System-Theoretic Model Order Reduction for Bilinear and Quadratic-Bilinear Systems

Dissertation
zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von **Pawan Kumar Goyal**

geb. am **01.07.1993** in Nokha, Rajasthan, India

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

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

Prof. Dr. Serkan Gugercin

eingereicht am: **28.11.2017**

Verteidigung am: **02.03.2018**

LIST OF PUBLICATIONS

Most of the material presented in this thesis is either published or submitted for publications.

Chapter 3 is an extended version of

[30]: P. Benner, P. Goyal, and M. Redmann, Truncated Gramians for bilinear systems and their advantages in model order reduction, in P. Benner, M. Ohlberger, T. Patera, G. Rozza, K. Urban (Eds.), Model Reduction of Parametrized Systems, MS&A - Modeling, Simulation and Applications, Springer International Publishing, Cham., vol. 17, pp. 285–300, 2017.

The material of Chapter 4 is available as preprint

[28]: P. Benner, P. Goyal, Balanced truncation model order reduction for quadratic-bilinear control systems, arXiv e-prints 1705.00160, 2017.

Chapter 5 is based on the preprint:

[29]: P. Benner, P. Goyal, and S. Gugercin, $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems, SIAM Journal on Matrix Analysis and Applications, 2018, to appear.

Chapter 6 is a combination of the following three published articles

[2]: M. I. Ahmad, P. Benner, and P. Goyal, Krylov subspace-based model reduction for a class of bilinear descriptor systems, J. Comput. Appl. Math., 315 (2017), pp. 303–318.

[27]: P. Benner and P. Goyal, Multipoint interpolation of Volterra series and $\mathcal{H}_2$-model reduction for a family of bilinear descriptor systems, Systems Control Lett., 97 (2016), pp. 1–11.

[71]: P. Goyal and P. Benner, An iterative model order reduction scheme for a special

class of bilinear descriptor systems appearing in constraint circuit simulation, in ECCOMAS Congress 2016, VII European Congress on Computational Methods in Applied Sciences and Engineering, vol. 2, 2016, pp. 4196–4212.

ACKNOWLEDGMENTS

Writing up this thesis is one of the biggest achievements of my life. This would not have been possible without having a great support from the people around me. I have been fortunate to be surrounded by people who have helped and encouraged me during the last four years, and I would like to take the opportunity to thank them.

First of all, I would like to thank my supervisor Prof. Dr. Peter Benner for giving me the opportunity to do my dissertation in his group *Computational Methods in Systems and Control Theory* at *Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany.* I am truly indebted to his great support and guidance, in helping me to improve not only my professional life but also my personal life in Germany. I wish to express my thank to Prof. Dr. Serkan Gugercin for agreeing to be the second referee for this thesis. Due to his expertise in the field of model reduction, especially in the area of $\mathcal{H}_2$-optimal model reduction, I think there could not have been a better choice.

Furthermore, I am grateful to many pleasant and friendly colleagues with whom I have got opportunities to work or had several scientific discussions. Although I am grateful to all of them, I would like to name a few of them explicitly. I am obliged to Jens Saak for clearing not only countless mathematical and model reduction related doubts but also problems related to LaTex/TikZ. I would like to thank Mian Ilyas Ahmad, Jan Heiland, Igor Pontes Duff Pereira, Ion Victor Gosea, and Patrick Kürschner. I further thank my officemates Jessica Bosch, Hamdullah Yücel, Martin Redmann, Carolin Penke, Sridhar Chellappa, Dimitrios Karachalios for providing a friendly and motivating atmosphere. Several of them have became my close friends and brightened up life at work as well as after work. I would like to thank my friends Franziska Baumert, Shahshank Bhandari, Milena Radoicic, Micheal Pieler along with my colleagues for making my life outside work in Magdeburg way better than I could imagine. I wish to express a special thank to my host family, Micheal and Jasmina Ritter, for accepting me as an integral part of their family and showing me the beautiful German and European cultures.

Last but not least, I would like to thank my family for supporting me at every possible moment and their encouragements. I cannot thank enough my father Bajrang

Lal, my mother Premlata, and my sisters Priya, Prity, Pinky. I endlessly appreciate it.

ABSTRACT

In this thesis, we study system-theoretic model reduction techniques for special classes of nonlinear systems, namely, bilinear and quadratic-bilinear (QB) control systems. There is a large variety of applications, where control systems can be modeled as one of the above-mentioned nonlinear systems, for example, boundary control problems, flow problems, neuronal dynamics. Our particular focus lies on balancing-type and $\mathcal{H}_2$-optimal model reduction problems of the latter nonlinear systems. In the first part of the thesis, we focus on balancing-type model reduction for bilinear and QB control systems. We begin by revisiting the connection between the Gramians and energy functions of bilinear systems and introduce a concept of truncated Gramians. We further study balanced truncation model reduction technique for QB systems by extending the idea of Gramians for bilinear systems and propose algebraic Gramians for the latter systems. We additionally establish connections between the proposed Gramians and the energy functionals for QB systems. Moreover, we discuss the usage of Gramians in the model reduction framework of QB systems. In the second part of the thesis, we turn our attention to interpolation-based $\mathcal{H}_2$-optimal model reduction. In this direction, we derive interpolation-based model reduction conditions for QB control systems, which aim at minimizing a system norm of the QB system, namely a truncated version of the $\mathcal{H}_2$-norm of the latter system. Based on these conditions, we propose an iterative scheme that *approximately* satisfies the derived optimality conditions. Lastly, we investigate interpolation-based model reduction for bilinear systems that are subject to algebraic constraints. We show how to extend the existing knowledge of model reduction for linear descriptor systems to interpolation-based model reduction for specially structured bilinear descriptor systems (DAEs). We also propose several modified iterative schemes, leading to locally $\mathcal{H}_2$-optimal reduced-order systems for the structured bilinear DAEs. By means of several numerical examples, we compare the efficiency of all the proposed model reduction schemes for bilinear and QB systems with the existing state-of-the-art methods.

ZUSAMMENFASSUNG

In dieser Arbeit untersuchen wir systemtheoretische Modellreduktionstechniken für spezielle Klassen nichtlinearer Systeme, insbesondere betrachten wir bilineare und quadratisch-bilineare (QB) Eiggangs/Ausgangs-Systeme. Es gibt eine große Vielfalt von Anwendungen, bei denen Eiggangs/Ausgangs-Systeme als eins der oben genannten nichtlinearen Systeme modelliert werden können, zum Beispiel Strömungsprobleme, Randsteuerungsprobleme und neuronale Dynamiken. Unser besonderer Fokus liegt dabei auf balancierenden und $\mathcal{H}_2$-optimalen Modellreduktionsproblemen zu letztgenannten nichtlinearen Systeme. Im ersten Teil der Arbeit konzentrieren wir uns auf die balancierende Modellreduktion für bilineare und QB Eiggangs/Ausgangs-Systeme. Wir beginnen damit, die Verbindung zwischen den Gramschen und den Energiefunktionalen bilinearer Systeme erneut aufzugreifen und ein Konzept von abgeschnittenen Gramschen einzuführen. Wir untersuchen weiterhin die Modellreduktionsmethode des balancierten Abschneidens für QB Systeme, indem wir die Idee der Gramschen für bilineare Systeme erweitern und algebraische Gramsche für die letztgenannten Systeme vorschlagen. Wir stellen außerdem die Verbindung zwischen den vorgeschlagenen Gramschen und den Energiefunktionalen für QB Systeme her. Darüber hinaus diskutieren wir die Verwendung von Gramschen in kontext der Modellreduktion für QB Systeme. Im zweiten Teil der Arbeit widmen wir uns der interpolationsbasierten $\mathcal{H}_2$-optimalen Modellreduktion. Diesbezüglich leiten wir interpolationsbasierte Modellreduktionsbedingungen für QB Steuersysteme ab, die darauf abzielen, eine gewisse Systemnorm des QB Systems zu minimieren, nämlich eine abgeschnittene Version der $\mathcal{H}_2$-Norm. Basierend auf diesen Bedingungen schlagen wir ein iteratives Verfahren vor, das die abgeleiteten Optimalitätsbedingungen *näherungsweise* erfüllt. Schließlich untersuchen wir interpolationsbasierte Modellreduktionsmethoden für bilineare Systeme mit algebraischen Nebenbedingungen. Wir zeigen, in der literature verfügbare verfahren zur Modellreduktion von linearen Deskriptorsystemen auf die interpolationsbasierte Modellreduktion für speziell strukturierte bilineare Deskriptorsysteme erweitert werden kann. Wir schlagen ebenso mehrere modifizierte iterative Schemata vor, die zu lokal $\mathcal{H}_2$-optimalen reduzierten Systemen für die strukturierten bilinearen DAEs führen. Anhand mehrerer numerischer Beispiele vergleichen wir die Effizienz aller vorgeschlagenen Modellreduktionsverfahren

für bilineare und QB Systeme mit der existierender Methoden.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS AND NOTATION

Acronyms and Abbreviations

ADI	alternating directions implicit
B-IRKA	bilinear iterative rational Krylov algorithm
BIBO	bounded-input-bounded-output
BT	balanced truncation
DAE	differential-algebraic equation
(D)EIM	(discrete) empirical interpolation method
EKSM	extended Krylov subspace method
IRKA	iterative rational Krylov algorithm
LTI	linear time-invariant
MIMO	multiple-input multiple-output
MATLAB®	software by The MathWorks Inc.
MOR	model order reduction
ODE	ordinary differential equation
PDE	partial differential equation
POD	proper orthogonal decomposition
QB	quadratic-bilinear
SISO	single-input single-output
SVD	singular value decomposition
TPWL	trajectory piecewise linear
TQB-IRKA	truncated QB iterative rational Krylov algorithm
Xeon®	processor series by Intel®

