,y)\right)^T\left(\frac{\partial}{\partial y}\mathbf{A}(y)\otimes\mathbf{I}\right)\operatorname{vec}\left(\mathbf{P}(y)\right).\end{aligned}$$

Again, the last line proves the second statement. □

Before we proceed, recall that the permutation $\mathbf{M}$ from Proposition 2.1.2 fulfills

$$\mathbf{M}^T \left(\mathbf{A} \otimes \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix} \right) \mathbf{M} = \begin{bmatrix} \mathbf{A}\otimes\mathbf{B} & \mathbf{A}\otimes\mathbf{C} \\ \mathbf{A}\otimes\mathbf{D} & \mathbf{A}\otimes\mathbf{E} \end{bmatrix}.$$

Finally, we are ready to derive the necessary conditions as follows.

$$\begin{aligned}
\frac{\partial J}{\partial \tilde{\mathbf{C}}_{ij}} &= 2\cdot \xi_p^T \left(\begin{bmatrix} \mathbf{0} & -\mathbf{e}_i\mathbf{e}_j^T \end{bmatrix} \otimes \mathbf{C}_{err} \right) \times \\
&\quad \left(- \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Lambda}} \end{bmatrix} \otimes \mathbf{I} - \mathbf{I}\otimes \mathbf{A}_{err} - \sum_{k=1}^m \begin{bmatrix} \mathbf{N}_k & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{N}}_k^T \end{bmatrix} \otimes \mathbf{N}_{err,k} \right)^{-1} \left(\begin{bmatrix} \mathbf{B} \\ \tilde{\mathbf{B}}^T \end{bmatrix} \otimes \mathbf{B}_{err} \right) \xi_m \\
&= 2\cdot \xi_p^T \left(-\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{C}_{err} \right) \mathbf{M}\mathbf{M}^T \times \\
&\quad \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I}\otimes \mathbf{A}_{err} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_{err} \right)^{-1} \mathbf{M}\mathbf{M}^T \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}_{err} \right) \xi_m \\
&= -2\cdot \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{C} \right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I}\otimes \mathbf{A} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B} \right) \xi_m \\
&\quad + 2\cdot \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \hat{\mathbf{C}} \right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I}\otimes \hat{\mathbf{A}} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}} \right) \xi_m .
\end{aligned}$$

Here, the last step is justified by the fact that $\mathbf{M}$ is a permutation matrix and, thus, $\mathbf{M}^T\mathbf{M} = \mathbf{I}$ and by the identities:

$$\begin{aligned}
\left(-\mathbf{e}_i\mathbf{e}_j^T \otimes \begin{bmatrix} \mathbf{C} & -\hat{\mathbf{C}} \end{bmatrix} \right) \mathbf{M} &= \begin{bmatrix} -\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{C} & \mathbf{e}_i\mathbf{e}_j^T \otimes \hat{\mathbf{C}} \end{bmatrix}, \\
\mathbf{M}^T \left(\tilde{\mathbf{B}} \otimes \begin{bmatrix} \mathbf{B} \\ \hat{\mathbf{B}} \end{bmatrix} \right) &= \begin{bmatrix} \tilde{\mathbf{B}} \otimes \mathbf{B} \\ \tilde{\mathbf{B}} \otimes \hat{\mathbf{B}} \end{bmatrix}.
\end{aligned}$$

Setting the resulting expression equal to zero reveals that $\hat{\mathbf{\Sigma}}$ has to satisfy:

$$\begin{aligned}
&\xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{C} \right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I}\otimes \mathbf{A} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B} \right) \xi_m \\
&\quad = \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \hat{\mathbf{C}} \right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I}\otimes \hat{\mathbf{A}} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}} \right) \xi_m .
\end{aligned} \tag{4.28}$$

In view of equation (3.7b) in the form of (4.18), we see that this demand naturally extends the interpolation-based condition known from the linear case. For the differentiation with respect to the poles of $\hat{\mathbf{A}}$, we use the second part of Lemma 4.3.1 in order to obtain

$$
\begin{aligned}
\frac{\partial J}{\partial \hat{\lambda}_i} &= 2 \cdot \xi_p^T \left(\begin{bmatrix} \mathbf{C} & -\tilde{\mathbf{C}} \end{bmatrix} \otimes \mathbf{C}_{err} \right) \times \\
&\quad \left(\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Lambda}} \end{bmatrix} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A}_{err} + \sum_{k=1}^{m} \begin{bmatrix} \mathbf{N}_k & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{N}}_k^T \end{bmatrix} \otimes \mathbf{N}_{err,k} \right)^{-1} \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{e}_i \mathbf{e}_i^T \end{bmatrix} \otimes \mathbf{I} \right) \times \\
&\quad \left(\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Lambda}} \end{bmatrix} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{A}_{err} + \sum_{k=1}^{m} \begin{bmatrix} \mathbf{N}_k & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{N}}_k^T \end{bmatrix} \otimes \mathbf{N}_{err,k} \right)^{-1} \left(\begin{bmatrix} \mathbf{B} \\ \tilde{\mathbf{B}}^T \end{bmatrix} \otimes \mathbf{B}_{err} \right) \xi_m \\
&= -2 \cdot \xi_p^T \left(\tilde{\mathbf{C}} \otimes \mathbf{C} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k \right)^{-1} \times \\
&\qquad \left(\mathbf{e}_i \mathbf{e}_i^T \otimes \mathbf{I} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B} \right) \xi_m \\
&\quad + 2 \cdot \xi_p^T \left(\tilde{\mathbf{C}} \otimes \hat{\mathbf{C}} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k \right)^{-1} \times \\
&\qquad \left(\mathbf{e}_i \mathbf{e}_i^T \otimes \mathbf{I} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}} \right) \xi_m.
\end{aligned}
$$

Once more, we find an interpolation-based condition generalizing (3.7c) in the form of (4.19) if we set the last expression equal to zero:

$$
\begin{aligned}
&\xi_p^T \left(\tilde{\mathbf{C}} \otimes \mathbf{C} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k \right)^{-1} \times \\
&\qquad \left(\mathbf{e}_i \mathbf{e}_i^T \otimes \mathbf{I} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B} \right) \xi_m \\
&\quad = \xi_p^T \left(\tilde{\mathbf{C}} \otimes \hat{\mathbf{C}} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k \right)^{-1} \times \\
&\qquad \left(\mathbf{e}_i \mathbf{e}_i^T \otimes \mathbf{I} \right) \left(-\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k \right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}} \right) \xi_m.
\end{aligned}
\tag{4.29}
$$

Finally, as a matter of careful analysis, we obtain similar optimality conditions when differentiating with respect to $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{N}}_k$, respectively:

$$\begin{aligned}\xi_p^T\left(\tilde{\mathbf{C}}\otimes\mathbf{C}\right)\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{A}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\mathbf{N}_k\right)^{-1}\left(\mathbf{e}_j\mathbf{e}_i^T\otimes\mathbf{B}\right)\xi_m\\ =\xi_p^T\left(\tilde{\mathbf{C}}\otimes\hat{\mathbf{C}}\right)\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\hat{\mathbf{A}}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\hat{\mathbf{N}}_k\right)^{-1}\left(\mathbf{e}_j\mathbf{e}_i^T\otimes\hat{\mathbf{B}}\right)\xi_m,\end{aligned}\tag{4.30}$$

$$\begin{aligned}\xi_p^T\left(\tilde{\mathbf{C}}\otimes\mathbf{C}\right)\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{A}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\mathbf{N}_k\right)^{-1}\times\\ \left(\mathbf{e}_j\mathbf{e}_i^T\otimes\mathbf{N}_k\right)\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{A}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\mathbf{N}_k\right)^{-1}\left(\tilde{\mathbf{B}}^T\otimes\mathbf{B}\right)\xi_m\\ =\xi_p^T\left(\tilde{\mathbf{C}}\otimes\hat{\mathbf{C}}\right)\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\hat{\mathbf{A}}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\hat{\mathbf{N}}_k\right)^{-1}\times\\ \left(\mathbf{e}_j\mathbf{e}_i^T\otimes\hat{\mathbf{N}}_k\right)\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\hat{\mathbf{A}}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\hat{\mathbf{N}}_k\right)^{-1}\left(\tilde{\mathbf{B}}^T\otimes\hat{\mathbf{B}}\right)\xi_m.\end{aligned}\tag{4.31}$$

Hence, the previous derivations can be summarized in the following theorem.

Theorem 4.3.1. *Let $\mathbf{\Sigma}_B$ denote a BIBO stable bilinear system. Assume that $\hat{\mathbf{\Sigma}}_B$ is a reduced bilinear system of dimension $\hat{n}$, locally minimizing the $\mathcal{H}_2$-norm of the error system among all bilinear systems of dimension $\hat{n}$. Then $\hat{\mathbf{\Sigma}}_B$ fulfills equations (4.28) – (4.31).*

Remark 4.3.1. *At this point one might wonder why it makes sense to denote the above conditions as being of interpolatory nature. Note that if the inverse*

$$\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{A}-\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\mathbf{N}_k\right)^{-1}$$

exists and the Volterra series converges, we can use the Neumann Lemma and obtain an infinite series of the form

$$\sum_{i=0}^\infty\left(\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{A}\right)^{-1}\left(\sum_{k=1}^m\tilde{\mathbf{N}}_k^T\otimes\mathbf{N}_k\right)\right)^i\left(-\hat{\mathbf{\Lambda}}\otimes\mathbf{I}-\mathbf{I}\otimes\mathbf{A}\right)^{-1}.$$

Now each term of this series corresponds to a term of the Volterra series representation *for bilinear control systems. For a more detailed insight, below we state a very interesting result from Flagg that can be found in [56].*

Theorem 4.3.2. *Let $\mathbf{\Sigma}_B$ be a SISO bilinear system of order n with finite $\mathcal{H}_2$-norm. Let $\hat{\mathbf{\Sigma}}_B = (\hat{\mathbf{A}}, \hat{\mathbf{N}}, \hat{\mathbf{b}}, \hat{\mathbf{c}})$ be an $\mathcal{H}_2$-optimal approximation of order $\hat{n}$. Then, $\hat{\mathbf{\Sigma}}$ satisfies the following multipoint Volterra series interpolation conditions*

$$\begin{aligned}&\sum_{k=1}^{\infty}\sum_{\ell_1=1}^{\hat{n}}\cdots\sum_{\ell_k=1}^{\hat{n}} \hat{\mathbf{\Phi}}_{\ell_1,\ldots,\ell_k} H_k(-\hat{\lambda}_1,\ldots,-\hat{\lambda}_k)\\ &\quad= \sum_{k=1}^{\infty}\sum_{\ell_1=1}^{\hat{n}}\cdots\sum_{\ell_k=1}^{\hat{n}} \hat{\mathbf{\Phi}}_{\ell_1,\ldots,\ell_k} \hat{H}_k(-\hat{\lambda}_1,\ldots,-\hat{\lambda}_k)\end{aligned}$$

and

$$\begin{aligned}&\sum_{k=1}^{\infty}\sum_{\ell_1=1}^{\hat{n}}\cdots\sum_{\ell_k=1}^{\hat{n}} \hat{\mathbf{\Phi}}_{\ell_1,\ldots,\ell_k} \left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j} H_k(-\hat{\lambda}_1,\ldots,-\hat{\lambda}_k)\right)\\ &\quad= \sum_{k=1}^{\infty}\sum_{\ell_1=1}^{\hat{n}}\cdots\sum_{\ell_k=1}^{\hat{n}} \hat{\mathbf{\Phi}}_{\ell_1,\ldots,\ell_k} \left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j} \hat{H}_k(-\hat{\lambda}_1,\ldots,-\hat{\lambda}_k)\right),\end{aligned}$$

where $\hat{\mathbf{\Phi}}_{\ell_1,\ldots,\ell_k}$ and $\hat{\lambda}_{\ell_i}$ are the (multivariate) residues and poles of the transfer functions $\hat{H}_k$ associated with $\hat{\mathbf{\Sigma}}_B$.

Note that Theorem 4.3.2 yields more explicit interpolation-based $\mathcal{H}_2$-optimality conditions corresponding to the Volterra series representation for bilinear systems. As is discussed in detail in [56, Chapter 4], our conditions and the ones from Theorem 4.3.2 are equivalent, but the latter ones benefit from being transferable to bilinear systems with infinite $\mathcal{H}_2$-norm.

4.3.3 Generalized Sylvester equations and BIRKA

Now that we have specified first order necessary conditions for $\mathcal{H}_2$-optimality, we propose two algorithms that iteratively construct a reduced-order system which locally minimizes the $\mathcal{H}_2$-error. We start with a procedure based on generalized Sylvester equations which in the linear case reduces to the concept discussed in [44]. For this, let us consider the following two matrix equations:

$$\mathbf{A}\mathbf{X} + \mathbf{X}\hat{\mathbf{A}}^T + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{X}\hat{\mathbf{N}}_k^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0}, \tag{4.32a}$$

$$\mathbf{A}^T\mathbf{Y} + \mathbf{Y}\hat{\mathbf{A}} + \sum_{k=1}^{m} \mathbf{N}_k^T\mathbf{Y}\hat{\mathbf{N}}_k - \mathbf{C}^T\hat{\mathbf{C}} = \mathbf{0}. \tag{4.32b}$$

Obviously, the solutions $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times\hat{n}}$ can be explicitly computed by vectorizing both sides and making use of the Kronecker product. However, this requires solving two linear systems of equations:

$$\left(-\mathbf{I}\otimes\mathbf{A} - \hat{\mathbf{A}}\otimes\mathbf{I} - \sum_{k=1}^{m}\hat{\mathbf{N}}_k\otimes\mathbf{N}_k\right)\operatorname{vec}(\mathbf{X}) = \operatorname{vec}\left(\mathbf{B}\hat{\mathbf{B}}^T\right),$$

$$\left(\mathbf{I}\otimes\mathbf{A}^T + \hat{\mathbf{A}}^T\otimes\mathbf{I} + \sum_{k=1}^{m}\hat{\mathbf{N}}_k^T\otimes\mathbf{N}_k^T\right)\operatorname{vec}(\mathbf{Y}) = \operatorname{vec}\left(\mathbf{C}^T\hat{\mathbf{C}}\right).$$

Throughout the rest of the thesis, we assume that there exist unique solutions satisfying these Sylvester equations. Due to the properties of the eigenvalue computation of Kronecker products, this certainly is satisfied if the eigenvalues of $\hat{\mathbf{A}}$ are located in $\mathbb{C}_-$ and the norms of $\hat{\mathbf{N}}_k$ are sufficiently small. However, in view of Theorem 4.2.1 we have already mentioned that this basically characterizes a stable bilinear system. Although in general this cannot be ensured by our proposed algorithms, we did not observe unstable reduced-order systems so far. For a similar discussion of the linear case we refer to [73]. Finally, we mention that, under appropriate assumptions, the solutions $\mathbf{X}$ and $\mathbf{Y}$ can be computed as the limits of infinite series of linear Sylvester equations.

Lemma 4.3.2. *Let $\mathcal{L}, \Pi : \mathbb{R}^{n\times\hat{n}} \to \mathbb{R}^{n\times\hat{n}}$ denote two linear operators defined by the bilinear systems $\mathbf{\Sigma}_B$ and $\hat{\mathbf{\Sigma}}_B$, with $\mathcal{L}(\mathbf{X}) := \mathbf{A}\mathbf{X} + \mathbf{X}\hat{\mathbf{A}}^T$ and $\Pi(\mathbf{X}) := \sum_{k=1}^{m}\mathbf{N}_k\mathbf{X}\hat{\mathbf{N}}_k^T$. If the spectral radius $\rho(\mathcal{L}^{-1}\Pi) < 1$, then the solution $\mathbf{X}$ of the generalized Sylvester equation (4.32a) is given as $\mathbf{X} = \lim\limits_{j\to\infty}\mathbf{X}_j$, with:*

$$\mathbf{A}\mathbf{X}_1 + \mathbf{X}_1\hat{\mathbf{A}}^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0},$$

$$\mathbf{A}\mathbf{X}_j + \mathbf{X}_j\hat{\mathbf{A}}^T + \sum_{k=1}^{m}\mathbf{N}_k\mathbf{X}_{j-1}\hat{\mathbf{N}}_k^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0}, \quad j > 1.$$

A dual statement obviously is true for equation (4.32b). Since the statement is a direct consequence of the theory of convergent splittings, we dispense with the proof and instead refer to [43] for an equivalent discussion on bilinear Lyapunov equations.

Remark 4.3.2. *Although the aforementioned splitting at least theoretically yields a possible way of solving the generalized Sylvester equation (4.32a), the procedure strongly depends on the spectral radius $\rho(\mathcal{L}^{-1}\Pi)$. Moreover, so far it seems hard to state properties of a bilinear control system that automatically ensure the desired convergence.*

Algorithm 4.3.1 Generalized Sylvester iteration

Input: $\mathbf{A}$, $\mathbf{N}_k$, $\mathbf{B}$, $\mathbf{C}$, $\hat{\mathbf{A}}$, $\hat{\mathbf{N}}_k$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ dimensioned as in (4.32)
Output: $\hat{\mathbf{A}}$, $\hat{\mathbf{N}}_k$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ locally minimizing $||\mathbf{\Sigma}_B - \hat{\mathbf{\Sigma}}_B||_{\mathcal{H}_2}$

```
while (not converged) do
    Solve $\mathbf{AX} + \mathbf{X}\hat{\mathbf{A}}^T + \sum_{k=1}^m \mathbf{N}_k\mathbf{X}\hat{\mathbf{N}}_k^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0}$.
    Solve $\mathbf{A}^T\mathbf{Y} + \mathbf{Y}\hat{\mathbf{A}} + \sum_{k=1}^m \mathbf{N}_k^T\mathbf{Y}\hat{\mathbf{N}}_k - \mathbf{C}^T\hat{\mathbf{C}} = \mathbf{0}$.
    $\mathbf{V} = \operatorname{orth}(\mathbf{X})$, $\mathbf{W} = \operatorname{orth}(\mathbf{Y})$, $\mathbf{Z} = \mathbf{W}(\mathbf{V}^T\mathbf{W})^{-1}$
    $\hat{\mathbf{A}} = \mathbf{Z}^T\mathbf{AV},\ \hat{\mathbf{N}}_k = \mathbf{Z}^T\mathbf{N}_k\mathbf{V},\ \hat{\mathbf{B}} = \mathbf{Z}^T\mathbf{B},\ \hat{\mathbf{C}} = \mathbf{CV}$
end while
```

Let us now focus on Algorithm 4.3.1 which in each step constructs a reduced system $\hat{\mathbf{\Sigma}}$ by a Petrov-Galerkin type projection $\mathbf{P} = \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T$, determined by the solutions of the generalized Sylvester equations associated with the preceding system matrices. Upon convergence, we have the following result.

Theorem 4.3.3. *Assume that Algorithm 4.3.1 converges with convergence tolerance zero, where the convergence criterion measures the change of the eigenvalues of* $\hat{\mathbf{A}}$. *Then,* $\hat{\mathbf{A}}$, $\hat{\mathbf{N}}_k$, $\hat{\mathbf{B}}$, $\hat{\mathbf{C}}$ *fulfill the necessary* $\mathcal{H}_2$*-optimality conditions (4.24).*

Proof. Let $\bar{\mathbf{A}}$, $\bar{\mathbf{N}}_k$, $\bar{\mathbf{B}}$, $\bar{\mathbf{C}}$ denote the matrices corresponding to the next to last iteration step. Due to convergence, $\hat{\mathbf{\Sigma}}_B$ is a state space transformation of $\bar{\mathbf{\Sigma}}_B$, i.e., there exists $\mathbf{T} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ nonsingular, such that

$$\bar{\mathbf{A}} = \mathbf{T}^{-1}\hat{\mathbf{A}}\mathbf{T},\ \bar{\mathbf{N}}_k = \mathbf{T}^{-1}\hat{\mathbf{N}}_k\mathbf{T},\ \bar{\mathbf{B}} = \mathbf{T}^{-1}\hat{\mathbf{B}},\ \bar{\mathbf{C}} = \hat{\mathbf{C}}\mathbf{T}.$$

Furthermore, according to step 4 in Algorithm 4.3.1, we have

$$\mathbf{V} = \mathbf{XF}, \quad \mathbf{W} = \mathbf{YG},$$

with $\mathbf{F}, \mathbf{G} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ nonsingular. Thus, it holds that

$$(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T = (\mathbf{G}^T\mathbf{Y}^T\mathbf{XF})^{-1}\mathbf{G}^T\mathbf{Y}^T = \mathbf{F}^{-1}(\mathbf{Y}^T\mathbf{X})^{-1}\mathbf{Y}^T.$$

From step 2, it follows that

$$\mathbf{A}\mathbf{X} + \mathbf{X}\bar{\mathbf{A}}^T + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{X}\bar{\mathbf{N}}_k^T + \mathbf{B}\bar{\mathbf{B}}^T = \mathbf{0}.$$

Hence, we have that

$$\underbrace{\mathbf{F}^{-1}(\mathbf{Y}^T\mathbf{X})^{-1}\mathbf{Y}^T}_{=(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T} \left(\mathbf{A}\mathbf{X} + \mathbf{X}\bar{\mathbf{A}}^T + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{X}\bar{\mathbf{N}}_k^T + \mathbf{B}\bar{\mathbf{B}}^T\right)\mathbf{F} = \mathbf{0},$$

which implies that

$$\hat{\mathbf{A}} + \mathbf{F}^{-1}\mathbf{T}^T\hat{\mathbf{A}}^T\mathbf{T}^{-T}\mathbf{F} + \sum_{k=1}^{m} \hat{\mathbf{N}}_k\mathbf{F}^{-1}\mathbf{T}^T\hat{\mathbf{N}}_k^T\mathbf{T}^{-T}\mathbf{F} + \hat{\mathbf{B}}\hat{\mathbf{B}}^T\mathbf{T}^{-T}\mathbf{F} = \mathbf{0}.$$

Finally, we end up with

$$\hat{\mathbf{A}}\mathbf{F}^{-1}\mathbf{T}^T + \mathbf{F}^{-1}\mathbf{T}^T\hat{\mathbf{A}}^T + \sum_{k=1}^{m} \hat{\mathbf{N}}\mathbf{F}^{-1}\mathbf{T}^T\hat{\mathbf{N}}_k^T + \hat{\mathbf{B}}\hat{\mathbf{B}}^T = \mathbf{0}.$$

From the last line and the fact that we assumed the reduced system to be stable, the solution of the generalized Lyapunov equation is unique and we conclude that $\mathbf{P}_{22} = \mathbf{F}^{-1}\mathbf{T}^T$, were $\mathbf{P}_{22}$ is the lower right block from the partitioning of $\mathbf{P}_{err}$ in (4.25). Similarly, we obtain

$$\mathbf{A}^T\mathbf{Y} + \mathbf{Y}\bar{\mathbf{A}} + \sum_{k=1}^{m} \mathbf{N}_k^T\mathbf{Y}\bar{\mathbf{N}}_k - \mathbf{C}^T\bar{\mathbf{C}} = \mathbf{0}.$$

This leads to

$$\mathbf{F}^T\mathbf{X}^T\left(\mathbf{A}^T\mathbf{Y} + \mathbf{Y}\bar{\mathbf{A}} + \sum_{k=1}^{m} \mathbf{N}_k^T\mathbf{Y}\bar{\mathbf{N}}_k - \mathbf{C}^T\bar{\mathbf{C}}\right)(\mathbf{X}^T\mathbf{Y})^{-1}\mathbf{F}^{-T} = \mathbf{0},$$

which can be transformed into

$$\hat{\mathbf{A}}^T + \left(\mathbf{F}^T\mathbf{X}^T\mathbf{Y}\mathbf{T}^{-1}\hat{\mathbf{A}} + \sum_{k=1}^{m} \hat{\mathbf{N}}_k^T\mathbf{F}^T\mathbf{X}^T\mathbf{Y}\mathbf{T}^{-1}\hat{\mathbf{N}}_k - \hat{\mathbf{C}}^T\hat{\mathbf{C}} \right) \mathbf{T}(\mathbf{X}^T\mathbf{Y})^{-1}\mathbf{F}^{-T} = \mathbf{0}.$$

Thus it follows that

$$-\hat{\mathbf{A}}^T\mathbf{F}^T\mathbf{X}^T\mathbf{Y}\mathbf{T}^{-1} - \mathbf{F}^T\mathbf{X}^T\mathbf{Y}\mathbf{T}^{-1}\hat{\mathbf{A}} - \sum_{k=1}^{m} \hat{\mathbf{N}}_k^T\mathbf{F}^T\mathbf{X}^T\mathbf{Y}\mathbf{T}^{-1}\hat{\mathbf{N}}_k + \hat{\mathbf{C}}^T\hat{\mathbf{C}} = \mathbf{0}.$$

Again, the unique solution of the generalized Lyapunov equation of the reduced system satisfies $\mathbf{Q}_{22} = -\mathbf{F}^T\mathbf{X}^T\mathbf{Y}\mathbf{T}^{-1}$, with $\mathbf{Q}_{22}$ as defined in (4.25). Moreover, due to the symmetry of the solution, it follows that $\mathbf{Q}_{22} = -\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}$. Finally, we need the solutions of the generalized Sylvester equations arising in (4.26). However, it holds that the identity

$$\mathbf{A}\mathbf{X} + \mathbf{X}\bar{\mathbf{A}}^T + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{X}\bar{\mathbf{N}}_k^T + \mathbf{B}\bar{\mathbf{B}}^T = \mathbf{0}$$

is equivalent to

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{T}^T\hat{\mathbf{A}}^T\mathbf{T}^{-T} + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{X}\mathbf{T}^T\hat{\mathbf{N}}_k^T\mathbf{T}^{-T} + \mathbf{B}\hat{\mathbf{B}}^T\mathbf{T}^{-T} = \mathbf{0},$$

yielding

$$\mathbf{A}\mathbf{X}\mathbf{T}^T + \mathbf{X}\mathbf{T}^T\hat{\mathbf{A}}^T + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{X}\mathbf{T}^T\hat{\mathbf{N}}_k^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0}.$$

Here, we make use of the unique solution of the generalized Sylvester equation. Thus, it follows that $\mathbf{P}_{12} = \mathbf{X}\mathbf{T}^T$. Since the argumentation for the dual Sylvester equation is completely analogous, we skip the derivation that leads to $\mathbf{Q}_{12} = \mathbf{Y}\mathbf{T}^{-1}$. Let us now show the optimality conditions (4.24)

$$\mathbf{Q}_{12}^T\mathbf{P}_{12} + \mathbf{Q}_{22}\mathbf{P}_{22} = \mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{T}^T - \mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}\mathbf{F}^{-1}\mathbf{T}^T = \mathbf{0},$$

$$\begin{aligned}\mathbf{Q}_{22}\hat{\mathbf{N}}_k\mathbf{P}_{22}+\mathbf{Q}_{12}^T\mathbf{N}_k\mathbf{P}_{12}&=-\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}\hat{\mathbf{N}}_k\mathbf{F}^{-1}\mathbf{T}^T+\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{N}_k\mathbf{X}\mathbf{T}^T\\&=-\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{N}_k\mathbf{V}\mathbf{F}^{-1}\mathbf{T}^T+\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{N}_k\mathbf{X}\mathbf{T}^T\\&=-\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}\mathbf{F}^{-1}(\mathbf{Y}^T\mathbf{X})^{-1}\mathbf{Y}^T\mathbf{N}_k\mathbf{X}\mathbf{F}\mathbf{F}^{-1}\mathbf{T}^T+\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{N}_k\mathbf{X}\mathbf{T}^T=\mathbf{0},\end{aligned}$$

$$\begin{aligned}\mathbf{Q}_{12}^T\mathbf{B}+\mathbf{Q}_{22}\hat{\mathbf{B}}&=\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{B}-\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}\hat{\mathbf{B}}\\&=\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{B}-\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{B}\\&=\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{B}-\mathbf{T}^{-T}\mathbf{Y}^T\mathbf{X}\mathbf{F}\mathbf{F}^{-1}(\mathbf{Y}^T\mathbf{X})^{-1}\mathbf{Y}^T\mathbf{B}=\mathbf{0},\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{C}}\mathbf{P}_{22}-\mathbf{C}\mathbf{P}_{12}&=\hat{\mathbf{C}}\mathbf{F}^{-1}\mathbf{T}^T-\mathbf{C}\mathbf{X}\mathbf{T}^T=\mathbf{C}\mathbf{V}\mathbf{F}^{-1}\mathbf{T}^T-\mathbf{C}\mathbf{X}\mathbf{T}^T\\&=\mathbf{C}\mathbf{X}\mathbf{F}\mathbf{F}^{-1}\mathbf{T}^T-\mathbf{C}\mathbf{X}\mathbf{T}^T=\mathbf{0}.\end{aligned}$$

□

Remark 4.3.3. *Note that the two main steps of Algorithm 4.3.1 consist of finding solutions to generalized Sylvester equation of the form*

$$\mathbf{A}\mathbf{X}+\mathbf{X}\hat{\mathbf{A}}^T+\sum_{k=1}^{m}\mathbf{N}_k\mathbf{X}\hat{\mathbf{N}}_k^T+\mathbf{B}\hat{\mathbf{B}}^T=\mathbf{0},$$

determined by the large size matrix $\mathbf{A}$ *from the original system and the small size matrix* $\hat{\mathbf{A}}$ *from the reduced system. Similar to the generalized Lyapunov equations arising for bilinear control systems, solving a matrix equation of this type might still pose a severe challenge. However, one might think of considering the explicit system of linear equations, given by the Kronecker formulation*

$$\left(-\mathbf{I}\otimes\mathbf{A}-\hat{\mathbf{A}}\otimes\mathbf{I}-\sum_{k=1}^{m}\hat{\mathbf{N}}_k\otimes\mathbf{N}_k\right)\operatorname{vec}(\mathbf{X})=-\operatorname{vec}\left(\mathbf{B}\hat{\mathbf{B}}^T\right),$$

which can be solved by means of an iterative Krylov subspace based solver. As a preconditioning technique, one naturally might think of the corresponding simplified Sylvester operator appearing in the linear case (i.e. $\mathbf{N}_k=\mathbf{0}$*) which can be efficiently applied by means of a Schur decomposition of* $\hat{\mathbf{A}}$*, see [25].*

Remark 4.3.4. *As in the linear case, a typical convergence criterion for Algorithm 4.3.1 is the relative change of the eigenvalues of the system matrix* $\hat{\mathbf{A}}$.

Remark 4.3.5. *Note that Algorithm 4.3.1 generalizes a Sylvester equation based algorithm for* $\mathcal{H}_2$*-optimality (see [44]) and thus does not require diagonalizability of* $\hat{\mathbf{A}}$.

Let us turn our attention to an interpolation-based approach that can be directly derived from Algorithm 4.3.1. For a similar derivation in the linear case, see, e.g., [73, 132]. Again, let $\hat{\mathbf{A}} = \mathbf{R}\hat{\mathbf{\Lambda}}\mathbf{R}^{-1}$ denote the eigenvalue decomposition of the reduced system. As already mentioned before, the explicit solution of equation (4.32a) in vectorized form reads:

$$\begin{aligned}\operatorname{vec}(\mathbf{X}) &= \left(-\mathbf{I}\otimes\mathbf{A} - \hat{\mathbf{A}}\otimes\mathbf{I} - \sum_{k=1}^{m}\hat{\mathbf{N}}_k\otimes\mathbf{N}_k\right)^{-1}\operatorname{vec}\left(\mathbf{B}\hat{\mathbf{B}}^T\right)\\ &= \left(-\mathbf{I}\otimes\mathbf{A} - \hat{\mathbf{A}}\otimes\mathbf{I} - \sum_{k=1}^{m}\hat{\mathbf{N}}_k\otimes\mathbf{N}_k\right)^{-1}\left(\hat{\mathbf{B}}\otimes\mathbf{B}\right)\xi_m\\ &= \left[(\mathbf{R}\otimes\mathbf{I})\left(-\mathbf{I}\otimes\mathbf{A} - \hat{\mathbf{\Lambda}}\otimes\mathbf{I} - \sum_{k=1}^{m}\mathbf{R}^{-1}\hat{\mathbf{N}}_k\mathbf{R}\otimes\mathbf{N}_k\right)(\mathbf{R}^{-1}\otimes\mathbf{I})\right]^{-1}\left(\hat{\mathbf{B}}\otimes\mathbf{B}\right)\xi_m\\ &= (\mathbf{R}\otimes\mathbf{I})\underbrace{\left(-\mathbf{I}\otimes\mathbf{A} - \hat{\mathbf{\Lambda}}\otimes\mathbf{I} - \sum_{k=1}^{m}\mathbf{R}^{-1}\hat{\mathbf{N}}_k\mathbf{R}\otimes\mathbf{N}_k\right)^{-1}\left(\mathbf{R}^{-1}\hat{\mathbf{B}}\otimes\mathbf{B}\right)\xi_m}_{\operatorname{vec}(\mathbf{V})}.\end{aligned}$$

From the last line, we can now conclude that

$$(\mathbf{R}\otimes\mathbf{I})^{-1}\operatorname{vec}(\mathbf{X}) = \operatorname{vec}(\mathbf{V}) \text{ and hence } \mathbf{X}\mathbf{R}^{-T} = \mathbf{V}.$$

Similarly, starting from equation (4.32b), we obtain:

$$\begin{aligned}\operatorname{vec}(\mathbf{Y}) &= \left(\mathbf{I}\otimes\mathbf{A}^T + \hat{\mathbf{A}}^T\otimes\mathbf{I} + \sum_{k=1}^{m}\hat{\mathbf{N}}_k^T\otimes N_k^T\right)^{-1}\operatorname{vec}\left(\mathbf{C}^T\hat{\mathbf{C}}\right)\\ &= \left(\mathbf{I}\otimes\mathbf{A}^T + \hat{\mathbf{A}}^T\otimes\mathbf{I} + \sum_{k=1}^{m}\hat{\mathbf{N}}_k^T\otimes\mathbf{N}_k^T\right)^{-1}\left(\hat{\mathbf{C}}^T\otimes\mathbf{C}^T\right)\xi_p^T\\ &= \left[(\mathbf{R}^{-T}\otimes\mathbf{I})\left(-\mathbf{I}\otimes\mathbf{A}^T - \hat{\mathbf{\Lambda}}\otimes\mathbf{I} - \sum_{k=1}^{m}\mathbf{R}^T\hat{\mathbf{N}}_k^T\mathbf{R}^{-T}\otimes\mathbf{N}_k^T\right)(-\mathbf{R}^T\otimes\mathbf{I})\right]^{-1}\left(\hat{\mathbf{C}}^T\otimes\mathbf{C}^T\right)\xi_p^T\\ &= \left(-\mathbf{R}^{-T}\otimes\mathbf{I}\right)\operatorname{vec}(\mathbf{W}).\end{aligned}$$

Once again, this leads to

$$\left(-\mathbf{R}^{-T}\otimes\mathbf{I}\right)^{-1}\operatorname{vec}(\mathbf{Y}) = \operatorname{vec}(\mathbf{W}) \text{ and } \mathbf{Y}(-\mathbf{R}) = \mathbf{W},$$

where

$$\operatorname{vec}(\mathbf{W}) := \left(-\mathbf{I}\otimes\mathbf{A}^T - \hat{\mathbf{\Lambda}}\otimes\mathbf{I} - \sum_{k=1}^{m}\mathbf{R}^T\hat{\mathbf{N}}_k^T\mathbf{R}^{-T}\otimes\mathbf{N}_k^T\right)^{-1}\left(\mathbf{R}^T\hat{\mathbf{C}}^T\otimes\mathbf{C}^T\right)\xi_p^T.$$

According to the proof of Theorem 4.3.3, as long as span$\{\mathbf{X}\} \subset \mathbf{V}$ and span$\{\mathbf{Y}\} \subset \mathbf{W}$, we can ensure that the reduced system satisfies the necessary $\mathcal{H}_2$-optimality conditions. Hence, we have found an equivalent method which obviously extends IRKA to the bilinear case, see Algorithm 4.3.2.

Algorithm 4.3.2 Bilinear IRKA (BIRKA)

```
Input: $\mathbf{A}, \mathbf{N}_k, \mathbf{B}, \mathbf{C}, \hat{\mathbf{A}}, \hat{\mathbf{N}}_k, \hat{\mathbf{B}}, \hat{\mathbf{C}}$
Output: $\hat{\mathbf{A}}, \hat{\mathbf{N}}_k, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ locally minimizing $||\mathbf{\Sigma}_B - \hat{\mathbf{\Sigma}}_B||_{\mathcal{H}_2}$
while (not converged) do
    $\mathbf{R}\hat{\mathbf{\Lambda}}\mathbf{R}^{-1} = \hat{\mathbf{A}},\ \tilde{\mathbf{B}} = \hat{\mathbf{B}}^T\mathbf{R}^{-T},\ \tilde{\mathbf{C}} = \hat{\mathbf{C}}\mathbf{R},\ \tilde{\mathbf{N}}_k = \mathbf{R}^T\hat{\mathbf{N}}_k^T\mathbf{R}^{-T}$
    $\mathrm{vec}\left(\mathbf{V}\right) = \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m$
    $\mathrm{vec}\left(\mathbf{W}\right) = \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A}^T - \sum_{k=1}^m \tilde{\mathbf{N}}_k \otimes \mathbf{N}_k^T\right)^{-1} \left(\tilde{\mathbf{C}}^T \otimes \mathbf{C}^T\right) \xi_p$
    $\mathbf{V} = \mathrm{orth}\left(\mathbf{V}\right),\ \mathbf{W} = \mathrm{orth}\left(\mathbf{W}\right),\ \mathbf{Z} = \mathbf{W}(\mathbf{V}^T\mathbf{W})^{-1}$
    $\hat{\mathbf{A}} = \mathbf{Z}^T\mathbf{A}\mathbf{V},\ \hat{\mathbf{N}}_k = \mathbf{Z}^T\mathbf{N}_k\mathbf{V},\ \hat{\mathbf{B}} = \mathbf{Z}^T\mathbf{B},\ \hat{\mathbf{C}} = \mathbf{C}\mathbf{V}$
end while
```

Finally, we point out the equivalence between the optimality conditions (4.24) and (4.28). For this, we need the following projection-based identity.