Notation

$\mathbb{R}$, $\mathbb{C}$	fields of real and complex numbers
$\mathbb{C}_+$, $\mathbb{C}_-$	open right/open left complex half plane
$\mathbb{R}^n$, $\mathbb{C}^n$	vector space of real/complex n-tuples
$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$	real/complex $m \times n$ matrices
$\lvert\xi\rvert$,	absolute value of real or complex scalar
$\jmath$	imaginary unit ($\jmath^2 = -1$)

x	vector $\in \mathbb{R}^n$
x_k	kth entry of x
A	matrix $\in \mathbb{R}^{n\times m}$
A_{ij}	the (i,j)th entry of A
$A(i:j,:)$, $A(:,k:\ell)$	rows $i,\dots,j$ of A and columns $k,\dots,\ell$ of A
$A(i:j,k:\ell)$	rows $i,\dots,j$ of columns $k,\dots,\ell$ of A
A^T	the transpose of A
A^{-1}	inverse of nonsingular A
A^{-T},	inverse of A^T
I_n, I	identity matrix of size $n \times n$, or of suitable size
$0_{n\times m}$, 0	zero matrix of size $n \times m$, or of suitable size
e_i^n	ith column of the identity matrix of size $n \times n$, i.e., I_n
$\operatorname{rank}(A)$	rank of a matrix A
$\operatorname{span}(A)$	subspace spanned by the columns of a matrix A
$\operatorname{orth}(A)$	orthonormal subspace spanned by the columns of a matrix A
Σ_l	linear control system .... Equation (2.1)
Σ_B	bilinear control system..................... Equation (3.1)
Σ_{QB}	quadratic-bilinear control system .... Equation (4.1)
$\mathcal{K}_q(A,b)$	Krylov subspace spanned by $\{b, Ab, \dots, A^{q-1}b\}$
$\mathbb{1}_r$	$:= (1,\dots,1)^T \in \mathbb{R}^r$
$\Lambda(A)$, $\Lambda(A,M)$	spectrum of matrix A/matrix pair (A,M)
$\sigma_{\max}(A)$, $\sigma_{\min}(A)$	the largest/smallest singular value of A
$\operatorname{tr}(A)$	$:= \sum_{i=1}^n a_{ii}$, trace of A
$\\|u\\|_p$	$:= \sqrt[p]{\sum\limits_{i=1}^{n} \|u_i\|^p}$ for $u \in \mathbb{C}^n$ and $1 \leq p < \infty$
$\\|u\\|_\infty$	the maximum norm ($\\|u\\|_\infty = \max_i \|u_i\|$)
$A \otimes B$	the Kronecker product of A and B....... (Definition 2.23)
$\operatorname{vec}(A)$	vectorization operator applied to matrix A (Definition 2.23)
$\mathcal{J}_m$	vectorized identity matrix of dimension m, i.e., $\mathcal{J}_m = \operatorname{vec}(I_m)$

CHAPTER 1

INTRODUCTION

Contents

1.1. Motivation of Model Order Reduction

In engineering studies such as control design, prediction, and optimization of dynamical systems, numerical simulations are considered to be one of the fundamental tools for studying various properties. These dynamical systems are generally governed by partial differential equations (PDEs), or ordinary differential equations (ODEs), or a combination of both. Furthermore, sometimes these dynamical systems are also subject to some constraints, coming from some practical considerations, or restrictions from the environments, or by the laws of physics. Thus, we obtain a set of differential equations along with algebraic equations. Such systems are called differential-algebraic equations (DAEs), or descriptor systems.

Highly accurate mathematical models, describing dynamical behaviors of the system are desirable for engineering design studies. However, these models are generally complex in nature, thus are computationally cumbersome. The word *complex* might have different interpretations. For instance, one way to think of complex systems is a system with a large number of degrees of freedom. As a result, we obtain a large number of equations, which is also called the state dimension. This can be viewed as a complex model since numerical simulations of such systems with a large number of states (large-scale systems) are very computationally expensive and might be inefficient, too.

Furthermore, to capture the dynamics of many real-life applications, nonlinear terms need to be added to describe the system dynamics accurately. These nonlinear terms make engineering studies difficult. Hence, such mathematical models also belong to the category of complex systems. Additionally, as mentioned before, sometimes algebraic constraints are also necessary to completely describe the dynamics, which also makes the numerical analysis very complicated. These models can also be described as complex systems.

In this thesis, we mainly focus on the dynamical systems that have high state-space dimensions and special structure of their nonlinear terms. These systems are controlled by external forces which are called control inputs. Formally, these systems are of the form

$$E\dot{x}(t) = f(x(t)) + g(x(t))u(t), \tag{1.1}$$

where $x(t) \in \mathbb{R}^n$ is state vector or solution trajectory of the system, $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are smooth nonlinear functions, $E \in \mathbb{R}^{n\times n}$ might be singular, and $u(t) : \mathbb{R} \to \mathbb{R}^m$ is the input vector at time t and is an L_2 bounded function. Additionally, the system (1.1) is generally a high fidelity model, i.e., $n \sim \mathcal{O}(10^5) - \mathcal{O}(10^6)$.

Furthermore, from a practical point of view, it is hardly possible to observe the whole state $x(t)$, instead we are interested either in very few state components or a function of the state vector. Therefore, we often have an output equation as well, which has of form

$$y(t) = h(x(t)) + k(x(t))u(t), \tag{1.2}$$

where $y(t) \in \mathbb{R}^p$ is an output vector and $h : \mathbb{R}^n \to \mathbb{R}^p$ and $k : \mathbb{R}^n \to \mathbb{R}^{p\times m}$ are smooth nonlinear functions. Although the dynamics of the system are governed by a large state vector $x(t) \in \mathbb{R}^n$, commonly, the numbers of the observed outputs and control inputs are relatively small, i.e., $m, p \ll n$. In most cases, we are interested in knowing how the control inputs influence the output of the system. For this, there exists a mapping $\mathcal{M} : \mathbb{R}^m \to \mathbb{R}^p$ which maps the input to the output of the system.

As noted before, numerical simulations of these complex systems are numerical inefficient and expensive. Furthermore, on several occasions, the storage of these large-scale systems can also cause some troubles. This inspires *model order reduction* (MOR), which aims at constructing simple and reliable surrogate models that approximate the input-output behavior of the original model. Precisely, our focus is to determine a surrogate model (reduced-order model) of the form

$$\begin{aligned} \widehat{E}\dot{\widehat{x}}(t) &= \widehat{f}(\widehat{x}(t)) + \widehat{g}(\widehat{x}(t))u(t), \\ \widehat{y}(t) &= \widehat{h}(\widehat{x}(t)) + \widehat{k}(\widehat{x}(t))u(t), \end{aligned} \tag{1.3}$$

where $\widehat{x}(t) \in \mathbb{R}^{\widehat{n}}$ is a reduced state vector or solution trajectory of the reduced-order system, $\widehat{f} : \mathbb{R}^{\widehat{n}} \to \mathbb{R}^{\widehat{n}}$, $\widehat{g} : \mathbb{R}^{\widehat{n}} \to \mathbb{R}^{\widehat{n}\times m}$, $\widehat{h} : \mathbb{R}^{\widehat{n}} \to \mathbb{R}^p$ and $\widehat{k} : \mathbb{R}^{\widehat{n}} \to \mathbb{R}^{\widehat{n}\times m}$ are reduced nonlinear functions, and $\widehat{E} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$. In a conventional MOR problem, we essentially, want to minimize $\|y(t) - \widehat{y}(t)\|$ while ensuring the order of the reduced-order system (1.3) is much less than the one of the original system (1.1), i.e., $\widehat{n} \ll n$. Additionally, the important properties such as stability and passivity of the original system are preserved in the reduced-order system (1.3). These surrogate models can then be used in engineering studies, which make numerical simulations faster and efficient.

In the past decades, numerous theoretical and computational aspects for MOR of linear systems have been developed; e.g., see [7, 15, 17, 31, 34, 115]. These systems are of the form

$$\begin{aligned} Ex(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1.4}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are state, input and output vectors, respectively, and all other system matrices are of appropriate sizes. These methods have been successfully applied in various fields, e.g., optimal control, PDE constrained optimization, uncertainty quantification, see, for example [31, 37, 89].

However, the dynamics of the several real-life applications, e.g., flow problems, are hard to capture using linear systems. Therefore, in recent years, MOR of nonlinear systems has gained significant interest with a goal of extending the MOR techniques based on systems theory from linear systems to nonlinear ones, for example, MOR techniques for linear systems such as balanced truncation [7, 103], or the iterative rational Krylov method (IRKA) [79]. The main obstacle in extending these methods to nonlinear systems is that it is hard to obtain an analytic expression or knowledge about the system dynamics. However, by considering special explicit expressions of nonlinear functions in (1.1) and (1.2), we can attempt to derive the analytic expressions, describing the dynamics of the nonlinear system. In this thesis, we, thus, consider two crucial classes of nonlinear systems: the first consists of bilinear systems which act as a bridge between fully nonlinear systems and linear systems, and the second one comprises the quadratic-bilinear systems which cover a large number of smooth nonlinear systems. We provide detailed descriptions of these nonlinear systems in their respective chapters; therefore, we refrain ourselves from giving details about these systems here.

1.2. Motivating Examples

In this section, we provide a couple of motivating examples, arising in different areas of science and engineering that illustrate practical usages of MOR.

Electrical circuits

As motivating examples, we consider systems or models arising from electrical engineering since model reduction of various electrical circuits are considered in this thesis. Depending on the consideration of electrical components, the resulting mathematical models can be different. For instance, we consider an electric circuit as shown in Figure 1.1, where a source of voltage is applied across a capacitor; but one can also

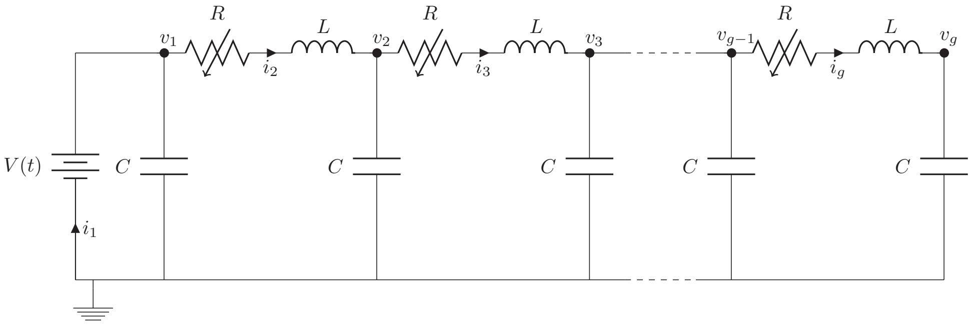


Figure 1.1.: An RLC circuit diagram.

consider a source of current being applied across a capacitor. Such electric circuits can be modeled, for example, by Kirchoff's law. In case all the electrical components are constant for the circuit shown in Figure 1.1, a mathematical model can be given by a linear descriptor system. However, in practice, there are several applications, where, for instance, resistors follow a nonlinear dynamics, thus leading to nonlinear models. In Chapters 4 and 5, we consider an electrical circuit ladder, where resistance dynamics are governed by a function containing exponential nonlinearity. However, it can be modeled as a quadratic-bilinear control system as well.

There are also engineering applications, where electric components are a function of parameters, e.g., the resistance R depends on a parameter p as $R(p) = R_0 + pR_p$, where R_0 and R_p are constants. This can lead to a linear parametric descriptor system. Although in this thesis, we particularly do not focus on model reduction for parametric systems, there is an interesting and strong connection between linear parametric systems and bilinear control systems; see, e.g., [20]. A detailed description of such a model is given in Chapter 6. We will consider model reduction of various electrical circuits in Chapters 4 to 6, where a detailed modeling of these circuits is explained.

The FitzHugh-Nagumo system

As a second motivating example, we consider a model which describes the activation and de-activation of a spiking neuron. This model, proposed by FitzHugh and Nagumo, is a simplified model of Hodgkin–Huxley model, which explains a spiking neural dynamics in a detailed manner. The FitzHugh-Nagumo (F-N) model is described by the following coupled nonlinear PDEs

$$\begin{aligned} \epsilon v_t(x,t) &= \epsilon^2 v_{xx}(t) + f(v(x,t)) - w(x,t) + g, && (1.5a)\\ w_t(x,t) &= hv(x,t) - \gamma w(x,t) + g, && (1.5b) \end{aligned}$$

where $f(v) = v(v-0.1)(1-v)$; ϵ, γ, h, g are constant terms, and the variables v and w represent a voltage and a recovery voltage related to the neuron subject to an

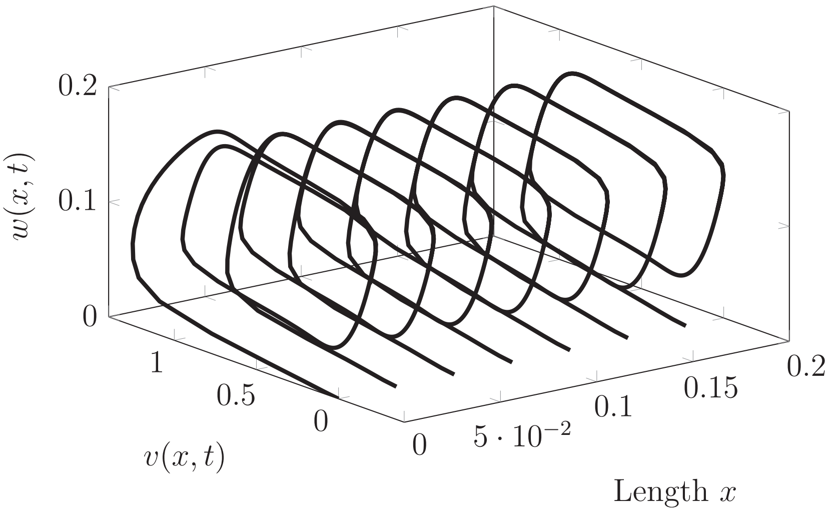


Figure 1.2.: Typical dynamical behavior for FitzHugh-Nagumo Model.

external excitation source. This model is an example of a relaxation oscillator. This is mainly due to the fact that the system will exhibit an excursion in a phase-space as the external excitation exceeds its threshold values, and after a while, the variables v and w are relaxed to their rest values. Such a phenomenon can be seen in Figure 1.2, which is a 3-dimensional figure, showing evolutions of the variables v and w with time. Even though discretizing these coupled PDEs (1.5) leads to a polynomial control system of order 3, rather than a bilinear or a quadratic-bilinear system, in Chapter 4, we will discuss how to rewrite such a polynomial control system equivalently into a quadratic-bilinear control systems. This allows us to employ model reduction techniques for quadratic-bilinear systems, developed in Chapters 4 and 5 of this thesis.

1.3. Outline of the Thesis

The main goal of this thesis is to study model reduction techniques for nonlinear systems via systems theory. For this, in Chapter 2, we first briefly provide basic concepts of linear control theory such as controllability, observability, and stability. For linear systems, we give primary ideas of projection-based model reduction methods, namely balanced truncation, and interpolation-based methods. Next, we revise the concepts of energy functionals and adjoint systems for nonlinear systems which play a critical role when we discuss balanced truncation for bilinear and quadratic-bilinear systems in Chapters 3 and 4, respectively. We also aim at providing necessary tensor theory tools. Related to it, we present various expressions and formulas related to Kronecker products, and matricizations concepts of a tensor which are heavily utilized mainly in Chapter 5.

In Chapter 3, we discuss balancing-based model reduction for bilinear control systems. We first provide control-theoretic concepts for bilinear systems. Subsequently, we review the balanced truncation method for bilinear systems as it is widely studied in the literature, see, e.g. [26]. We recall the connection of energy functionals and Gramians of the bilinear systems. Afterwards, we introduce a concept of truncated Gramians of bilinear systems and investigate their interpretations with respect to energy functionals of the latter systems. We also discuss the advantages of the truncated Gramians from the model reduction perspective. Lastly, we illustrate the efficiency of the reduced-order systems, obtained using the truncated Gramians by means of a couple of numerical examples.

In Chapter 4, we focus on extending the balancing-type model reduction framework to quadratic-bilinear (QB) control systems. We begin by illustrating the process of quadratic-bilinearization, allowing us to rewrite smooth nonlinear systems into the QB form. We then derive the Gramians for the latter class of nonlinear systems, extending the Gramians of linear or bilinear systems. We provide connections of the energy functionals and the proposed Gramians of QB systems. We discuss the usage of the Gramians in balancing QB systems, allowing us to construct reduced-order systems. We illustrate the proposed balancing method by means of various semi-discretized nonlinear PDEs.

Chapter 5 is devoted to the interpolation-based model reduction for the latter class of nonlinear systems. Regarding this, we first present a system norm, namely the $\mathcal{H}_2$-norm based on the Volterra series of QB systems, and we then aim at extending the idea of $\mathcal{H}_2$-optimal model reduction from linear or bilinear systems to QB systems. Subsequently, we derive the $\mathcal{H}_2$-optimality conditions, allowing us to propose an iterative scheme to construct reduced-order systems which satisfy these derived optimality conditions approximately. We numerically illustrate the efficiency of the proposed method for various semi-discretized nonlinear PDEs and compare it with earlier proposed balanced truncation approach in Chapter 4 as well as existing methods such as proper orthogonal decomposition (POD), one-sided and two-sided interpolatory projection methods for QB systems [25, 78].

In Chapter 6, we briefly turn our attention towards descriptor systems which certainly increase the complexity of the reduction methods. For this, we study interpolation-based model reduction of bilinear descriptor systems, having special index-1 and index-2 structures. We, in particular, investigate $\mathcal{H}_2$-optimal approximations of such systems, allowing us to propose the modified bilinear iterative Krylov algorithm (B-IRKA) for such specially structured bilinear descriptor systems. Several numerical examples illustrate the efficiency of these proposed algorithms.

We finally conclude in Chapter 7 with our contributions and provide an overview of possible research topics which are worthwhile investigating in the future.

CHAPTER 2

MATHEMATICAL FOUNDATIONS

Contents

In this chapter, we collect essential mathematical fundamentals which are well-known in the literature and are used extensively throughout the thesis. In the first section, we present important concepts or results of classical linear control theory, which can be found in any standard textbook on linear systems, e.g., [87]. Then, we introduce the ideas of two important system-theoretic model reduction techniques which are the main focus of this thesis, namely balanced truncation and interpolation-based methods. Subsequently, we also discuss energy functionals and a concept of dual systems for a nonlinear system which are crucial tools in studying balancing-type model reduction. In the end, we present some basic tools from the tensor theory which can be found in, e.g., [68, 72, 90, 95]. These play a useful role, especially in investigating $\mathcal{H}_2$-optimal model reduction techniques for quadratic-bilinear systems.

2.1. Linear Systems Theory

This section provides the fundamental concepts for linear time-invariant (LTI) systems from the control theory and linear algebra points of view. All these concepts can be found almost in every textbook, e.g., see [87, 123]. Since in this thesis we focus on continuous time-invariant systems, we provide only details for the latter systems. For this, let us consider an LTI system (ODE) in the *state-space* representation which is of the form

$$\Sigma_L : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \\ y(t) = Cx(t) + Du(t), \end{cases} \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$. The vector $x(t) \in \mathbb{R}^n$ is the *state* of the system; x_0 is an initial condition; $u(t)$ and $y(t)$ denote control *inputs* and measurable *outputs*, respectively. The length of the state vector $x(t)$ is called the order of the system, i.e., the system (2.1) is of order n. If $m = p = 1$, the system is said to be single-input single-output (SISO) system, otherwise we refer to it as a multi-input multi-output (MIMO) system. Furthermore, we denote the system (2.1) by Σ_L. First, we discuss the stability concept of the dynamical system (2.1).

The dynamical system Σ_L is internally stable if $\sigma(A) \subset \mathbb{C}_- \cup \imath\mathbb{R}$ with no repeated eigenvalues over the imaginary axis; otherwise, it is referred to as an internally unstable system. Moreover, the dynamical system Σ_L is internally asymptotically stable if all the eigenvalues of the matrix A lie in the left open complex plane ($\mathbb{C}_-$). In this thesis, in this thesis, whenever a system is called as an asymptotically stable, it means internally asymptotic stability of the system.