Lemma 4.3.3. *Let* $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times \hat{n}}$ *be matrices of full rank* $\hat{n}$.
(a) Let $\mathbf{z} \in \mathrm{span}\{\mathrm{vec}\left(\mathbf{V}\right)\}$. *Then* $\left(\mathbf{I} \otimes \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right)\mathbf{z} = \mathbf{z}$.
(b) Let $\mathbf{z} \in \mathrm{span}\{\mathrm{vec}\left(\mathbf{W}\right)\}$. *Then* $\mathbf{z}^T\left(\mathbf{I} \otimes \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) = \mathbf{z}^T$.

Proof. By assumption, there exists $\mathbf{x} \in \mathbb{R}^{n \cdot \hat{n}}$ s.t.

$$\begin{aligned}\left(\mathbf{I} \otimes \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right)\mathbf{z} &= \left(\mathbf{I} \otimes \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right)\mathrm{vec}\left(\mathbf{V}\right)\mathbf{x} \\ &= \mathrm{vec}\left(\mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{V}\right)\mathbf{x} = \mathrm{vec}\left(\mathbf{V}\right)\mathbf{x} = \mathbf{z}.\end{aligned}$$

The proof of the second statement is based on exactly the same arguments. □

Theorem 4.3.4. *Assume that Algorithm 4.3.2 converges. Then* $\hat{\mathbf{A}}, \hat{\mathbf{N}}_k, \hat{\mathbf{B}}, \hat{\mathbf{C}}$ *fulfill the necessary interpolation-based* $\mathcal{H}_2$*-optimality conditions.*

Proof. Since the only difference in proving the conditions (4.28) – (4.31) lies in using statement (b) of Lemma 4.3.3 and the combination of both (a) and (b), respectively, we restrict ourselves to showing the optimality condition (4.28):

$$\xi_p^T\left(\mathbf{e}_i\mathbf{e}_j^T \otimes \hat{\mathbf{C}}\right)\left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \hat{\mathbf{A}} - \sum_{k=1}^m \tilde{\mathbf{N}}_k^T \otimes \hat{\mathbf{N}}_k\right)^{-1}\left(\tilde{\mathbf{B}}^T \otimes \hat{\mathbf{B}}\right)\xi_m$$

$$
\begin{aligned}
&= \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{CV}\right) \times \\
&\quad \left[\left(\mathbf{I} \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right) (\mathbf{I} \otimes \mathbf{V})\right]^{-1} \times \\
&\quad \left(\tilde{\mathbf{B}}^T \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{B}\right) \xi_m \\
&= \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{CV}\right) \times \\
&\quad \left[\left(\mathbf{I} \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right) (\mathbf{I} \otimes \mathbf{V})\right]^{-1} \times \\
&\quad \left(\mathbf{I} \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right) \\
&\quad \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m \\
&\overset{(4.3.3a)}{=} \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{CV}\right) \times \\
&\quad \left[\left(\mathbf{I} \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right) (\mathbf{I} \otimes \mathbf{V})\right]^{-1} \times \\
&\quad \left(\mathbf{I} \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right) \times \\
&\quad \left(\mathbf{I} \otimes \mathbf{V}(\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m \\
&= \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{CV}\right) \left(\mathbf{I} \otimes (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\right) \times \\
&\quad \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m \\
&= \xi_p^T \left(\mathbf{e}_i\mathbf{e}_j^T \otimes \mathbf{C}\right) \left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m} \tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right)^{-1} \left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right) \xi_m.
\end{aligned}
$$

□

Remark 4.3.6. *Note that analogously to Lemma 4.3.2, it is also possible to construct the matrices appearing in Algorithm 4.3.2 as the limit of an infinite series of linear IRKA type computations. For this, in each iteration, one starts with*

$$
\mathbf{V}_i^1 = \left(-\hat{\lambda}_i \mathbf{I} - \mathbf{A}\right)^{-1} \mathbf{B}\tilde{\mathbf{B}}_i,
$$

and continues with

$$\mathbf{V}_i^j = \left(-\hat{\lambda}_i \mathbf{I} - \mathbf{A}\right)^{-1} \left(\sum_{k=1}^{m} \mathbf{N}_k \mathbf{V}^{j-1} (\tilde{\mathbf{N}}_k)_i \right).$$

The actual projection matrix $\mathbf{V}$ *then is given as* $\mathbf{V} = \sum_{j=1}^{\infty} \mathbf{V}^j$. *A dual derivation obviously yields the projection matrix* $\mathbf{W}$. *At this point, the interpolatory interpretation of the proposed algorithm is seen once more. The construction of each* $\mathbf{V}^j$ *in a way corresponds to the tangential interpolation framework appearing for linear dynamical systems with multiple inputs and multiple outputs. Furthermore, similar to the statement in Remark 4.3.3, another way of constructing the projection matrices is given by the use of an iterative solver. This solver might be implemented with a natural preconditioner determined by the simplified and underlying linear problem which can be easily tackled by IRKA. Since the latter method is computationally more efficient than the Schur decomposition based approach discussed in [25], the reformulation of Algorithm 4.3.1 into Algorithm 4.3.2 might turn out to be profitable for practicable computations and should be a topic of further research.*

A crucial observation is that if in Algorithm 4.3.2, the matrices $\mathbf{V}$ and $\mathbf{W}$ are replaced by the first q terms of the previously mentioned iteration, i.e., $\mathbf{V} = \sum_{j=1}^{q} \mathbf{V}_j$ and $\mathbf{W} = \sum_{j=1}^{q} \mathbf{W}_j$, one still can construct a reduced-order bilinear system that is optimal with respect to a slightly modified $\mathcal{H}_2$-norm corresponding to an underlying polynomial system of order q. Here, we refer to [56, Chapter 4] where a more detailed discussion on a numerically efficient algorithm (TB-IRKA) is given. In particular, as has been reported in [56], the latter algorithm usually outperforms Algorithm 4.3.2 for larger dimensions $\hat{n}$ of the reduced-order systems when it comes to computational efficiency.

Remark 4.3.7. *Note that the numerical efficiency of both Algorithm 4.3.1 and Algorithm 4.3.2 heavily depends on the number of iterations needed until the relative change of the eigenvalues of the system matrix* $\hat{\mathbf{A}}$ *approaches zero. As has already been shown for the linear case (cf. [73]), IRKA is a simplified Newton iteration where the Jacobian matrix is neglected. Obviously, this means that there exist a lot of examples where both algorithms diverge. Nevertheless, recently there have been some first convergence results for symmetric state space systems, see [58]. However, at this point it seems very hard to generalize those ideas to the bilinear case.*

4.3.4 Generalizations to other cases

Before we test the efficiency of our new method by means of several numerical examples, let us discuss differences that occur for other types of bilinear control systems. In particular, so far we have focused on standard continuous-time systems. Below, we present a few important details necessary for treating generalized state space systems as well as discrete-time systems. Similar to the theory from Chapter 3 for linear systems, both cases can be tackled by the previous tools without larger modifications.

Generalized state space systems

As is common for linear control systems, the spatial discretization of a nonlinear PDE often results in a mass matrix $\mathbf{E} \neq \mathbf{I}$ as well. As a consequence, we obtain a generalized state space system of the form

$$\boldsymbol{\Sigma}_B : \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \sum\limits_{k=1}^{m} \mathbf{N}_k \mathbf{x}(t) u_k(t) + \mathbf{B}\mathbf{u}(t), \\ \quad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \end{cases} \tag{4.33}$$

with dimensions as in (4.1). For the derivation of the corresponding $\mathcal{H}_2$-optimality conditions it obviously suffices to invert the mass matrix $\mathbf{E}$ and apply the results from the previous subsections. Since this in general will destroy the sparsity of the matrices, we certainly prefer to work with the original system data. As a first step, it is important to note that the computation of the $\mathcal{H}_2$-norm itself does not change, only the underlying generalized Lyapunov equations. To be precise, we have the following result.

Proposition 4.3.2. *Let $\boldsymbol{\Sigma}_B$ be a stable generalized bilinear system as in (4.33). Then it holds that*

$$||\boldsymbol{\Sigma}_B||_{\mathcal{H}_2}^2 = \operatorname{tr}\left(\mathbf{C}\mathbf{P}\mathbf{C}^T\right) = \operatorname{tr}\left(\mathbf{B}^T\mathbf{Q}\mathbf{B}\right),$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the solutions of the generalized Lyapunov equations

$$\mathbf{A}\mathbf{P}\mathbf{E}^T + \mathbf{E}\mathbf{P}\mathbf{A}^T + \sum_{k=1}^{m} \mathbf{N}_k\mathbf{P}\mathbf{N}_k^T + \mathbf{B}\mathbf{B}^T = \mathbf{0},$$
$$\mathbf{A}^T\mathbf{Q}\mathbf{E} + \mathbf{E}^T\mathbf{Q}\mathbf{A} + \sum_{k=1}^{m} \mathbf{N}_k^T\mathbf{Q}\mathbf{N}_k + \mathbf{C}^T\mathbf{C} = \mathbf{0}.$$

Proof. The first part immediately follows from the fact that $\mathbf{P}$ is the solution of

$$(\mathbf{E}^{-1}\mathbf{A})\mathbf{P} + \mathbf{P}(\mathbf{E}^{-1}\mathbf{A})^T + \sum_{k=1}^{m}(\mathbf{E}^{-1}\mathbf{N}_k)\mathbf{P}(\mathbf{E}^{-1}\mathbf{N}_k)^T + (\mathbf{E}^{-1}\mathbf{B})(\mathbf{E}^{-1}\mathbf{B})^T = \mathbf{0},$$

which is the Lyapunov equation arising for the standard state space system that is obtained after the theoretical inversion of $\mathbf{E}$.

For the second part, assume that $\tilde{\mathbf{Q}}$ is the solution of the dual Lyapunov equation for the standard state space system, i.e.,

$$(\mathbf{E}^{-1}\mathbf{A})^T\tilde{\mathbf{Q}} + \tilde{\mathbf{Q}}(\mathbf{E}^{-1}\mathbf{A}) + \sum_{k=1}^{m}(\mathbf{E}^{-1}\mathbf{N}_k)^T\tilde{\mathbf{Q}}(\mathbf{E}^{-1}\mathbf{N}_m) + \mathbf{C}^T\mathbf{C} = \mathbf{0}.$$

Hence, it holds $||\mathbf{\Sigma}_B||^2_{\mathcal{H}_2} = \operatorname{tr}\left(\mathbf{B}^T\mathbf{E}^{-T}\tilde{\mathbf{Q}}\mathbf{E}^{-1}\mathbf{B}\right)$. By introducing an artificial identity $\mathbf{I} = \mathbf{E}^{-1}\mathbf{E} = \mathbf{E}^T\mathbf{E}^{-T}$, the above equation can be rewritten as

$$(\mathbf{E}^{-1}\mathbf{A})^T\tilde{\mathbf{Q}}\mathbf{E}^{-1}\mathbf{E} + \mathbf{E}^T\mathbf{E}^{-T}\tilde{\mathbf{Q}}(\mathbf{E}^{-1}\mathbf{A}) + \sum_{k=1}^{m}(\mathbf{E}^{-1}\mathbf{N}_k)^T\tilde{\mathbf{Q}}(\mathbf{E}^{-1}\mathbf{N}_k) + \mathbf{C}^T\mathbf{C} = \mathbf{0},$$

which implies that $\mathbf{E}^{-T}\tilde{\mathbf{Q}}\mathbf{E}^{-1} = \mathbf{Q}$ by the uniqueness of the solution. □

Next, we state the generalized Wilson $\mathcal{H}_2$-optimality conditions for bilinear control systems in terms of the generalized system Gramians from Proposition 4.3.2.

Proposition 4.3.3. *Let $\mathbf{\Sigma}_B$ be a stable generalized bilinear system as in (4.33) and let $\hat{\mathbf{\Sigma}}_B$ be a locally $\mathcal{H}_2$-optimal stable reduced-order generalized bilinear system* $(\hat{\mathbf{E}}; \hat{\mathbf{A}}, \hat{\mathbf{N}}_k, \hat{\mathbf{B}}, \hat{\mathbf{C}})$. *Then it holds that*

$$\mathbf{Q}_{12}^T\mathbf{E}\mathbf{P}_{12} + \mathbf{Q}_{22}\hat{\mathbf{E}}\mathbf{P}_{22} = \mathbf{0}, \tag{4.35a}$$

$$\mathbf{Q}_{22}\hat{\mathbf{N}}_k\mathbf{P}_{22} + \mathbf{Q}_{12}^T\mathbf{N}_k\mathbf{P}_{12} = \mathbf{0}, \tag{4.35b}$$

$$\mathbf{Q}_{12}^T\mathbf{B} + \mathbf{Q}_{22}\hat{\mathbf{B}} = \mathbf{0}, \tag{4.35c}$$

$$\hat{\mathbf{C}}\mathbf{P}_{22} - \mathbf{C}\mathbf{P}_{12} = \mathbf{0}, \tag{4.35d}$$

where

$$\mathbf{P}_{err} = \begin{bmatrix}\mathbf{P}_{11} & \mathbf{P}_{12}\\ \mathbf{P}_{12}^T & \mathbf{P}_{22}\end{bmatrix}, \quad \mathbf{Q}_{err} = \begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{12}^T & \mathbf{Q}_{22}\end{bmatrix} \tag{4.36}$$

are the solutions of the generalized Lyapunov equations

$$\mathbf{A}_{err}\mathbf{P}_{err}\mathbf{E}_{err}^T + \mathbf{E}_{err}\mathbf{P}_{err}\mathbf{A}_{err}^T + \sum_{k=1}^{m}\mathbf{N}_{err,k}\mathbf{P}_{err}\mathbf{N}_{err,k}^T + \mathbf{B}_{err}\mathbf{B}_{err}^T = \mathbf{0}, \tag{4.37a}$$

$$\mathbf{A}_{err}^T\mathbf{Q}_{err}\mathbf{E}_{err} + \mathbf{E}_{err}^T\mathbf{Q}_{err}\mathbf{A}_{err} + \sum_{k=1}^{m}\mathbf{N}_{err,k}^T\mathbf{Q}_{err}\mathbf{N}_{err,k} + \mathbf{C}_{err}^T\mathbf{C}_{err} = \mathbf{0} \tag{4.37b}$$

and $\mathbf{E}_{err} = \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{E}} \end{bmatrix}$ *denotes the mass matrix of the error system.*

Proof. Due the proof of Proposition 4.3.2, we know that $\mathbf{P}_{err}$ and $\tilde{\mathbf{Q}}_{err} = \mathbf{E}_{err}^T \mathbf{Q}_{err} \mathbf{E}_{err}$ are the solutions of the Lyapunov equations for the standard state space error system that are obtained after inversion of $\mathbf{E}_{err}$. According to the theory for standard state space systems, the first optimality condition is

$$\tilde{\mathbf{Q}}_{12}^T \mathbf{P}_{12} + \tilde{\mathbf{Q}}_{22} \mathbf{P}_{22} = \mathbf{0}.$$

Taking into account the relation between $\tilde{\mathbf{Q}}$ and $\mathbf{Q}$, this now implies

$$\hat{\mathbf{E}}^T \mathbf{Q}_{12}^T \mathbf{E} \mathbf{P}_{12} + \hat{\mathbf{E}}^T \mathbf{Q}_{22} \hat{\mathbf{E}} \mathbf{P}_{22} = \mathbf{0}.$$

Since we assumed $\hat{\mathbf{\Sigma}}$ to be stable, it follows that $\hat{\mathbf{E}}^T$ is invertible, showing (4.35a). Similarly, one can prove the remaining optimality conditions. □

Remark 4.3.8. *For later purposes, it is important to note that (4.35a) can alternatively be replaced by*

$$\mathbf{Q}_{12}^T \mathbf{A} \mathbf{P}_{12} + \mathbf{Q}_{22} \hat{\mathbf{A}} \mathbf{P}_{22} = \mathbf{0}.$$

This is easily seen by computing $(4.35a)\hat{\mathbf{A}}^T + (4.35b)\hat{\mathbf{N}}^T + (4.35c)\hat{\mathbf{B}}^T$ *and using the fact that* $\mathbf{P}_{12}$ *and* $\mathbf{P}_{22}$ *are the solutions of the generalized Lyapunov and Sylvester equations, respectively.*

For readers with a background in linear control theory and $\mathcal{H}_2$-optimal model reduction, the extension of the optimality conditions to generalized state space systems certainly is not surprising. Nevertheless, especially for the subsequently following discrete-time case, the derivation of optimality conditions simplifies significantly. In summary, we keep in mind that even for generalized state space systems, we do not have to invert the mass matrix in order to construct a locally $\mathcal{H}_2$-optimal ROM. Since the necessary modifications for a suitable algorithm should be obvious by now, we refrain from a more detailed discussion. Instead, we turn our attention to discrete-time systems.

Discrete-time systems

As a further generalization of $\mathcal{H}_2$-theory, let us have a look at bilinear discrete-time control systems of the form

$$\mathbf{\Sigma}_{d,B}: \begin{cases} \mathbf{x}(k+1) = \mathbf{A}_d\mathbf{x}(k) + \sum_{k=1}^{m} \mathbf{N}_{d,k}\mathbf{x}(k)u_k(k) + \mathbf{B}_d\mathbf{u}(k), \\ \quad \mathbf{y}(k) = \mathbf{C}_d\mathbf{x}(k), \quad \mathbf{x}(0) = \mathbf{x}_0, \end{cases} \tag{4.38}$$

with dimensions again as in (4.1). While for continuous-time systems, the Volterra series representation plays a central role in analyzing system properties, in the discrete-time case, there exists an explicit solution formula as well, see [134]. However, the results are only of little importance for this thesis and we thus refer to the given reference. Instead, we give the definition of the h_2-norm from [23].

Definition 4.3.1. *Let $\mathbf{\Sigma}_{d,B}$ be a discrete-time bilinear system and let*

$$\mathbf{H}_i(z_1,\dots,z_i) = \mathbf{C}_d \left(\prod_{j=0}^{i-2} \mathbf{I}_{m^j} \otimes (z_{i-j}\mathbf{I} - \mathbf{A}_d)^{-1}\mathcal{N}_d \right) \left(\mathbf{I}_{m^{i-1}} \otimes (z_1\mathbf{I} - \mathbf{A}_d)^{-1}\mathbf{B}_d \right),$$

with $\mathcal{N} = \left[\mathbf{N}_{d,1},\dots,\mathbf{N}_{d,m}\right]$, denote its generalized j-th transfer function resulting from a multivariate Z-transform. Then we define

$$||\mathbf{\Sigma}||_{h_2}^2 = \operatorname{tr}\left(\sum_{j=1}^{\infty} \int_0^{2\pi} \cdots \int_0^{2\pi} \frac{1}{2\pi}^k \overline{\mathbf{H}_j(e^{i\theta_1},\dots,e^{i\theta_j})} \left(\mathbf{H}_j(e^{i\theta_1},\dots,e^{i\theta_j})\right)^T \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_j \right). \tag{4.39}$$

Along the theory for linear control systems, the computation via the solutions of special matrix equations again is possible and is summarized in the following Lemma, see [23].

Lemma 4.3.4. *Let $\mathbf{P}_d$ and $\mathbf{Q}_d$ be the solutions of the generalized Stein equations*

$$\mathbf{A}_d\mathbf{P}_d\mathbf{A}_d^T + \sum_{k=1}^{m} \mathbf{N}_{d,k}\mathbf{P}_d\mathbf{N}_{d,k}^T + \mathbf{B}_d\mathbf{B}_d^T = \mathbf{P}_d, \tag{4.40a}$$

$$\mathbf{A}_d^T\mathbf{Q}_d\mathbf{A}_d + \sum_{k=1}^{m} \mathbf{N}_{d,k}^T\mathbf{Q}_d\mathbf{N}_{d,k} + \mathbf{C}_d^T\mathbf{C}_d = \mathbf{Q}_d. \tag{4.40b}$$

Then the h_2-norm of $\mathbf{\Sigma}_{d,B}$ can be computed as

$$||\mathbf{\Sigma}_{d,B}||_{h_2}^2 = \operatorname{tr}\left(\mathbf{C}_d\mathbf{P}_d\mathbf{C}_d^T\right) = \operatorname{tr}\left(\mathbf{B}_d^T\mathbf{Q}_d\mathbf{B}_d\right).$$

With the above stated results, we can already derive necessary h_2-optimality conditions. As we have done in Chapter 3, the crucial trick is to transform the Stein operator into an equivalent Lyapunov type operator. For example, we have

$$\mathbf{A}_d\mathbf{P}_d\mathbf{A}_d^T + \sum_{k=1}^m \mathbf{N}_{d,k}\mathbf{P}_d\mathbf{N}_{d,k}^T - \mathbf{P}_d = \mathbf{A}\mathbf{P}_d\mathbf{E}^T + \mathbf{E}\mathbf{P}_d\mathbf{A}^T + \sum_{k=1}^m \mathbf{N}_{d,k}\mathbf{P}_d\mathbf{N}_{d,k}^T,$$

where $\mathbf{A} = \mathbf{A}_d - \mathbf{I}$ and $\mathbf{E} = \frac{1}{2}(\mathbf{A}_d + \mathbf{I})$. The same argument obviously holds true for $\mathbf{Q}_d$. Assume now that we are faced with a stable discrete-time bilinear system. While one again might argue about a precise definition of stability in context of bilinear control systems, here we restrict ourselves to the case where $\mathbf{A}_d$ has eigenvalues only in the interior of the unit disc $\mathbb{D}$. Moreover, we assume that the solutions $\mathbf{P}_d$ and $\mathbf{Q}_d$ are positive definite. For a more detailed discussion on stability issues, we additionally refer to [119]. Of course, discrete-time stability of $\mathbf{A}_d$ implies continuous-time stability of the transformed matrix pencil $(\mathbf{A}, \mathbf{E})$. Hence we have

$$||\mathbf{\Sigma}_{d,B}||_{h_2}^2 = ||\mathbf{\Sigma}_B||_{\mathcal{H}_2}^2$$

and we thus can apply $\mathcal{H}_2$-optimality theory for continuous-time generalized state space system. As a generalization of the discrete-time Wilson optimality conditions (3.9), we obtain the following.

Corollary 4.3.2. *Let $\mathbf{\Sigma}_{d,B}$ be a stable discrete-time bilinear system as in (4.38). Let $\hat{\mathbf{\Sigma}}_{d,B} = (\hat{\mathbf{A}}_d, \hat{\mathbf{N}}_{d,k}, \hat{\mathbf{B}}_d, \hat{\mathbf{C}}_d)$ be a locally h_2-optimal stable reduced-order discrete-time bilinear system. Then it holds that*

$$\mathbf{Q}_{d,12}^T\mathbf{P}_{d,12} + \mathbf{Q}_{d,22}\mathbf{P}_{d,22} = \mathbf{0}, \tag{4.41a}$$

$$\mathbf{Q}_{d,22}\hat{\mathbf{N}}_{d,k}\mathbf{P}_{d,22} + \mathbf{Q}_{d,12}^T\mathbf{N}_{d,k}\mathbf{P}_{d,12} = \mathbf{0}, \tag{4.41b}$$

$$\mathbf{Q}_{d,12}^T\mathbf{B}_d + \mathbf{Q}_{d,22}\hat{\mathbf{B}}_d = \mathbf{0}, \tag{4.41c}$$

$$\hat{\mathbf{C}}_d\mathbf{P}_{d,22} - \mathbf{C}_d\mathbf{P}_{d,12} = \mathbf{0}, \tag{4.41d}$$

where

$$\mathbf{P}_{d,err} = \begin{bmatrix} \mathbf{P}_{d,11} & \mathbf{P}_{d,12} \\ \mathbf{P}_{d,12}^T & \mathbf{P}_{d,22} \end{bmatrix}, \quad \mathbf{Q}_{d,err} = \begin{bmatrix} \mathbf{Q}_{d,11} & \mathbf{Q}_{d,12} \\ \mathbf{Q}_{d,12}^T & \mathbf{Q}_{d,22} \end{bmatrix}, \tag{4.42}$$

are the solutions of the generalized Stein equations

$$\mathbf{A}_{d,err}\mathbf{P}_{d,err}\mathbf{A}_{d,err}^T + \sum_{k=1}^{m}\mathbf{N}_{d,err,k}\mathbf{P}_{d,err}\mathbf{N}_{d,err,k}^T + \mathbf{B}_{d,err}\mathbf{B}_{d,err}^T = \mathbf{P}_{d,err}, \tag{4.43a}$$

$$\mathbf{A}_{d,errd}^T\mathbf{Q}_{d,err}\mathbf{A}_{d,err} + \sum_{k=1}^{m}\mathbf{N}_{d,err,k}^T\mathbf{Q}_{d,err}\mathbf{N}_{d,err,k} + \mathbf{C}_{d,err}^T\mathbf{C}_{d,err} = \mathbf{Q}_{d,err}. \tag{4.43b}$$

Proof. Due to the transformation, we know that h_2-optimality of $\mathbf{\Sigma}_{d,B}$ implies $\mathcal{H}_2$-optimality of $\mathbf{\Sigma}_B$. According to Proposition 4.3.3 and Remark 4.3.8, we know that if the transformed system $\mathbf{\Sigma}_B$ is locally $\mathcal{H}_2$-optimal, it follows that

$$\mathbf{Q}_{d,12}\mathbf{E}\mathbf{P}_{d,12} + \mathbf{Q}_{d,22}\hat{\mathbf{E}}\mathbf{P}_{d,22} = \mathbf{0}, \tag{4.44a}$$

$$\mathbf{Q}_{d,12}\mathbf{A}\mathbf{P}_{d,12} + \mathbf{Q}_{d,22}\hat{\mathbf{A}}\mathbf{P}_{d,22} = \mathbf{0}. \tag{4.44b}$$

Inserting $\mathbf{E} = \frac{1}{2}(\mathbf{A}_d + \mathbf{I})$ and $\mathbf{A} = \mathbf{A}_d - \mathbf{I}$ and computing $2(4.44a) - (4.44b)$ yields (4.41a). Conditions (4.41b) – (4.41d) immediately follow from the fact that $\mathbf{P}_{d,err}$ and $\mathbf{Q}_{d,err}$ are also the solutions for the transformed continuous-time error system. □

Finally, we briefly explain how to extend the interpolatory optimality conditions (4.28)-(4.31) to the discrete-time case. For the continuous-time case, we basically obtained terms of the form

$$\left(\tilde{\mathbf{C}} \otimes \mathbf{C}\right)\left(-\hat{\mathbf{\Lambda}} \otimes \mathbf{I} - \mathbf{I} \otimes \mathbf{A} - \sum_{k=1}^{m}\tilde{\mathbf{N}}_k^T \otimes \mathbf{N}_k\right)^{-1}\left(\tilde{\mathbf{B}}^T \otimes \mathbf{B}\right),$$

where the term $-\hat{\mathbf{\Lambda}}$ contains the mirror images of the reduced system poles with respect to the imaginary axis. Taking into account linear h_2-optimality theory, for discrete-time systems, we expect a similar term including a matrix $\hat{\mathbf{\Lambda}}^{-1}$, reflecting the mirror images with respect to the unit circle. Let us have a look at (4.41d). If we multiply the equation with $\hat{\mathbf{A}}_d^T\hat{\mathbf{C}}_d^T$, the second term (neglecting the sign) is of the form

$$\mathbf{C}_d\mathbf{P}_{d,12}\hat{\mathbf{A}}_d^T\hat{\mathbf{C}}_d^T,$$

which in vectorized notation reads

$$\left(\hat{\mathbf{C}}_d\hat{\mathbf{A}}_d \otimes \mathbf{C}_d\right)\operatorname{vec}\left(\mathbf{P}_{d,12}\right).$$

Making use of the explicit solution formula for $\mathbf{P}_{d,12}$, we obtain

$$(\hat{\mathbf{C}}_d\hat{\mathbf{A}}_d \otimes \mathbf{C}_d)\left(\mathbf{I}\otimes\mathbf{I} - \hat{\mathbf{A}}_d \otimes \mathbf{A}_d - \sum_{k=1}^{m} \hat{\mathbf{N}}_{d,k} \otimes \mathbf{N}_{d,k}\right)^{-1} \operatorname{vec}\left(\mathbf{B}_d\hat{\mathbf{B}}_d^T\right).$$

Assuming that $\hat{\mathbf{A}}_d = \mathbf{R}\hat{\mathbf{\Lambda}}\mathbf{R}^{-1}$ is the eigenvalue decomposition with $\hat{\mathbf{\Lambda}}$ being nonsingular, we conclude that

$$\begin{aligned}(\hat{\mathbf{C}}_d \otimes \mathbf{C}_d)&\left(\hat{\mathbf{A}}_d^{-1}\otimes\mathbf{I} - \mathbf{I} \otimes \mathbf{A}_d - \sum_{k=1}^{m} \hat{\mathbf{N}}_{d,k}\hat{\mathbf{A}}_d^{-1} \otimes \mathbf{N}_{d,k}\right)^{-1} \operatorname{vec}\left(\mathbf{B}_d\hat{\mathbf{B}}_d^T\right)\\ &= (\hat{\mathbf{C}}_d\mathbf{R} \otimes \mathbf{C}_d)\left(\hat{\mathbf{\Lambda}}^{-1}\otimes\mathbf{I} - \mathbf{I} \otimes \mathbf{A}_d - \sum_{k=1}^{m} \mathbf{R}^{-1}\hat{\mathbf{N}}_{d,k}\mathbf{R}\hat{\mathbf{\Lambda}}^{-1} \otimes \mathbf{N}_{d,k}\right)^{-1} \operatorname{vec}\left(\mathbf{B}_d\hat{\mathbf{B}}_d^T\mathbf{R}^{-T}\right).\end{aligned}$$

Hence, if we denote $\tilde{\mathbf{C}} = \hat{\mathbf{C}}_d\mathbf{R}$, $\tilde{\mathbf{B}}_d = \mathbf{B}_d^T\mathbf{R}^{-T}$ and $\tilde{\mathbf{N}}_k = \mathbf{R}^T\hat{\mathbf{N}}_{d,k}^T\mathbf{R}^{-T}$, we obtain the desired expression for the interpolation-based optimality conditions. Interestingly enough, in contrast to the continuous-time case, note that there is an additional term $\mathbf{\Lambda}^{-1}$ within the bracket. Moreover, for vanishing $\mathbf{N}_{d,k}$, we recognize the transfer function character that arises for linear systems. In particular, basically the terms include the evaluation of the transfer function at the mirror images of the reduced system poles with respect to the unit circle. Indeed, those are included in the matrix $\hat{\mathbf{\Lambda}}^{-1}$. At this point we do not discuss further details such as, e.g., suitable iterative algorithms since the main focus of this thesis is on continuous-time systems. Moreover, after the previous results it should be rather straightforward to extend the corresponding methods to the discrete-time case as well.

4.3.5 Numerical examples

In this section, we study several applications of bilinear control systems and discuss the performance of the approaches proposed above. As we have already mentioned, the method of balanced truncation for bilinear systems is connected to the generalized controllability Gramian and the reachability Gramian of the underlying system, respectively. Hence, similar to the linear case, we expect this method to yield reduced models with small relative $\mathcal{H}_2$-error as well and we thus use it for a comparison with our algorithms. Due to the theoretical equivalence of Algorithm 4.3.1 and Algorithm 4.3.2, we mainly report the results for the latter case. Nevertheless, we remark that if iterative solvers are included in numerical simulations, there might occur differences with respect to robustness and speed of convergence which will be subject of further studies. However, here we compute the projection matrices $\mathbf{V}$ and $\mathbf{W}$ by solving the large systems of linear equations explicitly instead of using more sophisticated iterative techniques.

Finally, all Lyapunov equations are solved by the method proposed in [43] which allows for solving medium-sized systems. However, in the next section, we give a more detailed insight into the method of balanced truncation for bilinear systems and propose some techniques that also allow solving very large-scale systems.

All simulations were performed on an Intel® Core$^{\mathrm{TM}}$i7 CPU 920, 8 MB cache, 12 GB RAM, openSUSE Linux 11.1 (x86_64), MATLAB Version 7.11.0.584 (R2010b) 64-bit (glnxa64).

An interconnected power system

The first application is a model for two interconnected power systems which can be described by a bilinear system of state dimension 17. The hydro unit as well as the steam unit each can be controlled by two input variations resulting in a system with 4 inputs and 3 outputs. Since we are only interested in the reduction process, we refer to [2] where a detailed derivation of the dynamics can be found. We have successively reduced the original model to systems varying from $\hat{n} = 1, \dots, 16$ state variables. A comparison of the associated relative $\mathcal{H}_2$-norm of the error system between our approaches and the method of balanced truncation is shown in Figure 4.3.

As one can see, except for the case $\hat{n} = 2$, we always obtain better results with the new technique. The initialization of Algorithm 4.3.1 and Algorithm 4.3.2 is done completely at random, using arbitrary reduced-order models, interpolations points and tangential directions, respectively. For both algorithms we use the same initialization and, as shown in Figure 4.3, obtain the exact same results. This underscores the theoretical equivalence and thus justifies to concentrate on Algorithm 4.3.2. As indicated in Figure 4.4, for system dimensions $\hat{n} = 5, 10, 14$, the algorithm does not always converge in a few steps. On the other hand, we see that the relative $\mathcal{H}_2$-error stagnates very fast. Hence, the stopping criterion, which is chosen as the relative change of the norm of the poles of the reduced system, becomes smaller than $\sqrt{\epsilon}$, where ϵ denotes machine precision, might be too restrictive. Again, finding appropriate criteria seems to be a reasonable topic of further research.

Fokker-Planck equation

The second example is the stochastic control example from Chapter 1. Recall that we can describe the dynamics by its underlying probability distribution function such that

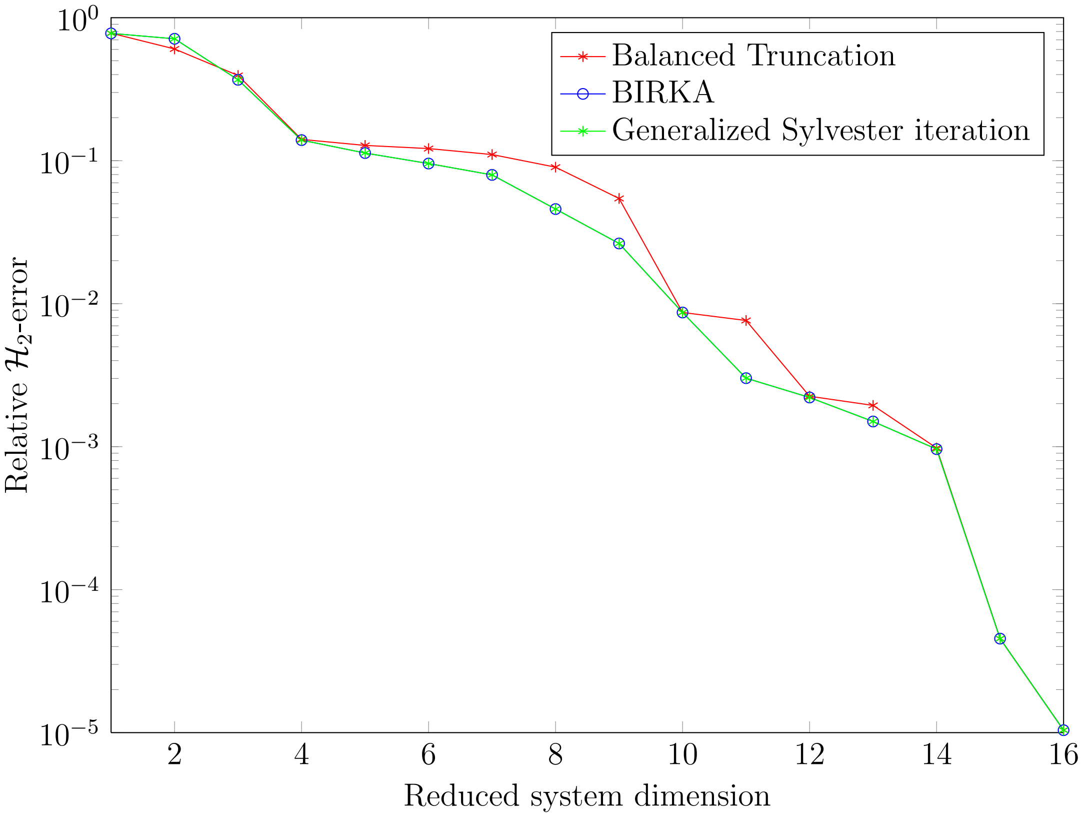


Figure 4.3: Power system. Comparison of relative $\mathcal{H}_2$-error between the method of balanced truncation and BIRKA.

we obtain the Fokker-Planck equation

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= \sigma \Delta \rho + \nabla \cdot (\rho \nabla V), && (x,t) \in (a,b) \times (0,T], \\
0 &= \sigma \nabla \rho + \rho \nabla V, && (x,t) \in \{a,b\} \times [0,T], \\
\rho_0 &= \rho, && (x,t) \in (a,b) \times 0.
\end{aligned}$$

After a finite difference scheme consisting of 500 nodes in the interval $[-2,2]$, we obtain a SISO bilinear control system, where we choose the output matrix $\mathbf{C}$ to be the discrete characteristic function of the interval $[0.95, 1.05]$. Since we only pointed out the most important parameters of the model, for a more detailed insight into this topic, we once more refer to [77]. In Figure 4.5, we again compare the relative $\mathcal{H}_2$-errors between balanced truncation and BIRKA for varying system dimensions. We observe convergence for all reduced system dimensions and our new method clearly outperforms the method of balanced truncation.