For a given initial condition x_0 and the control input $u(t)$, the solution of the system (2.1) at any time instant t is characterized by

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau,$$

thus,

$$y(t) = Ce^{At}x_0 + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t).$$

Next, we define the so-called *impulse response* of the system, which is the system response of the input signal $u_k(t) = \delta(t)$, where $u_k(t)$ is the kth component of the input vector $u(t)$, and $\delta(t)$ is the Dirac delta distribution. Assuming a zero initial condition $x_0 = 0$, we obtain the impulse response as follows:

$$G_L(t) = Ce^{At}B + D\delta(t). \tag{2.2}$$

Subsequently, we discuss the concepts of controllability and observability of the system Σ_L. They play important roles in solving many problems arising in control theory, including the model order reduction problem. We begin by noting the controllability of an LTI system.

Definition 2.1 (e.g., [7]):
Consider a linear system Σ_L. The system is said to be controllable in the time interval $[t_0, t_f]$, if there exists an input function $u(t)$ of finite energy such that the system can be steered from any initial state $x(t_0)$ to any final state $x(t_f)$. ♢

The follow-up question is how to check weather the system Σ_L is controllable or not. This is clarified in the following proposition.

Proposition 2.2:
For a given dynamical system Σ_L, define the controllability matrix of Σ_L as

$$\mathcal{R}(A, B) = \left[B, AB, \ldots, A^{n-1}B\right]. \tag{2.3}$$

Then, the linear system Σ_L is controllable if and only if $\operatorname{rank}(\mathcal{R}(A, B)) = n$. ♢

The range of the controllability matrix is also known as the nth-order block Krylov subspace generated by the matrices A and B, i.e., $\operatorname{range}(\mathcal{R}(A, B)) = \mathcal{K}_n(A, B)$.

Furthermore, note that generally it is almost impossible to measure the full state in practice, especially in large-scale dynamical systems; we rather collect some measurement at selected points $y(t)$, which are given by linear combinations of state vectors. Thus, it would be interesting to study the observable subspace of Σ_L, and for this, in the following we define first the observability concept of the system Σ_L.

Definition 2.3:
Given a dynamical system Σ_L. The system is said to observable in $[t_0, t_f]$, if for a given input $u(t)$, the initial condition can be uniquely determined from the given output $y(t)$ or observations. ♢

Remark 2.4:
If a linear system is internally stable, then the system is also a bounded-input bounded-output stable. However, conversely, a bounded-input bounded-output stable system is internally stable as well if the system is controllable and observable. ♢

The concepts of controllability and observability are dual in nature; therefore, the observability of a system Σ_L can be checked by investigating the controllability of the pair (A^T, C^T). Similar to the controllable system, the system is completely observable if and only if $\mathcal{R}(A^T, C^T)$ is of the full rank. These concepts of controllability and observability of a system play a crucial role in model order reduction as we discuss in the next section.

The controllability and observability concepts can also be defined by a means of a certain type of matrix equations. For this, we define the controllability and observability Gramians of an asymptotically stable system Σ_L, respectively, as follows:

$$P = \int_0^\infty e^{At} B B^T e^{A^T t} dt \tag{2.4}$$

and

$$Q = \int_0^\infty e^{A^T t} C^T C e^{At} dt. \tag{2.5}$$

It is clear from the definitions of P and Q that they are symmetric and positive semi-definite matrices. Next, we present the relation between these Gramians and the solutions of the continuous-time algebraic Lyapunov equation [7, Section 4.3].

Proposition 2.5:
For a given asymptotically stable dynamical system Σ_L, let the controllability and observability Gramians be defined as in (2.4) and (2.5), respectively. Then, P and Q are the unique solutions of the Lyapunov equations

$$AP + PA^T + BB^T = 0, \tag{2.6a}$$

$$A^T Q + QA + C^T C = 0. \tag{2.6b}$$

$\diamondsuit$

These Gramians also give an interpretation whether the system is controllable and observable or not. If P and Q are positive definite matrices, i.e., $P > 0$ and $Q > 0$, then the system is completely controllable and observable. Furthermore, as we see in the subsequent section, these Gramians also play a crucial role in balancing-type model order reduction method.

Another important tool to analyze the characteristics of LTI systems is to establish the input-output relation in the *frequency domain*, instead of the state-space representation (2.1). This is obtained by utilizing the *Laplace transformation*, which is defined in the following.

Definition 2.6:
Consider a locally integrable function $f(t)$, which is defined for all $t \geq 0$. Then, the unilateral Laplace transformation $F(s)$ of $f(t)$ is defined as

$$F(s) := \mathcal{L}(f(t))(s) = \int_0^\infty e^{-st} f(t) dt,$$

where $s \in \mathbb{C}$, and the above integral exists. $\diamondsuit$

Performing the Laplace transformation to the system (2.1) yields

$$sX(s) - x_0 = AX(s) + BU(s), \tag{2.7a}$$

$$Y(s) = CX(s) + DU(s), \tag{2.7b}$$

where $X(s)$, $U(s)$, and $Y(s)$ are the Laplace transformations of $x(t)$, $u(t)$, and $y(t)$, respectively. Next, we determine the explicit expression for $X(s)$ using (2.7a) and substitute it in (2.7b), leading to an expression for $Y(s)$ in terms of the input and the initial condition:

$$Y(s) = \left(C\left(sI - A\right)^{-1} B + D\right) U(s) + C\left(sI - A\right)^{-1} x_0.$$

Assuming the zero initial condition, i.e., $x_0 = 0$, we obtain a direct relation between input and output

$$Y(s) = G_L(s)U(s),$$

where

$$G_L(s) = C\left(sI - A\right)^{-1} B + D.$$

The function $G_L(s)$, via which the input-output mapping is defined, is called the *transfer function* of Σ_L. The transfer function contains all the important information about the system dynamical behavior. Moreover, an advantage of the transfer function is also that it is independent of the input. Just as a remark, the transfer function $G_L(s)$ can also be derived by taking the Laplace transformation of the impulse response $G_L(t)$ of Σ_L.

In the SISO case, the transfer function $G_L(s)$ is a rational function of degree n, i.e., $G_L(s) = \frac{d(s)}{n(s)}$, where $d(s)$ and $n(s)$ are polynomial functions in the variable s. Moreover, the zeros of $n(s)$ are called the poles of Σ_L. Assuming the matrix A in Σ_L has all distinct eigenvalues, we can write the transfer function $G_L(s)$ in the *pole-residue formulation*, which is a very useful representation in particular while studying optimal model reduction problems. This formulation of $G_L(s)$ can be given by

$$G_L(s) = C(sI - A)^{-1}B + D = \sum_{j=1}^{n} \frac{r_j}{s - \lambda_j} + D, \tag{2.8}$$

where

$$r_j = \lim_{s \to \lambda_j} G_L(s)(s - \lambda_j),$$

and λ_j, $j \in \{1, \ldots, n\}$ are the eigenvalues of A. Next, we turn our attention to system norms. Of particular interest for this thesis, we focus on the $\mathcal{H}_2$-norm and $\mathcal{H}_\infty$-norm which are defined as follows for a linear system:

Definition 2.7:
Consider an asymptotically stable linear dynamical system (2.1) with $D = 0$ and assume $G_L(s)$ to be the transfer function of the system. Then, the $\mathcal{H}_2$-norm of the system is defined by

$$\|G_L\|_{\mathcal{H}_2} := \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \|G_L(i\omega)\|_F^2 d\omega\right)^{\frac{1}{2}},$$

and the $\mathcal{H}_\infty$-norm of the system is defined as

$$\|G_L\|_{\mathcal{H}_\infty} = \sup_{\omega\in\mathbb{R}} \sigma_{\max}(G_L(\imath\omega)),$$

where $\sigma_{\max}(\cdot)$ denotes the maximum singular value of a matrix. ◊

The above definition of the $\mathcal{H}_2$-norm of the system is rather theoretical; however, an alternative way to compute the $\mathcal{H}_2$-norm of a linear system is by using the system Gramians, which we summarize next.

Lemma 2.8 ([79]):
Consider the asymptotically stable linear system (2.1) with $D = 0$, and let P and Q be the controllability and observability Gramians as defined in (2.6a) and (2.6b), respectively. Then, the $\mathcal{H}_2$-norm can be computed as

$$\|G_L\|_{\mathcal{H}_2} = \sqrt{\operatorname{tr}\left(CPC^T\right)} = \sqrt{\operatorname{tr}\left(B^TQB\right)}. \tag{2.9}$$

$\Diamond$

Another important formula for the $\mathcal{H}_2$-norm can be given in terms of the pole-residue form of the transfer function (2.8). We outline this in the following lemma.

Lemma 2.9 ([79]):
For a given asymptotically stable SISO linear system with $D = 0$, let $\lambda_i \in \mathbb{C}_-$, $i \in \{1, \dots, n\}$, denote its simple poles. Then,

$$\|G_L\|_{\mathcal{H}_2} = \sum_{j=1}^{n} G_L(-\lambda_j) \operatorname{res}\left[G_L(s), \lambda_j\right]. \tag{2.10}$$

$\Diamond$

Having defined the system norms, one can measure the quality of a reduced-order system using these norms. As we will see later, depending on the model reduction method, we can construct an $\mathcal{H}_2$-optimal or $\mathcal{H}_\infty$-optimal reduced-order system.

All above discussed properties can easily be extended to linear systems, having the generalized state-space representation as follows:

$$E\dot{x}(t) = Ax(t) + Bu(t), \tag{2.11a}$$
$$y(t) = Cx(t) + Du(t), \tag{2.11b}$$

where the matrix E is nonsingular. However, when a linear system is subject to some constraints, then it leads to a linear descriptor system (DAEs). In such a case, the matrix E is singular, and the theoretic concepts become much more complicated. Nevertheless, if it is assumed that the matrix pencil $\alpha E - \beta A$ is regular, that is

$$\det(\alpha E - \beta A) \neq 0, \quad \text{for some} \quad (\alpha, \beta) \in \mathbb{C}^2,$$

then we can obtain the transfer function of (2.11) by taking the Laplace transform even when E is singular:

$$G_L^{(E)}(s) = C(sE - A)^{-1}B + D.$$

The transfer function $G_L^{(E)}(s)$ is called a proper transfer function if $\lim\limits_{s\to\infty} G_L^{(E)}(s) < \infty$ and strictly proper in case $\lim\limits_{s\to\infty} G_L^{(E)}(s) = 0$; otherwise, it is called an improper transfer function. Furthermore, when $G_L^{(E)}(s)$ is not proper, then one can additively decompose $G_L^{(E)}(s)$ into two parts as:

$$G_L^{(E)}(s) = G_{sp}(s) + P(s), \tag{2.12}$$

where $G_{sp}(s)$ and $P(s)$ are referred to as the strictly proper part and the polynomial part of $G_L^{(E)}(s)$. For more details, we refer to, e.g., [125].

2.2. Model Reduction via Projections

Having revised important control theoretic concepts of linear systems, in this section, we provide a short review of the important model order reduction (MOR) methods for linear systems which can be found, e.g., in [7, 8, 34, 62, 104, 115]. As already discussed, the purpose of MOR is to construct a reduced-order system, replicating the dynamical behavior of the large-scale dynamical system. More precisely, we want to replace the dynamical system (2.1) by a reduced-order system of the following form:

$$\widehat{\Sigma} : \begin{cases} \dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t) + \widehat{B}u(t), \\ \widehat{y}(t) = \widehat{C}\widehat{x}(t) + \widehat{D}u(t), \quad \widehat{x}(0) = 0, \end{cases} \tag{2.13}$$

where $\widehat{A} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$, $\widehat{B} \in \mathbb{R}^{\widehat{n}\times m}$, $\widehat{C} \in \mathbb{R}^{p\times\widehat{n}}$, and $\widehat{D} \in \mathbb{R}^{p\times m}$. Our goal is to ensure $\widehat{n} \ll n$ and that the error $\|y - \widehat{y}\|$ is small in an appropriate norm. Depending on the norm of the error, many methods have been proposed in the literature, which can be found in the aforementioned references.

To have the error $y - \widehat{y}$ as an output of a system, we define the so-called *error system* as follows:

$$\Sigma_l^{(e)} : \begin{cases} \dot{x}^{(e)}(t) &= A^{(e)}x^{(e)}(t) + B^{(e)}u(t), \\ y^{(e)}(t) &= C^{(e)}x^{e}(t) + D^{(e)}u(t), \qquad x^{(e)}(t) = 0, \end{cases} \tag{2.14}$$

where

$$A^{(e)} = \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix}, \quad B^{(e)} = \begin{bmatrix} B \\ \widehat{B} \end{bmatrix}, \quad C^{(e)} = \begin{bmatrix} C & -\widehat{C} \end{bmatrix}, \quad D^{(e)} = D - \widehat{D}.$$

It can be easily verified that the output of the error system (2.14) is the difference between the output of the original and reduced-order systems, i.e., $y^{(e)} = y - \widehat{y}$. Moreover, it can also be confirmed that the transfer function $G_L^{(e)}(s)$ of the error system $\Sigma^{(e)}$ is given by $G_L^{(e)}(s) := G_L(s) - \widehat{G}_L(s)$, where $G_L(s)$ and $\widehat{G}_L(s)$ are the transfer functions of the original and reduced-order systems.

Next, we focus on constructing reduced-order systems. In this thesis, we aim at obtaining such systems via a projection-type framework. In other words, the reduced-order systems are obtained by projecting the high-fidelity systems onto a lower dimensional subsystem. We begin by recalling the properties of the projection matrices which can be found e.g., in [18].

Definition 2.10:

- A matrix $\mathcal{P}$ is called a projection matrix or projection if $\mathcal{P}^2 = \mathcal{P}$.
- If $\operatorname{range}(\mathcal{P}) = \mathcal{V}$, then $\mathcal{P}$ is said to be the projector onto the subspace $\mathcal{V}$.
- If the projector $\mathcal{P}$ is symmetric, i.e., $\mathcal{P} = \mathcal{P}^T$, then $\mathcal{P}$ is an orthogonal projection *(Galerkin projection)*, else it is referred to as an oblique projector *(Petrov-Galerkin projection)*.
- Consider a matrix $Z \in \mathbb{R}^{n\times n}$ with eigenvalue spectrum $\Lambda(Z) = \Lambda_1 \cup \Lambda_2, \Lambda_1 \cap \Lambda_2 = \emptyset$. Moreover, assume that $\mathcal{V}_1$ is the right Z-invariant subspace corresponding to Λ_1. Then, a projector to the subspace $\mathcal{V}_1$ is referred to as a spectral projector. ♢

Next, we outline some valuable properties of projectors. We refer to [18, 112] for more details.

Lemma 2.11:
For a given projector $\mathcal{P} \in \mathbb{R}^{n\times n}$, the following assertions hold:

1. The matrix $I_n - \mathcal{P}$ is also a projection, called complementary projector.
2. $\ker \mathcal{P} = \operatorname{range}(I_n - \mathcal{P})$ and $\ker(I_n - \mathcal{P}) = \operatorname{range}(\mathcal{P})$.
3. If $\mathcal{P}$ is a projector onto $\mathcal{V}$, then $\mathcal{P}$ is the identity operator on $\mathcal{V}$, i.e., $\mathcal{P}v = v, \;$ for all $v \in \mathcal{V}$.
4. Consider an orthonormal basis matrix $V = [v_1, \ldots, v_{\widehat{n}}] \in \mathbb{R}^{n\times\widehat{n}}$, and let $\mathcal{V}$ be the subspace spanned by the columns of V, then $\mathcal{P} = VV^T$ is an orthogonal projector onto $\mathcal{V}$.
5. Furthermore, let $\mathcal{W} \in \mathbb{R}^{n\times\widehat{n}}$ be another subspace, spanned by the columns of the basis matrix $W = \left[w_1, \ldots, w_{\widehat{n}}\right]$, and assume that W^TV is invertible. Then, $\mathcal{P} = V(W^TV)^{-1}W^T$ is an oblique projector onto $\mathcal{V}$ along $\mathcal{W}$. ♢

Coming back to the MOR problem, the state vector $x(t)$ in the system (2.1) is approximating by an oblique projection $\mathcal{P} = V(W^TV)^{-1}W^T$, i.e., $x(t) \approx \mathcal{P}x(t)$, leading to

$$\mathcal{P}\dot{x}(t) = A\mathcal{P}x(t) + Bu(t) + \epsilon,$$

where ϵ can be seen as a residual after the approximation. Next, we impose a Petrov-Galerkin condition by choosing a subspace $\mathcal{W}$ which is orthogonal to the residual, i.e., $\epsilon \perp \mathcal{W}$, resulting in

$$(W^TV)^{-1}W^T\left(\mathcal{P}\dot{x}(t) - A\mathcal{P}x(t) - Bu(t)\right) = 0. \tag{2.15}$$

Defining a reduced state $\widehat{x} = (W^TV)^{-1}W^Tx$, Eq. (2.15) yields

$$\dot{\widehat{x}} = \widehat{A}\widehat{x}(t) + \widehat{B}u(t), \tag{2.16}$$

where $\widehat{A} = (W^TV)^{-1}W^TAV$ and $\widehat{B} = (W^TV)^{-1}W^T$. Furthermore, using the output equation in the system (2.1), we obtain an approximated output using the reduced state $\widehat{x}$ as

$$y \approx \widehat{y} = \underbrace{CV}_{\widehat{C}}\widehat{x}.$$

The next follow up question is a choice of subspaces $\mathcal{V}$ and $\mathcal{W}$, minimizing the error $\|y - \widehat{y}\|$. In the following, we present two prominent MOR methods based on different ideas that allow us to construct subspaces $\mathcal{V}$ and $\mathcal{W}$.

2.2.1. Balanced truncation for linear systems

Balancing-based model reduction is a well-known system-theoretic approach. The main idea of this method is to identify the states which are *important* and *less important* with respect to the input-output behavior of the system. A less important state can be defined as a state which is hard to control as well as hard to observe. In other words, a less important state requires a lot of *input energy* to reach and yet produces very little *output energy*. In the following, we provide the formal definitions of these energy functionals.

Definition 2.12 (e.g., [7]):
The *controllability energy functional* is defined as the minimum amount of energy required to steer the system from $x(-\infty) = 0$ to $x(0) = x_0$:

$$E_c(x_0) = \min_{\substack{u \in L_2(-\infty,0] \\ x(-\infty)=0,\ x(0)=x_0}} \frac{1}{2}\int_{-\infty}^{0} \|u(t)\|^2 dt.$$

The *observability energy functional* is defined as the energy generated by the nonzero initial condition $x(0) = x_0$ with zero control input:

$$E_o(x_0) = \frac{1}{2}\int_0^\infty \|y(t)\|^2 dt. \qquad \diamond$$

Coming back to the linear systems (2.1), these energy functions for a given state can be easily computed using the system Gramians P and Q as defined in (2.6a) and (2.6b). Precisely, assuming a controllable system, the minimum amount of input energy needed to reach x_0 from the zero initial condition is given by

$$E_c(x_0) = \frac{1}{2} x_0^T P^{-1} x_0.$$

Moreover, for an uncontrolled system ($u \equiv 0$), the energy observable with an initial condition x_0 is given by

$$E_o(x_0) = \frac{1}{2} x_0^T Q x_0.$$

As said, the objective is to find the states which are hard to control and hard to observe. For this, let us consider a linear system (2.1), which is a *balanced* one that is defined as follows:

Definition 2.13:
A linear system is called a balanced system if the system Gramians, namely controllability and observability Gramians, are equal and diagonal, i.e., $P = Q = \Sigma$, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ and $\sigma_i \geq \sigma_{i+1}$. Moreover, the σ_i's are called the *Hankel singular values* (HSV) of the system (2.1). ♢