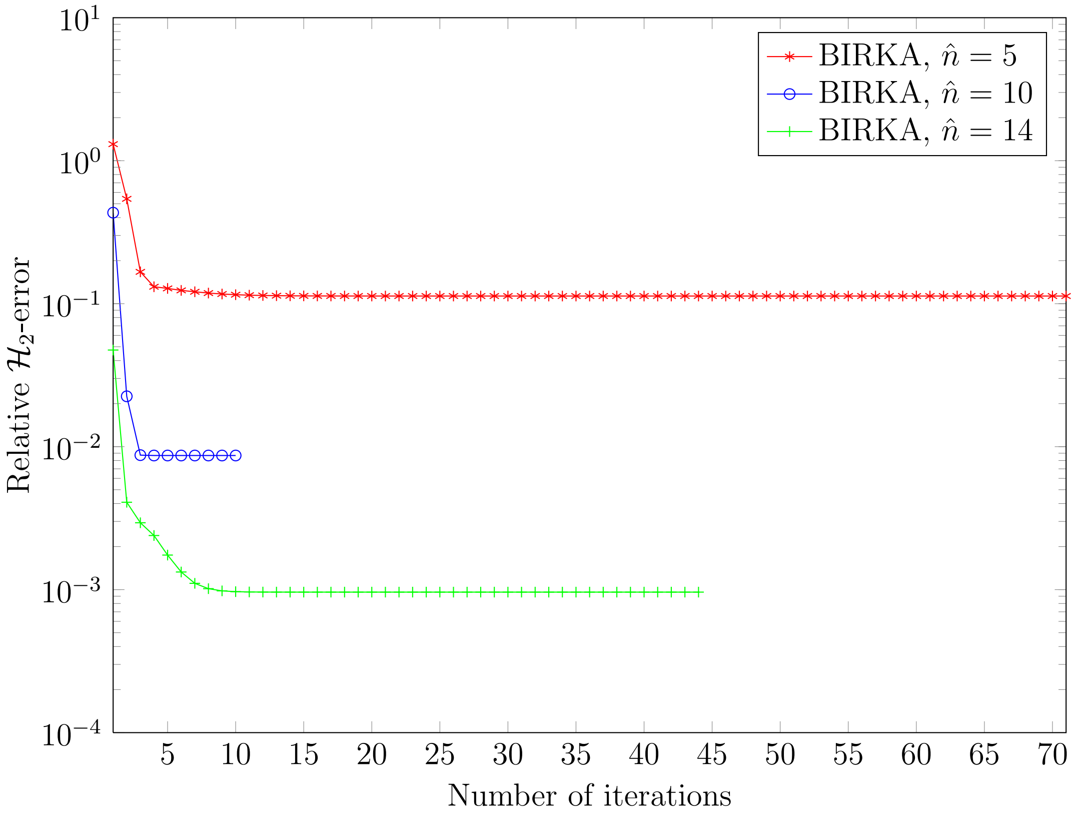


Figure 4.4: Power system. Convergence history of the relative $\mathcal{H}_2$-error.

Viscous Burgers equation

Next, let us consider the viscous Burgers equation

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = \nu\frac{\partial^2 v}{\partial x^2}, \quad (x,t) \in (0,1) \times (0,T),$$

subject to the initial and boundary conditions

$$v(x,0) = 0, \quad x \in [0,1], \qquad v(0,t) = u(t), \quad v(1,t) = 0, \quad t \geq 0.$$

Following [33], after a spatial semi-discretization of this nonlinear partial differential equation using k nodes in a finite difference scheme, we end up with an ordinary differential equation including a quadratic nonlinearity. According to the beginning of this chapter, we can approximate this system by means of the Carleman linearization technique. Here, we use a second-order approximation that yields a bilinearized system of dimension $n = k + k^2$. For the simulations, we use $\nu = 0.1$ and $k = 30$. The measurement vector $\mathbf{C}$ is chosen as the spatial average value for the quantity v. As shown in Figure 4.6,

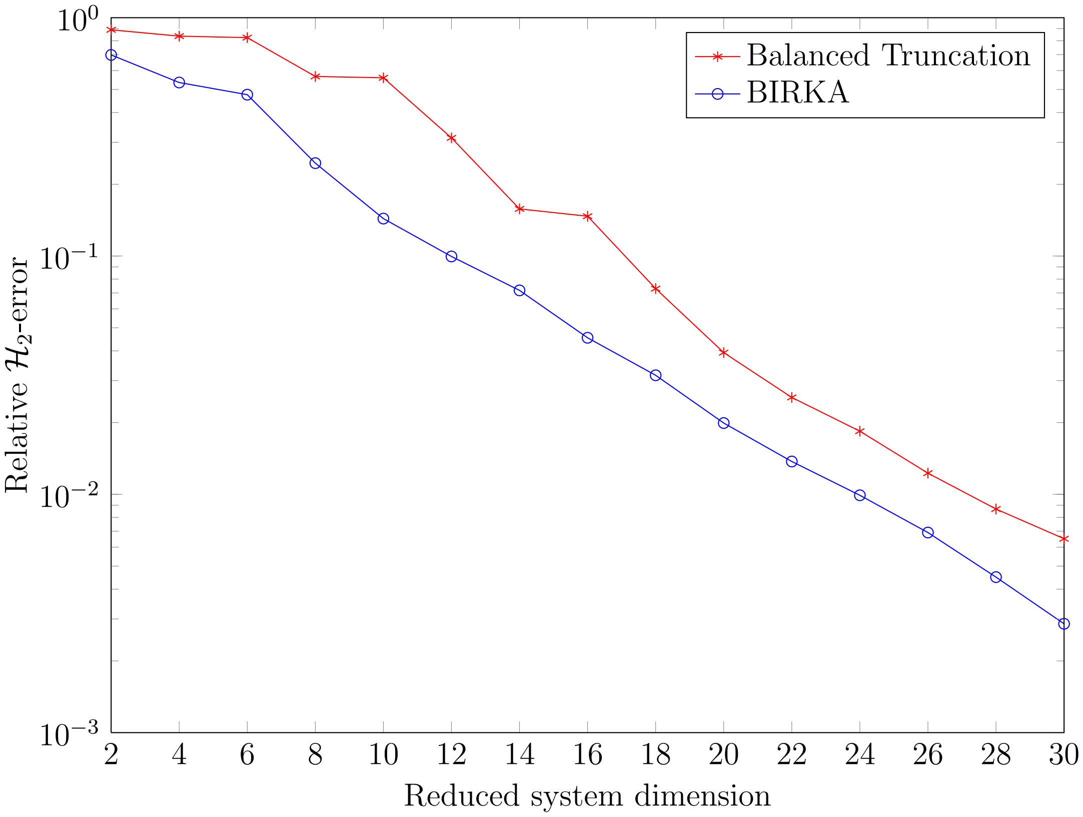


Figure 4.5: Fokker-Planck equation. Comparison of relative $\mathcal{H}_2$-error between balanced truncation and BIRKA.

in all cases the relative $\mathcal{H}_2$-error for the reduced-order systems constructed by BIRKA is smaller than that resulting from balanced truncation. Moreover, once more there are no convergence problems at all although we again use random data for the initialization.

A heat transfer model

Finally, we study another standard bilinear test example resulting from a boundary controlled heat transfer system, see, e.g., [31]. Formally, the dynamics are described by the heat equation subject to Dirichlet and Robin boundary conditions, i.e.,

$$\begin{aligned} x_t &= \Delta x && \text{in } (0,1)\times(0,1),\\ n\cdot\nabla x &= 0.75\cdot u_{1,2,3}(x-1) && \text{on } \Gamma_1,\Gamma_2,\Gamma_3,\\ x &= u_4 && \text{on } \Gamma_4, \end{aligned}$$

where $\Gamma_1, \Gamma_2, \Gamma_3$ and Γ_4 denote the boundaries of Ω. Hence, a spatial discretization using k^2 grid points yields a bilinear system of dimension $n = k^2$, with 4 inputs and 1 output,

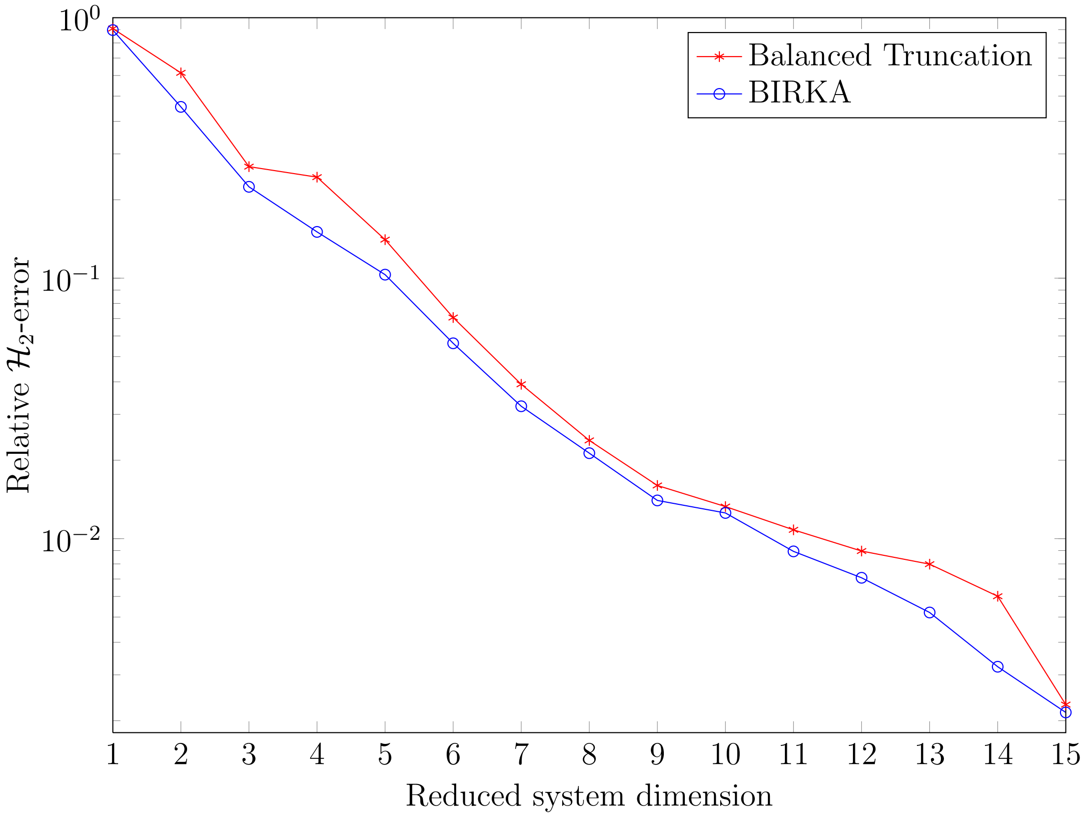


Figure 4.6: Burgers' equation. Comparison of relative $\mathcal{H}_2$-error between the method of balanced truncation and BIRKA.

which is chosen to be the average temperature on the grid. In order to show that our algorithm also works in large-scale settings, we implement the above system with 10 000 grid points. The results for reduced system dimensions $\hat{n} = 2, \ldots, 30$, are given in Figure 4.7 and demonstrate that we can improve the approximation quality in the $\mathcal{H}_2$-norm with our abstract interpolation-based framework. In order to show the superiority of the new approach we further plot the results for the reduced systems obtained by IRKA and those generated by the new interpolation framework together with some clever, but non-optimal interpolation points. This means that we use real equi-distributed as well as Chebyshev interpolations points between the smallest and largest real part of the mirror images of the eigenvalues of the system matrix $\mathbf{A}$ and stop Algorithm 4.3.2 after the first iteration step. However, the relative $\mathcal{H}_2$-error is only computed when the corresponding reduced systems are stable, leading to positive definite solutions of the Gramians of the error systems. Moreover, as we show in Figure 4.7, IRKA only converges for reduced system dimensions up to $\hat{n} = 18$.

So far, most bilinear reduction methods have been evaluated by a comparison of the relative error for outputs corresponding to typical system inputs. For this reason, we compute the transient response to an input of the form $u_k(t) = \cos(k\pi t),\ k = 1, 2, 3, 4.$

The results are plotted in Figure 4.8, where we test the performance for an original bilinear system of order $n = 2500$ and different scaling values γ. This means that the matrices $\mathbf{N}_k$ and $\mathbf{B}$, respectively are multiplied with γ, while the input signal $u_k(t)$ is replaced by $\frac{1}{\gamma}u_k(t)$. Similar experiments are studied in [24]. Interestingly enough, while the convergence results for BIRKA do not change significantly, the relative error is smaller for smaller values of γ. However, all tested values γ can certainly compete with the approximation quality obtained from the method of balanced truncation.

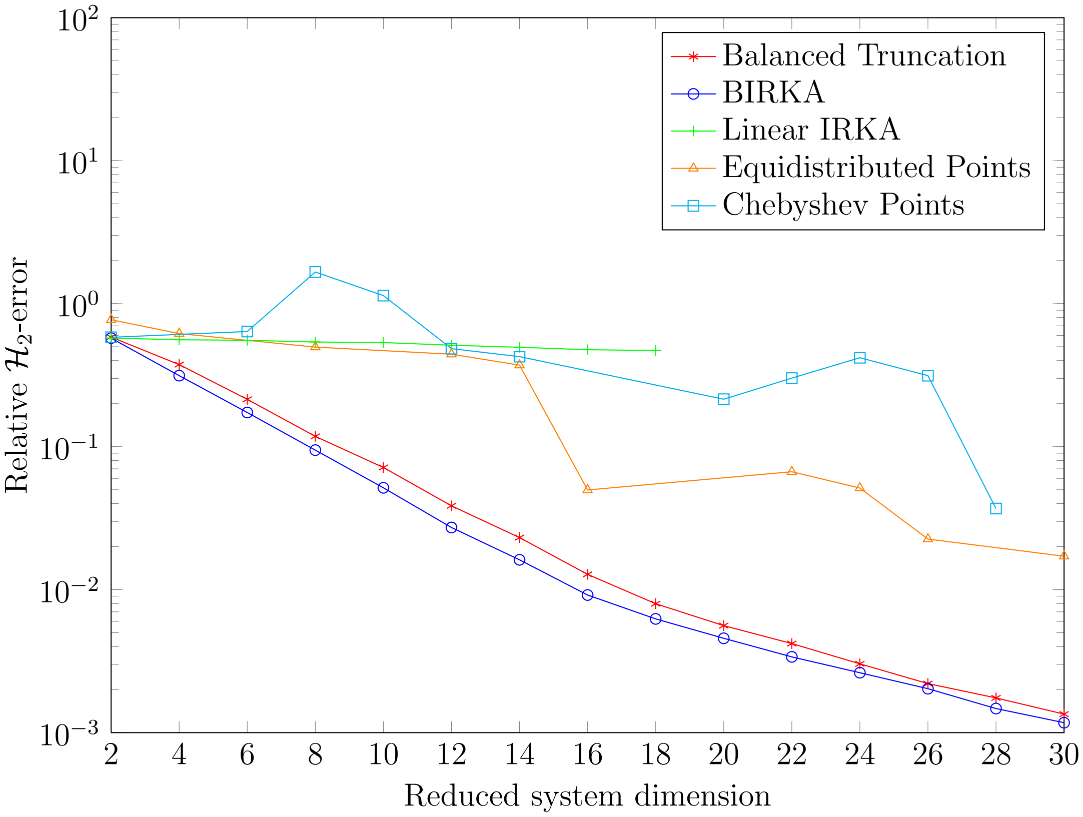


Figure 4.7: Heat transfer model. Comparison of relative $\mathcal{H}_2$-error between balanced truncation and BIRKA.

4.3.6 Conclusions

Before we turn our attention to the solution of large-scale generalized matrix equations arising within the method of balanced truncation for bilinear systems, let us briefly recapitulate the main results from the foregoing discussion. Based on the generalization of the $\mathcal{H}_2$-norm from [133], we have derived first order necessary conditions for optimality. As has been shown, these can be interpreted as an extension of those obtained for the linear case and lead to a generalization of IRKA. We have further proposed an equivalent iterative procedure that requires solving certain generalized Sylvester equations.

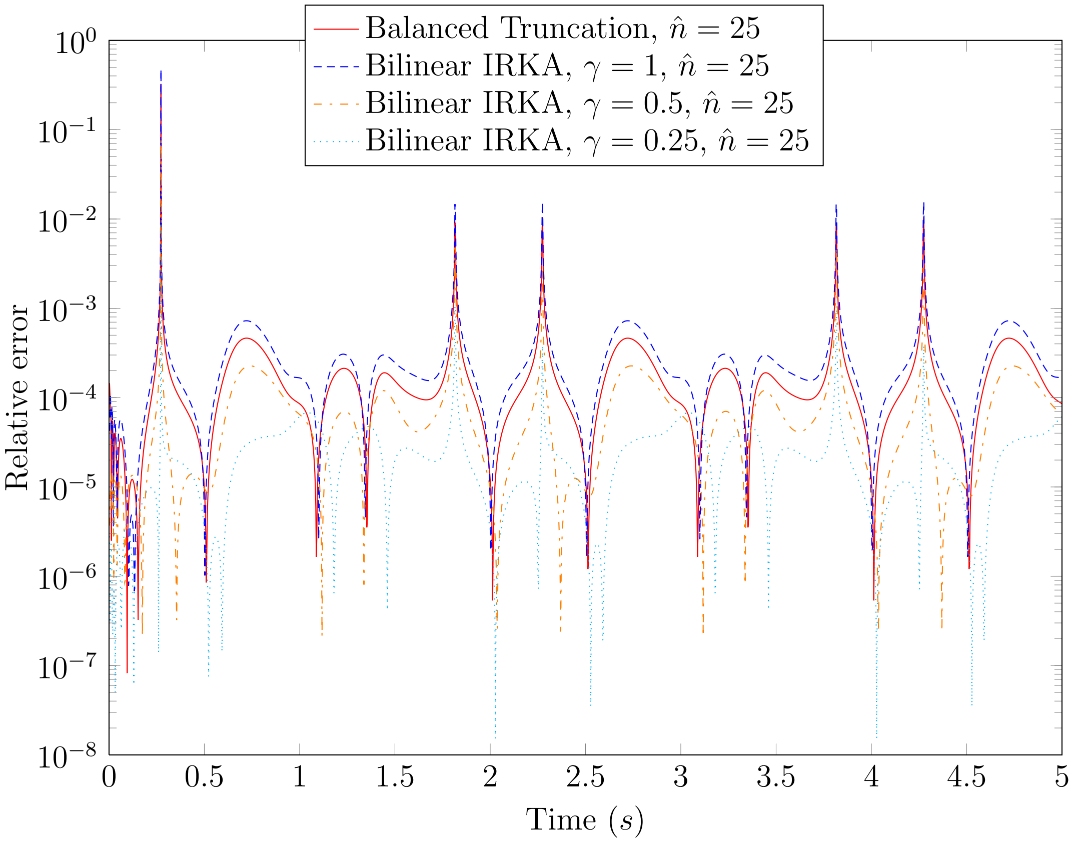


Figure 4.8: Heat transfer model. Comparison of relative error to an input of the form $u_k(t) = \cos(k\pi t)$ for a bilinear system of order $n = 2500$ between balanced truncation and BIRKA for varying scaling factors γ.

The efficiency of our approaches has been evaluated by several bilinear test examples for which they yield better results than the popular method of balanced truncation. Finally, it was shown that the new method can additionally compete when the approximation quality is measured in terms of the transient response in time domain. As a topic of further research, one should mention the possible effect of choosing reasonable initial data in order to improve convergence rates of the algorithms as well as efficient solution techniques for the special generalized Sylvester equations one has to solve in each iteration step.

4.4 Solving large-scale matrix equations arising for balanced truncation

Let us now focus on balancing-based methods for bilinear control systems. First discussed in [81] and, later on, picked up in [2, 24, 40, 43], the concept of a balanced realization of a bilinear system of the form (4.1) allows to truncate states that contribute less to the input-output behavior than others. As in the linear case discussed in Chapter 2, the main idea relies on the fact that a balanced realization can be obtained by solving the generalized Lyapunov equations (4.12). Although in the bilinear case the meaning of the system Gramians is not as clear as in the linear case, the resulting model reduction approach still allows to construct very accurate reduced-order models and, so far, has been the first method of choice. Unfortunately, computing the solutions $\mathbf{P}$ and $\mathbf{Q}$ for dimensions $n > 10^4$ is infeasible since the sole storage of the matrices already becomes a rather non-trivial task. In contrast to the linear case, where we made use of the very fast singular value decay of $\mathbf{P}$ and $\mathbf{Q}$ in order to construct low rank approximation techniques, for bilinear systems, except in the case that the $\mathbf{N}_k$ commute with $\mathbf{A}$, little is known about the singular value decay of $\mathbf{P}$ and $\mathbf{Q}$. However, as already observed in [24], the solutions $\mathbf{P}$ and $\mathbf{Q}$ still seem to exhibit similar properties as in the *standard case* associated with a linear system. Moreover, in the context of high-dimensional eigenvalue problems, in [91], the authors already have proposed some methods for the d-dimensional case in which the solution $\mathbf{P}$ possesses good low rank approximations. The goal of this section is to give a theoretical explanation of this phenomenon for some special cases.

4.4.1 Existence of low rank approximations

For showing the existence of low rank approximations for equations of the form

$$\left(\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes\mathbf{I}+\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j\right)\operatorname{vec}\left(\mathbf{P}\right)=-\operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right), \tag{4.45}$$

it makes sense to consider the explicit system of linear equations

$$\mathcal{A}\operatorname{vec}\left(\mathbf{P}\right):=\left(\mathcal{L}+\Pi\right)\mathbf{p}=\mathcal{B}, \tag{4.46}$$

with $\mathcal{L}=\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes\mathbf{I}$, $\Pi=\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j$ and $\mathcal{B}=-\operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right)$. As already indicated in Chapter 2, an important tool in constructing low rank approximations is given by the integral representation of the inverse of $\mathcal{A}$. In particular, according to [66], for a stable matrix $\mathcal{A}$, we have that

$$\mathcal{A}\left(-\int_0^\infty \exp(t\mathcal{A})\mathrm{d}t\right)=-\int_0^\infty\frac{\partial}{\partial t}\exp(t\mathcal{A})\ \mathrm{d}t=\exp(0\cdot\mathcal{A})=\mathbf{I},$$

implying that $\mathcal{A}^{-1}=-\int_0^\infty\exp(t\mathcal{A})\ \mathrm{d}t$. Hence, constructing an approximation to the inverse of $\mathcal{A}$ can be realized by approximating the latter integral with a suitable quadrature formula similar to the one used in Theorem 2.2.1.

Lemma 4.4.1. *([66]) Let $\mathbf{G}$ be a matrix with spectrum $\sigma(\mathbf{G})$ contained in the strip $\boldsymbol{\Omega} := -[2,\Lambda] \oplus i[-\mu,\mu] \subseteq \mathbb{C}_-$. Let Γ denote the boundary of $-[1,\Lambda+1] \oplus i[-\mu-1,\mu+1]$. Let $k \in \mathbb{N}$ and define the quadrature weights and points according to Theorem 2.2.1. Then there exists C_{st} s.t. for an arbitrary matrix norm, we have*

$$\left|\left|\int_0^\infty \exp(t\mathbf{G})\mathrm{d}t - \sum_{j=-k}^{k} w_j \exp(t_j \mathbf{G})\right|\right| \leq \\ \frac{C_{st}}{2\pi}\exp\left(\frac{\mu+1}{\pi} - \pi\sqrt{k}\right)\oint_\Gamma ||(\lambda\mathbf{I}-\mathbf{G})^{-1}||\mathrm{d}_\Gamma\lambda.$$

In case that $\mathbf{G}$ is symmetric, this simplifies to

$$\left|\left|\int_0^\infty \exp(t\mathbf{G})\mathrm{d}t - \sum_{j=-k}^{k} w_j \exp(t_j \mathbf{G})\right|\right| \leq \frac{C_{st}}{2\pi}\exp\left(\frac{1}{\pi} - \pi\sqrt{k}\right)(4+2\Lambda).$$

Keeping the above result in mind, let us come back to equations of the form (4.45). For a better understanding of the problems that occur in showing the existence of low rank approximations, let us have a look at the main aspects used in the case of the usual Lyapunov equation (2.13) which we from now on refer to as the *standard case*. As we have already mentioned in Chapter 2, one way of constructing low rank approximations is based on the possibility of alternatively considering the approximation of the function

$$f(x_1,x_2) = \frac{1}{x_1+x_2}.$$

This equivalence is easily seen as follows. Assume that an eigenvalue decomposition $\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1}$ is given. Then for the *standard* Lyapunov equation we have

$$(\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes\mathbf{I})\,\mathrm{vec}\,(\mathbf{P}) = -\,\mathrm{vec}\left(\mathbf{B}\mathbf{B}^T\right)$$

which is the same as

$$(\mathbf{Q}\otimes\mathbf{Q})\,(\mathbf{I}\otimes\boldsymbol{\Lambda}+\boldsymbol{\Lambda}\otimes\mathbf{I})\left(\mathbf{Q}^{-1}\otimes\mathbf{Q}^{-1}\right)\mathrm{vec}\,(\mathbf{P}) = -\,\mathrm{vec}\left(\mathbf{B}\mathbf{B}^T\right).$$

However, this means that we can solve the transformed linear system of equations

$$(\mathbf{I}\otimes\boldsymbol{\Lambda}+\boldsymbol{\Lambda}\otimes\mathbf{I})\,\mathrm{vec}\left(\tilde{\mathbf{P}}\right) = -\,\mathrm{vec}\left(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T\right), \tag{4.47}$$

with $\mathrm{vec}\left(\tilde{\mathbf{P}}\right) = (\mathbf{Q}^{-1} \otimes \mathbf{Q}^{-1})\,\mathrm{vec}\,(\mathbf{P})$ and $\tilde{\mathbf{B}} = \mathbf{Q}^{-1}\mathbf{B}$. In (4.47), we have to invert a diagonal matrix leading to expressions of the form $\frac{1}{\lambda_i+\lambda_j}$.

Obviously, to obtain an at least similar structure in the bilinear case, one has to impose severe restrictions on the matrices $\mathbf{A}$ and $\mathbf{N}_j$. Indeed, what one needs is a simultaneous diagonalization as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ and $\mathbf{N}_j = \mathbf{Q}\mathbf{\Gamma}_j\mathbf{Q}^{-1}$. As it is well-known, see, e.g., [80], this means that $\mathbf{A}$ and $\mathbf{N}_j$ must commute which in practice is almost never the case.

Hence, let us consider what happens if we want to make use of the integral representation from Lemma 4.4.1. For the inverse of the matrix $\mathcal{A}$, we conclude that the inverse

$$\mathcal{A}^{-1} = -\int_0^\infty \exp\,(t\mathcal{A})\;\mathrm{d}t,$$

can be approximated by

$$\sum_{i=-k}^{k} w_i \exp\,(t_i\mathcal{A}), \tag{4.48}$$

with the quadrature points t_i and weights w_i from Theorem 2.2.1. Once more, in the *standard case* the computation of the above matrix exponentials (see [80]) boils down to

$$\exp\,(t_i(\mathbf{I}\otimes\mathbf{A} + \mathbf{A}\otimes\mathbf{I})) = \exp\,(t_i\mathbf{A}) \otimes \exp\,(t_i\mathbf{A}).$$

This in turn means that the approximate inverse of the matrix $\mathbf{M}$ is of tensor rank $2k+1$, leading to an approximative solution $\mathrm{vec}\,(\mathbf{P})$ of tensor rank or, equivalently, of column rank $(2k+1)\cdot m$, where m is the number of columns of $\mathbf{B}$. Again, for the bilinear case there arise some problems. Here, we end up with expressions of the form

$$\exp\left(t_i\left(\mathbf{I}\otimes\mathbf{A} + \mathbf{A}\otimes\mathbf{I} + \sum_{j=1}^{m} \mathbf{N}_j\otimes\mathbf{N}_j\right)\right), \tag{4.49}$$

where we can neither make an assertion on their tensor ranks nor on the column rank of the solution $\mathbf{P}$. As we can see, the crucial point is that the matrix exponential in general cannot be split up into its components if the matrices do not commute, i.e.,

$$\exp\left(t_i\Bigl(\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes+\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j\Bigr)\right)$$
$$\neq\left(\exp\left(t_i\mathbf{A}\right)\otimes\exp\left(t_i\mathbf{A}\right)\right)\exp\left(t_i\Bigl(\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j\Bigr)\right).$$

However, in case of commutativity and additional low rank structure of the matrices $\mathbf{N}_j$, we obtain a first simple result.

Proposition 4.4.1. *Let $\mathbf{A}$, $\mathbf{N}_j\in\mathbb{R}^{n\times n}$ be diagonalizable and assume they commute. Further assume that $r_j=\operatorname{rank}\left(\mathbf{N}_j\right)$, $r=\sum_{j=1}^{m}r_j<n$ and that the spectrum of $\mathcal{A}=\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes\mathbf{I}+\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j$ is contained in the strip $\mathbf{\Omega}:=-[2,\Lambda]\oplus i[-\mu,\mu]\subseteq\mathbb{C}_-$. Let Γ denote the boundary of $-[1,\Lambda+1]\oplus i[-\mu-1,\mu+1]$. Then there exists a matrix $\tilde{\mathcal{A}}$ of tensor rank $(2k+1)\cdot(r+1)$ s.t. for an arbitrary matrix norm it holds*

$$||\mathcal{A}^{-1}-\tilde{\mathcal{A}}||\leq\frac{C_{st}}{2\pi}\exp\left(\frac{\mu+1}{\pi}-\pi\sqrt{k}\right)\oint_{\Gamma}||(\lambda\mathbf{I}-\mathcal{A})^{-1}||\mathrm{d}_{\Gamma}\lambda.$$

In case that $\mathcal{A}$ is symmetric, this simplifies to

$$||\mathcal{A}^{-1}-\tilde{\mathcal{A}}||\leq\frac{C_{st}}{2\pi}\exp\left(\frac{1}{\pi}-\pi\sqrt{k}\right)(4+2\Lambda).$$

Proof. The approximation error directly follows from Lemma 4.4.1. It only remains to show that the tensor rank of $\tilde{\mathcal{A}}=\sum_{i=-k}^{k}w_i\exp\left(t_i\mathcal{A}\right)$ does not exceed $(2k+1)\cdot(r+1)$. First, due to commutativity of the matrices, it holds that

$$\exp\left(t_i\mathcal{A}\right)=\left(\exp\left(t_i\mathbf{A}\right)\otimes\exp\left(t_i\mathbf{A}\right)\right)\exp\left(t_i\Bigl(\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j\Bigr)\right).$$

Thus, we only need to check the tensor rank of the latter term. Since we assumed commutativity, all $\mathbf{N}_j = \mathbf{T}\mathbf{D}_j\mathbf{T}^{-1}$ can be diagonalized simultaneously, leading to

$$\begin{aligned}\exp\left(t_i\left(\sum_{j=1}^{m}\mathbf{N}_j\otimes\mathbf{N}_j\right)\right) &= (\mathbf{T}\otimes\mathbf{T})\exp\left(t_i\left(\sum_{j=1}^{m}\mathbf{D}_j\otimes\mathbf{D}_j\right)\right)(\mathbf{T}\otimes\mathbf{T})^{-1}\\ &= (\mathbf{T}\otimes\mathbf{T})\exp\left(t_i\left(\sum_{j=1}^{m}\sum_{k=1}^{r_j} d_{j_{kk}}\,\mathbf{e}_{j_{kk}}\mathbf{e}_{j_{kk}}^T\otimes\mathbf{D}_j\right)\right)(\mathbf{T}\otimes\mathbf{T})^{-1},\end{aligned}$$

with j_{kk} denoting the index of the k-th nonzero diagonal entry of $\mathbf{D}_j$. The assertion now trivially follows by the definition of the matrix exponential and the fact that $\mathbf{e}_{j_{kk}}\mathbf{e}_{j_{kk}}^T$ is an idempotent matrix. □

Remark 4.4.1. *Similar to Theorem 2.2.1, for the symmetric case one could exploit the results from [89] for a better error bound depending on* $\exp(-k)$ *instead of* $\exp(-\sqrt{k})$*. However, since we already discussed the rareness of commutative matrices in practice, the result merely is of theoretical interest anyway.*

Proposition 4.4.1 not only explains the singular value decay of the solution $\mathbf{P}$ of the generalized Lyapunov equation (4.12), but yields an approximation of low tensor rank to the inverse $\mathcal{A}^{-1}$ as well. Obviously, in general this is more complicated than showing the singular value decay of $\mathbf{P}$. However, for our purposes it suffices to show the property for $\mathbf{P}$. Let us now assume that the matrices $\mathbf{N}_j$ have a low rank representation given by matrices $\mathbf{U}_j,\ \mathbf{V}_j\in\mathbb{R}^{n\times r_j}$ s.t. $\mathbf{N}_j = \mathbf{U}_j\mathbf{V}_j^T$. As discussed in [43], we can make use of the splitting (4.46) in order to apply the Sherman-Morrison-Woodbury formula which helps us to prove our main result of this section.

Theorem 4.4.1. *Let* $\mathcal{A}$ *denote a matrix of tensor product structure as in (4.46) with right-hand side* $\mathcal{B} = -\operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right)$*. Assume that the spectrum of* $\mathcal{L}$ *is contained in the strip* $\mathbf{\Omega} := -[\lambda_{\min},\lambda_{\max}]\oplus i[-\mu,\mu]\subseteq\mathbb{C}_-$ *and let* Γ *denote the boundary of* $-[1, 2\lambda_{\max}/\lambda_{\min}+1]\oplus i[-2\mu/\lambda_{\min}-1, 2\mu/\lambda_{\min}+1]$*. Let further* $\mathbf{N}_j = \mathbf{U}_j\mathbf{V}_j^T$*, with* $\mathbf{U}_j,\mathbf{V}_j\in\mathbb{R}^{n\times r_j}$*,* $r=\sum_{j=1}^m r_j$*,* $\mathbf{U} = \left[\mathbf{U}_1\otimes\mathbf{U}_1,\ldots,\mathbf{U}_m\otimes\mathbf{U}_m\right]$*, and* $\mathbf{V} = \left[\mathbf{V}_1\otimes\mathbf{V}_1,\ldots,\mathbf{V}_m\otimes\mathbf{V}_m\right]$*. Then, the solution* $\mathbf{p}$ *to* $\mathcal{A}\mathbf{p}=\mathcal{B}$ *can be approximated by a vector of tensor rank* $(2\cdot k+1)\cdot(m+r)$ *of the form*

$$\tilde{\mathbf{p}} := -\sum_{\ell=-k}^{k}\frac{2w_\ell}{\lambda_{\min}}\left(\exp\left(\frac{2t_\ell}{\lambda_{\min}}\mathbf{A}\right)\otimes\exp\left(\frac{2t_\ell}{\lambda_{\min}}\mathbf{A}\right)\right)\left[\mathcal{B}\quad -\mathbf{U}\mathcal{Y}\right], \tag{4.50}$$

where $\mathcal{Y}$ *is the solution of*

$$\left(\mathbf{I}+\mathbf{V}^T\mathcal{L}^{-1}\mathbf{U}\right)\mathcal{Y} = \mathbf{V}^T\mathcal{L}^{-1}\mathcal{B} \tag{4.51}$$

and w_ℓ*,* t_ℓ *are the quadrature weights and points from Theorem 2.2.1. The corresponding* *approximation error is given as*

$$\begin{aligned}||\mathbf{p}-\tilde{\mathbf{p}}||_2 \leq \frac{C_{st}}{\pi\lambda_{\min}} \exp\left(\frac{2\mu\lambda_{\min}^{-1}+1}{\pi} - \pi\sqrt{k}\right) \oint_{\Gamma} \left|\left|\left(\lambda\mathbf{I} - 2\frac{\mathcal{L}}{\lambda_{\max}}\right)^{-1}\right|\right|_2 \mathrm{d}_{\Gamma}\lambda \\ \times \left|\left|\mathbf{B}\mathbf{B}^T + \sum_{j=1}^{m} \mathbf{U}_j \operatorname{vec}^{-1}\left(\mathcal{Y}_{r_j}\right) \mathbf{U}_j^T\right|\right|_F,\end{aligned} \tag{4.52}$$

where $\mathcal{Y}_{r_j}$ denotes the r_j^2 elements of $\mathcal{Y}$ ranging from $\sum_{i=1}^{j-1} r_i^2 + 1$ to $\sum_{i=1}^{j} r_i^2$.