The balanced system immediately suggests that states, requiring a lot of input energy to reach, yield very little output energy. Therefore, they are less important for the input-output behavior. Hence, they can be eliminated, thus resulting in a reduced-order system. However, it is rather unlikely that a given linear system is a balanced system. Therefore, the first step in balanced truncation MOR is to transform a given linear system into a balanced one via an appropriate balancing transformation T. This can be achieved by, e.g., the *square root balancing* method proposed in [67]. Assuming $P > 0$, $Q > 0$, we determine the Cholesky factor of P and Q, i.e, $P =: S^T S$ and $Q =: R^T R$. Then, the transformation matrix T is given by $T = D^{-\frac{1}{2}} Z^T R$ and its inverse $T^{-1} = S^T U D^{-\frac{1}{2}}$, where $UDZ^T := SR^T$. Let us consider a balanced linear system (2.1), which is partitioned as follows:

$$\begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_1^T(t) & x_2^T(t) \end{bmatrix}^T, \end{aligned} \tag{2.17}$$

where $x_1(t) \in \mathbb{R}^{\widehat{n}}, x_2(t) \in \mathbb{R}^{n-\widehat{n}}$, and all other matrices are of appropriate dimensions. By setting $x_2(t) \equiv 0$, we obtain a reduced-order system $\widehat{\Sigma}_l = (A_{11}, B_1, C_1)$. Moreover, it can be shown that the reduced-order system $\widehat{\Sigma}_l$ is a balanced system also, i.e., $\widehat{P} = \widehat{Q} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\widehat{n}})$, where $\widehat{P}$ and $\widehat{Q}$ are the Gramians of a reduced-order system after truncation [7, 103]. In the Petrov-Galerkin formulation, the projection $\mathcal{P} = VW^T$, where

$$V = S^T U_{\widehat{n}} D_{\widehat{n}}^{-\frac{1}{2}}, \quad W = \left(D_{\widehat{n}}^{-\frac{1}{2}} Z_{\widehat{n}}^T R \right)^T,$$

and $D_{\widehat{n}} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\widehat{n}})$, and $U_{\widehat{n}}$ $Z_{\widehat{n}}$ denote, respectively, the first $\widehat{n}$ columns of matrices U and Z. Furthermore, the error between the original system Σ_l and the reduced-order system $\widehat{\Sigma}_l$ can be quantified as follows:

$$\|G_L - \widehat{G}_L\|_{\mathcal{H}_\infty} \leq 2 \sum_{k=\widehat{n}+1}^{n} \sigma_k,$$

where $G_L(s)$ and $\widehat{G}_L(s)$ are the transfer functions of the original and reduced-order systems. Thus, the above error bounds allows us to construct a reduced-order system of the appropriate order, satisfying the desired *tolerance*.

2.2.2. Interpolation-based model reduction

Here, we present the fundamental idea of interpolation-based MOR. For a linear SISO system, its transfer function $G_L(s)$ is a rational function of degree n, where n is the order of the system. So, the interpolation-based model reduction seeks to determine another rational function, but of smaller degree, which interpolates the original system at the predefined interpolation points within the complex plane. Precisely, we aim at constructing a transfer function $\widehat{G}_L(s)$ such that

$$G_L(\lambda_i) = \widehat{G}_L(\lambda_i), \quad i \in \{1, \ldots, \widehat{n}\},$$

where $\lambda_i \in \mathbb{C}$ are interpolation points. The next task is to construct an interpolating reduced-order system. For this, the projector $\mathcal{P}$ can be constructed by solving a certain type of shifted linear systems, see, e.g., [77, 129].

Theorem 2.14:
Consider a linear SISO system, and let $\{\sigma_i\}_{i=1}^{\widehat{n}} \in \mathbb{C}$ and $\{\mu_i\}_{i=1}^{\widehat{n}} \in \mathbb{C}$ be two sets of interpolation points such that $\sigma_i, \mu_i \notin \sigma(A)$. Assume the projection matrices V and W are constructed as follows:

$$\operatorname{range}(V) = \operatorname{span}\left((\sigma_1 I - A)^{-1}B, \ldots, (\sigma_{\widehat{n}} I - A)^{-1}B\right), \tag{2.18a}$$

$$\operatorname{range}(W) = \operatorname{span}\left((\mu_1 I - A)^{-T}C^T, \ldots, (\mu_{\widehat{n}} I - A)^{-1}C^T\right). \tag{2.18b}$$

Assuming W^TV is invertible, if a reduced-order system is constructed as

$$\widehat{A} = (W^TV)^{-1}WAV, \quad \widehat{B} = (W^TV)^{-1}WB, \quad \widehat{C} = CV,$$

then

$$G_L(\sigma_i) = \widehat{G}_L(\sigma_i), \qquad G_L(\mu_i) = \widehat{G}_L(\mu_i).$$

Moreover, if $\sigma_i = \mu_i$, then

$$G_L(\sigma_i) = \widehat{G}_L(\sigma_i), \qquad G_l'(\sigma_i) = \widehat{G}_l'(\sigma_i). \qquad \Diamond$$

The above interpolation conditions can also be achieved for MIMO systems; however, the numbers of columns of the projection matrices V and W might increase very fast, depending on the numbers of control inputs and outputs. In order to keep a small number of columns in the projection matrices, the idea of *tangential interpolation* can be employed, see, e.g., [65]. For this, let us consider left and right tangential directions $\widetilde{b}_i \in \mathbb{R}^m$, $i \in \{1, \dots, \widehat{n}\}$ and $\widetilde{c}_i \in \mathbb{R}^p$, $i \in \{1, \dots, \widehat{n}\}$ along with interpolation points $\sigma_i \in \mathbb{C}$, $\mu_i \in \mathbb{C}$, $i \in \{1, \dots, \widehat{n}\}$. Then, our goal is to derive a reduced-order system, satisfying

$$\begin{aligned} G_L(\sigma_i)\widetilde{b}_i &= \widehat{G}_L(\sigma_i)\widetilde{b}_i, \quad i \in \{1, \dots, \widehat{n}\}, \\ \widetilde{c}_i^T G_L(\mu_i) &= \widetilde{c}_i^T \widehat{G}_L(\mu_i), \quad i \in \{1, \dots, \widehat{n}\}. \end{aligned} \tag{2.19}$$

The tangential interpolation conditions can be achieved if the projection matrices V and W are computed as follows:

$$\operatorname{range}(V) = \operatorname{span}\left((\sigma_1 I - A)^{-1} B\widetilde{b}_1, \dots, (\sigma_{\widehat{n}} I - A)^{-1} B\widetilde{b}_{\widehat{n}}\right), \tag{2.20a}$$

$$\operatorname{range}(W) = \operatorname{span}\left((\mu_1 I - A)^{-T} C^T\widetilde{c}_1, \dots, (\mu_{\widehat{n}} I - A)^{-1} C^T\widetilde{c}_{\widehat{n}}\right). \tag{2.20b}$$

Clearly, the quality of a reduced-order system highly depends on the choice of interpolation points and its tangential directions. Several methods are proposed in the literature, showing how to choose a good set of interpolation points and tangential directions, see, e.g., [7, 43, 66, 54, 56, 65, 77, 79, 129, 44].

Remark 2.15:
Note that the interpolation point λ_i and tangential directions $\widetilde{b}_i$ and $\widetilde{c}_i$ should be closed under conjugation. This means that if the interpolation point λ_j is a complex number, then $\bar{\lambda}_j$ should also be an interpolation points. Moreover, if $\widetilde{b}_j$ and $\widetilde{c}_j$ are the tangential directions corresponding to the interpolation point λ_j, then $\bar{\widetilde{b}}_j$ and $\bar{\widetilde{c}}_j$ are the tangential directions corresponding to the interpolation point $\bar{\lambda}_j$. $\Diamond$

In this thesis, we particularly focus on $\mathcal{H}_2$-optimal model reduction problems. In the following, we define the $\mathcal{H}_2$ model reduction problem for a linear system.

Definition 2.16:
Given a transfer function $G_L(s)$, a reduced-order system $\widehat{G}_L$ of order $\widehat{n}$ is said to the $\mathcal{H}_2$-optimal if it satisfies

$$\|G_L - \widehat{G}_L\|_{\mathcal{H}_2} = \min_{\substack{\dim(\widetilde{G}_L) = \widehat{n} \\ \widetilde{G}_L \text{ stable}}} \|G_L - \widetilde{G}_L\|_{\mathcal{H}_2}. \qquad \Diamond$$

The above problem for a SISO linear system was initially considered in [101]. Therein, it was shown that $\widehat{G}_L$ is a locally $\mathcal{H}_2$-optimal reduced-order system if it interpolates the original transfer function $G_L(s)$ and its derivative $G'_L(s)$ at the poles of the reduced order system reflected across the imaginary axis. In other words, the optimality conditions for a SISO system are :

$$G_L(-\lambda_i) = \widehat{G}_L(-\lambda_i), \qquad G'_l(-\lambda_i) = \widehat{G}'_l(-\lambda_i), \qquad i \in \{1, \dots, \widehat{n}\},$$

where λ_i are the simple poles of the reduced-order system. This problem was later considered in [79] for MIMO linear systems. The transfer function of a MIMO system is a matrix-value rational function, and generally tangential interpolation of the transfer function is enforced. Using this information, a locally $\mathcal{H}_2$-optimal reduced-order system fulfills

$$\widetilde{c}_i^T G_L(-\lambda_i) = \widetilde{c}_i^T \widehat{G}_L(-\lambda_i), \qquad i \in \{1, \dots, \widehat{n}\}, \tag{2.21a}$$
$$G_L(-\lambda_i)\widetilde{b}_i = \widehat{G}_L(-\lambda_i)\widetilde{b}_i, \qquad i \in \{1, \dots, \widehat{n}\}, \tag{2.21b}$$
$$\widetilde{c}_i^T G'_l(-\lambda_i)\widetilde{b}_i = \widetilde{c}_i^T \widehat{G}'_l(-\lambda_i)\widetilde{b}_i, \qquad i \in \{1, \dots, \widehat{n}\}, \tag{2.21c}$$

where $\operatorname{diag}(\lambda_1, \dots, \lambda_{\widehat{n}}) = R^{-1}\widehat{A}R$, $\left[\widetilde{b}_1, \dots, \widetilde{b}_{\widehat{n}}\right] = \widehat{B}^T R^{-T}$ and $\left[\widetilde{c}_1, \dots, \widetilde{c}_{\widehat{n}}\right] = \widehat{C}R$.

We have already discussed in this section how to construct an interpolating reduced-order system for a given set of interpolation points and tangential directions. Moreover, we know that in order to determine an $\mathcal{H}_2$-optimal reduced-order system, we need to interpolate at the mirror images of the poles of the reduced-order system across the imaginary axis, using the tangential directions which are determined by spectral decomposition of the reduced-order system. However, the problem we encounter now is that we do not know a priori these interpolation points and tangential directions since these qualities require the reduced-order system which we want to construct. To overcome this issue, the authors in [79] have proposed a fixed point iterative scheme, the so-called *iterative rational Krylov algorithm* (*IRKA*), see Algorithm 2.1, which upon convergence yields a reduced-order system, satisfying (2.21) at a modest cost.

2.3. Energy Functionals and Adjoint Systems for Nonlinear Systems

Here, we discuss two concepts, which are important while extending balanced-type model reduction from linear systems to special classes of nonlinear systems. We begin with recapitulating energy functionals for nonlinear systems and their relations with partial differential equations.

Algorithm 2.1: Iterative rational Krylov algorithm (IRKA) [79].

```
Input: $A, B, C$ and $tol$.
Make an initial guess of the interpolation points $\{\lambda_i\}_{i=1}^{\widehat{n}}$ and the tangential
  directions $\{\widetilde{b}_i\}_{i=1}^{\widehat{n}}$ and $\{\widetilde{c}_i\}_{i=1}^{\widehat{n}}$.
while relative change in $\{\lambda_i\} > tol$ do
    Choose $V$ and $W$ such that
        $\text{range}\,(V) = \text{span}\left((\lambda_1 I - A)^{-1}B\widetilde{b}_1, \dots, (\lambda_{\widehat{n}} I - A)^{-1}B\widetilde{b}_{\widehat{n}}\right)$,
        $\text{range}\,(W) = \text{span}\left((\lambda_1 I - A)^{-T}C^T\widetilde{c}_1, \dots, (\lambda_{\widehat{n}} I - A)^{-T}C^T\widetilde{c}_{\widehat{n}}\right)$.
    Compute reduced-order system matrices:
        $\widehat{A} = (W^TV)^{-1}W^TAV$, $\widehat{B} = (W^TV)^{-1}W^TB$, $\widehat{C} = CV$.
    Determine the spectral decomposition of $\widehat{A} =: R\Lambda R^{-1}$, and assign
     $\lambda_i \leftarrow -\Lambda_i$, where $\Lambda_i$ is the $i$th diagonal entry of $\Lambda$.
    Compute directions $\left[\widetilde{b}_1, \dots, \widetilde{b}_{\widehat{n}}\right] \leftarrow R^{-1}\widehat{B}$ and $\left[\widetilde{c}_1, \dots, \widetilde{c}_{\widehat{n}}\right] \leftarrow \widehat{C}R$.
return $\widehat{A}, \widehat{B}, \widehat{C}$.
```

2.3.1. Relation of energy functionals with partial differential equations

From Subsection 2.2.1, we now know that energy functionals, namely controllability and observability energy functionals, of a system are the main ingredients in balancing-type model order reduction. We thus first discuss these energy functionals for nonlinear systems. For this, let us consider in the following sufficiently smooth, for example, C^∞, nonlinear asymptotically stable input-affine nonlinear system of the form

$$\dot{x}(t) = f(x) + g(x)u(t), \tag{2.22a}$$

$$y(t) = h(x), \qquad x(0) = 0, \tag{2.22b}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ are the state, input and output vectors of the system, respectively, and also $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are smooth nonlinear functions. Without loss of generality, we assume that 0 is a stable equilibrium of the system (2.22). The controllability and observability energy functionals for a general nonlinear system have been studied in the literature; see, e.g., [74, 114]. In the following, we state the definitions of controllability and observability energy functionals for the system (2.22).

Definition 2.17 ([74, 114]):
The controllability energy functional is defined as the minimum amount of energy

required to steer the system from $x(-\infty) = 0$ to $x(0) = x_0$:

$$E_c(x_0) = \min_{\substack{u \in L_2(-\infty,0] \\ x(-\infty)=0,\ x(0)=x_0}} \frac{1}{2}\int_{-\infty}^{0} \|u(t)\|^2 dt. \qquad \diamondsuit$$

As can be seen, the controllability energy functional for nonlinear systems has a similar definition as we have in the case of linear systems. Furthermore, initially, the observability energy functionals are also defined analogous to the linear case as follows:

Definition 2.18 ([114]):
The observability energy functional for the system (2.22) can be defined as the energy generated by the nonzero initial condition $x(0) = x_0$ with zero control input:

$$E_o(x_0) = \frac{1}{2}\int_{0}^{\infty} \|y(t)\|^2 dt. \qquad \diamondsuit$$

However, while proposing the above definition for the observability energy functional, the author in [114] has assumed the nonlinear system (2.22) is zero-state observable. This means that if $u(t) = 0$ and $y(t) = 0$ for $t \geq 0$, then $x(t) = 0\ \forall t \geq 0$. As discussed in [74], for a nonlinear system such a condition can be very strong. As a result, therein, it is shown how this condition can be relaxed in the context of general input balancing, and a new definition for the observability functionals is provided as follows:

Definition 2.19 ([74]):
The observability energy functional is defined as the energy generated by a nonzero initial condition $x(0) = x_0$ and by applying an L_2-bounded input:

$$E_o(x_0) = \max_{\substack{u \in L_2(0,\infty), \|u\|_{L_2} \leq \alpha \\ x(0)=x_0, x(\infty)=0}} \frac{1}{2}\int_{0}^{\infty} \|y(t)\|^2 dt. \qquad \diamondsuit$$

In an abstract way, the main idea of introducing Definition 2.19 is to find the state component that contributes least from a state-to-output point of view for all possible L_2-bounded inputs. The connections between these energy functionals and the solutions of the partial differential equations are established in [74, 114], which are outlined in the following theorem.

Theorem 2.20 ([74, 114]):
Consider the nonlinear system (2.22), having $x = 0$ as an asymptotically stable equilibrium in a neighborhood W_o of 0. Then, for all $x \in W_o$, the observability energy functional $E_o(x)$ can be determined by the following partial differential equation:

$$\frac{\partial E_o}{\partial x} f(x) + \frac{1}{2} h^T(x) h(x) - \frac{1}{2}\mu^{-1}\frac{\partial E_o}{\partial x} g(x) g(x)^T \frac{\partial^T E_o}{\partial x} = 0, \quad E_o(0) = -\frac{1}{2}\mu, \qquad (2.23)$$

assuming that there exists a smooth solution $\bar{E}_o$ on W_o, and 0 is an asymptotically stable equilibrium of $\bar{f} := (f - \mu^{-1} g g^T \frac{\partial^T \bar{E}_o}{\partial x})$ on W_o with the negative real number $\mu := -\|g^T(\phi)\frac{\partial^T \bar{E}_o}{\partial x}(\phi)\|_{L_2}$, and $\dot{\phi} = \bar{f}(\phi)$ with $\phi(0) = x$. Moreover, for all $x \in W_c$, the controllability energy functional $E_c(x)$ is a unique smooth solution of the following Hamilton-Jacobi equation:

$$\frac{\partial E_c}{\partial x} f(x) + f(x) \frac{\partial E_c}{\partial x} + \frac{\partial E_c}{\partial x} g(x) g^T(x) \frac{\partial^T E_c}{\partial x} = 0, \quad E_c(0) = 0 \tag{2.24}$$

under the assumption that (2.24) has a smooth solution $\bar{E}_c$ on W_c, and 0 is an asymptotically stable equilibrium of $-\left(f(x) + g(x) g^T(x) \frac{\partial \bar{E}_c(x)}{\partial x}^T\right)$ on W_c. ◊

Note that in Definition 2.19, the zero-state condition is relaxed by considering an input that is only L_2-bounded. However, an alternative way to relax the zero-state observable condition is by considering an input which is not only L_2-bounded but also L_∞-bounded.

We thus propose a new definition of the observability energy functional as follows:

Definition 2.21:
The observability energy functional can be defined as the energy generated by the nonzero initial condition $x(0) = x_0$ and by applying an L_2-bounded and L_∞-bounded input:

$$E_o(x_0) = \max_{\substack{u \in \mathcal{B}_{(\alpha,\beta)} \\ x(0)=x_0, x(\infty)=0}} \frac{1}{2} \int_0^\infty \|y(t)\|^2 dt,$$

where $\mathcal{B}_{\alpha,\beta} \overset{\text{def}}{=} \{u \in L_2^m[0,\infty), \|u\|_{L_2} \leq \alpha, \|u\|_{L_\infty} \leq \beta\}$. ◊

In this thesis, we use the above definition to determine the observability energy functional for quadratic-bilinear control systems in Chapter 4.

2.3.2. Hilbert adjoint operator for nonlinear systems

The importance of the adjoint operator (dual system) can be seen, particularly, in the computation of the observability energy functional or Gramian. For general nonlinear systems, a duality between controllability and observability energy functionals is shown in [64] with the help of state-space realizations for nonlinear adjoint operators. In what follows, we briefly outline the state-space realizations for nonlinear adjoint operators of nonlinear systems. For this, we consider a nonlinear system of the form

$$\Sigma_{NL} := \begin{cases} \dot{x}(t) = \mathcal{A}(x,u,t)x(t) + \mathcal{B}(x,u,t)u(t), \\ y(t) = \mathcal{C}(x,u,t)x(t) + \mathcal{D}(x,u,t)u(t), \qquad x(0) = 0 \end{cases} \tag{2.25}$$

in which $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ are the state, input and output vectors of the system, respectively, and $\mathcal{A}(x,u,t)$, $\mathcal{B}(x,u,t)$, $\mathcal{C}(x,u,t)$ and $\mathcal{D}(x,u,t)$ are appropriately sized matrices. Also, we assume that the origin is a stable equilibrium of the system. The Hilbert adjoint operators for the general nonlinear systems have been investigated in [64]. Therein, they also discussed the connection between the state-space realization of the adjoint operators and the port-control Hamiltonian systems, leading to the state-space characterization of the nonlinear Hilbert adjoint operators of $\Sigma_{NL} : L_2^m[0,\infty) \to L_2^p[0,\infty)$. In the following lemma, we summarize the state-space realization of the Hilbert adjoint operator of the nonlinear system.