Proof. Let us consider the tensor structure

$$\Big(\underbrace{\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes\mathbf{I}}_{\mathcal{L}} + \underbrace{\sum_{j=1}^{m} \mathbf{N}_j\otimes\mathbf{N}_j}_{\mathbf{U}\mathbf{V}^T}\Big)\mathbf{p} = \mathcal{B}.$$

Making use of the low rank structure and the Sherman-Morrison-Woodbury formula, the computation of the inverse of $\mathcal{A}$ simplifies to

$$\mathcal{A}^{-1} = \mathcal{L}^{-1} - \mathcal{L}^{-1}\mathbf{U}\left(\mathbf{I}+\mathbf{V}^T\mathcal{L}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^T\mathcal{L}^{-1}.$$

Hence, solving $\mathcal{A}\mathbf{p} = \mathcal{B}$ is equivalent to solving

$$\left(\mathbf{I}\otimes\mathbf{A}+\mathbf{A}\otimes\mathbf{I}\right)\mathbf{p} = \mathcal{B} - \mathbf{U}\underbrace{\left(\mathbf{I}+\mathbf{V}^T\mathcal{L}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^T\mathcal{L}^{-1}\mathcal{B}}_{\mathcal{Y}}.$$

However, the last equation is a standard Lyapunov equation for which we can apply the results from Theorem 2.2.1. Nevertheless, for the assertion on the tensor rank of $\tilde{\mathbf{p}}$, it remains to show that the tensor rank of $\mathcal{B} - \mathbf{U}\mathcal{Y}$ is $m + r$. This is easily seen by the definition of $\mathbf{U} = \left[\mathbf{U}_1\otimes\mathbf{U}_1,\ldots,\mathbf{U}_m\otimes\mathbf{U}_m\right]$. In fact, what we obtain is

$$\begin{aligned}\operatorname{vec}^{-1}\left(\mathbf{U}\mathcal{Y}\right) &= \operatorname{vec}^{-1}\left(\left[\mathbf{U}_1\otimes\mathbf{U}_1,\ldots,\mathbf{U}_m\otimes\mathbf{U}_m\right]\mathcal{Y}\right) \\ &= \sum_{j=1}^{m}\mathbf{U}_j\underbrace{\operatorname{vec}^{-1}\left(\mathcal{Y}_{r_j}\right)\mathbf{U}_j^T}_{:=Y_j^T} = \sum_{j=1}^{m}\sum_{i=1}^{r_j}\mathbf{U}_{j,i}\mathbf{Y}_{j,i}^T.\end{aligned}$$

Consequently, it follows that

$$\mathbf{U}\mathcal{Y} = \sum_{j=1}^{m}\sum_{i=1}^{r_j} \mathbf{Y}_{j,i} \otimes \mathbf{U}_{j,i},$$

where the second subscript i denotes the i-th column of the matrices. By assumption, the r_j sum up to r, leading to a tensor rank of $(2 \cdot k + 1) \cdot (m + r)$. The approximation error follows by the same inversion of the $\operatorname{vec}(\cdot)$-operator and applying the results from [66] for a modified right-hand side $-\operatorname{vec}\left(\mathbf{B}\mathbf{B}^T\right) - \mathbf{U}\mathcal{Y}$. □

Remark 4.4.2. *We point out that we do not claim that Theorem 4.4.1 provides an error bound useful for an estimation of the true error of the proposed approximation. The result rather yields a theoretical evidence for the often observed fast singular value decay of generalized Lyapunov equations of the form (4.12). Moreover, the numerical techniques we propose later on are of different nature and do not approximate the integral of $\mathcal{A}^{-1}$. Since at this point we simply are not aware of a suitable generalization of error bounds known for the standard case, we refer to Theorem 4.4.1 that makes the search for numerical methods reasonable.*

Remark 4.4.3. *Obviously, there exist special cases where the $\mathbf{N}_j$ are full-rank matrices and we still can expect a strong singular value decay of the solution $\mathbf{P}$. Here, one might think of*

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0},$$

or the even easier case

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T + \mathbf{P} + \mathbf{B}\mathbf{B}^T = \mathbf{0}.$$

Both of the above equations reduce to a modified linear Lyapunov equation with right-hand side of rank m. However, this is not surprising since $\mathbf{N} = \mathbf{A}$ and $\mathbf{N} = \mathbf{I}$ both obviously commute with $\mathbf{A}$. Nevertheless, so far it remains an open question if it is possible to extend decay results for a more general setting as well. The numerical results we show later on indicate that there seem to be conditions for low rank properties also in other cases.

Although for the higher dimensional case

$$\underbrace{\left(\sum_{i=1}^{d} \mathbf{I} \otimes \cdots \otimes \mathbf{I} \otimes \mathbf{A}_i \otimes \mathbf{I} \otimes \cdots \otimes \mathbf{I} + \sum_{j=1}^{k} \mathbf{N}_{j_1} \otimes \cdots \otimes \mathbf{N}_{j_d}\right)}_{\mathcal{A}_d} \operatorname{vec}(\mathbf{P}) = \bigotimes_{i=1}^{d} \mathbf{b}_i, \qquad (4.53)$$

the tensor rank increases exponentially with the dimensions, it might be worth noting that we can still expect low rank approximations as stated in the following corollary. For this, let

$$\mathcal{L}_d = \sum_{i=1}^{d} \mathbf{I} \otimes \cdots \otimes \mathbf{I} \otimes \mathbf{A}_i \otimes \mathbf{I} \otimes \cdots \otimes \mathbf{I}.$$

Corollary 4.4.1. *Let $\mathcal{A}_d$ denote a matrix of tensor product structure as in (4.53) with tensor right-hand side $\mathcal{B} = \bigotimes_{i=1}^{d} \mathbf{b}_i$ and $\mathbf{N}_{j_\ell} = \mathbf{N}_j$, with $\operatorname{rank}(\mathbf{N}_j) = r_j$. Assume that the sum of the spectra of the $\mathbf{A}_i$ is contained in the strip $\mathbf{\Omega} := -[\lambda_{\min}, \lambda_{\max}] \oplus i[-\mu, \mu] \subseteq \mathbb{C}_-$ and let Γ denote the boundary of $-[1, 2\lambda_{\max}/\lambda_{\min} + 1] \oplus i[-2\mu/\lambda_{\min} - 1, 2\mu/\lambda_{\min} + 1]$. Let further $\mathbf{N}_j = \mathbf{U}_j \mathbf{V}_j^T$, with $\mathbf{U}_j, \mathbf{V}_j \in \mathbb{R}^{n \times r_j}$, $r = \sum_{j=1}^{m} r_j$, $\mathbf{U} = \left[\bigotimes_{i=1}^{d} \mathbf{U}_1, \ldots, \bigotimes_{i=1}^{d} \mathbf{U}_m\right]$, and $\mathbf{V} = \left[\bigotimes_{i=1}^{d} \mathbf{V}_1, \ldots, \bigotimes_{i=1}^{d} \mathbf{V}_m\right]$. Then, the solution $\mathbf{p}$ to $\mathcal{A}_d \mathbf{p} = \mathcal{B}$ can be approximated by a vector of tensor rank $(2 \cdot k + 1) \cdot (m + r^{d-1})$ of the form*

$$\tilde{\mathbf{p}} := -\sum_{\ell=-k}^{k} \frac{2w_\ell}{\lambda_{\min}} \bigotimes_{i=1}^{d} \exp\left(\frac{2t_\ell}{\lambda_{\min}} \mathbf{A}_i\right) \left[\mathcal{B} \quad -\mathbf{U}\mathcal{Y}\right], \tag{4.54}$$

where $\mathcal{Y}$ is the solution of

$$\left(\mathbf{I}_{r^d} + \mathbf{V}^T \mathcal{L}_d^{-1} \mathbf{U}\right) \mathcal{Y} = \mathbf{V}^T \mathcal{L}_d^{-1} \mathcal{B} \tag{4.55}$$

and w_ℓ, t_ℓ are the weights from Theorem 2.2.1. The corresponding approximation error is given as

$$\begin{aligned} \|\mathbf{p} - \tilde{\mathbf{p}}\|_2 \leq \frac{C_{st}}{\pi \lambda_{\min}} \exp\left(\frac{2\mu\lambda_{\min}^{-1} + 1}{\pi} - \pi\sqrt{k}\right) \oint_{\Gamma} \left\| \left(\lambda \mathbf{I} - 2\frac{\mathcal{L}_d}{\lambda_{\min}}\right)^{-1} \right\|_2 \mathrm{d}_{\Gamma}\lambda \\ \times \left\| \mathcal{B} + \sum_{j=1}^{m} \left(\bigotimes_{i=1}^{d} \mathbf{U}_j\right) \mathcal{Y} \right\|_2 . \end{aligned} \tag{4.56}$$

Proof. The assertion on the tensor rank easily follows by iteratively applying the procedure from the proof of Theorem 4.4.1 to the terms $\left(\bigotimes_{i=1}^{d} \mathbf{U}_j\right) \mathcal{Y}$, e.g. for $d = 3$, we obtain

$$\left(\mathbf{U}_j \otimes \mathbf{U}_j \otimes \mathbf{U}_j\right) \mathcal{Y} = \operatorname{vec}\left(\left[\mathbf{U}_{j_1} \otimes \left(\mathbf{U}_j \otimes \mathbf{U}_j\right) \mathcal{Y}_1, \ldots, \mathbf{U}_{j_r} \otimes \left(\mathbf{U}_j \otimes \mathbf{U}_j\right) \mathcal{Y}_r\right]\right).$$

Since each of the terms $\left(\mathbf{U}_j \otimes \mathbf{U}_j\right) \mathcal{Y}_i$ is of tensor rank r, it is clear that $\left(\mathbf{U}_j \otimes \mathbf{U}_j \otimes \mathbf{U}_j\right) \mathcal{Y}$ is of tensor rank at most r^2. All other results can be proved analogously as before. □

Remark 4.4.4. *Though the rank of the approximation increases exponentially with d,* *so does the maximum possible tensor rank which is n^{d-1}. Hence, the ratio between full and approximate solution is $\sim \left(\frac{r}{n}\right)^{d-1}$.*

4.4.2 Low rank solution methods

Now that we have seen that we indeed can expect a reasonably fast singular value decay of the solution matrix $\mathbf{P}$ of (4.12), we want to discuss possible extensions of existing linear low rank Lyapunov solvers that have been proven to yield accurate low rank approximations $\mathbf{L}\mathbf{L}^T \approx \mathbf{P}$. Here, we point out the LRCF-ADI iteration, KPIK together with the more general rational Krylov framework and finally approaches that solve the explicit linear system in tensorized form by iterative methods like, e.g., BiCGstab. As has been pointed out in [43], for the generalized Lyapunov equation

$$\underbrace{\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T}_{\mathcal{L}} + \underbrace{\sum_{j=1}^{m} \mathbf{N}_j\mathbf{P}\mathbf{N}_j^T}_{\Pi} + \mathbf{B}\mathbf{B}^T = \mathbf{0},$$

it makes sense to demand that the spectral radius satisfies $\rho\left(\mathcal{L}^{-1}\Pi\right) < 1$, since otherwise we cannot ensure that $\mathbf{P}$ is positive definite. However, at least in the bilinear case there exist a lot of interesting applications that lead to indefinite solution matrices $\mathbf{P}$ and we therefore address problems that might occur in these cases.

The low rank ADI iteration

Let us now focus on the low rank version of the ADI iteration we mentioned in the beginning of Chapter 3. In general, the main idea is that for any parameter $p > 0$, the Lyapunov operator $\mathcal{L}$ can be shifted according to

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^T = \frac{1}{2p}\left((\mathbf{A} + p\mathbf{I})\mathbf{P}(\mathbf{A} + p\mathbf{I})^T - (\mathbf{A} - p\mathbf{I})\mathbf{P}(\mathbf{A} - p\mathbf{I})^T\right). \tag{4.57}$$

In [43], for a given set of shift parameters $\{p_0, p_1, \dots\}$, this circumstance is used to solve (4.12) via the following fixed-point iteration

$$\begin{aligned}\mathbf{P}_{k+1} = (\mathbf{A} - p_k\mathbf{I})^{-1}(\mathbf{A} + p_k\mathbf{I})\mathbf{P}_k(\mathbf{A} + p_k\mathbf{I})^T(\mathbf{A} - p_k\mathbf{I})^{-T} \\ + 2p_k(\mathbf{A} - p_k\mathbf{I})^{-1}\left(\sum_{j=1}^{m} \mathbf{N}_j\mathbf{P}_k\mathbf{N}_j^T + \mathbf{B}\mathbf{B}^T\right)(\mathbf{A} - p_k\mathbf{I})^{-T}.\end{aligned}$$

However, for dimensions n larger than 10^3 the above scheme is infeasible since in each step we have to solve a linear system with a matrix right-hand side which might easily become too expensive. Moreover, for even larger dimensions, the simple storing of the generally dense matrix $\mathbf{P}_k$ already causes serious memory problems. On the other hand, we can expect that the solution matrix $\mathbf{P}$ is symmetric and, according to the previous section, tends to have a strong singular value decay as well. For this reason, as in the *standard case*, suggested in [26, 97, 109], instead of the full-rank version, it is reasonable to start with a symmetric initial guess, e.g. $\mathbf{P}_0 = \mathbf{B}\mathbf{B}^T$, and then only compute the low rank factors $\mathbf{Z}_k$ according to

$$\mathbf{Z}_{k+1} = \Big[(\mathbf{A} - p_k\mathbf{I})^{-1}(\mathbf{A} + p_k\mathbf{I})\mathbf{Z}_k, \sqrt{2p_k}(\mathbf{A} - p_k\mathbf{I})^{-1}\mathbf{N}_1\mathbf{Z}_k, \dots, \\ \sqrt{2p_k}(\mathbf{A} - p_k\mathbf{I})^{-1}\mathbf{N}_m\mathbf{Z}_k, \sqrt{2p_k}(\mathbf{A} - p_k\mathbf{I})^{-1}\mathbf{B}\Big].$$

Obviously, the advantage is that we now only have to solve $2 + m$ systems of linear equations with low rank right-hand side. In the *standard case*, it has been shown, see [96], that the iteration can be rewritten in such a way that $\mathbf{Z}_{k+1} = \begin{bmatrix}\mathbf{Z}_k & \mathbf{V}_k\end{bmatrix}$, with $\mathbf{V}_k \in \mathbb{R}^{n\times m}$, making an appropriate algorithm much cheaper to execute. Unfortunately, due to the non-commutativity of $\mathbf{A}$ and $\mathbf{N}_j$, in our case this is not possible. If we assume that the iterate $\mathbf{Z}_k$ consists of r columns, at least theoretically $\mathbf{Z}_{k+1}$ consists of $(m+1)\cdot r + m$ columns. However, we often obtain a deflation in the column spaces such that a column compression can prevent a too strong column increase. Another problem might arise in case of the already mentioned absence of a convergent splitting which is quite common for real-life examples of bilinear control systems. Here, it should be noted that the ADI iteration will not converge and we therefore recommend the use of one of the other low rank solvers which we discuss in the next subsections.

Choice of shift parameters

For the *standard case*, a very important point in the competitiveness of the ADI iteration is the choice of the shift parameters p_k. If good shift parameters are known, the iteration tends to converge very fast to an accurate approximation. On the other hand, for bad shift parameters the iteration might stagnate. Moreover, the computation of such parameters often can be one of the most expensive tasks for this approach, see, e.g., [27, 97]. It is known, see [128], that for the *standard case* a set of q optimal parameters is given by the solution to the rational min-max problem

$$\min_{\{p_1,\dots,p_q\}} \max_{\lambda\in\sigma(\mathbf{A})} \prod_{\ell=1}^{q} \left|\frac{\lambda + p_\ell}{\lambda - p_\ell}\right|, \tag{4.58}$$

where $\sigma(\mathbf{A})$ denotes the spectrum of $\mathbf{A}$. For the generalized version considered here, the situation becomes more complicated. In what follows, for the ease of presentation we assume that $m = 1$, i.e., we consider

$$\mathbf{AP} + \mathbf{PA}^T + \mathbf{NPN}^T + \mathbf{bb}^T = \mathbf{0}.$$

Moreover, let us focus on real parameters p_k. According to the shifting (4.57), for the solution $\mathbf{P}$ it holds that

$$\begin{aligned}\mathbf{P} = &(\mathbf{A} - p_k\mathbf{I})^{-1}(\mathbf{A} + p_k\mathbf{I})\mathbf{P}(\mathbf{A} + p_k\mathbf{I})^T(\mathbf{A} - p_k\mathbf{I})^{-T}\\ &+ 2p_k(\mathbf{A} - p_k\mathbf{I})^{-1}\left(\mathbf{NPN}^T + \mathbf{bb}^T\right)(\mathbf{A} - p_k\mathbf{I})^{-T}.\end{aligned}$$

Hence, for the iterate $\mathbf{P}_{k+1}$ we can compute

$$\begin{aligned}\mathbf{P}_{k+1} - \mathbf{P} = &(\mathbf{A} - p_k\mathbf{I})^{-1}(\mathbf{A} + p_k\mathbf{I})(\mathbf{P}_k - \mathbf{P})(\mathbf{A} + p_k\mathbf{I})^T(\mathbf{A} - p_k\mathbf{I})^{-T}\\ &+ 2p_k(\mathbf{A} - p_k\mathbf{I})^{-1}\mathbf{N}(\mathbf{P}_k - \mathbf{P})(\mathbf{A} - p_k\mathbf{I})^{-T}.\end{aligned}$$

In other words, if we use the Kronecker product notation, iteratively applying the latter equation yields

$$\operatorname{vec}(\mathbf{P}_{k+1} - \mathbf{P}) = \prod_{i=1}^{k} \mathbf{G}_i \operatorname{vec}(\mathbf{P}_0 - \mathbf{P}),$$

with

$$\mathbf{G}_i = (\mathbf{A} - p_i\mathbf{I})^{-1} \otimes (\mathbf{A} - p_i\mathbf{I})^{-1}\Big((\mathbf{A} + p_i\mathbf{I}) \otimes (\mathbf{A} + p_i\mathbf{I}) + 2p_i\mathbf{N} \otimes \mathbf{N}\Big).$$

Obviously, this means that minimizing the error implies minimizing the spectral radius of $\prod_{i=1}^{k} \mathbf{G}_i$. Unfortunately, for general $\mathbf{A}$ and $\mathbf{N}$ this is by far more complicated than solving the min-max problem (4.58). On the other hand, if we assume that $\mathbf{A}$ and $\mathbf{N}$ commute, they can be simultaneously diagonalized and we conclude that for q optimal shift parameters, we have to solve

$$\min_{\{p_1,\ldots,p_q\}} \max_{\substack{\lambda_i,\lambda_j \in \sigma(\mathbf{A})\\ \mu_i,\mu_j \in \sigma(\mathbf{N})}} \left| \frac{(\lambda_i + p_\ell)(\lambda_j + p_\ell) + 2p_\ell\mu_i\mu_j}{(\lambda_i - p_\ell)(\lambda_j - p_\ell)} \right|, \tag{4.59}$$

where $\sigma(\mathbf{A})$ and $\sigma(\mathbf{N})$ again denote the spectrum of $\mathbf{A}$ and $\mathbf{N}$, respectively. Obviously, even the assumption of commutativity still leads to a more complex minimization problem for which a discussion of solution methods is beyond the scope of this thesis.

On the other hand, for the linear setting, it has recently been shown that so-called $\mathcal{H}_2$-optimal shifts take a special position among ADI shift parameters. As the authors discuss in [45, 59], $\mathcal{H}_2$-optimal shifts share the property that the ADI method in this case yields exactly the same results as the rational Krylov subspace method, meaning that both methods are equivalent in this setting. Moreover, in Chapter 3, for the special case of a symmetric matrix $\mathbf{A}$, we have seen that the corresponding subspaces for these shifts yield optimal solutions with respect to the naturally induced energy norm of the Lyapunov operator. Hence, we can can say that for the *standard case*, these parameters are a reliable alternative to the optimal ones that solve problem (4.58). Since in the first part of this chapter, we studied the $\mathcal{H}_2$-optimal model reduction problem for bilinear systems, we make use of the corresponding theory later on. However, instead of optimal interpolation points, we use so-called pseudo-optimal points, i.e., points that are constructed by a one-sided projection. As a consequence, except in the symmetric case, these points only fulfill a part of the presented optimality conditions. Nevertheless, these interpolation points have a positive effect on the convergence rate of the bilinear ADI iteration as well.

Low rank solutions by projection

In Chapter 3, we already extensively discussed the idea of obtaining low rank approximate solutions by projecting on certain (rational) Krylov subspaces. Although not aiming at an optimal approximation for a given rank, we mentioned that a fast and reliable approach is given by the Krylov-Plus-Inverted-Krylov (KPIK) method from [120]. Recall that here we have to compute the two (block)-Krylov subspaces

$$\mathcal{K}_q(\mathbf{A},\mathbf{B}), \quad \mathcal{K}_q(\mathbf{A}^{-1},\mathbf{A}^{-1}\mathbf{B})$$

and then construct $\mathbf{V}$ as an orthonormal basis of the union of the corresponding column spaces. Alternatively, this may be achieved by the following iterative procedure

$$\mathbf{V}_1 = [\mathbf{B},\mathbf{A}^{-1}\mathbf{B}], \quad \mathbf{V}_k = [\mathbf{A}\mathbf{V}_{k-1},\mathbf{A}^{-1}\mathbf{V}_{k-1}], \quad k \leq q.$$

Usually, the above subspaces are generated by a modified Gram-Schmidt process which leads to orthonormal bases in each step. In order to extend the approach to our generalized setting, we suggest to proceed as follows

$$\mathbf{V}_1 = [\mathbf{B},\mathbf{A}^{-1}\mathbf{B}], \quad \mathbf{V}_k = [\mathbf{A}\mathbf{V}_{k-1},\mathbf{A}^{-1}\mathbf{V}_{k-1},\mathbf{N}_j\mathbf{V}_{k-1}], \quad k \leq q,\ j = 1,\ldots,m.$$

Again, the Galerkin condition demands an orthogonal $\mathbf{V}$, such that we have $\mathbf{V} := \operatorname{orth}(\mathbf{V}_q)$. Moreover, similar to the ADI iteration, one should perform a column compression which keeps the rank increase in each step at a compatible level. Analog to the discussions of the *standard case* given in [83, 84, 116, 120], one can use the nestedness of the subspaces generated during the process to simplify the computation of the residual.

Theorem 4.4.2. *Let* $\mathbf{R}_k := \mathbf{A}\mathbf{P}_k + \mathbf{P}_k\mathbf{A}^T + \sum_{j=1}^m \mathbf{N}_j\mathbf{P}_k\mathbf{N}_j^T + \mathbf{B}\mathbf{B}^T$ *denote the residual associated with the approximate solution* $\mathbf{P}_k = \mathbf{V}_k\hat{\mathbf{P}}_k\mathbf{V}_k^T$, *where* $\hat{\mathbf{P}}_k$ *is the solution of the reduced Lyapunov equation*

$$\mathbf{V}_k^T\mathbf{A}\mathbf{V}_k\hat{\mathbf{P}}_k + \hat{\mathbf{P}}_k\mathbf{V}_k^T\mathbf{A}^T\mathbf{V}_k + \sum_{j=1}^m \mathbf{V}_k^T\mathbf{N}_j\mathbf{V}_k\hat{\mathbf{P}}_k\mathbf{V}_k^T\mathbf{N}_j^T\mathbf{V}_k + \mathbf{V}_k^T\mathbf{B}\mathbf{B}^T\mathbf{V}_k = \mathbf{0}.$$

Then, it holds that $\operatorname{range}(\mathbf{R}_k) = \operatorname{range}(\mathbf{V}_{k+1})$ *and* $||\mathbf{R}_k|| = ||\mathbf{V}_{k+1}^T\mathbf{R}_k\mathbf{V}_{k+1}||$, *where* $||\cdot||$ *may denote the Frobenius norm or the spectral norm, respectively.*

Proof. The first assertion follows from the fact that, due to the iterative construction of $\mathbf{V}_{k+1}$, we have

$$\mathbf{V}_k \subset \mathbf{V}_{k+1}, \ \mathbf{A}\mathbf{V}_k \subset \mathbf{V}_{k+1}, \ \mathbf{N}_j\mathbf{V}_k \subset \mathbf{V}_{k+1}.$$

Moreover, with the same argument and the orthonormality of $\mathbf{V}_{k+1}$, it holds

$$\mathbf{R}_k = \mathbf{V}_{k+1}\mathbf{V}_{k+1}^T\mathbf{R}_k\mathbf{V}_{k+1}\mathbf{V}_{k+1}^T.$$

This implies $||\mathbf{R}_k|| = ||\mathbf{V}_{k+1}^T\mathbf{R}_k\mathbf{V}_{k+1}||$. □

Note that in contrast to the *standard case* it seems to be impossible to further simplify the expression for the residual. The problem is that the Hessenberg structure of the projected system matrix $\mathbf{T} = \mathbf{V}_k^T\mathbf{A}\mathbf{V}_k$ is lost.

Also, so far we are not aware of a possible generalization of usable and, more importantly, a priori computable error bounds as the ones specified in [16]. Although it seems to be a complicate issue to extend the concepts presented therein to the setting (4.12), we think that this is certainly an interesting topic of further research.

Iterative linear solvers

Finally, let us address the possibility of efficiently solving the tensorized linear system of equations (4.45) by iterative solvers like CG (symmetric case) or BiCGstab (unsymmetric case). The crucial point is to note that we can incorporate the to-expected low rank structure of $\mathbf{P}$ into the algorithm which allows to reduce the complexity significantly.

The symmetric case

Since a quite similar discussion for more general tensorized linear systems can be found in [90], we follow the notations therein and only briefly discuss how to adapt the main concepts to our purposes. Assuming that the matrices $\mathbf{A}$ and $\mathbf{N}_j$ are symmetric, we can modify the preconditioned CG method. For this, let us have a look at Algorithm 4.4.1 which has already been studied in [90] in the context of solving equations of the form (2.15). The application of the matrix function $\mathcal{A}$ to a matrix $\mathbf{P}$ here should denote the operation $\mathbf{AP}+\mathbf{PA}^T+\sum_{j=1}^m \mathbf{N}_j\mathbf{PN}_j^T$. As a preconditioner $\mathcal{M}^{-1}$ we use the low rank version of the bilinear ADI iteration which we studied before, whereas the truncation operator $\mathcal{T}$ should be understood as a simple column compression as described in e.g. [90]. The only point to clarify is that we indeed can ensure a decomposition $\mathbf{P}_{\hat{n}} = \mathbf{G}_k\mathbf{D}_k\mathbf{G}_k^T$, with diagonal matrix $\mathbf{D}_k$, in each step of the algorithm. We start with $\mathbf{R}_0 = \mathcal{B} = -\mathbf{BB}^T$ which obviously can be decomposed into $\mathbf{R}_0 = \mathbf{G}_{\mathbf{R}_0}\mathbf{D}_{\mathbf{R}_0}\mathbf{G}_{\mathbf{R}_0}^T$ by setting $\mathbf{G}_{\mathbf{R}_0} = \mathbf{B}$ and $\mathbf{D}_{\mathbf{R}_0} = -\mathbf{I}_m$. Next, we note that the bilinear ADI iteration is not restricted to a factorization of the form $\mathbf{ZZ}^T$ but can also be applied to low rank decompositions $\mathbf{GDG}^T$, cp. [27]. This is easily seen as follows. Recalling the iteration procedure, we formally assume that $\mathbf{Z}_k = \mathbf{G}_k\sqrt{\mathbf{D}_k}$ and obtain the new iterate

Algorithm 4.4.1 Preconditioned CG method

```
Input: Matrix functions $\mathcal{A},\mathcal{M} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, low rank factor $\mathbf{B}$ of right-hand side
    $\mathcal{B} = -\mathbf{BB}^T$. Truncation operator $\mathcal{T}$ w.r.t. relative accuracy $\epsilon_{rel}$.
Output: Low rank approximation $\mathbf{P}_{\hat{n}} = \mathbf{GDG}^T$ with $||\mathcal{A}(\hat{\mathbf{P}}) - \mathcal{B}||_F \le$ tol.
$\mathbf{P}_0 = \mathbf{0}$, $\mathbf{R}_0 = \mathcal{B}$, $\mathbf{Z}_0 = \mathcal{M}^{-1}(\mathbf{R}_0)$, $\mathbf{P}_0 = \mathbf{Z}_0$, $\mathbf{Q}_0 = \mathcal{A}(\mathbf{P}_0)$, $\xi_0 = \langle \mathbf{P}_0, \mathbf{Q}_0\rangle, k = 0$
while $||\mathbf{R}_k||_F >$ tol do
    $\omega_k = \frac{\langle \mathbf{R}_k,\mathbf{P}_k\rangle}{\xi_k}$
    $\mathbf{P}_{k+1} = \mathbf{P}_k + \omega_k\mathbf{P}_k$,                          $\mathbf{P}_{k+1} \leftarrow \mathcal{T}(\mathbf{P}_{k+1})$
    $\mathbf{R}_{k+1} = \mathcal{B} - \mathcal{A}(\mathbf{P}_{k+1})$,          Optionally: $\mathbf{R}_{k+1} \leftarrow \mathcal{T}(\mathbf{R}_{k+1})$
    $\mathbf{Z}_{k+1} = \mathcal{M}^{-1}(\mathbf{R}_{k+1})$
    $\beta_k = -\frac{\langle \mathbf{Z}_{k+1},\mathbf{Q}_k\rangle}{\xi_k}$
    $\mathbf{P}_{k+1} = \mathbf{Z}_{k+1} + \beta_k\mathbf{P}_k$,                      $\mathbf{P}_{k+1} \leftarrow \mathcal{T}(\mathbf{P}_{k+1})$
    $\mathbf{Q}_{k+1} = \mathcal{A}(\mathbf{P}_{k+1})$,                  Optionally: $\mathbf{Q}_{k+1} \leftarrow \mathcal{T}(\mathbf{Q}_{k+1})$
    $\xi_{k+1} = \langle \mathbf{P}_{k+1}, \mathbf{Q}_{k+1}\rangle$
    $k = k + 1$
end while
$\mathbf{P}_{\hat{n}} = \mathbf{P}_k$
```

$$\mathbf{Z}_{k+1} = (\mathbf{A}-p_k\mathbf{I})^{-1}\Big[(\mathbf{A}+p_k\mathbf{I})\mathbf{G}_k\sqrt{\mathbf{D}}_k, \sqrt{2p_k}\mathbf{N}_1\mathbf{G}_k\sqrt{\mathbf{D}}_k, \ldots, \\ \sqrt{2p_k}\mathbf{N}_m\mathbf{G}_k\sqrt{\mathbf{D}}_k, \sqrt{2p_k}\mathbf{G}\sqrt{\mathbf{D}}\Big],$$

where $\mathbf{G}\sqrt{\mathbf{D}}$ is the initial input to the ADI iteration. Forming the product $\mathbf{Z}_{k+1}\mathbf{Z}_{k+1}^T$, it is clear that we can replace the step by setting

$$\begin{aligned}\mathbf{G}_{k+1} &= (\mathbf{A}-p_k\mathbf{I})^{-1}\left[(\mathbf{A}+p_k\mathbf{I})\mathbf{G}_k, \sqrt{2p_k}\mathbf{N}_1\mathbf{G}_k, \ldots, \sqrt{2p_k}\mathbf{N}_m\mathbf{G}_k, \sqrt{2p_k}\mathbf{G}\right],\\ \mathbf{D}_{k+1} &= \operatorname{blkdiag}(\mathbf{D}_k, \mathbf{D}_k, \ldots, \mathbf{D}_k, \mathbf{D}),\end{aligned}$$

where we used the MATLAB notation blkdiag$(\cdot)$ for a block diagonal matrix. Now we only have to check for a possible decomposition of the matrix that is returned after applying the matrix function $\mathcal{A}$ to a factorized matrix $\mathbf{GDG}^T$. By the definition of $\mathcal{A}$, it follows that

$$\begin{aligned}\mathcal{A}(\mathbf{GDG}^T) &= \mathbf{AGDG}^T + \mathbf{GDG}^T\mathbf{A}^T + \sum_{j=1}^m \mathbf{N}_j\mathbf{GDG}^T\mathbf{N}_j^T\\ &= \underbrace{\left[\mathbf{AG}, \mathbf{G}, \mathbf{N}_1\mathbf{G}, \ldots, \mathbf{N}_m\mathbf{G}\right]}_{\hat{\mathbf{G}}}\underbrace{\begin{bmatrix}\mathbf{0} & \mathbf{D} & \mathbf{0}\\ \mathbf{D} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}_m\otimes\mathbf{D}\end{bmatrix}}_{\hat{\mathbf{D}}}\underbrace{\left[\mathbf{AG}, \mathbf{G}, \mathbf{N}_1\mathbf{G}, \ldots, \mathbf{N}_m\mathbf{G}\right]^T}_{\hat{\mathbf{G}}^T}.\end{aligned}$$

Since $\hat{\mathbf{D}}$ is symmetric, it follows that $\hat{\mathbf{G}}\hat{\mathbf{D}}\hat{\mathbf{G}}^T$ is also symmetric and thus can be factorized as $\tilde{\mathbf{G}}\tilde{\mathbf{D}}\tilde{\mathbf{G}}^T$, where $\tilde{\mathbf{D}}$ again is diagonal. All other computations in Algorithm 4.4.1 do not influence the diagonal structure of $\mathbf{D}$ and thus allow to preserve the desired factorization and solely operate on the low rank factors $\mathbf{G}$ and $\mathbf{D}$, respectively.

The unsymmetric case

Similarly, one might implement more sophisticated algorithms, which are also applicable in the case that $\mathbf{A}$ and $\mathbf{N}_j$ are unsymmetric. Obviously, there are numerous possible iterative solvers which can be used. However, in this thesis we restrict ourselves to the BiCGstab algorithm. Again, we refer to [90], for a similar discussion of Algorithm 4.4.2. Once more, the only difference is that our version here is dedicated to solving equations of the form (4.12) which has to be taken care of in evaluating $\mathcal{A}$ and the special preconditioner $\mathcal{M}^{-1}$ given by the bilinear ADI iteration. As it has been discussed in [50, 51] for the standard case, unsymmetric matrices might also be tackled by a low rank variant of the GMRES method together with a suitable preconditioning technique.

Just as solving the Lyapunov equation by a projection onto a smaller subspace, the use of an iterative linear solver has the advantage that we do not need the assumption $\sigma\left(\mathcal{L}^{-1}\Pi\right) < 1$ as long as we refrain from preconditioning by the bilinear ADI iteration which in case of $\sigma\left(\mathcal{L}^{-1}\Pi\right) \geq 1$ will not converge. For $\sigma\left(\mathcal{L}^{-1}\Pi\right) \geq 1$, we can still precondition with a number of linear ADI iterations which we assume to be at least a rough approximation to the inverse of the bilinear Lyapunov operator, see also the discussion in [43] and the following examples.

Algorithm 4.4.2 Preconditioned BiCGstab method

Input: Matrix functions $\mathcal{A}, \mathcal{M} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, low rank factor $\mathbf{B}$ of right-hand side $\mathcal{B} = -\mathbf{B}\mathbf{B}^T$. Truncation operator $\mathcal{T}$ w.r.t. relative accuracy ϵ_{rel}.

Output: Low rank approximation $\mathbf{P}_{\hat{n}} = \mathbf{G}\mathbf{D}\mathbf{G}^T$ with $||\mathcal{A}(\mathbf{P}) - \mathcal{B}||_F \leq$ tol.

```
$\mathbf{P}_0 = \mathbf{0}$, $\mathbf{R}_0 = \mathcal{B}$, $\tilde{\mathbf{R}} = \mathcal{B}$, $\rho_0 = \langle \tilde{\mathbf{R}}, \mathbf{R}_0 \rangle$, $\mathbf{P}_0 = \mathbf{R}_0$, $\hat{\mathbf{P}}_0 = \mathcal{M}^{-1}(\mathbf{P}_0)$, $\mathbf{V}_0 = \mathcal{A}(\hat{\mathbf{P}}_0)$, $k = 0$
while $||\mathbf{R}_k||_F >$ tol do
    $\omega_k = \frac{\langle \tilde{\mathbf{R}}, \mathbf{R}_k \rangle}{\langle \tilde{\mathbf{R}}, \mathbf{V}_k \rangle}$,
    $\mathbf{S}_k = \mathbf{R}_k - \omega_k \mathbf{V}_k$                          Optionally: $\mathbf{S}_k \leftarrow \mathcal{T}(\mathbf{S}_k)$
    $\hat{\mathbf{S}}_k = \mathcal{M}^{-1}(\mathbf{S}_k)$,                          Optionally: $\hat{\mathbf{S}}_k \leftarrow \mathcal{T}(\hat{\mathbf{S}}_k)$
    $\mathbf{T}_k = \mathcal{A}(\hat{\mathbf{S}}_k)$,                              Optionally: $\mathbf{T}_k \leftarrow \mathcal{T}(\mathbf{T}_k)$
    if $||\mathbf{S}_k||_F \leq$tol then
        $\mathbf{P}_{\hat{n}} = \mathbf{P}_k + \omega_k \hat{\mathbf{P}}_k$,
        return,
    end if
    $\xi_k = \frac{\langle \mathbf{T}_k, \mathbf{S}_k \rangle}{\langle \mathbf{T}_k, \mathbf{T}_k \rangle}$,
    $\mathbf{P}_{k+1} = \mathbf{P}_k + \omega_k \hat{\mathbf{P}}_k + \xi_k \hat{\mathbf{S}}_k$,                       $\mathbf{P}_{k+1} \leftarrow \mathcal{T}(\mathbf{P}_{k+1})$
    $\mathbf{R}_{k+1} = \mathcal{B} - \mathcal{A}(\mathbf{P}_{k+1})$,                    Optionally: $\mathbf{R}_{k+1} \leftarrow \mathcal{T}(\mathbf{R}_{k+1})$
    if $||\mathbf{R}_{k+1}||_F \leq$tol then
        $\mathbf{P}_{\hat{n}} = \mathbf{P}_k$,
        return,
    end if
    $\rho_{k+1} = \langle \tilde{\mathbf{R}}, \mathbf{R}_{k+1} \rangle$,
    $\beta_k = \frac{\rho_{k+1}}{\rho_k} \frac{\omega_k}{\xi_k}$,
    $\mathbf{P}_{k+1} = \mathbf{R}_{k+1} + \beta_k (\mathbf{P}_k - \xi_k \mathbf{V}_k)$,                $\mathbf{P}_{k+1} \leftarrow \mathcal{T}(\mathbf{P}_{k+1})$
    $\hat{\mathbf{P}}_{k+1} = \mathcal{M}^{-1}(\mathbf{P}_{k+1})$,                     Optionally: $\hat{\mathbf{P}}_{k+1} \leftarrow \mathcal{T}(\hat{\mathbf{P}}_{k+1})$
    $\mathbf{V}_{k+1} = \mathcal{A}(\hat{\mathbf{P}}_{k+1})$,                        Optionally: $\mathbf{V}_{k+1} \leftarrow \mathcal{T}(\mathbf{V}_{k+1})$
    $k = k + 1$
end while
$\mathbf{P}_{\hat{n}} = \mathbf{P}_k$
```

4.4.3 Numerical examples

We now study the performance of the proposed methods by means of some standard numerical test examples. The first and the second benchmark examples fulfill the assumptions stated in Theorem 4.4.1, meaning that the bilinear coupling matrix $\mathbf{N}$ is of low rank compared to the system dimension n. Hence, we know that we can indeed expect low rank approximations of the generalized Lyapunov equations as well. However, the third benchmark contains a coupling matrix $\mathbf{N}$ which has full rank. Nevertheless, we show that there still seems to be a significant singular value decay in the solution matrix $\mathbf{P}$ which allows for low rank approximations. All simulations were generated on an Intel®Xeon®Westmere X5650 with 2.66GHz, 48GB DDR3 RAM and MATLAB® Version 7.11.0.584 (R2010b) 64-bit (glnxa64).

Heat equation

The first example we want to discuss is the heat equation we already discussed in the context of $\mathcal{H}_2$-optimal model order reduction of bilinear systems. This time, we consider a setting with slightly different boundary conditions of the form

$$\begin{aligned} x_t &= \Delta x && \text{in } \Omega = (0,1)\times(0,1),\\ n\cdot\nabla x &= 0.5\cdot u(x-1) && \text{on } \Gamma_1,\\ x &= 0 && \text{on } \Gamma_2,\Gamma_3,\Gamma_4. \end{aligned}$$

The reason for changing the boundary conditions is to obtain only one bilinear coupling matrix $\mathbf{N}$ that additionally is of low rank. Recall that for the case of $\mathcal{H}_2$-optimality, we were particularly interested in the MIMO case which here would destroy the low rank character of the solution matrix $\mathbf{P}$. As it is shown in Figure 4.9, we solve the generalized Lyapunov equation up to a system dimension of $n = 562\,500$, corresponding to a grid consisting of 750 grid points in each direction. As a stopping criterion we choose a relative residual of 10^{-8}. All truncation steps that aim at keeping the rank of the iterates small are performed with a tolerance of 10^{-10}. For the bilinear extension of the ADI iteration, we test several choices of the shift parameters. As previously indicated, we compare the interpolation points resulting from a locally $\mathcal{H}_2$-optimal reduced-order model obtained by Algorithm 4.3.2 with the optimal shifts for the *standard case* derived by Wachspress, see [129]. Interestingly, for $\mathcal{H}_2$-shifts, the residual significantly decreases within the first few iteration steps and then almost stagnates. In particular, for 6 and 8 $\mathcal{H}_2$-shifts, respectively, the stopping criterion is not reached after 100 iteration steps. On the other hand, for 10 $\mathcal{H}_2$-shifts, the bilinear ADI iteration stops after 43 iterations. Moreover, we see that the Wachspress shifts seem to be a very good alternative, leading to a constant decrease in the residual and leading to an accurate approximation after 49 iteration steps. Furthermore, in Figure 4.9, we see that the approximations obtained by using a low rank implementation of the CG method perform the best for this specific example.

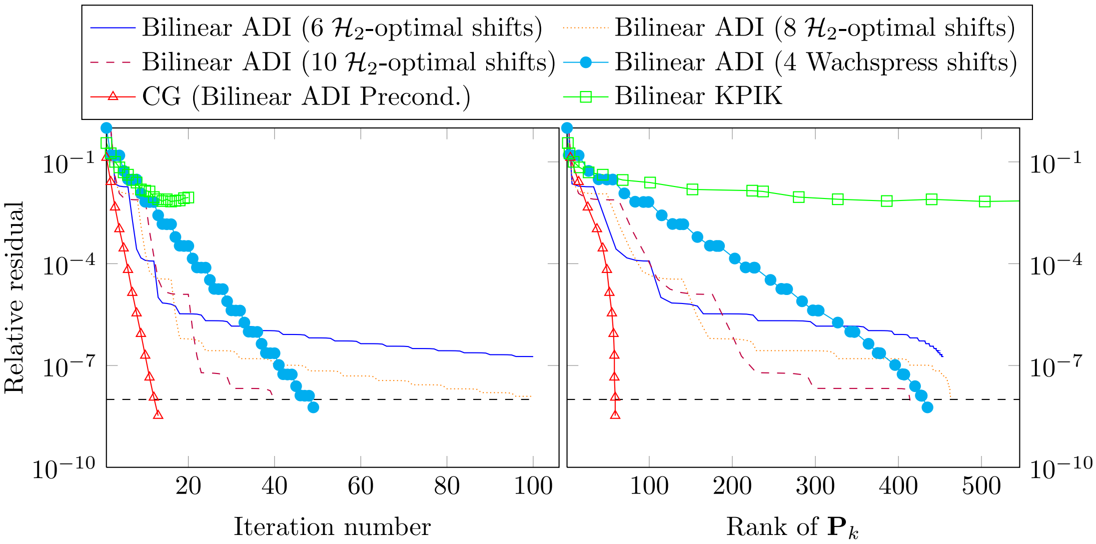


Figure 4.9: Heat equation. Comparison of low rank solution methods for $n = 562\,500$.

Here, as a preconditioner we use three steps of the bilinear ADI iteration. Note that the rank of the final iterate is only 59, while the corresponding relative residual is smaller than 10^{-8}. On the other hand, the extension of the KPIK method stagnates at a relative residual of the order 10^{-2} and for that reason is stopped after 20 iterations.

A nonlinear RC circuit

Our second example is the scalable RC ladder from Section 4.2. As we have seen, for this example, the bilinearization process leads to a bilinear coupling matrix $\mathbf{N}$ which is only of rank k, where k denotes the number of resistors in the system. Here, the computations were done for $k = 500$ and consequently $n = 250\,500$. Moreover, we scale the matrix $\mathbf{N}$ by a factor of 0.5 in order to ensure a positive (semi-)definite solution $\mathbf{P}$ of the associated generalized Lyapunov equation. We use the same stopping criterion and truncation error as above. In Figure 4.10, we again compare the performance of the bilinear ADI iteration for two different sets of shift parameters. Similar to the previous example, using $\mathcal{H}_2$-optimal shift parameters leads to a very fast decrease of the residual within a few iteration steps before the speed of convergence becomes drastically slower. Moreover, we see that the residual curve is not monotone and exhibits several peaks starting from iteration number 40. Again, the optimal linear parameters proposed by Wachspress outperform a larger set of $\mathcal{H}_2$-optimal shifts due to the fact that the residual decreases linearly without being slowed down. Furthermore, in Figure 4.10, we see the results for two different preconditioners for the low rank implementation of the BiCGstab method. The first one is the low rank version of the bilinear ADI iteration which we have previously discussed in detail and for which we compute the first four iterates in each step of the BiCGstab algorithm. The second preconditioner is the standard low rank ADI iteration which we expect to approximate only the inverse of the standard Lyapunov operator. As we

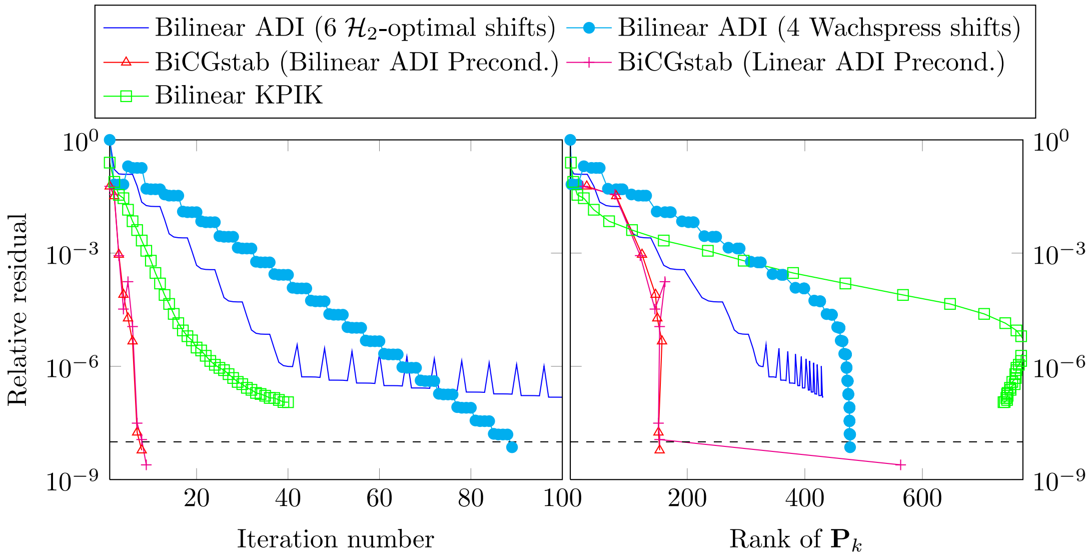


Figure 4.10: RC circuit. Comparison of low rank solution methods for $n = 250\ 000$.

see in Figure 4.10, there is no visible advantage which might allow recommending the first method since both methods perform similarly. However, for the latter approach it takes an additional iteration compared to the bilinear ADI preconditioned method before the stopping criterion is reached. Interestingly, within that last step the rank of the iterate almost triples. Finally, the extension of the KPIK method converged to a relative residual of 10^{-7}. Note that the ranks of the approximations decrease after several steps of the algorithm. This is due to the fact that we solve the reduced Lyapunov equation by means of the bilinear ADI iteration (with residual tolerance 10^{-14}) as well so that in some cases the ranks of the solutions can be further reduced.

Fokker-Planck equation

In order to show that in some cases one might obtain a fast singular value decay even if the bilinear coupling matrix is of full rank, as a final example we once more consider the Fokker-Planck equation. Here, we use $\sigma = \frac{1}{2}$ and spatially discretize the underlying probability distribution function with $n = 10\ 000$ points. As shown in [77], this setting leads to a bilinear matrix $\mathbf{N}$ of rank 10 000. Here, as a stopping criterion we choose a relative residual of 10^{-9}. Accordingly, the truncations are performed with a tolerance of 10^{-11}. In Figure 4.11, we see the convergence history for the bilinear ADI iteration using 2 (pseudo) $\mathcal{H}_2$-optimal shifts. After only 5 iteration steps, the stopping criterion is fulfilled and the rank of the iterate is only 17. On the other hand, the Wachspress shifts require 23 iterations and lead to an approximation of rank 34. Both cases indicate that the full solution $\mathbf{P}$ indeed exhibits a very strong singular value decay. Finally, also the low rank implementation of the BiCGstab method as well as the bilinear KPIK variant converged to approximations with the desired residual of 10^{-9}, although the

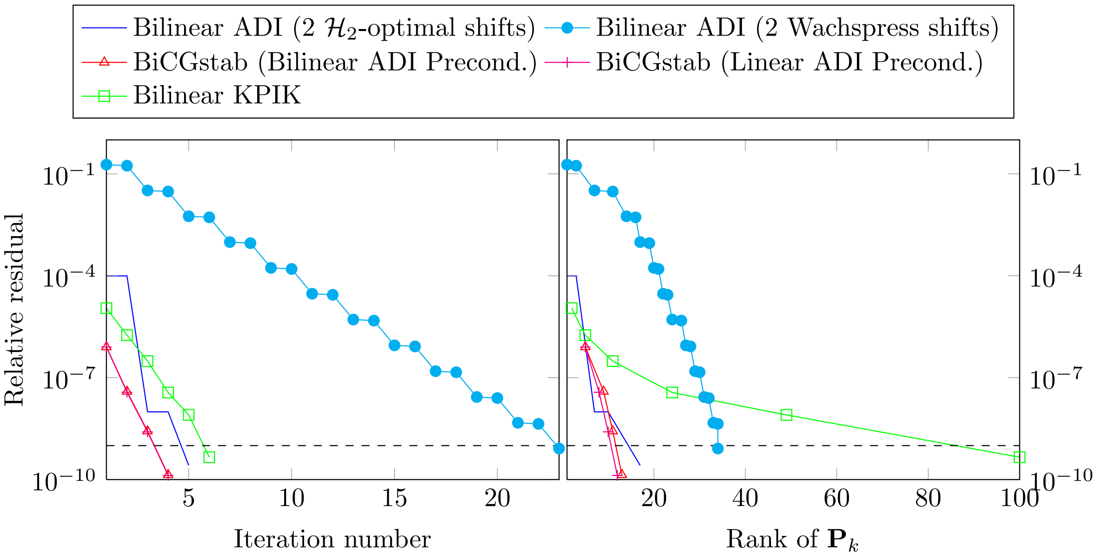


Figure 4.11: Fokker-Planck. Comparison of low rank solution methods for $n = 10\,000$.

approximation of the latter one resulted in having the largest rank.

Remarks on the computational complexity

Based on the above results, it seems reasonable to recommend the use of an iterative linear solver since the number of iterations as well as the rank of the final approximation often is the smallest. However, when choosing a numerical algorithm, the computational complexity clearly has to be taken into account. Unfortunately, a rigorous complexity analysis of our algorithms is hardly possible. This is due to the fact that if the theoretical costs were actually reached, all our algorithms would become infeasible. Let us for example consider the bilinear ADI iteration. We have already seen that in each step we have to solve $(2 + m)$ systems of linear equations. The point is that the corresponding right-hand side theoretically grows from size k up to size $(m+1) \cdot k + m$, where m is the number of inputs. Hence, performing the truncation operation that keeps the growth of the low rank approximation at a decent level, becomes more and more expensive. Nevertheless, the actual growth of the iterates cannot be specified in general and usually is much smaller than the theoretical expectation. Furthermore, the computation of good shift parameters is even more complicated than in the standard case such that the total costs might exceed those of the other methods, depending on the speed of convergence.

Regarding the costs of an iterative solver like CG or BiCGstab, one has to keep in mind that using an appropriate preconditioner is essential to obtain a small iteration number. Since we proposed to precondition with a few steps of the bilinear ADI iteration, the complexity also depends on the rank of the current iterate. To be more specific, we can record that the major costs result from the truncation operator and, in case of the projection-based approach, from the necessary orthogonalization by a modified Gram-

	Heat equation	RC circuit	Fokker-Planck
Bilinear ADI 2 $\mathcal{H}_2$ shifts	-	-	1.733 (1.578)
Bilinear ADI 6 $\mathcal{H}_2$ shifts	144065 (2274)	20900 (3091)	-
Bilinear ADI 8 $\mathcal{H}_2$ shifts	135711 (3177)	-	-
Bilinear ADI 10 $\mathcal{H}_2$ shifts	33051 (4652)	-	-
Bilinear ADI 2 Wachspress shifts	-	-	6.617 (4.562)
Bilinear ADI 4 Wachspress shifts	41883 (2500)	18046 (308)	-
CG (Bilinear ADI precond.)	15640	-	-
BiCG (Bilinear ADI precond.)	-	16131	11.581
BiCG (Linear ADI precond.)	-	12652	9.680
KPIK	7093	19778	8.555

Table 4.1: Comparison of computation times in seconds for different low rank methods. Values in brackets denote the time needed for computing the shift parameters.

Schmidt process of the generated Krylov subspaces.

In order to provide a clear picture at least for the examples studied here, we list the total computation times for all low rank methods in Table 4.1. For the bilinear ADI iteration, we always include the time needed for the computation of the shift parameters. For the first two large-scale examples, we conclude that the low rank implementations of the iterative solvers perform the best. Recall that the small computation time for KPIK in the case of the heat equation is due to the fact that the residual stagnated and the method was stopped after 20 iteration steps. Although choosing 10 $\mathcal{H}_2$-optimal shift parameters leads to faster convergence than 4 Wachspress shifts, none of the methods can compete with the low rank CG implementation. A similar conclusion can be drawn for the RC circuit. However, here it is important to note that preconditioning with the bilinear ADI iteration does not seem to pay off and preconditioning with the linear ADI iteration thus can be recommended. Finally, for the Fokker-Planck equation, the bilinear ADI iteration implemented with 2 (pseudo) $\mathcal{H}_2$-optimal shifts performs the best. After 1.733 seconds an approximation of the solution is computed. On the other hand, for this example the iterative solvers cannot compete with the other techniques. Still, based on Table 4.1, it seems reasonable to recommend a low rank implementation of an iterative solver as a first method of choice for very large-scale generalized Lyapunov equations. Nevertheless, almost all methods allow to solve these equations approximately up to a dimension of 10^6 in MATLAB in less than one day.

4.4.4 Conclusions

In conclusion, we have studied a class of generalized Lyapunov equations which naturally arise in the context of model order reduction of bilinear control systems as well as for the stability analysis of linear stochastic differential equations. Under certain low rank assumptions on the involved matrices, we have shown that one can expect a rapid decrease of the singular values of the solutions, justifying the construction of low rank approximations of the form $\mathbf{P} = \mathbf{Z}\mathbf{Z}^T$ and $\mathbf{P} = \mathbf{G}\mathbf{D}\mathbf{G}^T$, respectively. We have further proposed some extensions of successful linear low rank approximation procedures and have investigated their usefulness by means of certain large-scale numerical test examples. While the performance is quite good and allows for solving generalized Lyapunov equations of up to the order 562 500, some problems are still open. Here, we think of the solution of the generalized rational Zolotarev problem which in the *standard case* leads to optimal shift parameters for the ADI iteration. Moreover, it seems to be an interesting topic of further research to study the observed fast singular value decay of the solution matrix $\mathbf{P}$ in cases when the bilinear coupling matrix $\mathbf{N}$ is of full rank.

4.5 Applications to parametric model order reduction

Besides being large-scale, dynamical processes often are additionally subject to certain parameter variations resulting from, e.g., optimization of geometry and topology in micro-electro-mechanical systems (MEMS) design. MOR techniques usually should allow multiple simulations for varying parameter values in design studies or optimization algorithms. In the case of parameters being constant during one simulation cycle, there exist several generalizations of linear MOR methods like moment-matching, balanced truncation and rational interpolation. For a detailed overview on this research topic, we refer to [13] and references therein. However, the situation becomes rather complicated if the parameters vary with time. Here, efficient reduction methods are still an open question. Let us have a look at so-called linear parameter-varying (LPV) systems

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \sum_{i=1}^{d} p_i(t)\mathbf{A}_i\,\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \qquad \mathbf{y}(t) \;= \mathbf{C}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{x}_0, \tag{4.60}$$

where $\mathbf{A}, \mathbf{A}_i \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{p\times n}$. The crucial observation, see also [18], is that the above structure almost coincides with the structure of bilinear control systems of the form (4.1). In fact, setting $\tilde{\mathbf{u}}(t) = [\mathbf{u}(t)^T, p_1(t), \ldots, p_d(t)]^T$, $\tilde{\mathbf{B}} = \begin{bmatrix}\mathbf{B}, \mathbf{0}, \ldots, \mathbf{0}\end{bmatrix}$, $\tilde{\mathbf{N}}_k = 0$ for $k \leq m$ and $\tilde{\mathbf{N}}_{m+k} = \mathbf{A}_k$ for $k = 1, \ldots, d$, we can interpret the LPV system (4.60) as a special type of a bilinear system (4.1) with $m + d$ input variables. Hence, if we now reduce this bilinear system we certainly preserve its parametric character and can return to the LPV structure by re-interpretation of the reduced bilinear system as LPV system. Obviously, the tradeoff we have to accept is that instead of a parameter-dependent system we are now faced with a parameter-independent system which exhibits a bilinear nonlinearity. However, as we have seen throughout this section, there exist several reliable techniques that can be used for the construction of a reduced model. Another more subtle issue is that one usually knows a parameter range of interest that may be used for the reduction. As the author has already discussed in [56], the interpretation as a bilinear system does not take care of this fact and may not be the optimal reduction method. On the other hand, at least for time-varying parameters, there so far do not really exist successful alternatives. Moreover, note that the quality of a reduction can be influenced by scaling the parameter values. For example, if we use

$$p_i(t)\mathbf{A}_i = \left(\frac{1}{\gamma}p_i(t)\right)(\gamma\mathbf{A}_i),$$

we can increase or decrease the impact of $\mathbf{A}_i$ on the model reduction procedure. Unfortunately, so far we are not aware of a rigorous analysis of the optimal choice of the parameter values which still might be an interesting point of further research.

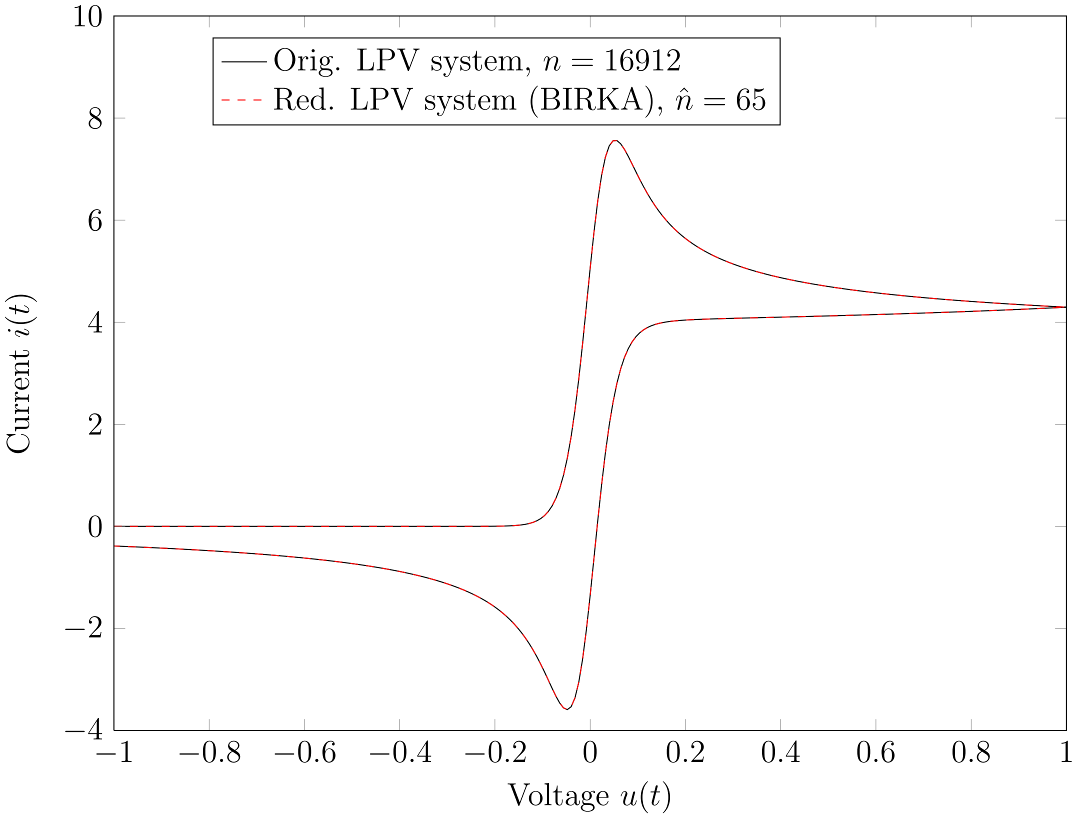


Figure 4.12: Results for cyclic voltammetry for voltage $\frac{du}{dt} = \pm 0.5$.

In order to show that the above idea still may be useful in parametric model reduction, we present a numerical example which typically arises in electrochemical microscopy and has been studied in detail in [54]. The original model is a time-varying system with non-zero initial condition which can be interpreted as a linear parametric system, see [54]. The results for the original model with $n = 16912$ are compared with the ones produced by a reduced-order system with $\hat{n} = 65$ as shown in 4.12. Here, we use the optimal $\mathcal{H}_2$-MOR technique specified in Algorithm 4.3.2. As one can see, for a common excitation, our method leads to a very accurate reproduction of the system dynamics.

CHAPTER 5

NONLINEAR SYSTEMS

Contents

5.1 Introduction

Now that we have carefully studied model order reduction for bilinear control systems, we come to an even more general class of nonlinear control affine systems of the form

$$\mathbf{\Sigma}_N : \quad \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}u(t), \\ y(t) = \mathbf{c}^T\mathbf{x}(t), \quad \mathbf{x}_0 = \mathbf{0}, \end{cases} \tag{5.1}$$

where $\mathbf{f}:\mathbb{R}^n \rightarrow \mathbb{R}^n$ is a nonlinear state evolution function and $\mathbf{b}$, $\mathbf{c} \in \mathbb{R}^n$ are the usual input and output vectors, respectively. While assuming a zero initial condition again is no necessary assumption for the following concepts to be valid, we want to focus on the simpler case of a SISO system which can be seen as a restriction. However, the basic ideas also hold true in the MIMO case although the derivations will become rather cumbersome.

As already mentioned in the previous chapters, if the state dimension n of the system becomes too large, one usually is interested in a reduced-order model of the same structure

$$\hat{\mathbf{\Sigma}}_N : \quad \begin{cases} \dot{\hat{\mathbf{x}}}(t) = \hat{\mathbf{f}}(\hat{\mathbf{x}}(t)) + \hat{\mathbf{b}}u(t), \\ \hat{\mathbf{y}}(t) = \hat{\mathbf{c}}^T\hat{\mathbf{x}}(t), \quad \hat{\mathbf{x}}_0 = \mathbf{0}, \end{cases} \tag{5.2}$$

with $\hat{\mathbf{f}}:\mathbb{R}^{\hat{n}} \rightarrow \mathbb{R}^{\hat{n}}$ and $\hat{\mathbf{b}}$, $\hat{\mathbf{c}} \in \mathbb{R}^{\hat{n}}$. In contrast to linear systems, one of the main difficulties clearly is the construction of a reduced evolution function $\hat{\mathbf{f}}$. Trajectory-based methods like proper orthogonal decomposition (POD) rely on a Galerkin projection $\mathbf{P} = \mathbf{V}\mathbf{V}^T$ and compute $\hat{\mathbf{f}} = \mathbf{V}^T\mathbf{f}(\mathbf{V}\hat{\mathbf{x}})$. While this definitely preserves the nonlinear structure of the original system, it also displays a major bottleneck of the classical POD approach. To be more precise, note that the function $\mathbf{f}$ still has to be evaluated on the original state space $\mathbb{R}^n$, making the simulation of the reduced-order system too expensive. However, there exist several ways to circumvent this problem, e.g., the empirical interpolation method (EIM), see [11], missing point estimation (MPE), see [7], best points interpolation method (BPIM), see [105], and the discrete empirical interpolation method (DEIM), see [39]. For a detailed discussion on POD, we refer to e.g. [35, 39, 92, 93]. Motivated by the same idea, the reduced basis method is a further popular and successful approach in the context of nonlinear model order reduction, see e.g. [11, 70].

Another way is to replace the nonlinearity by a weighted combination of linear systems which then can be efficiently treated by well-known linear reduction methods like balanced truncation or interpolation. For a more detailed insight into the resulting trajectory piecewise linear (TPWL) method, the reader is referred to [114], where more information can be found.

So far, the above mentioned methods all share the common drawback of input dependency, i.e., in order to construct a reduced-order model one at first needs some snapshots of a given or computed solution trajectory of the original model. If this has been done, one indeed can get very accurate approximations of the system. However, as soon as the input function is varied, which is common in control, optimization and design problems, no rigorous assertions on the error for the new dynamics can be specified. In this chapter, we pick up a method which extends the concept of interpolation or moment matching, respectively, discussed for linear systems in, e.g., [71]. The main idea was introduced in [72], where the author shows how to transform a specific class of nonlinear control systems into a system of so-called quadratic-bilinear differential algebraic equations (QBDAEs). For those systems, in [72] an approximation procedure based on generalized moment matching about the interpolation point 0 was discussed and evaluated by means of some typical numerical test examples in the context of nonlinear model reduction. Basically, the method can be seen as a suitable extension of techniques that have been discussed in [9, 33, 112, 111]. The main advantage of this approach is that it tries to construct a reduced-order model that aims at capturing the input-output behavior of the underlying system, making it input independent and thus allowing to use the reduced-order model for varying controls.

The structure of the chapter is as follows. In the next section, we review the basic idea from [72] and further state the main properties of quadratic-bilinear differential algebraic equations. This includes a recapitulation of the concept of variational analysis which allows to replace the nonlinear system by a nested sequence of pseudo-linear subsystems and subsequently opens up the possibility to derive generalized transfer functions. In Section 5.3, we discuss how the computation of a reduced-order model can be efficiently realized by means of basic tensor theoretic tools from Chapter 2. The main result is proven in Section 5.4, where we show how to construct appropriate two-sided moment matching methods for quadratic-bilinear differential algebraic equations that extend existing concepts from the literature. Finally, we carefully implement and test the method for some numerical examples in Section 5.5 and underline its advantages and disadvantages, respectively. We conclude with a summary and an outlook for topics of further research.

5.2 Quadratic-bilinear DAEs

In this section, we review the basic properties of QBDAEs. These systems are of the form

$$\Sigma_Q : \quad \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{H}\mathbf{x}(t) \otimes \mathbf{x}(t) + \mathbf{N}\mathbf{x}(t)u(t) + \mathbf{b}u(t), \\ \quad y(t) = \mathbf{c}^T\mathbf{x}(t), \quad \mathbf{x}_0 = \mathbf{0}, \end{cases} \tag{5.3}$$

where $\mathbf{E}, \mathbf{A}, \mathbf{N} \in \mathbb{R}^{n\times n}$, $\mathbf{H} \in \mathbb{R}^{n\times n^2}$, and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$. Analog to more general nonlinear systems of the form (5.1), here $u(t), y(t) \in \mathbb{R}$ are input and output variables, respectively. Note that the matrix $\mathbf{H}$ which denotes the Hessian of the right hand side exhibits a special symmetric structure. To be more precise, for two arbitrary vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we can always ensure that it holds

$$\mathbf{H}\left(\mathbf{u} \otimes \mathbf{v}\right) = \mathbf{H}\left(\mathbf{v} \otimes \mathbf{u}\right).$$

Since this might not be obvious, let us study a simple example that illustrates the main point.

Example 5.2.1. *Let us consider a two-dimensional purely quadratic control systems of* *the form*

$$\dot{\mathbf{x}}(t) = \mathbf{H}\mathbf{x}(t) \otimes \mathbf{x}(t), \qquad \textit{with} \qquad \mathbf{H} = \begin{bmatrix} a & b & c & d \\ e & f & g & h \end{bmatrix}.$$

Writing down the dynamics explicitly, we obtain

$$\begin{aligned} \dot{x}_1(t) &= ax_1(t)^2 + bx_1(t)x_2(t) + cx_2(t)x_1(t) + dx_2(t)^2, \\ \dot{x}_2(t) &= ex_1(t)^2 + fx_1(t)x_2(t) + gx_2(t)x_1(t) + hx_2(t)^2. \end{aligned}$$

Using $j = \frac{b+c}{2}$ *and* $k = \frac{f+g}{2}$, *the above system is equivalent to*

$$\begin{aligned} \dot{x}_1(t) &= ax_1(t)^2 + jx_1(t)x_2(t) + jx_2(t)x_1(t) + dx_2(t)^2, \\ \dot{x}_2(t) &= ex_1(t)^2 + kx_1(t)x_2(t) + kx_2(t)x_1(t) + hx_2(t)^2. \end{aligned}$$

Hence, we can replace $\mathbf{H}$ *by* $\tilde{\mathbf{H}} = \begin{bmatrix} a & j & j & d \\ e & k & k & h \end{bmatrix}$. *One now easily observes that for arbitrary* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, *it holds that*

$$\tilde{\mathbf{H}}\left(\mathbf{u} \otimes \mathbf{v}\right) = \tilde{\mathbf{H}}\left(\mathbf{v} \otimes \mathbf{u}\right) = \begin{bmatrix} au_1v_1 + ju_1v_2 + ju_2v_1 + ku_2v_2 \\ eu_1v_1 + ku_1v_2 + ku_2v_1 + fu_2v_2 \end{bmatrix}.$$

Obviously, the above also holds true for $n > 2$.

5.2.1 Quadratic-bilinerization of nonlinear systems

As we have already mentioned, a large class of smooth nonlinear control affine systems can be transformed into a system of QBDAEs. This is done via introducing new state variables for the occurring nonlinearities of the underlying control system. The new dynamics then can be derived by symbolic differentiation or adding algebraic constraints. For example, if the dynamics of a nonlinear system are given via

$$\begin{aligned} \dot{x}_1 &= \exp(-x_2) \cdot \sqrt{x_1^2 + 1}, \\ \dot{x}_2 &= -x_2 + u, \end{aligned}$$

we can introduce two new state variables $x_3 := \exp(-x_2)$ and $x_4 := \sqrt{x_1^2+1}$. As a result, we can transform the above system as follows

$$\begin{aligned}
\dot{x}_1 &= x_3 x_4,\\
\dot{x}_2 &= -x_2 + u,\\
\dot{x}_3 &= -\exp(-x_2)\dot{x}_2 = -x_3 x_2 + x_3 u,\\
\dot{x}_4 &= \frac{1}{2(\sqrt{x_1^2+1})} 2x_1\dot{x}_1 = x_1 x_3.
\end{aligned}$$

Hence, we have found a quadratic bilinear system of dimension 4 whose solution is also a solution of the original nonlinear system. In a similar way, we may proceed for common nonlinear functions such as $\sin(x), \cos(x), x^\beta, \frac{x}{k+x}$, see [72]. In general, the transformation is done in two steps. First, one tries to polynomialize the system by suitable variable changes before in the second step, the polynomial system is iteratively simplified to a quadratic bilinear system. As it has been discussed in [72], instead of computing the Lie derivative of the artificially introduced state variables, in special situations it might be advantageous to add the algebraic constraints resulting from the introduction of the variables. However, for our purposes this is not of particular interest and we thus refer to [72] for a discussion on this topic.

Of course, the transformation to a set of QBDAEs is not unique in general and the question arises if there exists a *minimal* transformation. So far, this issue has not been considered and there does not seem to be a trivial answer to that question. Moreover, note that for the transformation of the original system it is desirable to have nonlinearities that are given by (a composition of) uni-variable functions. Especially for systems of ODEs that result from the semi-discretization of an underlying PDE, this is often fulfilled, rendering this approach quite promising.

It should be mentioned that the idea of the above transformation has already been known as McCormick-Relaxation for several years, see [98]. The fact that the idea has not been used for model reduction purposes might be surprising. On the other hand, at a first glance it seems counterintuitive to first increase the state dimension of a control system which actually is to be reduced.

Before we proceed with the concepts of variational analysis for these systems, we mention some differences to the theory discussed in [72]. There, the author includes a term of the form

$$\mathbf{L}(\mathbf{x}(t)\otimes\mathbf{x}(t))u(t), \quad \mathbf{L}\in\mathbb{R}^{n\times n^2}.$$

Although it might further increase the state dimension of a transformed system, it should be emphasized that by introducing a new state variable $\mathbf{z}(t) := \mathbf{x}(t)\otimes\mathbf{x}(t)$, the nonlinearity becomes purely bilinear, i.e. $\mathbf{L}\mathbf{z}(t)u(t)$. Since this simplifies the structure of the transfer functions that we introduce in the following, we always assume that the system under consideration does not contain multiplicative couplings of quadratic and bilinear variables. Moreover, in [72], the systems are denoted as quadratic-linear since the state variable $\mathbf{x}(t)$ appears quadratically while the input variable appears linearly. Since one can interpret system (5.3) as a combination of a purely quadratic system and a bilinear control system we use the notation QBDAE.

5.2.2 Variational analysis for nonlinear systems

Let us now turn our attention to the analysis of QBDAEs. The key idea in the analysis of nonlinear systems is to express the solution by means of a Volterra series analog to the one for bilinear systems in (4.4). Although this concept can also be applied for more general linear-analytic systems (see [115, Section 3.4]), we illustrate the approach for the special case of QBDAEs. Here, we follow the discussion in [115, Section 3.4] and present the *variational analysis* approach. As a first step, we want to assume that the system (5.3) is forced by an input of the form $\alpha u(t)$. Due to the fact that a system of QBDAEs belongs to the class of linear-analytic systems, we may assume that the solution $\mathbf{x}(t)$ of (5.3) exhibits an analytic representation and thus can be written as

$$\mathbf{x}(t) = \alpha \mathbf{x}_1(t) + \alpha^2 \mathbf{x}_2(t) + \alpha^3 \mathbf{x}_3(t) + \dots \tag{5.4}$$

where $\mathbf{x}_i \in \mathbb{R}^n$. Next, we insert the above expressions for input and response, respectively, into the state space representation of $\mathbf{\Sigma}_Q$. Hence, for the state equation (5.3), we obtain

$$\begin{aligned}\mathbf{E}\left(\alpha\dot{\mathbf{x}}_1(t) + \alpha^2\dot{\mathbf{x}}_2(t) + \dots\right) = \mathbf{A}&\left(\alpha\mathbf{x}_1(t) + \alpha^2\mathbf{x}_2(t) + \dots\right)\\ &+ \mathbf{H}\left(\alpha\mathbf{x}_1(t) + \alpha^2\mathbf{x}_2(t) + \dots\right) \otimes \left(\alpha\mathbf{x}_1(t) + \alpha^2\mathbf{x}_2(t) + \dots\right)\\ &+ \mathbf{N}\left(\alpha\mathbf{x}_1(t) + \alpha^2\mathbf{x}_2(t) + \dots\right)\alpha u(t) + \mathbf{b}\alpha u(t).\end{aligned}$$

Finally, if we collect all terms α^i corresponding to powers of α, we obtain dynamical equations for each of the state variables, i.e.,

$$\begin{aligned}\mathbf{E}\dot{\mathbf{x}}_1(t) &= \mathbf{A}\mathbf{x}_1(t) + \mathbf{b}u(t),\\ \mathbf{E}\dot{\mathbf{x}}_2(t) &= \mathbf{A}\mathbf{x}_2(t) + \mathbf{H}\mathbf{x}_1(t) \otimes \mathbf{x}_1(t) + \mathbf{N}\mathbf{x}_1(t)u(t),\\ \mathbf{E}\dot{\mathbf{x}}_3(t) &= \mathbf{A}\mathbf{x}_3(t) + \mathbf{H}\left(\mathbf{x}_1(t) \otimes \mathbf{x}_2(t) + \mathbf{x}_2(t) + \mathbf{x}_1(t)\right) + \mathbf{N}\mathbf{x}_2(t)u(t),\\ &\vdots\end{aligned}$$

The advantage of this approach is that the solution $\mathbf{x}(t)$ can be derived by solving a series of nonlinearly coupled linear systems. In particular, this means that we can start by integrating the first *subsystem* in order to get

$$\mathbf{x}_1(t) = \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{b}u(\tau)\mathrm{d}\tau.$$

If we consider this expression as a pseudo-input for the second equation, we can easily derive an expression for $\mathbf{x}_2(t)$. Continuing in this manner, we finally arrive at the desired Volterra series representation for $\mathbf{\Sigma}_Q$. As already mentioned, the basic idea is well-known and its origin goes back to works by Euler, Cauchy and Poincaré, see [115]. Another derivation of the variational expansion approach is discussed in detail in [63].

5.2.3 Generalized transfer functions of QBDAEs

A similar technique allows an input-output characterization in the frequency domain. Again, we just recapitulate the presentation from [115, Section 3.5]. Here, the essential idea is motivated by the following property of a stable linear continuous time-invariant control system. Let us assume that such a system is driven by an input signal $u(t) = e^{\lambda t}, \lambda > 0$. Due to the explicit solution formula, we know that it holds

$$y(t) = \int_0^\infty h(\sigma)e^{(t-\sigma)\lambda}\mathrm{d}\sigma = \int_0^\infty h(\sigma)e^{-\lambda\sigma}\mathrm{d}\sigma \; e^{\lambda t} = H(\lambda)e^{\lambda t},$$

where $H(\lambda) = \mathbf{c}^T(\lambda\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$ denotes the transfer function of the linear system. Hence, the output signal to a *growing exponential* signal is simply scaled by the transfer function. Moreover, for a linear combination of growing exponentials, i.e.,

$$u(t) = \sum_{i=1}^p \alpha_i e^{\lambda_i t}, \quad \lambda_1, \ldots, \lambda_p > 0,$$

the output is given by

$$y(t) = \sum_{i=1}^p \alpha_i H(\lambda_i)e^{\lambda_i t}.$$

In order to keep the derivations clear, in the following we restrict ourselves to the computation of the first two transfer functions for the system $\mathbf{\Sigma}_Q$. Since we denoted the Hessian of the system by $\mathbf{H}$, we use a slightly different notation for the resulting transfer functions $G_1(s_1)$ and $G_2(s_1, s_2)$. Let us now consider an input of the form $u(t) = e^{s_1 t} + e^{s_2 t}$ which is supposed to yield a transient response

$$\mathbf{x}(t) = \mathbf{G}_{10}e^{s_1 t} + \mathbf{G}_{01}e^{s_2 t} + \mathbf{G}_{20}e^{2s_1 t} + \mathbf{G}_{02}e^{2s_2 t} + \mathbf{G}_{11}e^{(s_1+s_2)t}.$$

Inserting this expression into the state equation (5.3) and comparing the coefficients then leads to the first two generalized *symmetric* transfer functions

$$
\begin{aligned}
G_1(s_1) &= \mathbf{c}^T \underbrace{(s_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{b}}_{\mathbf{F}(s_1)}, \\
G_2(s_1, s_2) &= \frac{1}{2}\mathbf{c}^T \left((s_1+s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{H}\left(\mathbf{F}(s_1)\otimes\mathbf{F}(s_2) + \mathbf{F}(s_2)\otimes\mathbf{F}(s_1)\right) \\
&\quad + \frac{1}{2}\mathbf{c}^T \left((s_1+s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{N}\left(\mathbf{F}(s_1) + \mathbf{F}(s_2)\right) \\
&= \mathbf{c}^T \left((s_1+s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{H}\left(\mathbf{F}(s_1)\otimes\mathbf{F}(s_2)\right) \\
&\quad + \frac{1}{2}\mathbf{c}^T \left((s_1+s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{N}\left(\mathbf{F}(s_1) + \mathbf{F}(s_2)\right).
\end{aligned}
$$

Similarly, one can derive higher order transfer functions, see e.g. [72, 115]. As we can see, the first two transfer functions of $\mathbf{\Sigma}_Q$ generalize the theory for linear control systems. However, similarly to the case of bilinear control systems, a meaningful interpretation of the frequency variables s_1 and s_2 cannot be given. The approach thus should rather be considered as an abstract theoretical tool for interpolation-based model order reduction. Especially, it is important to realize that $G_1(s_1)$ and $G_2(s_1, s_2)$ formally describe the input-output relationship of $\mathbf{\Sigma}_Q$ in the frequency domain.

5.3 Computation of a reduced-order model

In this section, we analyze how to efficiently construct a reduced-order quadratic-bilinear system. So far, we have not explicitly stated how to obtain the reduced dynamics. Analog to the previous cases, the assumption that the solution $\mathbf{x}(t)$ can be approximated in a low order subspace of dimension $\hat{n}$, can formally be written as $\mathbf{x}(t) \approx \mathbf{V}\hat{\mathbf{x}}(t)$, with $\mathbf{V} \in \mathbb{R}^{n\times\hat{n}}$ and $\hat{\mathbf{x}} \in \mathbb{R}^{\hat{n}}$. Consequently, we get

$$
\begin{aligned}
\mathbf{E}\mathbf{V}\dot{\hat{\mathbf{x}}}(t) &\approx \mathbf{A}\mathbf{V}\hat{\mathbf{x}}(t) + \mathbf{H}\left(\mathbf{V}\hat{\mathbf{x}}(t)\otimes\mathbf{V}\hat{\mathbf{x}}(t)\right) + \mathbf{N}\mathbf{V}\hat{\mathbf{x}}(t)u(t) + \mathbf{b}u(t), \\
\hat{\mathbf{y}}(t) &= \mathbf{c}^T\mathbf{V}\hat{\mathbf{x}}(t).
\end{aligned}
$$

Assuming that a left projection subspace $\mathbf{W} \in \mathbb{R}^{n\times\hat{n}}$ is given such that it is orthogonal to the residual from the state equation, i.e.,

$$
\mathbf{W} \perp \Big(\mathbf{E}\mathbf{V}\dot{\hat{\mathbf{x}}}(t) - \mathbf{A}\mathbf{V}\hat{\mathbf{x}}(t) + \mathbf{H}\left(\mathbf{V}\hat{\mathbf{x}}(t)\otimes\mathbf{V}\hat{\mathbf{x}}(t)\right) + \mathbf{N}\mathbf{V}\hat{\mathbf{x}}(t)u(t) + \mathbf{b}u(t)\Big),
$$

we arrive at a Petrov-Galerkin type reduced-order model of the form

$$\begin{aligned}\hat{\mathbf{E}}\dot{\hat{\mathbf{x}}}(t) &= \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \hat{\mathbf{H}}\hat{\mathbf{x}}(t) \otimes \hat{\mathbf{x}}(t) + \hat{\mathbf{N}}\hat{\mathbf{x}}(t)u(t) + \hat{\mathbf{b}}u(t), \qquad (5.5)\\ \hat{\mathbf{y}}(t) &= \hat{\mathbf{c}}^T\hat{\mathbf{x}}(t), \qquad (5.6)\end{aligned}$$

with $\hat{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{V},\ \hat{\mathbf{A}} = \mathbf{W}^T\mathbf{A}\mathbf{V},\ \hat{\mathbf{H}} = \mathbf{W}^T\mathbf{H}(\mathbf{V}\otimes\mathbf{V}),\ \hat{\mathbf{b}} = \mathbf{W}^T\mathbf{b},\ \hat{\mathbf{c}} = \mathbf{V}^T\mathbf{c}$.

Although formally analog, in contrast to linear and bilinear control systems, computing the reduced system matrices is quite a tricky task. Recall that even for originally sparse systems, the reduced quantities in general are dense. Hence, we clearly do not ever want to form the projection matrix $\mathbf{V}\otimes\mathbf{V}$ since already storing such a huge matrix might be an unrealizable task even on modern computer architectures. Note that the complexity would be of $\mathcal{O}(n^2\cdot\hat{n}^2)$. Instead, we have to think about alternatives. Due to the properties of the Kronecker product, an obvious way is given by the splitting

$$\mathbf{V}\otimes\mathbf{V} = (\mathbf{V}\otimes\mathbf{I})(\mathbf{I}\otimes\mathbf{V}).$$

This still requires a computational storage complexity of $\mathcal{O}(n^2\cdot\hat{n} + n\cdot\hat{n}^2)$. The special structure of the Hessian matrix $\mathbf{H}\in\mathbb{R}^{n\times n^2}$ is the key to overcome this problem. From Chapter 2, we know that we can interpret $\mathbf{H}$ as the matricization of a tensor $\mathcal{H}\in\mathbb{R}^{n^3}$. Moreover, we already discussed that $\mathbf{H}$ exhibits a special symmetric structure that can always be ensured. Hence, if we assume that $\mathbf{H}$ is the 1-matricization of the 3-tensor $\mathcal{H}$, we can conclude that the remaining matricizations $\mathcal{H}^{(2)}$ and $\mathcal{H}^{(3)}$ coincide. This also implies that

$$\mathbf{w}^T\mathbf{H}(\mathbf{u}\otimes\mathbf{v}) = \mathbf{u}^T\mathcal{H}^{(2)}(\mathbf{v}\otimes\mathbf{w}) = \mathbf{u}^T\mathcal{H}^{(3)}(\mathbf{v}\otimes\mathbf{w}),$$

for arbitrary vectors $\mathbf{u},\mathbf{v},\mathbf{w}\in\mathbb{R}^n$. If we now want to compute the Hessian $\hat{\mathbf{H}}$ of the reduced-order model, we can proceed as follows. We start by a left multiplication with the projection matrix $\mathbf{W}^T$. The result is a matrix $\mathbf{H}_{\mathbf{W}} = \mathbf{W}^T\mathbf{H}$ which still exhibits the symmetric structure of $\mathbf{H}$. Also, we can assume that $\mathbf{H}_{\mathbf{W}}$ is the 1-matricization of a tensor $\mathcal{H}_{\mathbf{W}}\in\mathbb{R}^{\hat{n}\cdot n^2}$. Hence, it follows that

$$\mathbf{H}_{\mathbf{W}}(\mathbf{V}\otimes\mathbf{V}) = \mathbf{V}^T\mathcal{H}_{\mathbf{W}}^{(2)}(\mathbf{I}\otimes\mathbf{V}).$$

Since the above computation leads to a matrix $\mathbf{H}_{\mathbf{VW}}\in\mathbb{R}^{\hat{n}\times\hat{n}\cdot n}$, we can make use of the matricization concept a final time in order to construct $\mathbf{V}^T\mathcal{H}_{\mathbf{VW}}^{(3)}\in\mathbb{R}^{\hat{n}\times\hat{n}\cdot\hat{n}}$. Eventually, we transform the result by reshaping it into the 1-matricization and end up with $\mathbf{W}^T\mathbf{H}(\mathbf{V}\otimes\mathbf{V})$. In summary, this allows to compute the reduced system Hessian without ever explicitly forming the matrix $\mathbf{V}\otimes\mathbf{V}$, leading to a total storage complexity of only

$\mathcal{O}(n \cdot \hat{n})$. One might argue that the matrix $\mathbf{H_W}$ that is obtained after the first projection step is dense and thus the storage complexity still is $\mathcal{O}(n^2 \cdot \hat{n})$. However, for common PDE related problems, this is in general not the case. Note that the columns of the matrix $\mathbf{H}$ correspond to terms of the form $\mathbf{x}_i \cdot \mathbf{x}_j$. Hence, for typical discretizations with homogeneous nonlinearities, most cross terms vanish due to the local nature of FEM or finite difference techniques. Consequently, the associated columns in $\mathbf{H}$ are zero columns and the first left multiplication with $\mathbf{W}^T$ does not create fill in, retaining a sparse structure of $\mathbf{H_W}$. Still, in a worst case scenario where all nonlinearities are coupled among each other, a reduction of the storage complexity is impossible since the $\mathbf{H_W}$ would already be a dense matrix. Fortunately, many typical real-life applications can be handled by the previous procedure allowing for a more efficient computation of the reduced-order model.

5.4 Multimoment matching for QBDAEs

In this section, we discuss how to construct a reduced-order model that approximates the dynamics of the original system by means of interpolating the values and derivatives of its first two transfer functions. In order to emphasize the close connection to known linear concepts, we start with a very brief survey of moment matching methods for SISO linear control systems. If the mass matrix $\mathbf{E}$ of the system under consideration is different from the identity, the transfer function can be computed as $G_1(s_1) = \mathbf{c}^T(s_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{b}$. Assume now that $\sigma \in \mathbb{C}$ is given such that σ is not an eigenvalue of the matrix pencil $(\mathbf{E}, \mathbf{A})$. We now can rewrite the transfer function as follows

$$\begin{aligned} G_1(s_1) &= \mathbf{c}^T(s_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{b} = \mathbf{c}^T((s_1 - \sigma)\mathbf{E} + \sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{b} \\ &= \mathbf{c}^T(\mathbf{I} - (-1)(s_1 - \sigma)(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E})^{-1}(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{b} \end{aligned}$$

Due to the Neumann Lemma, see, e.g., [130, Satz II.1.11], for σ sufficiently close to s we can expand the first inverse and thus locally represent the transfer function as

$$G_1(s_1) = \sum_{i=0}^{\infty} (-1)^i \mathbf{c}^T \left((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}\right)^i (\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{b}\ (s_1 - \sigma)^i. \tag{5.7}$$

For these terms, we introduce the following notation to keep the expressions straightforward.

Definition 5.4.1. *Let $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}, j \in \mathbb{N}$ and $\sigma \in \mathbb{C}$ s.t. $(\sigma\mathbf{E} - \mathbf{A})^{-1}$ exists. Then we set*

$$\mathcal{A}^j_{\mathbf{E},\sigma} := \left((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}\right)^j (\sigma\mathbf{E} - \mathbf{A})^{-1}$$

and

$$\mathcal{A}^{T,j}_{\mathbf{E},\sigma} := \left((\sigma \mathbf{E}^T - \mathbf{A}^T)^{-1}\mathbf{E}^T\right)^j (\sigma \mathbf{E}^T - \mathbf{A}^T)^{-1}.$$

Moreover, later on we make use of certain rational Krylov subspaces that we denote as specified in the following definition.

Definition 5.4.2. *Let* $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{q} \in \mathbf{N}$ *and* $\sigma \in \mathbb{C}$ *Then we define the associated* rational Krylov subspace *as*

$$\mathcal{K}_q\left(\mathbf{E}, \mathbf{A}, \mathbf{b}, \sigma\right) := \mathcal{K}_q\left((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}, (\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{b}\right).$$

Let us come back to the expression (5.7). By means of the previous definitions, we now conclude that it locally holds that $G_1(s_1) = \sum_{i=0}^{\infty} m_i \ \mathbf{b} \ (s_1 - \sigma)^i$, where $m_i = (-1)^i \cdot \mathbf{c}^T \mathcal{A}^i_{\mathbf{E},\sigma}\mathbf{b}$ are the *moments* of the transfer function. The idea of moment matching is to preserve a specified number q of these moments in a reduced-order model. As has been shown in, e.g., [3, 71], this can be achieved by a projection $\mathbf{V}$ with $\mathcal{K}_q(\mathbf{E}, \mathbf{A}, \mathbf{b}, \sigma) \subset \operatorname{span}(\mathbf{V})$. In other words, since the moments are the derivatives of G_1 at σ, we can obtain an interpolation-based reduced-order model by projecting onto a rational Krylov subspace such that it holds

$$\frac{\partial^i G_1}{\partial s_1^i}(\sigma) = \frac{\partial^i \hat{G}_1}{\partial s_1^i}(\sigma), \quad i = 0, \dots, q-1.$$

A similar approach can be derived for quadratic-bilinear system of the form (5.3). As we have seen, the concept of (generalized) transfer functions also exist for QBDAEs. Hence, analog to what we have seen above, we can also expand the second transfer function

$$\begin{aligned} G_2(s_1, s_2) &= \frac{1}{2}\mathbf{c}^T \left((s_1 + s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{H}\left(\mathbf{F}(s_1) \otimes \mathbf{F}(s_2) + \mathbf{F}(s_2) \otimes \mathbf{F}(s_1)\right) \\ &\quad + \frac{1}{2}\mathbf{c}^T \left((s_1 + s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{N}\left(\mathbf{F}(s_1) + \mathbf{F}(s_2)\right) \\ &= \mathbf{c}^T \left((s_1 + s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{H}\left(\mathbf{F}(s_1) \otimes \mathbf{F}(s_2)\right) \\ &\quad + \frac{1}{2}\mathbf{c}^T \left((s_1 + s_2)\mathbf{E} - \mathbf{A}\right)^{-1} \mathbf{N}\left(\mathbf{F}(s_1) + \mathbf{F}(s_2)\right). \end{aligned}$$

Although the existence of two frequency variables s_1 and s_2 results in a lot of freedom in choosing a pair (σ_1, σ_2) of interpolations points, here we stick to the case were both points coincide, i.e., $\sigma_1 = \sigma_2 = \sigma$. Since the physical meaning of the frequency variables s_1 and s_2 is ambiguous anyway, this is not a too severe restriction. Moreover, in the procedure described in Theorem 5.4.1, this assumption allows to recycle vectors for certain Krylov subspaces and thus reduces the required complexity of the resulting algorithm. Accordingly, we then obtain the following multivariate Taylor expansion of the second transfer function

$$\begin{aligned} G_2(s_1,s_2) = &\sum_{i,j,k} m_{i,j,k}(s_1+s_2-2\sigma)^i(s_1-\sigma)^j(s_2-\sigma)^k \\ &+ \sum_{i,\ell_1,\ell_2} m_{i,\ell_1,\ell_2}(s_1+s_2-2\sigma)^i\left((s_1-\sigma)^l+(s_2-\sigma)^m\right), \end{aligned}$$

with *multimoments* given as

$$\begin{aligned} m_{i,j,k} &= (-1)^{i+j+k+1}\cdot\frac{1}{2}\mathbf{c}^T\mathcal{A}^i_{\mathbf{E},2\sigma}\ \mathbf{H}\left(\mathcal{A}^j_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^k_{\mathbf{E},\sigma}\mathbf{b}+\mathcal{A}^k_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^j_{\mathbf{E},\sigma}\mathbf{b}\right), \\ m_{i,\ell_1,\ell_2} &= (-1)^{i+\ell_1}\cdot\frac{1}{2}\mathbf{c}^T\mathcal{A}^i_{\mathbf{E},2\sigma}\ \mathbf{N}\mathcal{A}^{\ell_1}_{\mathbf{E},\sigma}\mathbf{b}+(-1)^{i+\ell_2}\frac{1}{2}\cdot\mathbf{c}^T\mathcal{A}^i_{\mathbf{E},2\sigma}\ \mathbf{N}\mathcal{A}^{\ell_2}_{\mathbf{E},\sigma}\mathbf{b}. \end{aligned}$$

Analog to the transfer function G_1 of the linear subsystem, it is easily seen that $m_{i,j,k}$ and m_{i,ℓ_1,ℓ_2} basically determine the partial derivatives of the second transfer function G_2. Hence, it seems reasonable to construct a reduced-order system in such a way that for a given pair of interpolation points (σ,σ), the derivatives of $\hat{G}_2$ coincide with those of the original transfer function up to a certain order q. The following result now states how to choose an appropriate sequence of nested Krylov subspaces that extends the known results for the special case of a one-sided projection about 0, see [72].

Algorithm 5.4.1 Two-sided multimoment matching for QBDAEs

```
Input: $\mathbf{A}$, $\mathbf{H}$, $\mathbf{N}$, $\mathbf{b}$, $\mathbf{c}$, $\sigma \in \mathbb{C}$, $q_1, q_2 \in \mathbb{N}$, with $q_2 \leq q_1$.
Output: $\hat{\mathbf{A}}$, $\hat{\mathbf{H}}$, $\hat{\mathbf{N}}$, $\hat{\mathbf{b}}$, $\hat{\mathbf{c}}$
$\mathbf{V}_1 = \mathcal{K}_{q_1}(\mathbf{E},\mathbf{A},\mathbf{b},\sigma),\ \mathbf{W}_1 = \mathcal{K}_{q_1}(\mathbf{E}^T,\mathbf{A}^T,\mathbf{c},2\sigma)$
for $i = 1,\dots,q_2$ do
    $\mathbf{V}_2^i = \mathcal{K}_{q_2-i+1}(\mathbf{E},\mathbf{A},\mathbf{N}\mathbf{V}_1(:,i),2\sigma)$
    $\mathbf{W}_2^i = \mathcal{K}_{q_2-i+1}(\mathbf{E}^T,\mathbf{A}^T,\mathbf{N}^T\mathbf{W}_1(:,i),\sigma)$
    for $j = 1,\dots,\min(q_2-i,i)$ do
        $\mathbf{V}_3^{i,j} = \mathcal{K}_{q_2-i-j+2}(\mathbf{E},\mathbf{A},\mathbf{H}\mathbf{V}_1(:,i)\otimes\mathbf{V}_1(:,j),2\sigma)$
        $\mathbf{W}_3^{i,j} = \mathcal{K}_{q_2-i-j+2}(\mathbf{E}^T,\mathbf{A}^T,\mathcal{H}^{(2)}\mathbf{V}_1(:,i)\otimes\mathbf{W}_1(:,j),\sigma)$
    end for
end for
$\mathbf{V} = \operatorname{orth}\left(\operatorname{span}(\mathbf{V}_1)\cup\bigcup_i\operatorname{span}(\mathbf{V}_2^i)\cup\bigcup_{i,j}\operatorname{span}(\mathbf{V}_3^{i,j})\right),$
$\mathbf{W} = \operatorname{orth}\left(\operatorname{span}(\mathbf{W}_1)\cup\bigcup_i\operatorname{span}(\mathbf{W}_2^i)\cup\bigcup_{i,j}\operatorname{span}(\mathbf{W}_3^{i,j})\right)$
$\hat{\mathbf{E}} = \mathbf{W}^T\mathbf{E}\mathbf{V},\ \hat{\mathbf{A}} = \mathbf{W}^T\mathbf{A}\mathbf{V},\ \hat{\mathbf{H}} = \mathbf{W}^T\mathbf{H}\mathbf{V}\otimes\mathbf{V},\ \hat{\mathbf{N}} = \mathbf{W}^T\mathbf{N}\mathbf{V},\ \hat{\mathbf{b}} = \mathbf{W}^T\mathbf{b},\ \hat{\mathbf{c}} = \mathbf{V}^T\mathbf{c}$
```

Theorem 5.4.1. *Let $\boldsymbol{\Sigma} = (\mathbf{E}, \mathbf{A}, \mathbf{H}, \mathbf{N}, \mathbf{b}, \mathbf{c})$ denote a system of quadratic-bilinear differential algebraic equations of dimension n. Let $q_1, q_2 \in \mathbf{N}$ with $q_2 \leq q_1$. Assume that a reduced QBDAE system is constructed by Algorithm 5.4.1. Then, it holds:*

$$\begin{aligned} \frac{\partial^i G_1}{\partial s_1^i}(\sigma) &= \frac{\partial^i \hat{G}_1}{\partial s_1^i}(\sigma), \quad \frac{\partial^i G_1}{\partial s_1^i}(2\sigma) = \frac{\partial^i \hat{G}_1}{\partial s_1^i}(2\sigma), && i = 0, \dots, q_1 - 1, \\ \frac{\partial^{i+j}}{\partial s_1^i s_2^j} G_2(\sigma, \sigma) &= \frac{\partial^{i+j}}{\partial s_1^i s_2^j} \hat{G}_2(\sigma, \sigma), && i + j \leq 2q_2 - 1. \end{aligned}$$

Proof. The assertion for the first transfer function $\hat{G}_1$ immediately follows from known moment matching results for linear systems, see e.g. [3, 71]. Hence, we only have to consider the second transfer function $\hat{G}_2$. Here, it suffices to focus on the contributions of the quadratic part of the system. For the bilinear contributions, we refer to e.g. [33], where two-sided multimoment matching for these systems is studied. Using once more that $\frac{\partial}{\partial y}\left(\mathbf{A}(y)^{-1}\right) = -\mathbf{A}(y)^{-1} \frac{\partial \mathbf{A}(y)}{\partial y} \mathbf{A}(y)^{-1}$, aside from constant factors, we thus have to concentrate on terms of the form

$$\mathbf{c}^T \mathcal{A}^j_{\mathbf{E},2\sigma} \mathbf{H} \left(\mathcal{A}^k_{\mathbf{E},\sigma} \, \mathbf{b} \otimes \mathcal{A}^l_{\mathbf{E},\sigma} \, \mathbf{b} \right),$$

with $j + k + l \leq 2q_2 - 1$ and, w.l.o.g., $k \geq l$. From the results for the first transfer function, we know that

$$\mathbf{V} \hat{\mathcal{A}}^i_{\hat{\mathbf{E}},\sigma} \hat{\mathbf{b}} = \mathcal{A}^i_{\mathbf{E},\sigma} \mathbf{b}, \quad \mathbf{W} \hat{\mathcal{A}}^{T,i}_{\hat{\mathbf{E}}^T,2\sigma} \hat{\mathbf{c}} = \mathcal{A}^{T,i}_{\mathbf{E}^T,2\sigma} \mathbf{c}, \tag{5.8}$$

for $i = 0, \dots, q_1 - 1$. This yields the statement for $j, k, l \leq q_2 - 1$. Let us now assume that $j = 2q_2 - 1, k = l = 0$. Note that we have

$$\mathbf{V}\mathbf{V}^T \mathcal{A}^0_{\mathbf{E},2\sigma} \mathbf{H} \left(\mathcal{A}^0_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^0_{\mathbf{E},\sigma} \mathbf{b} \right) = \mathcal{A}^0_{\mathbf{E},2\sigma} \mathbf{H} \left(\mathcal{A}^0_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^0_{\mathbf{E},\sigma} \mathbf{b} \right). \tag{5.9}$$

This follows from the construction of span $(\mathbf{V})$ and the property of $\mathbf{V}$ being orthonormal.

Next, it holds that

$$\begin{aligned}
\mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right) &= \mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^{T}\mathbf{H}\left(\mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right)\\
&= \mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^{T}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right)\\
&= \mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^{T}\left(\mathcal{A}^{0}_{\mathbf{E},2\sigma}\right)^{-1}\mathcal{A}^{0}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right)\\
&= \mathbf{V}\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^{T}\left(\mathcal{A}^{0}_{\mathbf{E},2\sigma}\right)^{-1}\mathbf{V}\mathbf{V}^{T}\mathcal{A}^{0}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right)\\
&= \mathbf{V}\mathbf{V}^{T}\mathcal{A}^{0}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right)\\
&= \mathcal{A}^{0}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right).
\end{aligned}$$

With the same arguments, one can iteratively show that

$$\mathbf{V}\hat{\mathcal{A}}^{i}_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right) = \mathcal{A}^{i}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right), \tag{5.10}$$

for $i = 0, \ldots, q_2 - 1$. Hence, let us consider

$$\hat{\mathbf{c}}^{T}\hat{\mathcal{A}}^{2q_2-1}_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right).$$

By Definition 5.4.1, we have

$$\begin{aligned}
\hat{\mathcal{A}}^{2q_2-1}_{\hat{\mathbf{E}},2\sigma} &= \left((2\sigma\hat{\mathbf{E}}-\hat{\mathbf{A}})^{-1}\hat{\mathbf{E}}\right)^{q_2-1}\left((2\sigma\hat{\mathbf{E}}-\hat{\mathbf{A}})^{-1}\hat{\mathbf{E}}\right)\left((2\sigma\hat{\mathbf{E}}-\hat{\mathbf{A}})^{-1}\hat{\mathbf{E}}\right)^{q_2-1}(2\sigma\hat{\mathbf{E}}-\hat{\mathbf{A}})^{-1}\\
&= \hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^{T}\mathbf{E}\mathbf{V}\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},2\sigma}.
\end{aligned}$$

Thus, it follows

$$\hat{\mathbf{c}}^{T}\hat{\mathcal{A}}^{2q_2-1}_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right) = \hat{\mathbf{c}}^{T}\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^{T}\mathbf{E}\mathbf{V}\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right).$$

From (5.8) and (5.10), we can conclude that this is equal to

$$\mathbf{c}^{T}\mathcal{A}^{q_2-1}_{\mathbf{E},2\sigma}\mathbf{E}\mathcal{A}^{q_2-1}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right).$$

However, this is the same as

$$\mathbf{c}^{T}\mathcal{A}^{2q_2-1}_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{0}_{\mathbf{E},\sigma}\mathbf{b}\right).$$

In the following, we assume that $k = 2q_2 - 1, j = l = 0$. Analog to (5.9), one easily obtains

$$\mathbf{W}\mathbf{W}^T \mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} c \right) = \mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right).$$

Again, this is true since

$$\mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right) \in \operatorname{span}(\mathbf{W})$$

and $\mathbf{W}^T\mathbf{W} = \mathbf{I}$. With this in mind, we subsequently observe that

$$\begin{aligned}
&\mathbf{W} \hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},\sigma} \mathbf{V}^T \mathcal{H}^{(2)} \left(\mathbf{V} \hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma} \hat{\mathbf{b}} \otimes \mathbf{W} \hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},2\sigma} \hat{\mathbf{c}} \right) \\
&\quad = \mathbf{W} \hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},\sigma} \mathbf{V}^T \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right) \\
&\quad = \mathbf{W} \hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},\sigma} \mathbf{V}^T \left(\mathcal{A}^{T,0}_{\mathbf{E},\sigma} \right)^{-1} \mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right) \\
&\quad = \mathbf{W} \hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},\sigma} \mathbf{V}^T \left(\mathcal{A}^{T,0}_{\mathbf{E},\sigma} \right)^{-1} \mathbf{W}\mathbf{W}^T \mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right) \\
&\quad = \mathbf{W}\mathbf{W}^T \mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right) \\
&\quad = \mathcal{A}^{T,0}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right).
\end{aligned}$$

Iteratively using the above arguments, we finally get

$$\mathbf{W} \hat{\mathcal{A}}^{T,i}_{\hat{\mathbf{E}},\sigma} \mathbf{V}^T \mathcal{H}^{(2)} \left(\mathbf{V} \hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma} \hat{\mathbf{b}} \otimes \mathbf{W} \hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},2\sigma} \hat{\mathbf{c}} \right) = \mathcal{A}^{T,i}_{\mathbf{E},\sigma} \mathcal{H}^{(2)} \left(\mathcal{A}^{0}_{\mathbf{E},\sigma} \mathbf{b} \otimes \mathcal{A}^{T,0}_{\mathbf{E},2\sigma} \mathbf{c} \right), \tag{5.11}$$

for $i = 0, \ldots, q_2 - 1$. What we have to consider for $k = 2q_2 - 1, j = l = 0$ is the term

$$\hat{\mathbf{c}}^T \hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},2\sigma} \hat{\mathbf{H}} \left(\hat{\mathcal{A}}^{2q_2-1}_{\hat{\mathbf{E}},\sigma} \hat{\mathbf{b}} \otimes \hat{\mathcal{A}}^{0}_{\hat{\mathbf{E}},\sigma} \hat{\mathbf{b}} \right).$$

According to Definition 5.4.1 and (5.8) and (5.10), this term is rewritten as

$$\begin{aligned}
&\hat{\mathbf{c}}^T\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},\sigma}\mathbf{W}^T\mathbf{E}\mathbf{V}\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right)\\
&\quad=\hat{\mathbf{c}}^T\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{H}}\left(\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},\sigma}\mathbf{W}^T\mathbf{E}\mathcal{A}^{q_2-1}_{\mathbf{E},\sigma}\mathbf{b}\otimes\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right)\\
&\quad=\hat{\mathbf{c}}^T\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},2\sigma}\mathbf{W}^T\mathbf{H}\left(\mathbf{V}\hat{\mathcal{A}}^{q_2-1}_{\hat{\mathbf{E}},\sigma}\mathbf{W}^T\mathbf{E}\mathcal{A}^{q_2-1}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathbf{V}\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\right)\\
&\quad=\mathbf{b}^T\mathcal{A}^{T,q_2-1}_{\mathbf{E},\sigma}\mathbf{E}^T\mathbf{W}\hat{\mathcal{A}}^{T,q_2-1}_{\hat{\mathbf{E}},\sigma}\mathbf{V}^T\mathcal{H}^{(2)}\left(\mathbf{V}\hat{\mathcal{A}}^0_{\hat{\mathbf{E}},\sigma}\hat{\mathbf{b}}\otimes\mathbf{W}\hat{\mathcal{A}}^{T,0}_{\hat{\mathbf{E}},2\sigma}\hat{\mathbf{c}}\right)\\
&\quad=\mathbf{b}^T\mathcal{A}^{T,q_2-1}_{\mathbf{E},\sigma}\mathbf{E}^T\mathcal{A}^{T,q_2-1}_{\mathbf{E},\sigma}\mathcal{H}^{(2)}\left(\mathcal{A}^0_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^{T,0}_{\mathbf{E},2\sigma}\mathbf{c}\right)\\
&\quad=\mathbf{c}^T\mathcal{A}^0_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{q_2-1}_{\mathbf{E},\sigma}\mathbf{E}\mathcal{A}^{q_2-1}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^0_{\mathbf{E},\sigma}\mathbf{b}\right)\\
&\quad=\mathbf{c}^T\mathcal{A}^0_{\mathbf{E},2\sigma}\mathbf{H}\left(\mathcal{A}^{2q_2-1}_{\mathbf{E},\sigma}\mathbf{b}\otimes\mathcal{A}^0_{\mathbf{E},\sigma}\mathbf{b}\right).
\end{aligned}$$

Since the previous extremal cases contain the essential ideas, we omit a detailed derivation for the remaining combinations j, k, l with $j + k + l \leq 2q_2 - 1$. □

To sum up, we have seen that we indeed can construct two-sided projection methods for systems of QBDAEs. At least theoretically, the new approach doubles the number of interpolated derivatives of the first two transfer functions and thus should lead to more accurate reduced-order models. However, as has already been indicated in [33], in context of nonlinear model reduction, the benefit of matching more multimoments might come along with a loss of numerical stability and thus has to be treated with caution.

5.5 Numerical examples

In this section, we now want to study the procedure specified in Theorem 5.4.1 by means of some numerical examples. Besides the already discussed scalable RC circuit, we investigate different large-scale ODEs resulting from the semi-discretization of several nonlinear partial differential equations. Here, we refrain from sophisticated finite element discretizations and instead use a simple finite difference scheme for all test cases.

In general, moment matching type methods only allow to make an assertion on the approximation of the input-output behavior of a dynamical system. However, one can often reconstruct the full state vector $\mathbf{x} \approx \mathbf{V}\hat{\mathbf{x}}$ by a prolongation with the projection matrix $\mathbf{V}$. Moreover, for some problems one might only be interested in the steady state behavior without controlling the process. In this context, we investigate the approximation quality for two uncontrolled systems with nonzero initial condition.

All simulations were generated on an Intel® Dual-Core CPU E5400, 2 MB cache, 3 GB RAM, Ubuntu Linux 10.04 (i686), MATLAB Version 7.11.0 (R2010b) 32-bit (glnx86).

5.5.1 A nonlinear RC circuit

The first example we want to study is the scalable nonlinear transmission line circuit which we already studied in the context of model reduction for Carleman linearized large-scale bilinear systems in Chapter 4. In order to demonstrate the advantage of transforming a nonlinear system into an equivalent quadratic-bilinear one, we give a review of the transformation from [72]. Recall from Chapter 4 that the dynamics of the corresponding nonlinear control system are given as follows:

$$\begin{aligned}\dot{\mathbf{v}}(t) &= \mathbf{f}(\mathbf{v}(t)) + \mathbf{b}u(t),\\ \mathbf{y}(t) &= \mathbf{c}^T\mathbf{v}(t)),\end{aligned}$$

where

$$f(v) = f\left(\begin{bmatrix}\mathbf{v}_1\\ \mathbf{v}_2\\ \vdots\\ \mathbf{v}_k\\ \vdots\\ \mathbf{v}_N\end{bmatrix}\right) = \begin{bmatrix}-g(\mathbf{v}_1) - g(\mathbf{v}_1 - \mathbf{v}_2)\\ g(\mathbf{v}_1 - \mathbf{v}_2) - g(\mathbf{v}_2 - \mathbf{v}_3)\\ \vdots\\ g(\mathbf{v}_{k-1} - \mathbf{v}_k) - g(\mathbf{v}_k - \mathbf{v}_{k+1})\\ \vdots\\ g(\mathbf{v}_{N-1} - \mathbf{v}_N)\end{bmatrix}, \quad \mathbf{b} = \mathbf{c} = \begin{bmatrix}1\\0\\ \vdots\\0\end{bmatrix},$$

and $g(\mathbf{v}) = e^{40\mathbf{v}} + \mathbf{v} - 1$ describes the nonlinear input-voltage characteristics of each resistor. As has been discussed in [72], a transformation to quadratic-bilinear form is easily obtained by introducing additional state variables $x_i = e^{40v_i}$ and $z_i = e^{-40v_i}$. However, the resulting system will have a state dimension $3 \cdot N$. On the other hand, if we first rewrite the system by defining new state variables as $x_1 = v_1$ and $x_i = v_i - v_{i+1}$, followed by introducing additional state variables $z_1 = e^{40v_1-1}$ and $z_i = e^{40x_i}$, it is even possible to construct an equivalent quadratic-bilinear system of dimension $2 \cdot N$. To be precise, the final system is determined by

$$\begin{aligned}\dot{z}_1 &= 40(z_1+1)(-x_1 - x_2 - z_1 - z_2 + u(t)),\\ \dot{z}_2 &= 40(z_2+1)(-x_1 - 2x_2 + x_3 - z_1 - 2z_2 + z_3 + u(t)),\\ \dot{z}_i &= 40(z_i+1)(x_{i-1} - 2x_i + x_{i+1} + z_{i-1} - 2z_i + z_{i+1}),\\ \dot{z}_N &= 40(z_N+1)(x_{N-1} - 2x_N + z_{N-1} - 2z_N).\end{aligned}$$

At this point, we see that the non uniqueness of the transformation indeed can lead to rather different results with different complexity.

In Figure 5.1, we see a comparison between the new method discussed here and the classical one-sided method discussed in [72] for a state dimension $n = 2 \cdot 1000$. More-

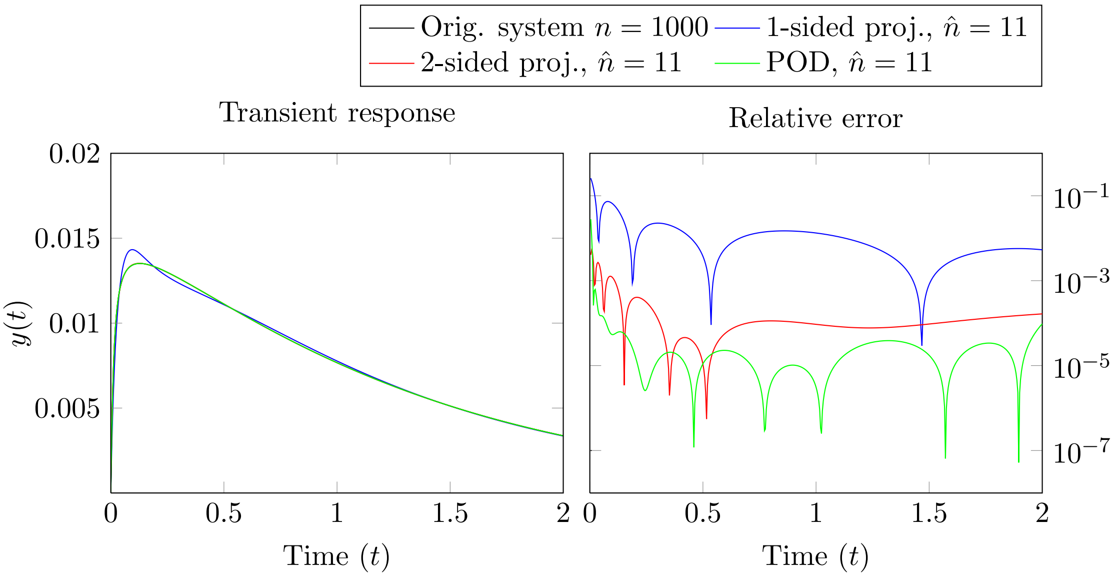


Figure 5.1: A nonlinear RC circuit. Comparison of moment matching methods and POD subject to boundary control $u(t) = e^{-t}$.

over, we compute a POD-based approximation by taking 100 snapshots of the original solution for the input excitation $u(t) = e^{-t}$. Obviously, the POD reduced-order model performs the best. However, the two-sided method exhibits a comparable approximation quality while the one-sided approach performs the worst. All reduced-order models are of dimension $\hat{n} = 11$. The moment matching based techniques are generated according to Theorem 5.4.1 with values $\sigma = 1,\ q_1 = 5,\ q_2 = 2$.

In order to test our method with respect to input variations, in Figure 5.2, we show the approximations for the input signal $u(t) = (\cos(2\pi\frac{t}{10}) + 1)/2$. Clearly, the POD approximation shows a significant deviation from the original output. On the other side, the two-sided method still reflects the dynamics very accurately and also outperforms the one-sided technique as well.

5.5.2 Burgers' equation

Next, let us consider the one-dimensional viscous Burgers' equation on $\Omega = (0,1)\times(0,T)$, leading to the following set of equations

$$v_t + v \cdot v_x = \nu \cdot v_{xx}, \quad \text{in } (0,1)\times(0,T), \tag{5.12a}$$

$$\alpha v(0,\cdot) + \beta v_x(0,\cdot) = u(t), \quad \text{in } (0,T), \tag{5.12b}$$

$$v_x(1,\cdot) = 0, \quad \text{in } (0,T), \tag{5.12c}$$

$$v(x,0) = v_0(x),\ \text{in } (0,1), \tag{5.12d}$$

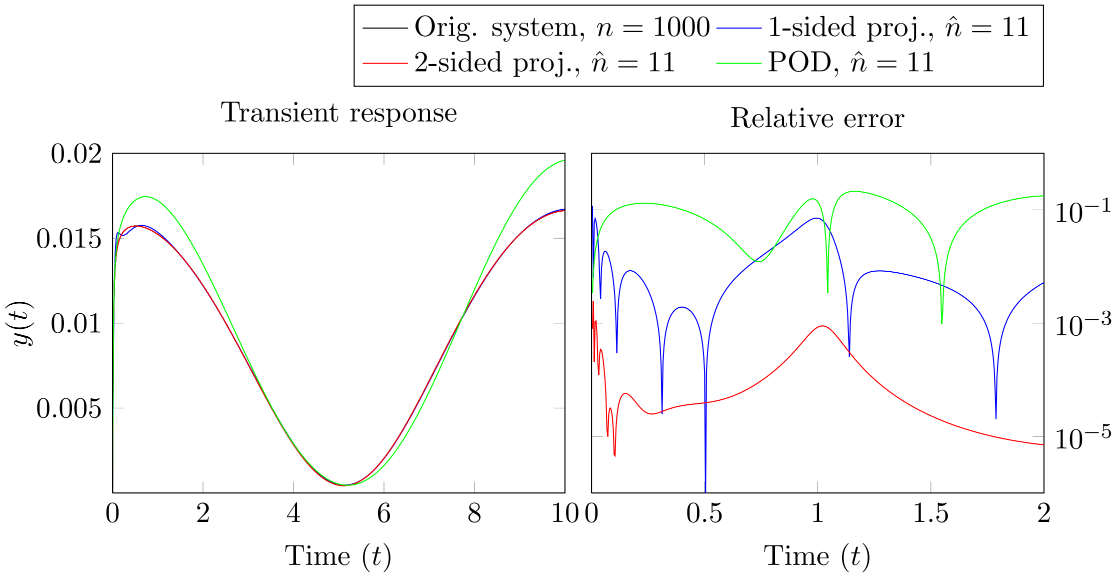


Figure 5.2: A nonlinear RC circuit. Comparison of moment matching methods and POD subject to boundary control $u(t) = (\cos(2\pi\frac{t}{10}) + 1)/2$.

where ν is the viscosity parameter and $v_0(x)$ denotes the initial condition of the system. This equation can be seen as a standard numerical test example for nonlinear model reduction and optimal control, respectively, and has already been extensively studied in e.g. [92, 93]. In the context of this paper, the above PDE is of particular interest since a semi-discretization automatically leads to a quadratic-bilinear control system of the form (5.3).

Boundary control

Let us assume that the equation is subject to a boundary control on the left side of the interval, i.e. $\alpha = 1$ and $\beta = 0$. Furthermore, we assume the initial state of the system to be zero, i.e. $v_0(x) = 0$. For the viscosity parameter ν we start by choosing the value 0.02. However, while for larger values of ν, the accuracy of the reduced-order models often becomes better, decreasing ν makes the model more difficult to reduce.

In Figure 5.3, we show the results for the reduction of system (5.12) which was spatially discretized using $n = 1000$ points and $T = 10$. The reduced-order models are of dimension $\hat{n} = 9$ and are generated by Algorithm 5.4.1 with $\sigma = 0.0288, q_1 = 4$ and $q_2 = 2$. The specific interpolation point σ is computed by IRKA applied to the linearized system. For the one-sided projection method, we simply set $\mathbf{W} = \mathbf{V}$. The measured output of the system is assumed to be the value of the right boundary, leading to an output vector $\mathbf{c} = \left[0, \ldots, 0, 1\right]^T$. Besides a comparison between one-sided and two-sided projection, we compute a POD approximation by using the SVD of the solution matrix of the original problem over the whole interval range. As can be seen in Figure 5.3, for the control

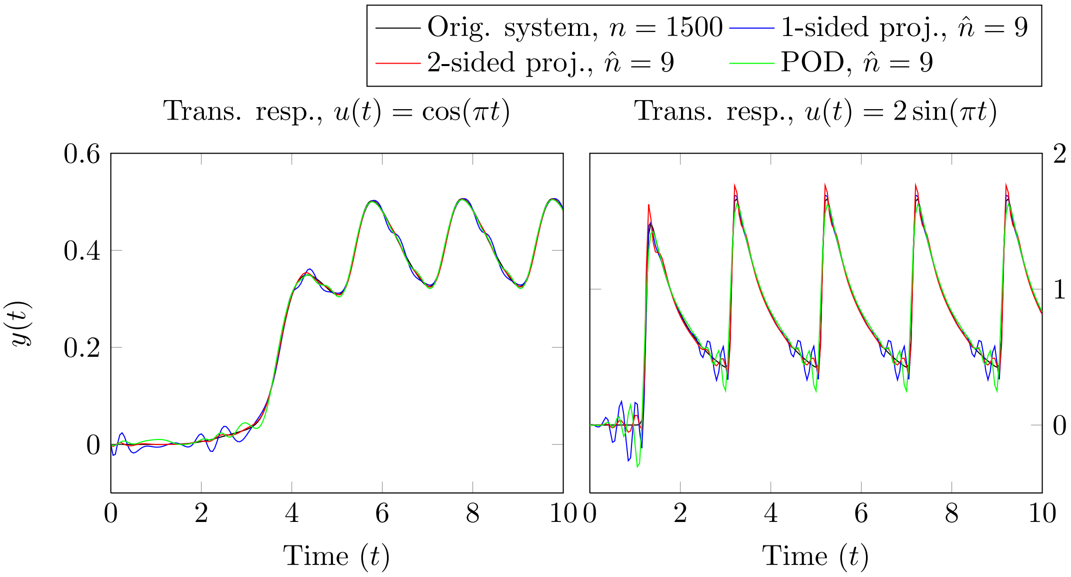


Figure 5.3: Burgers' equation. Comparison of moment matching methods and POD subject to boundary control ($\nu = 0.02$).

$u(t) = \cos(\pi t)$, all approaches faithfully reproduce the dynamics of the original system although the one-sided approach exhibits some smaller oscillations. In order to investigate the methods with regard to robustness to input variations, we slightly change the control to $u(t) = 2\sin(\pi t)$. Increasing the amplitude of $u(t)$ seems to make the reduction process more difficult. For the POD approximation, we use the projection subspace derived by the first input signal. As expected, we see that this results in a less accurate reduced-order model indicating the input dependency of POD. On the other hand, for the two-sided projection we observe overshoots at the sharper fronts of the curve. Nevertheless, altogether for this parameter configuration of σ, q_1, q_2, we can conclude that the new method performs well and seems to outperform the one-sided projection. It has to be mentioned though that for the two-sided approach many of the parameter constellations lead to unstable reduced-order models. A similar observation already was discussed in [33]. Hence, a reasonable choice of the interpolation points together with the order of the matched derivatives seems to be an important aspect of further research.

The uncontrolled case

In order to test the efficiency of the reduction method, we also want to investigate the performance when the system under consideration exhibits a non-zero initial condition. In view of (5.12), we use $\alpha = 0, \beta = 1$ and $v_0(x) = 1 + \sin((2x+1)\pi)$. After a semi-discretization with $n = 1000$, the system is rewritten to a system with zero initial condition, leading to a single-input and single-output (SISO) QBDAE system with constant input vector $u(t)$. The viscosity parameter is $\nu = 0.01$ while we choose $T = 2$. In contrast to the previous example, we now consider the entire state $\mathbf{x}$. Since

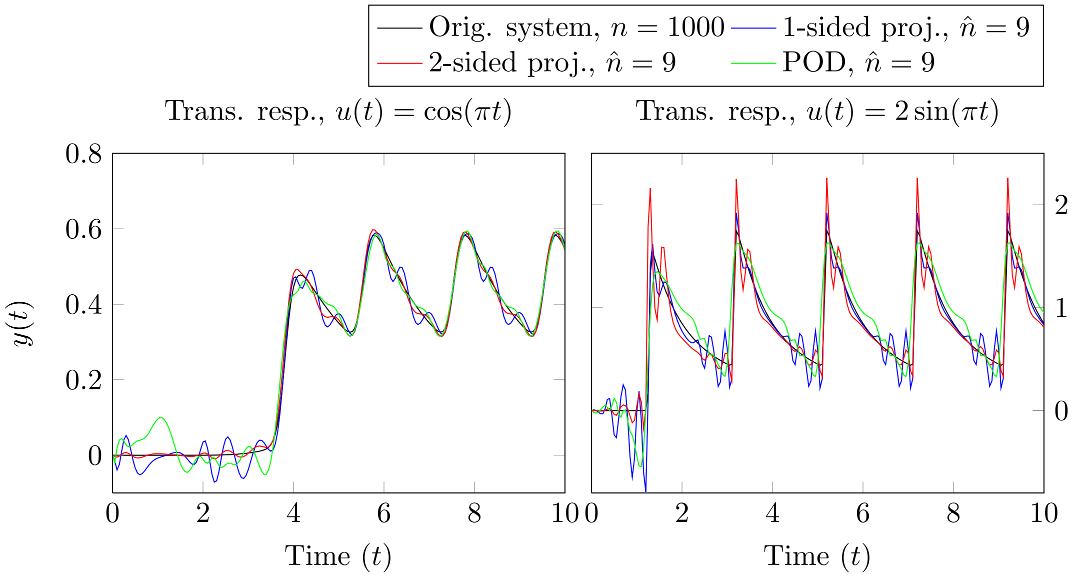
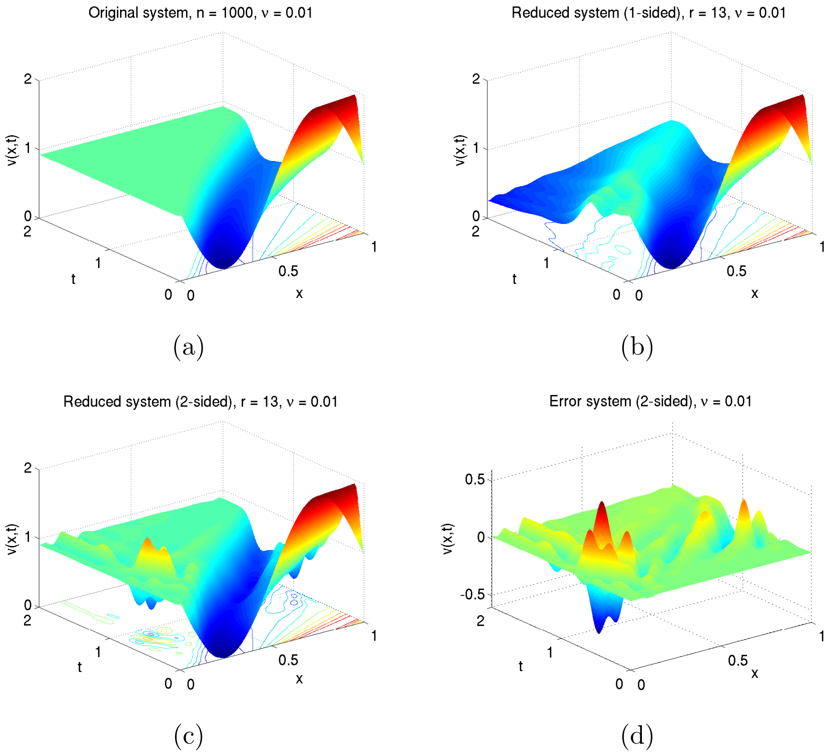


Figure 5.4: Burgers' equation. Comparison of moment matching methods and POD subject to boundary control ($\nu = 0.01$).

we want compare to the results for a two-sided reduction method, we artificially have to choose a certain output matrix $\mathbf{c}$ such that we can run Algorithm 5.4.1. Here, we use $\mathbf{c} = \frac{1}{k}\left[1, \ldots, 1\right]^T$, i.e., the average value of $v(x,t)$ on the interval $(0,1)$.

In Figure 5.5, we show the different steady state solutions for the original system, see Figure 5.5(a), the reduced-order system obtained by an orthogonal projection, see Figure 5.5(b), and the reduced-order system resulting from an oblique projection, see Figure 5.5(c). For the reduction process we choose $\sigma = 5$, $q_1 = 10$ and $q_2 = 2$, leading to reduced-order models of dimension $\hat{n} = 13$. Here, the interpolation point now is chosen as the one performing the best among several random choices. Obviously, the one-sided approach deviates significantly from the original solution, while the two-sided method produces some undesired peaks. However, one still has to keep in mind that we cannot make a theoretical assertion on the reconstruction of a state vector but only on the input-output behavior of the system. If we keep this in mind, the approximations still might be appropriate for the analysis of the uncontrolled dynamics. Note that we do not compare the results with POD at this point since we do have a specific constant input which does not vary. Hence, it is clear that POD will outperform the moment matching approaches due to its intrinsic properties. Recall that for a given input which is not subject to variation, the approximation given by POD is optimal due the properties of the singular value decomposition.



(a)

(b)

(c)

(d)

Figure 5.5: Burgers' equation. Comparison of uncontrolled solutions.

5.5.3 Chafee-Infante equation

Next, we consider the one-dimensional Chafee-Infante equation. For more details on this nonlinear PDE, we refer to [37, 76]. The equation exhibits a cubic nonlinearity and is subject to similar initial and boundary conditions as the Burgers' equation, namely

$$\begin{aligned} v_t + v^3 &= v_{xx} + v && \text{in } (0,1) \times (0,T), && \text{(5.13a)}\\ \alpha v(0,\cdot) + \beta v_x(0,\cdot) &= u(t), && \text{in } (0,T), && \text{(5.13b)}\\ v_x(1,\cdot) &= 0, && \text{in } (0,T), && \text{(5.13c)}\\ v(x,0) &= v_0(x), && \text{in } (0,1). && \text{(5.13d)} \end{aligned}$$

Following the discussion in [76], we once more use a finite difference scheme for the spatial discretization. The resulting system of nonlinear ODEs then has to be transformed to quadratic-bilinear structure. This is done by introducing a new state variable $w_i = v_i^2$. Computing the derivate of w_i leads to $\dot{w}_i = 2v_i\dot{v}_i$ which can be rewritten in the desired QBDAE form (5.3).

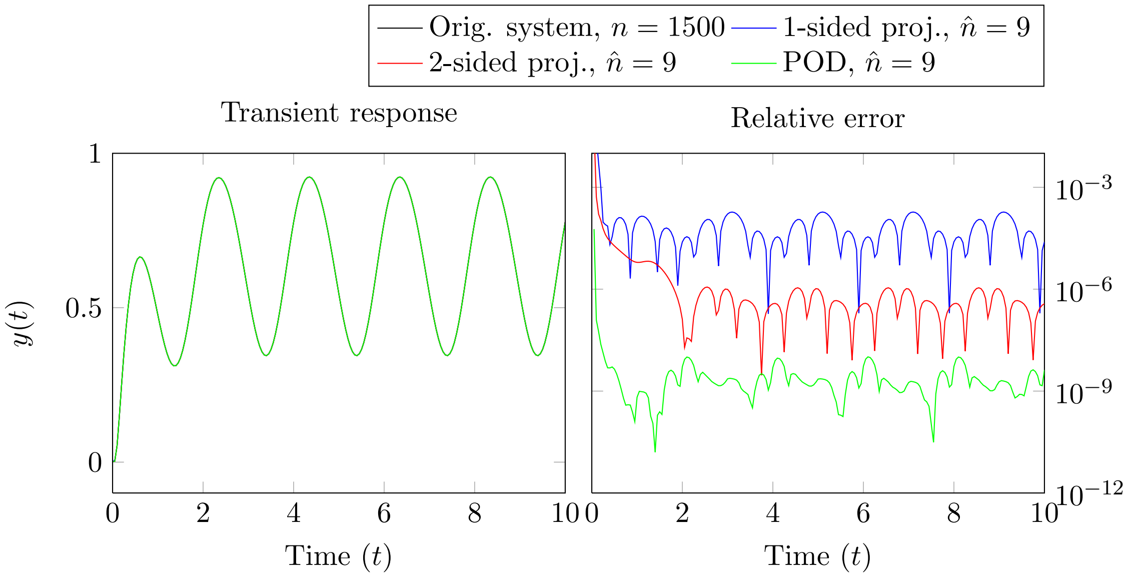


Figure 5.6: Chafee-Infante equation. Comparison of moment matching methods and POD subject to boundary control $u(t) = (1 + \cos(\pi t))/2$.

Boundary control

Completely analogously to Section 5.5.2, we start with the boundary controlled equation with $T = 10$ and a zero initial condition $v_0(x) = 0$. We further use the same output, i.e., the value at the right boundary, leading to an output vector $\mathbf{c} = \left[0, \ldots, 0, 1\right]^T$. The discretization was done with $n = 750$ points. Hence, after transformation to QBDAE form, the system consists of $2 \cdot 750$ states.

The reductions are generated with $\sigma = 1$, $q_1 = 4$ and $q_2 = 3$, yielding systems of dimension $\hat{n} = 9$. Similar to the Burgers' equation, we run IRKA in order to get an $\mathcal{H}_2$-optimal interpolation point for the linearized system, leading to the specific choice $\sigma = 1$. Again, in Figure 5.6, we visualize the approximations of our new method and compare them with a one-sided projection as well as POD. For the input $u(t) = (1 + \cos(\pi t))/2$, we see that the new approach clearly outperforms the one-sided projection. On the other hand, it cannot compete with POD.

But if we change the input signal to $u(t) = 25 \cdot (1 + \sin(\pi t))/2$, the corresponding results are given in Figure 5.7. Though a bit surprising, we observe that the reduced-order model for the one-sided approach completely fails in reproducing the original dynamics. Once more, we do not vary the projection subspace of POD but simply use the one for the first test signal specified above. Here, we now see that POD indeed also has problems in the approximation of the maxima of the transient response which is not the case for the two-sided approach.

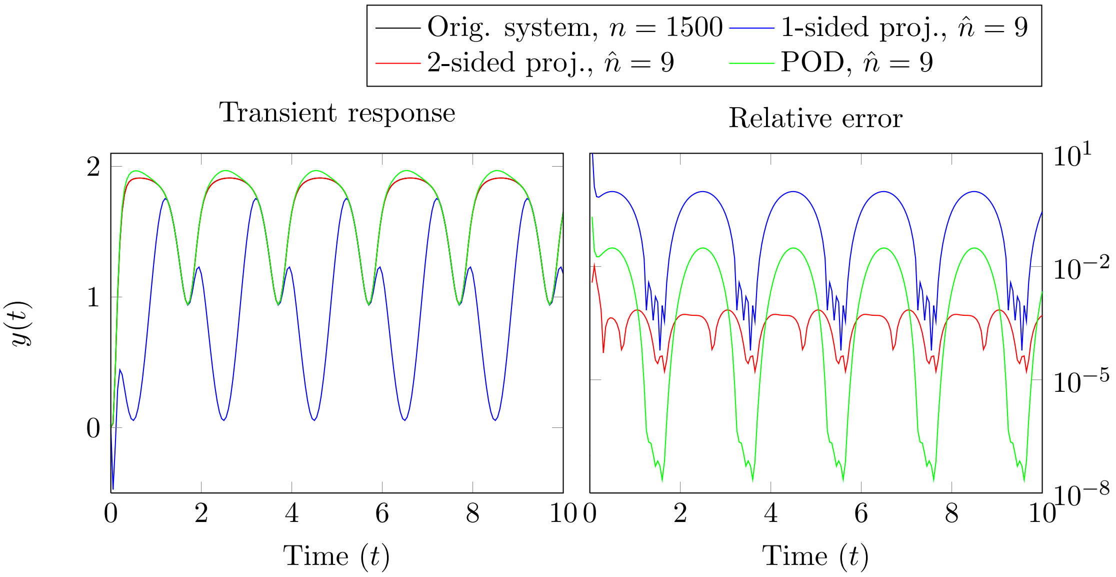


Figure 5.7: Chafee-Infante equation. Comparison of moment matching methods and POD subject to boundary control $u(t) = 25 \cdot (1 + \sin(\pi t))/2$.

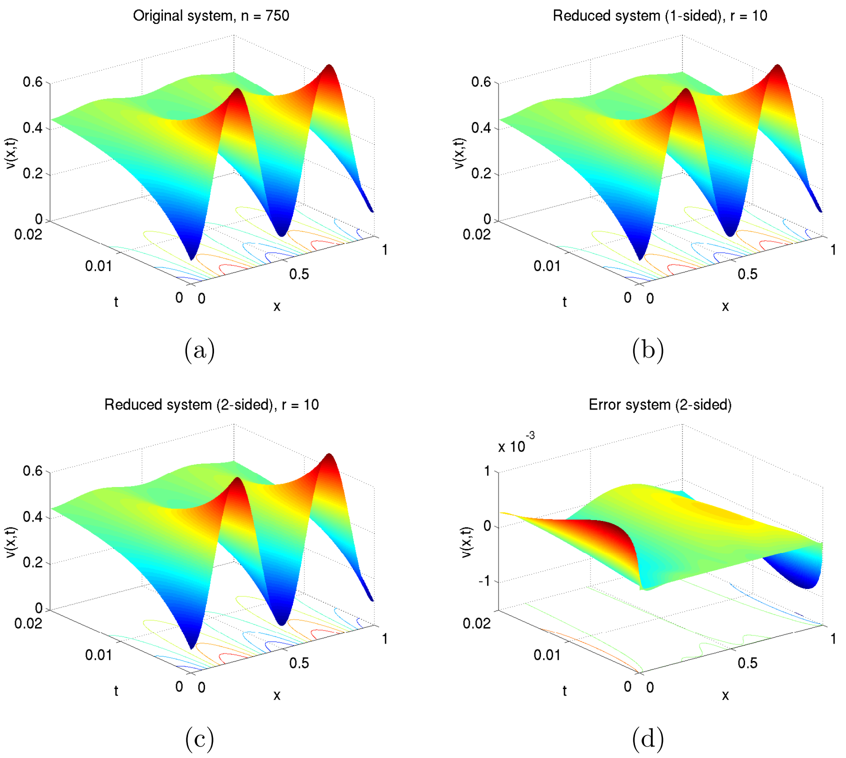
The uncontrolled case

For the uncontrolled case, we set $\alpha = 0$, $\beta = 1$ and implement a non-zero initial condition which was already discussed in [76]. To be more precise, we have $v_0(x) = \frac{1}{10} + \frac{7}{10} \cdot \sin^2((2 \cdot x + 1)\pi)$. In Figure 5.8, we compare the full state vector for the time interval $T = 0.02$ for a semi-discretization with $n = 750$. The reduced-order systems are of dimension $\hat{n} = 10$ and result from the model reduction parameters $\sigma = 3$, $q_1 = 3$, $q_2 = 3$, which basically are chosen at random. As we can see, both approaches yield very accurate reconstructions. However, due to several parameter studies, it seems that the one-sided projection method performs more robust with respect to stability issues of the reduced-order model.

5.5.4 FitzHugh-Nagumo system

Finally, as a last example we study the FitzHugh-Nagumo system modeling activation and deactivation dynamics of a spiking neuron. This model has been under consideration in the context of POD-based model reduction in [39]. Formally, the model is described by the following coupled system of nonlinear PDEs

$$\epsilon v_t(x,t) = \epsilon^2 v_{xx}(x,t) + f(v(x,t)) - w(x,t) + g, \tag{5.14a}$$
$$w_t(x,t) = hv(x,t) - \gamma w(x,t) + g, \tag{5.14b}$$



(a) (b) (c) (d)

Figure 5.8: Chafee-Infante equation. Comparison of uncontrolled solutions.

with $f(v) = v(v-0.1)(1-v)$ and initial and boundary conditions

$$\begin{aligned} v(x,0) &= 0, & w(x,0) &= 0, & x &\in [0,1], && \text{(5.15a)}\\ v_x(0,t) &= -i_0(t), & v_x(1,t) &= 0, & t &\geq 0, && \text{(5.15b)} \end{aligned}$$

where $\epsilon = 0.015,\ h = 0.5,\ \gamma = 2,\ g = 0.05,\ i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$. Again, one can easily use a finite difference scheme, resulting in a system of cubic ODEs. Similar to the Chafee-Infante equation, introducing an additional dynamical variable $z_i = v_i^2$ allows to reformulate the dynamics as a system of QBDAEs of dimension $3 \cdot n$, where n is the number of degrees of freedom used in the finite difference scheme. However, in contrast to the first two examples, the system no longer is of SISO type since the constant parameter g as well as the stimulus $i_0(t)$ have to be incorporated within the modeling process. In order to apply the previously discussed reduction techniques, we run the corresponding algorithm once for each column of the input vector.

Here, we follow the setting in [39] and use a discretization with $n = 1000$ points. In Figure 5.9, we show the reduction results measured in terms of the limit cycle behavior which is a typical phenomenon when modeling neuronal dynamics. For the comparison between one-sided and two-sided projections, we assume the output matrix $\mathbf{C} \in \mathbb{R}^{2\times 3n}$ to sort out the values $v(0,t)$ and $w(0,t)$, i.e. the limit cycle at the left boundary. The

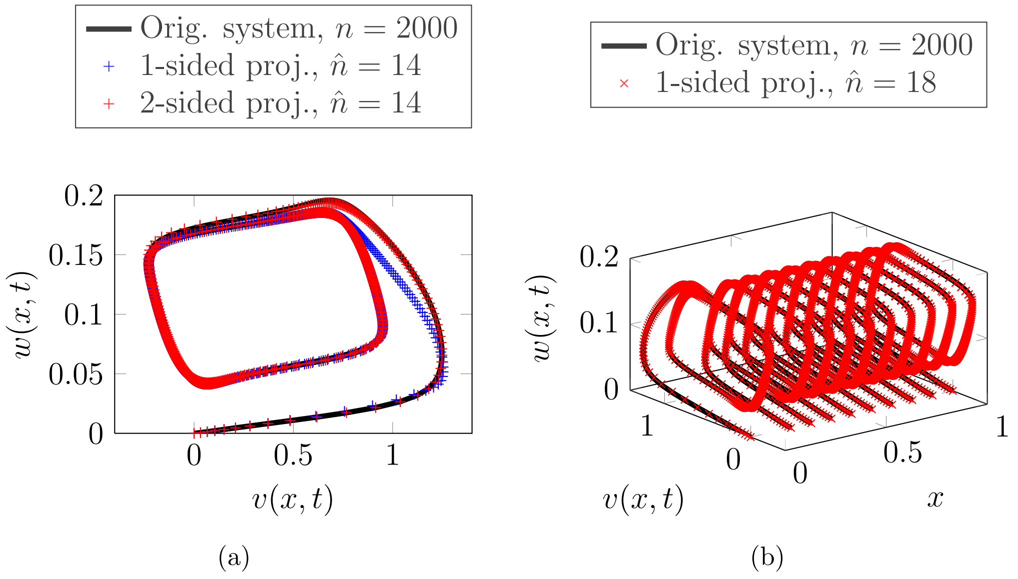


Figure 5.9: FitzHugh-Nagumo system. Limit cycles for original and reduced systems.

results shown in Figure 5.9(a) are constructed with parameter values $\sigma = 100$, $q_1 = 2$, $q_2 = 2$ and the reduced-order models both are of dimension $\hat{n} = 14$. Although the approximation of the two-sided reduced model performs better, based on this specific example we cannot recommend using the new approach. This is due to the fact that nearly all generated reduced-order models become unstable and it does not seem to be obvious how to circumvent this significant drawback. On the other hand, most reductions obtained by using random interpolation points yield accurate approximations for the one-sided technique. For example, in Figure 5.9(b) we plot the limit cycle behavior similar to the one studied in [39] for a discretization of $n = 1000$ and the parameter setting $\sigma = 14$, $q_1 = 5$, $q_2 = 2$ and a reduced-order system of dimension $\hat{n} = 18$. Although the results are not as accurate as in [39], where a sufficient reduction to a system of dimension $\hat{n} = 10$ is obtained, we are certainly able to construct an appropriate reduced-order model.

5.6 Conclusions

In this chapter, we have studied a recently introduced new approach for model order reduction of nonlinear control systems. In contrast to other methods in this field, the technique relies on generalized moment matching and thus is input independent. Besides a slight extension of existing results for the case of $\sigma = 0$, we have shown how the sequence of nested Krylov subspaces has to be chosen in order to interpolate at arbitrary interpolation points $\sigma \neq 0$. Moreover, we used some basic tools and properties known from tensor theory to show how one can improve the efficiency of the necessary projection step leading to the reduced-order system. In particular, we have seen that one can avoid building up the matrix $\mathbf{V} \otimes \mathbf{V}$ which easily might exceed given memory capacity. The main contribution of this chapter was the construction of an appropriate two-sided projection method which theoretically allows to double the number of interpolated derivatives of the first two transfer functions. However, here one has to be careful in applying the new method since the gain of accuracy sometimes destroys the stability of the underlying system making a reduction unreliable. Nevertheless, by means of several nonlinear partial differential equations, we have proven that the moment matching approach indeed seems to have potential and even allows to reconstruct typical dynamics observed in fluid mechanics and neuron modeling. Moreover, for two examples we could show that the new method can compete with POD and, in some cases, might be advantageous if the input signal is known to exhibit larger variations. Hence, it might be an interesting field of further research. In particular, the study of optimal interpolation points seems to be an important issue. Similarly, investigating structure preserving methods which ensure stable reduced-order models should be one of the major challenges in order to improve the applicability of the new method.

CHAPTER 6

CONCLUSIONS AND OUTLOOK

Contents

6.1 Summary and conclusions

In this thesis, we have studied the topic of model order reduction for large-scale dynamical control systems. Most of the results and numerical approaches are based on the concept and the extension of rational interpolation, previously studied for the purely linear case in [71]. In particular, a special focus of this work has been on the $\mathcal{H}_2$-optimal model order reduction problem, initially investigated in [99, 131] and, later on, picked up in [36, 73]. Starting out from the special case of linear control systems, we have developed and extended several interpolation-based concepts to nonlinear systems. A large part of this thesis has been dedicated to a special class of nonlinear control systems, so-called bilinear control systems, that can be seen as the connection between linear and fully nonlinear systems. Based on the previous works in [24, 42, 66, 97, 120, 133], we have discussed numerically and computationally efficient model order reduction techniques for these systems. The results have been theoretically explained and interpreted and have been, in several numerical simulations, practically verified. For an even more general class of smooth nonlinear control-affine systems, we have enhanced a rather recently introduced method from [72] by means of basic tools from tensor theory.

In Chapter 3, we have considered linear control systems and the associated problem of approximately solving large-scale matrix equations. For the symmetric case, we have shown that the Riemannian optimization technique from [125] can alternatively be realized by solving the $\mathcal{H}_2$-optimal model reduction problem, i.e., by the iterative rational Krylov algorithm from [73]. In particular, the latter algorithm generates subspaces that, by means of the rational Krylov subspace method, lead to approximations that are locally optimal with respect to a certain energy norm naturally induced by the corresponding Lyapunov operator. Moreover, we have derived an extension of this theory that can be applied to unsymmetric linear systems and, in this case, locally minimizes the residual of the Lyapunov equation. Moreover, in view of the results from Chapter 4, we can interpret this technique as a modified $\mathcal{H}_2$-model reduction problem, closely related to the one for bilinear control systems. Finally, we have studied the case of the Sylvester equation and derived interpolation-based optimality conditions that lead to an algorithm which constructs approximations that are optimal with respect to the induced energy norm of the Sylvester operator.

In the first part of Chapter 4, the problem of $\mathcal{H}_2$-optimal model reduction for bilinear control systems has been discussed. We have shown how to generalize the interpolation-based optimality conditions from [73, 99] and proposed an iterative algorithm (BIRKA) that, upon convergence, yields an optimal reduced-order model fulfilling these conditions. Moreover, we have proven the equivalence between the new conditions and the ones obtained in [133]. Motivated by the latter work, we have implemented a further iterative algorithm that relies on the solution of certain generalized Sylvester equations. This approach has been shown to be theoretically as well as numerically equivalent to BIRKA.

The second part of Chapter 4 has been dedicated to the solution of large-scale generalized Lyapunov equations arising in the method of balanced truncation for bilinear control systems. By means of several results from [66, 67] and properties of tensors, we have given a theoretical explanation for the in [24, 43] observed fast singular value decay of the solution matrix. Here, we have made the additional assumption that the bilinear coupling matrices are of low rank. Despite that restriction, to the best of authors' knowledge, this is the first result that theoretically underscores the observations in [24, 43]. Moreover, based on the foundations of projection-based and ADI-based low rank solvers for the standard case, we have extended the basic concepts to our setting and implemented different low rank methods that easily allow to solve these generalized matrix equations up to the order of 10^5.

In Chapter 5, we have studied the topic of model order reduction for a very general class of nonlinear systems. The results and concepts strongly depend on the fact that, according to [72], smooth nonlinear control-affine systems can be transformed into an equivalent system of quadratic-bilinear differential algebraic equations. Based on similar techniques for bilinear control systems, we have proposed an oblique projection method that essentially allows to double the number of matched multimoments when compared to the original method from [72]. Since we have discussed the equivalence between these multimoments and the derivatives of the generalized transfer functions, this approach can again be interpreted as an interpolation-based technique. Besides this extension, a further contribution has been the efficient computation of a projection-based reduced-order model. This has been achieved by some basic tools from tensor theory. Moreover, we have studied different numerical applications and compared the results with the common reduction techniques for nonlinear systems, POD. The main advantage of the new method is that it is input independent, a feature that no other known reduction method shares. Although several points still have to be discussed further, we can conclude that the new method seems to be a promising new field of research.

6.2 Future research perspectives

The thesis has revealed some new aspects within the area of model order reduction of linear, bilinear and nonlinear control systems. Though, there are several questions and problems that remained open and should be discussed in future research. In particular, we have seen that bilinear as well as quadratic-bilinear control systems share some concepts that so far only have been used for linear control systems, such as, e.g., transfer functions and system Gramians. However, due to the complex nature of the corresponding extensions, a computationally efficient realization often is the bottleneck of the methods and can prevent from reducing very large-scale systems with dimensions $n > 10^5$. In the following, we present some open problems that deserve further attention.

When it comes to model order reduction of linear systems, a common misconception clearly is that there is nothing else to be investigated. As we have seen in Chapter 3, despite the fact that there exist numerous different low rank techniques for approximating solutions of large-scale Lyapunov equations, most of them are closely related in some sense and so far have been considered independently from each other. However, there still remains the interesting question of a possible connection between the method of balanced truncation and rational interpolation. To be more precise, it would be desirable to find the interpolation points that determine a balanced reduced-order model. On the other hand, for rational Krylov methods like IRKA, besides the property of being locally optimal, little is known about a priori error bounds, neither with respect to the $\mathcal{H}_2$-norm nor the $\mathcal{H}_\infty$-norm. Moreover, a better understanding between the Riemannian optimization method and IRKA might give insight into the construction of more efficient descent techniques that yield $\mathcal{H}_2$-optimal reduced order models.

In the context of $\mathcal{H}_2$-model reduction for bilinear systems, especially the issue of a computationally efficient implementation is of great interest. So far, the proposed iterative algorithms such as, e.g., BIRKA solve the generalized Sylvester equations

$$\mathbf{A}\mathbf{X} + \mathbf{X}\hat{\mathbf{A}}^T + \sum_{k=1}^{m} \mathbf{N}_k \mathbf{X} \hat{\mathbf{N}}_k^T + \mathbf{B}\hat{\mathbf{B}}^T = \mathbf{0},$$

by means of the explicit system of linear equations, requiring a theoretical complexity of $\mathcal{O}(n^3\hat{n}^3)$ for the LU decomposition. Since the sparsity of the matrix $\mathbf{A}$ and $\mathbf{N}_k$ is carried over to the Kronecker system, an iterative solver might only require a complexity of $\mathcal{O}(n\cdot \hat{n})$ to obtain the solution and should be a topic of further research. Moreover, in contrast to the linear case, even for symmetric systems nothing is known about the convergence of the algorithms. Further, suitable initialization techniques might be interesting in order to improve the speed of convergence of the method. Of course, due to the similarity to linear systems, for the symmetric case a lower bound property of the bilinear $\mathcal{H}_2$-norm of the error system, analog to the linear one, might be proven. This would immediately allow to make an assertion on the optimality of the projection subspaces in terms of the energy norm of the generalized Lyapunov operator

$$\mathcal{A} = \mathbf{I} \otimes \mathbf{A} + \mathbf{A} \otimes \mathbf{I} + \sum_{k=1}^{m} \mathbf{N}_k \otimes \mathbf{N}_k.$$

Regarding the method of balanced truncation and the associated generalized Lyapunov equations, we think that the mathematical area of tensor theory holds a lot of interesting ideas that can improve the efficiency even of the proposed low rank techniques by far. Nevertheless, for methods like the generalized ADI iteration, we have seen that the choice of shift parameters is a crucial point and certainly requires more attention. In particular, even for commutative matrices $\mathbf{A}$ and $\mathbf{N}$, the min-max problem

$$\min_{\{p_1,\dots,p_q\}} \max_{\substack{\lambda_i,\lambda_j \in \sigma(\mathbf{A}) \\ \mu_i,\mu_j \in \sigma(\mathbf{N})}} \left| \frac{(\lambda_i + p_\ell)(\lambda_j + p_\ell) + 2p_\ell \mu_i \mu_j}{(\lambda_i - p_\ell)(\lambda_j - p_\ell)} \right|,$$

is an interesting topic that could yield important insights. As we mentioned in Chapter 4, real-life applications often are accompanied with a high-dimensional parameter space, leading to the *curse of dimensionality.* Here, appropriate and optimal model order reduction techniques are still desirable and allow for several research topics.

Of course, the theory for nonlinear model order reduction is known the least. Several minor and major issues have to be resolved. Starting with the characterization of minimal or optimal transformations to QBDAE structure and automatic differentiation tools that might help to derive the desired structure by a black box technique. Further, the meaning of the transformation itself is not very clear. Perhaps there is a relation to known nonlinear model order reduction techniques such as, e.g., DEIM. For more general nonlinear systems, the transformation might lead to a system of DAEs, making all proposed model reduction techniques by far more complicated. However, an entire classification of QBDAEs probably cannot be given and the approach has to be studied for individual examples instead. As we have seen in Chapter 5, the model reduction procedure itself heavily depends on the choice of the interpolation points. Although this is common for interpolation-based reduction techniques, we observed in several examples that for QBADEs this choice is particularly important to ensure stability of the resulting reduced-order model. Finally, a possible extension of balancing-based or balancing-related techniques for these types of systems certainly would be a significant contribution to nonlinear model order reduction.

THESES

1. This thesis deals with the problem of interpolation-based model order reduction for linear, bilinear and quadratic-bilinear control systems. A special emphasis lies on the $\mathcal{H}_2$-optimal model order reduction problem.

2. Several well-known and important reduction techniques for linear systems are reviewed. This contains a discussion on $\mathcal{H}_2$-model reduction as well as balanced truncation. For nonlinear systems, the common POD technique is compared with a recently introduced interpolation-based approach.

3. For symmetric state space systems and the associated Lyapunov equations, the Riemannian optimization from Vandereycken is shown to be realizable by means of the iterative rational Krylov algorithm (IRKA) from Gugercin/Antoulas/Beattie. In particular, the already practically observed phenomenon of accurate low rank approximations is theoretically explained.

4. Within the same context, for unsymmetric systems, a new connection between the Frobenius norm of low rank approximations and the $\mathcal{H}_2$-norm of the associated error system is presented.

5. An abstract extension of IRKA, minimizing the Lyapunov residual of low rank approximations, is proposed. The approach can be interpreted as a modified $\mathcal{H}_2$-model reduction problem, similarly arising for bilinear control systems. Moreover, the method connects another Riemannian optimization technique from Vandereycken with the concept of rational interpolation.

6. The more general case of a Sylvester equation is considered. Optimality conditions in terms of interpolation of transfer functions are derived and an iterative algorithm for symmetric Sylvester equations is discussed and tested by means of a numerical example.

7. New necessary $\mathcal{H}_2$-optimality conditions for bilinear control systems are derived.

An interpretation of their meaning and relation to the known interpolation-based conditions for linear systems is given. The new conditions are shown to be equivalent to existing ones.

8. Two iterative algorithms generalizing IRKA are presented and tested in detailed by means of different numerical examples. The new approach is shown to outperform the method of balanced truncation for bilinear systems with respect to the bilinear $\mathcal{H}_2$-norm.

9. For large-scale bilinear control systems, the applicability of the method of balanced truncation is investigated. The often observed fast singular value decay of the bilinear system Gramians is theoretically explained by the use of tensor theoretic tools and quadrature formulas for the matrix exponential.

10. Three known low rank techniques for the standard Lyapunov equation have been extended to the more general case arising for bilinear control systems. For the extension of the Krylov-Plus-Inverted-Krylov (KPIK) method, an efficient computation of the Lyapunov residual that in a way automatically arises within the iteration is proposed. For the generalized low rank ADI method, the choice of (optimal) shift parameters is discussed and the use of $\mathcal{H}_2$-optimal interpolation points is proposed. Finally, iterative linear solvers such as CG and BiCGstab are implemented in a way that allows to make use of the low rank structure of the iterates in each step.

11. A new and alternative method for nonlinear model order reduction is reviewed and precisely tested by means of interesting real-life applications. The approach extends existing interpolation-based techniques proposed for linear and bilinear control systems and thus is input independent.

12. The computation of a projection-based reduced-order model is efficiently realized by making use of tensor theoretic concepts such as matricizations.

13. A new two-sided projection method is derived and theoretically shown to double the number of moments of the first two transfer functions that are matched by a one-sided projection method. Numerical examples underscore the potential of the new method.

BIBLIOGRAPHY

[1] S.A. Al-Baiyat and M. Bettayeb, *A new model reduction scheme for k-power bilinear systems*, in Proceedings of the 32nd IEEE Conference on Decision and Control, IEEE, 1993, pp. 22–27. 73, 74

[2] S.A. Al-Baiyat, A.S. Farag, and M. Bettayeb, *Transient approximation of a bilinear two-area interconnected power system*, Electric Power Systems Research, 26 (1993), pp. 11–19. 105, 111

[3] A.C. Antoulas, *Approximation of Large-Scale Dynamical Systems*, SIAM Publications, Philadelphia, PA, 2005. 8, 11, 15, 16, 17, 23, 24, 27, 28, 145, 147

[4] ——, *A new result on passivity preserving model reduction*, Systems & Control Letters, 54 (2005), pp. 361–374. 27

[5] A.C. Antoulas, D.C. Sorensen, and S. Gugercin, *A survey of model reduction methods for large-scale systems*, Contemporary Mathematics, 280 (2001), pp. 193–220. 24, 38

[6] A.C. Antoulas, D.C. Sorensen, and Y. Zhou, *On the decay rate of Hankel singular values and related issues*, Systems & Control Letters, 46 (2002), pp. 323–342. 18

[7] P. Astrid, S. Weiland, K. Willcox, and T. Backx, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Transactions on Automatic Control, 53 (2008), pp. 2237–2251. 136

[8] M. Athans, *The matrix minimum principle*, Information and control, 11 (1967), pp. 592–606. 12

[9] Z. Bai, *Krylov subspace techniques for reduced-order modeling of nonlinear dynamical systems*, Applied Numerical Mathematics, 43 (2002), pp. 9–44. 71, 137

[10] Z. Bai and D. Skoogh, *A projection method for model reduction of bilinear dynamical systems*, Linear Algebra and its Applications, 415 (2006), pp. 406–425. 71

[11] M. Barrault, Y. Maday, N.C. Nguyen, and A.T. Patera, *An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations*, Comptes Rendus Mathematique, 339 (2004), pp. 667– 672. 136

[12] R.H. Bartels and G.W. Stewart, *Solution of the matrix equation $AX+XB = C$: Algorithm 432*, Communications of the ACM, 15 (1972), pp. 820–826. 37

[13] U. Baur, C.A. Beattie, P. Benner, and S. Gugercin, *Interpolatory projection methods for parameterized model reduction*, SIAM Journal on Scientific Computing, 33 (2011), pp. 2489–2518. 133

[14] U. Baur and P. Benner, *Factorized solution of Lyapunov equations based on hierarchical matrix arithmetic*, Computing, 78 (2006), pp. 211–234. 38

[15] ——, *Gramian-based model reduction for data-sparse systems*, SIAM Journal on Scientific Computing, 31 (2008), pp. 776–798. 38

[16] B. Beckermann, *An error analysis for rational Galerkin projection applied to the Sylvester equation*, SIAM Journal on Numerical Analysis, 49 (2011), p. 2430. 124

[17] P. Benner, *Partial stabilization of descriptor systems using spectral projectors*, in Numerical Linear Algebra in Signals, Systems and Control, P. Van Dooren, P.S. Bhattacharyya, R.H. Chan, V. Olshevsky, and A. Routray, eds., vol. 80 of Lecture Notes in Electrical Engineering, Springer Netherlands, 2011, pp. 55–76. 25

[18] P. Benner and T. Breiten, *On $\mathcal{H}_2$-model reduction of linear parameter-varying systems*, in Proceedings in Applied Mathematics and Mechanics, vol. 11, 2011, pp. 805–806. 133

[19] ——, *On optimality of interpolation-based low-rank approximations of large-scale matrix equations*, MPI Magdeburg Preprints MPIMD/11-10, 2011. Available from `http://www.mpi-magdeburg.mpg.de/preprints/abstract.php?nr=11-10&year=2011`. i

[20] ——, *Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems*, SIAM Journal on Matrix Analysis and Applications, 33 (2012), pp. 859–885. i, 12

[21] ——, *Two-sided moment matching methods for nonlinear model reduction*, MPI Magdeburg Preprints MPIMD/12-12, 2012. Available from `http://www.mpi-magdeburg.mpg.de/preprints/2012/12/`. i

[22] ——, *Low rank methods for a class of generalized Lyapunov equations and related issues*, Numerische Mathematik, to appear (2013). i

[23] P. Benner, T. Breiten, and T. Damm, *Generalised tangential interpolation for model reduction of discrete-time MIMO bilinear systems*, International Journal of Control, 84 (2011), pp. 1398–1407. 101

[24] P. Benner and T. Damm, *Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems*, SIAM Journal on Control and Optimization, 49 (2011), pp. 686–711. 70, 71, 75, 110, 111, 112, 163, 164

[25] P. Benner, M. Köhler, and J. Saak, *Sparse-dense Sylvester equations in $\mathcal{H}_2$-model order reduction*, MPI Magdeburg Preprints MPIMD/11-11, 2011. Available from `http://www.mpi-magdeburg.mpg.de/preprints/abstract.php?nr=11-11&year=2011`. 59, 93, 97

[26] P. Benner, J.-R. Li, and T. Penzl, *Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems*, Numerical Linear Algebra with Applications, 15 (2008), pp. 755–777. 38, 121

[27] P. Benner, R.C. Li, and N. Truhar, *On the ADI method for Sylvester equations*, Journal of Computational and Applied Mathematics, 233 (2009), pp. 1035–1045. 38, 121, 125

[28] P. Benner, V. Mehrmann, and D.C. Sorensen, *Dimension Reduction of Large-Scale Systems*, vol. 45 of Lecture Notes in Computational Science and Engineering, Springer-Verlag, Berlin/Heidelberg, Germany, 2005. 24, 38

[29] P. Benner and E.S. Quintana-Orti, *Solving stable generalized Lyapunov equations with the matrix sign function*, Numerical Algorithms, 20 (1999), pp. 75–100. 38

[30] P. Benner and J. Saak, *Efficient numerical solution of the LQR-problem for the heat equation*, in Proceedings in Applied Mathematics and Mechanics, vol. 4, 2004, pp. 648–649. 38, 64

[31] ——, *Linear-quadratic regulator design for optimal cooling of steel profiles*, Tech. Report SFB393/05-05, Sonderforschungsbereich 393 *Parallele Numerische Simulation für Physik und Kontinuumsmechanik*, TU Chemnitz, 09107 Chemnitz, FRG, 2005. Available from `http://www.tu-chemnitz.de/sfb393`. 38, 108

[32] T. Breiten, *Krylov Subspace Methods for Model Order Reduction of Bilinear Control Systems*, Diplomarbeit, Department of Mathematics, Technical University of Kaiserslautern, 2009. 73, 77

[33] T. Breiten and T. Damm, *Krylov subspace methods for model order reduction of bilinear control systems*, Systems & Control Letters, 59 (2010), pp. 443–450. 71, 107, 137, 147, 150, 154

[34] C. Bruni, G. DiPillo, and G. Koch, *On the mathematical models of bilinear systems*, Automatica, 2 (1971), pp. 11–26. 70

[35] T. Bui-Thanh, M. Damodaran, and K. Willcox, *Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition*, AIAA Journal, 42 (2004), pp. 1505–1516. 136

[36] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek, h_2-

norm optimal model reduction for large scale discrete dynamical MIMO systems, Journal of Computational and Applied Mathematics, 233 (2010), pp. 1202–1216. 24, 27, 36, 46, 163

[37] N. Chafee and E.F. Infante, *A bifurcation problem for a nonlinear partial differential equation of parabolic type*, Applicable Analysis, 4 (1974), pp. 17–37. 156

[38] Y. Chahlaoui and P. Van Dooren, *A collection of benchmark examples for model reduction of linear time invariant dynamical systems*, SLICOT Working Note 2002–2, 2002. 3

[39] S. Chaturantabut and D.C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM Journal on Scientific Computing, 32 (2010), pp. 2737–2764. 136, 158, 159, 160

[40] M. Condon and R. Ivanov, *Krylov subspaces from bilinear representations of nonlinear systems*, The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, 26 (2007), pp. 11–26. 71, 77, 111

[41] P. D'Alessandro, A. Isidori, and A. Ruberti, *Realization and structure theory of bilinear dynamical systems*, SIAM Journal on Control and Optimization, 12 (1974), pp. 517–535. 73

[42] T. Damm, *Rational Matrix Equations in Stochastic Control*, PhD thesis, Fachbereich 3, Universität Bremen, 2002. 70, 75, 163

[43] ——, *Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations*, Numerical Linear Algebra with Applications, 15 (2008), pp. 853–871. 70, 75, 89, 105, 111, 116, 120, 126, 164

[44] P. Van Dooren, K.A. Gallivan, and P.A. Absil, *$\mathcal{H}_2$-optimal model reduction of MIMO systems*, Appl. Math. Lett., 21 (2008), pp. 1267–1273. 27, 88, 93

[45] V. Druskin, L. Knizhnerman, and V. Simoncini, *Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation*, SIAM Journal on Numerical Analysis, 49 (2011), pp. 1875–1898. 38, 65, 123

[46] V. Druskin and V. Simoncini, *Adaptive rational Krylov subspaces for large-scale dynamical systems*, Systems & Control Letters, 60 (2011), pp. 546–560. 27

[47] E.V. Dulov and N.A. Andrianova, *On differentiability of the matrix trace operator and its applications*, Journal of Applied Mathematics and Computing, 8 (2001), pp. 97–109. 12

[48] D.L.R. Elliott, *Bilinear control systems: matrices in action*, vol. 169, Springer Verlag, 2009. 71, 72

[49] D.F. Enns, *Model reduction with balanced realizations: An error bound and a fre-*

quency weighted generalization, in 23rd IEEE Conference on Decision and Control, vol. 23, IEEE, 1984, pp. 127–132. 28

[50] A. Eppler and M. Bollhöfer, *An alternative way of solving large Lyapunov equations*, in Proceedings in Applied Mathematics and Mechanics, vol. 10, 2010, pp. 547–548. 38, 126

[51] ——, *A structure preserving FGMRES method for solving large Lyapunov equations*, in Progress in Industrial Mathematics at ECMI 2010, M. Günther, A. Bartel, M. Brunk, S. Schöps, and M. Striebel, eds., vol. 17 of Mathematics in Industry, Springer Verlag, Berlin/Heidelberg, 2012, pp. 131–136. 38, 126

[52] P. Feldmann and R.W. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 14 (1995), pp. 639–649. 27

[53] L. Feng and P. Benner, *A note on projection techniques for model order reduction of bilinear systems*, in Numerical Analysis and Applied Mathematics, AIP Conference Proceedings, vol. 936, 2007, pp. 208–211. 71

[54] L. Feng, D. Koziol, E.B. Rudnyi, and J.G. Korvink, *Parametric model reduction for fast simulation of cyclic voltammograms*, Sensor Letters, 4 (2006), pp. 165–173. 134

[55] G. Flagg, *$\mathcal{H}_2$-optimal interpolation: New properties and applications*, 2010. Talk given at the 2010 SIAM Annual Meeting, Pittsburgh (PA). 38

[56] G.M. Flagg, *Interpolation Methods for the Model Reduction of Bilinear Systems*, PhD thesis, Virginia Polytechnic Institute and State University, 2012. 88, 97, 133

[57] G. Flagg, C. Beattie, and S. Gugercin, *Interpolatory H-infinity model reduction*, tech. report, 2011. submitted, available as arXiv:1107.5364. 25

[58] G. Flagg, C.A. Beattie, and S. Gugercin, *Convergence of the iterative rational Krylov algorithm*, Systems & Control Letters, 61 (2012), pp. 688–691. 43, 97

[59] G. Flagg and S. Gugercin, *On the ADI method for the Sylvester equation and the optimal-$\mathcal{H}_2$ points*, Applied Numerical Mathematics, 64 (2013), pp. 50–58. 38, 65, 123

[60] R.W. Freund, *Passive reduced-order modeling via Krylov-subspace methods*, in IEEE International Symposium on Computer-Aided Control System Design, 2000, pp. 261–266. 27

[61] ——, *Model reduction methods based on Krylov subspaces*, Acta Numerica, 12 (2003), pp. 267–319. 24

[62] K. Gallivan, A. Vandendorpe, and P. Van Dooren, *Model reduction of MIMO systems via tangential interpolation*, SIAM Journal on Matrix Analysis and

Applications, 26 (2004), pp. 328–349. 27, 62

[63] E. Gilbert, *Functional expansions for the response of nonlinear differential systems*, IEEE Transactions on Automatic Control, 22 (1977), pp. 909 – 921. 141

[64] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^{∞}-error bounds*, International Journal of Control, 39 (1984), pp. 1115–1193. 28

[65] G.H. Golub and C.F. Van Loan, *Matrix computations*, vol. 3, Johns Hopkins University Press, 1996. 11, 12

[66] L. Grasedyck, *Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure*, Computing, 72 (2004), pp. 247–265. 19, 20, 21, 112, 113, 118, 163, 164

[67] ——, *Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation*, Numerical Linear Algebra with Applications, 11 (2004), pp. 371–389. 21, 38, 164

[68] ——, *Hierarchical singular value decomposition of tensors*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2029–2054. 8, 11, 13

[69] L. Grasedyck and W. Hackbusch, *A multigrid method to solve large scale sylvester equations*, SIAM Journal on Matrix Analysis and Applications, 29 (2007), pp. 870–894. 38

[70] M.A. Grepl, Y. Maday, N.C. Nguyen, and A.T. Patera, *Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 41 (2007), pp. 575–605. 136

[71] E.J. Grimme, *Krylov Projection Methods For Model Reduction*, PhD thesis, University of Illinois, 1997. 8, 26, 27, 136, 145, 147, 163

[72] C. Gu, *QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 30 (2011), pp. 1307 – 1320. 9, 136, 137, 139, 140, 142, 146, 151, 163, 164

[73] S. Gugercin, A.C. Antoulas, and S. Beattie, *$\mathcal{H}_2$ model reduction for large-scale dynamical systems*, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 609–638. 8, 21, 22, 27, 33, 34, 36, 68, 75, 89, 94, 97, 163, 164

[74] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, vol. 42 of Springer Series in Computational Mathematics, Springer-Verlag. 8

[75] S.J. Hammarling, *Numerical solution of the stable, non-negative definite Lyapunov equation*, IMA Journal of Numerical Analysis, 2 (1982), pp. 303–323. 37

[76] E. Hansen, F. Kramer, and A. Ostermann, *A second-order positivity preserving scheme for semilinear parabolic problems*, Applied Numerical Mathematics,

62 (2012), pp. 1428–1435. 156, 158

[77] C. Hartmann, A. Zueva, and B. Schäfer-Bung, *Balanced model reduction of bilinear systems with applications to positive systems*, (2010). submitted. 5, 106, 130

[78] D. Hinrichsen and A.J. Pritchard, *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*, vol. 1, Springer Verlag, 2005. 7, 11, 14

[79] D. Hohlfeld and H. Zappe, *An all-dielectric tunable optical filter based on the thermo-optic effect*, Journal of Optics A: Pure and Applied Optics, 6 (2004), pp. 504–511. 65

[80] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990. 8, 11, 114

[81] C.S. Hsu, U.B. Desai, and C.A. Crawley, *Realization algorithms and approximation methods of bilinear systems*, in The 22nd IEEE Conference on Decision and Control, vol. 22, 1983, pp. 783–788. 111

[82] A. Isidori, *Nonlinear Control Systems*, vol. 1, Springer Verlag, 1995. 7, 71, 72

[83] I.M. Jaimoukha and E.M. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM Journal on Numerical Analysis, 31 (1994), pp. 227–251. 37, 50, 124

[84] K. Jbilou and A. J. Riquet, *Projection methods for large Lyapunov matrix equations*, Linear Algebra and its Applications, 415 (2006), pp. 344–358. 37, 124

[85] D.L. Kleinman, *Suboptimal design of linear regulator systems subject to computer storage limitations*, PhD thesis, Massachusetts Institute of Technology, 1967. 12

[86] H.W. Knobloch and H. Kwakernaak, *Lineare Kontrolltheorie*, Springer-Verlag, Berlin, 1985. In German. 14

[87] T.G. Kolda and B.W. Bader, *Tensor decompositions and applications*, SIAM Review, 51 (2009), pp. 455–500. 8, 11, 13

[88] A.J. Krener, *Linearization and bilinearization of control systems*, in Proceedings of the 12th Annual Allerton Conference on Circuit and System Theory, vol. 834, Monticello, 1974. 75

[89] D. Kressner and C. Tobler, *Krylov subspace methods for linear systems with tensor product structure*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 1688–1714. 13, 19, 21, 116

[90] ——, *Low-rank tensor Krylov subspace methods for parametrized linear systems*, SIAM Journal on Matrix Analysis and Applications, 32 (2011), pp. 1288–1316. 38, 125, 126

[91] ——, *Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems*, Computational Methods in Applied Mathematics, 11 (2011), pp. 363–381. 112

[92] K. Kunisch and S. Volkwein, *Control of the Burgers equation by a reduced-order approach using proper orthogonal decomposition*, Journal of Optimization Theory and Applications, 102 (1999), pp. 345–371. 136, 153

[93] ——, *Proper orthogonal decomposition for optimality systems*, ESAIM: Mathematical Modelling and Numerical Analysis, 42 (2008), pp. 1–23. 136, 153

[94] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, Orlando, 2nd ed., 1985. 8, 47

[95] J.M. Landsberg, *Tensors: Geometry and applications*, vol. 128, American Mathematical Society, 2011. 8

[96] J.-R. Li, *Model Reduction of Large Linear Systems via Low Rank System Gramians*, PhD thesis, Massachusettes Institute of Technology, September 2000. 38, 121

[97] J.-R. Li and J. White, *Low rank solution of Lyapunov equations*, SIAM Journal on Matrix Analysis and Applications, 24 (2002), pp. 260–280. 38, 121, 163

[98] G.P. McCormick, *Computability of global solutions to factorable nonconvex programs: Part I – Convex underestimating problems*, Mathematical Programming, 10 (1976), pp. 147–175. 139

[99] L. Meier and D.G. Luenberger, *Approximation of linear constant systems*, IEEE Transactions on Automatic Control, 12 (1967), pp. 585–588. 8, 34, 68, 163, 164

[100] R.R. Mohler, *Bilinear Control Processes*, New York Academic Press, 1973. 7, 70, 71

[101] ——, *Nonlinear Systems (vol. 2): Applications to Bilinear Control*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1991. 70

[102] ——, *Natural bilinear control processes*, IEEE Transactions on Systems Science and Cybernetics, 6 (2007), pp. 192–197. 70

[103] B.C. Moore, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Transactions on Automatic Control, AC-26 (1981), pp. 17–32. 27, 28

[104] C. Mullis and R. Roberts, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Transactions on Circuits and Systems, 23 (1976), pp. 551–562. 27

[105] N.C. Nguyen, A.T. Patera, and J. Peraire, *A 'best points' interpolation method for efficient approximation of parametrized functions*, International Journal

for Numerical Methods in Engineering, 73 (2008), pp. 521–543. 136

[106] G. Obinata and B.D.O. Anderson, *Model Reduction for Control System Design*, Springer Verlag, 2000. 24

[107] D.W. Peaceman and H.H. Rachford, *The numerical solution of parabolic and elliptic differential equations*, Journal of the Society for Industrial and Applied Mathematics, 3 (1955), pp. 28–41. 38

[108] T. Penzl, *A cyclic low rank Smith method for large, sparse Lyapunov equations with applications in model reduction and optimal control*, Tech. Report SFB393/98-6, Fakultät für Mathematik, TU Chemnitz, 09107 Chemnitz, FRG, 1998. Available from `http://www.tu-chemnitz.de/sfb393/sfb98pr.html`. 38

[109] ——, *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case*, Systems & Control Letters, 40 (2000), pp. 139–144. 18, 121

[110] ——, *Algorithms for model reduction of large dynamical systems*, Linear Algebra and its Applications, 415 (2006), pp. 322 – 343. Special Issue on Order Reduction of Large-Scale Systems. 38, 65

[111] J.R. Phillips, *Projection frameworks for model reduction of weakly nonlinear systems*, in Proceedings of Design Automatic Conference, 2000, pp. 184–189. 71, 77, 137

[112] ——, *Projection-based approaches for model reduction of weakly nonlinear, time-varying systems*, IEEE Trans. Circuits and Systems, 22 (2003), pp. 171–187. 71, 77, 137

[113] J.W. Polderman and J.C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, vol. 26, Springer Verlag, 1998. 14

[114] M.J. Rewienski, *A Trajectory Piecewise-Linear Approach to Model Order Reduction of Nonlinear Dynamical Systems*, PhD thesis, Massachusetts Institute of Technology, 2003. 136

[115] W.J. Rugh, *Nonlinear System Theory*, The Johns Hopkins University Press, 1982. 7, 70, 71, 72, 73, 75, 76, 140, 141, 142

[116] Y. Saad, *Numerical solution of large Lyapunov equation*, in Signal Processing, Scattering, Operator Theory and Numerical Methods, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Birkhauser, 1990, pp. 503–511. 37, 40, 124

[117] ——, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, PA, USA, 2003. 8, 25, 54, 55

[118] W.H.A. Schilders, H.A. Van Der Vorst, and J. Rommes, *Model Order Reduction: Theory, Research Aspects and Applications*, vol. 13, Springer Verlag, 2008. 24

[119] H. Schwarz, *Stability of discrete-time equivalent homogeneous bilinear systems*,

Control Theory and Advanced Technology, 3 (1987), pp. 263–269. 102

[120] V. Simoncini, *A new iterative method for solving large-scale Lyapunov matrix equations*, SIAM Journal on Scientific Computing, 29 (2007), pp. 1268–1288. 37, 123, 124, 163

[121] T. Siu and M. Schetzen, *Convergence of Volterra series representation and BIBO stability of bilinear systems*, International Journal of Systems Science, 22 (1991), pp. 2679–2684. 73

[122] E.D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, vol. 6, Springer Verlag, 1998. 14

[123] D.C. Sorensen and Y. Zhou, *Bounds on eigenvalue decay rates and sensitivity of solutions of Lyapunov equations*, Tech. Report 7, Rice University, 2002. 19

[124] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, vol. 20 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1993. 21

[125] B. Vandereycken, *Riemannian and multilevel optimization for rank-constrained matrix problems*, PhD thesis, Department of Computer Science, Katholieke Universiteit Leuven, 2010. 8, 38, 39, 64, 66, 68, 163

[126] B. Vandereycken and S. Vandewalle, *A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2553–2579. 39, 50

[127] C. De Villemagne and R.E. Skelton, *Model reduction using a projection formulation*, International Journal of Control, 46 (1987), pp. 2141–2169. 26, 27

[128] E.L. Wachspress, *Iterative solution of the Lyapunov matrix equation*, Applied Mathematics Letters, 107 (1988), pp. 87–90. 121

[129] ——, *The ADI model problem*, 1995. Available from the author. 38, 128

[130] D. Werner, *Funktionalanalysis*, Springer Verlag, 2005. 144

[131] D.A. Wilson, *Optimum solution of model-reduction problem*, Proceedings of the Institution of Electrical Engineers, 117 (1970), pp. 1161–1165. 33, 163

[132] Y. Xu and T. Zeng, *Optimal $\mathcal{H}_2$ model reduction for large scale MIMO systems via tangential interpolation*, International Journal of Numerical Analysis and Modeling, 8 (2011), pp. 174–188. 94

[133] L. Zhang and J. Lam, *On $\mathcal{H}_2$ model reduction of bilinear systems*, Automatica, 38 (2002), pp. 205–216. 53, 54, 74, 75, 82, 110, 163, 164

[134] L. Zhang, J. Lam, B. Huang, and G. Yang, *On gramians and balanced truncation of discrete-time bilinear systems*, International Journal of Control, 76 (2003), pp. 414–427. 71, 101

[135] Y. Zhou and D. Sorensen, *Approximate implicit subspace iteration with alternating directions for LTI system model reduction*, Numerical Linear Algebra with Applications, 15 (2008), pp. 873–886. 38

SCHRIFTLICHE ERKLÄRUNG

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Insbesondere habe ich nicht die Hilfe einer kommerziellen Promotionsberatung in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation, Diplom- oder ähnliche Prüfungsarbeit eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

(Ort, Datum)

(Unterschrift)