Lemma 2.22 ([64]):
Consider the system (2.25) with the initial condition $x(0) = 0$, and assume that the input-output mapping $u \to y$ is denoted by the operator $\Sigma_{NL} : L_2^m[0,\infty) \to L_2^p[0,\infty)$. Then, the state-space realization of the nonlinear Hilbert adjoint operator $\Sigma^* : L_2^{m+p}[0,\infty) \to L_2^m[0,\infty)$ is given by

$$\Sigma_{NL}^*(u_d, u) := \begin{cases} \dot{x}(t) = \mathcal{A}(x,u,t)x(t) + \mathcal{B}(x,u,t)u(t), & x(0) = 0, \\ \dot{x_d}(t) = -\mathcal{A}^T(x,u,t)x_d(t) - \mathcal{C}^T(x,u,t)u_d(t), & x_d(\infty) = 0, \\ y_d(t) = \mathcal{B}^T(x,u,t)x_d(t) + \mathcal{D}^T(x,u,t)u_d(t), & \end{cases} \tag{2.26}$$

where $x_d \in \mathbb{R}^n$, $u_d \in \mathbb{R}^p$ and $y_d \in \mathbb{R}^m$ can be interpreted as the dual state, dual input and dual output vectors of the system, respectively. ◊

We will see the importance of dual systems in determining the observability energy functional or observability Gramian for bilinear systems and quadratic-bilinear systems mainly in Chapter 4.

2.4. Tensor Theory Concepts

In this section, we review some basic concepts from tensor algebra which can be found e.g., in [68, 72, 90, 95] We begin by defining the Kronecker product of matrices and vectorization of matrices.

Definition 2.23 ([68]):
Consider two matrices $X = \begin{bmatrix} x_1, \ldots, x_m \end{bmatrix} \in \mathbb{R}^{n \times m}$ and $Y \in \mathbb{R}^{p \times q}$. Then, $\operatorname{vec}(X)$ is determined by stacking the columns of X on top of each other, i.e.,

$$\operatorname{vec}(X) = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}. \tag{2.27}$$

The Kronecker product of two matrices X and Y is determined by

$$X \otimes Y = \begin{bmatrix} x_{11}Y & \cdots & x_{1m}Y \\ \vdots & & \vdots \\ x_{n1}Y & \cdots & x_{nm}Y \end{bmatrix}, \tag{2.28}$$

where x_{ij} is the (i, j)th entry of the matrix X. ◊

Next, we note the following important properties of the $\operatorname{vec}(\cdot)$ operator and the Kronecker product.

Proposition 2.24:
Let $X \in \mathbb{R}^{n\times m}$, $Y \in \mathbb{R}^{m\times q}$, $Z \in \mathbb{R}^{q\times r}$, $T \in \mathbb{R}^{r\times s}$. Then,

$$\operatorname{tr}(XY) = \left(\operatorname{vec}\left(X^T\right)\right)^T \operatorname{vec}(Y) = (\operatorname{vec}(Y))^T \operatorname{vec}(X), \tag{2.29a}$$

$$\operatorname{vec}(XYZ) = (Z^T \otimes X)\operatorname{vec}(Y), \tag{2.29b}$$

$$(X \otimes Z)(Y \otimes T) = (XY \otimes ZT). \tag{2.29c}$$

◊

Now we turn our attention towards tensors. Tensors are a natural extension of the concept of a matrix and consist of several matrices. Vectors and matrices can also be interpreted as tensors of order 1 and 2, respectively. However, in the following we define a tensor of a general order ℓ.

Definition 2.25:
A K-order tensor $\mathcal{X} \in \mathbb{R}^{n_1\times\cdots\times n_K}$ is a K-dimensional array of entries $\mathcal{X}_{i_1,\ldots,i_K} \in \mathbb{R}$, where $i_j \in \{1, \ldots n_j\}$ and $j \in \{1, \ldots, K\}$. ◊

Next, we review the concept of matricization of a tensor. Similar to how rows and columns are defined for a matrix, one can define a fiber of $\mathcal{X}$ by fixing all indices except for one, e.g., $\mathcal{X}(:, i, j), \mathcal{X}(i, :, j)$ and $\mathcal{X}(i, j, :)$. The mathematical operations, involving tensors, are easier to perform using its corresponding matrix representations. Therefore, there exists a very well-known process of unfolding a tensor into a matrix, called *matricization* of a tensor.

Definition 2.26 (e.g., [95]):
By $X^{(k)}$, we denote the matrix that is obtained by unfolding the K-order tensor $\mathcal{X} \in \mathbb{R}^{I_1\times I_2\times\cdots\times I_K}$ along the *kth* dimension, $k \in 1, 2, \ldots, K$. This *k-matricization* is formally obtained via the mapping of the tensor indices $(i_1, i_2, \ldots, i_K)$ onto the matrix indices (i_k, j) via

$$j = 1 + \sum_{l=1, l\neq k}^{K} (i_l - 1)J_l, \quad \text{where} \quad J_l := \prod_{m=1, m\neq l}^{l-1} I_m.$$

In this thesis, we only focus on 3-order tensors, i.e. $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, and show the tensor properties using the 3-order tensor. However, these properties can be extended to a general K-order tensor. For a 3-order tensor, there are 3 different ways to unfold it, depending on its mode-μ fibers that are used for the unfolding. In the following example, we illustrate how a 3rd order tensor $\mathcal{X} \in \mathbb{R}^{n \times n \times n}$ can be unfolded into different matrices.

Example 2.1:
Consider a 3rd order tensor $\mathcal{X}^{n \times n \times n}$ whose frontal slices are given by matrices $X_i \in \mathbb{R}^{n \times n}$ as shown in Figure 2.1.

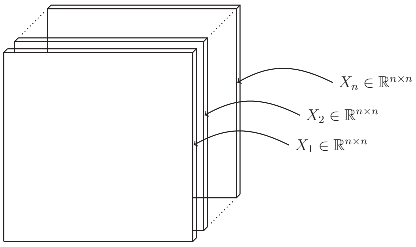

$X_n \in \mathbb{R}^{n \times n}$
$X_2 \in \mathbb{R}^{n \times n}$
$X_1 \in \mathbb{R}^{n \times n}$

Figure 2.1.: Representation of a tensor using frontal slices [95].

Then, its mode-μ matricizations, $\mu \in \{1, 2, 3\}$, are given by

$$\mathcal{X}^{(1)} = [X_1, X_2, \ldots, X_n], \quad \mathcal{X}^{(2)} = [X_1^T, X_2^T, \ldots, X_n^T], \text{ and}$$
$$\mathcal{X}^{(3)} = [\text{vec}\,(X_1)\,, \text{vec}\,(X_2)\,, \ldots, \text{vec}\,(X_n)]^T. \qquad \Diamond$$

Similar to the matrix-matrix product, one can also perform a tensor-matrix or tensor-tensor multiplication. Of particular interest for this thesis are tensor-matrix multiplications, which can be performed by means of matricizations; e.g., see [95]. For a given tensor $\mathcal{X} \in \mathbb{R}^{n \times n \times n}$ and a matrix $A \in \mathbb{R}^{n_1 \times n}$, the μ-mode matrix product is denoted by $\mathcal{X} \times_\mu A =: \mathcal{Y}$, i.e., $\mathcal{Y} \in \mathbb{R}^{n_1 \times n \times n}$ for $\mu = 1$. In the case of the μ-mode matrix multiplication, the mode-μ fiber is multiplied with the matrix A, which can be written as

$$\mathcal{Y} = \mathcal{X} \times_\mu A \Leftrightarrow \mathcal{Y}^{(\mu)} = A\mathcal{X}^{(\mu)}.$$

Furthermore, if a tensor is given as

$$\mathcal{Z} = \mathcal{X} \times_1 A \times_2 B \times_3 C, \tag{2.30}$$

where $A \in \mathbb{R}^{n_1 \times n}$, $B \in \mathbb{R}^{n_2 \times n}$ and $C \in \mathbb{R}^{n_3 \times n}$, then the mode-$\mu$ matriciziations of $\mathcal{Z}$ satisfy:

$$\mathcal{Z}^{(1)} = A\mathcal{X}^{(1)}(C \otimes B)^T, \ \mathcal{Z}^{(2)} = B\mathcal{X}^{(2)}(C \otimes A)^T, \ \mathcal{Z}^{(3)} = C\mathcal{X}^{(3)}(B \otimes A)^T. \tag{2.31}$$

Using these properties of the tensor products, we now introduce our first result on tensor matricizations.

Lemma 2.27:
Consider tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n\times n\times n}$ and let $\mathcal{X}^{(i)}$ and $\mathcal{Y}^{(i)}$ denote, respectively, their mode-i matricizations. Then,

$$\operatorname{tr}\left(\mathcal{X}^{(1)}(\mathcal{Y}^{(1)})^T\right) = \operatorname{tr}\left(\mathcal{X}^{(2)}(\mathcal{Y}^{(2)})^T\right) = \operatorname{tr}\left(\mathcal{X}^{(3)}(\mathcal{Y}^{(3)})^T\right). \qquad \Diamond$$

Proof. We begin by denoting the ith frontal slice of $\mathcal{X}$ and $\mathcal{Y}$ by X_i and Y_i, respectively; see Figure 2.1. Thus,

$$\begin{aligned}\operatorname{tr}\left(\mathcal{X}^{(1)}(\mathcal{Y}^{(1)})^T\right) &= \operatorname{tr}\left(\left[X_1, X_2, \ldots, X_n\right]\left[Y_1, Y_2, \ldots, Y_n\right]^T\right)\\ &= \sum_{i=1}^{n} \operatorname{tr}\left(X_i Y_i^T\right) = \sum_{i=1}^{n} \operatorname{tr}\left(X_i^T Y_i\right)\\ &= \operatorname{tr}\left(\left[X_1^T, X_2^T, \ldots, X_n^T\right]\left[Y_1^T, Y_2^T, \ldots, Y_n^T\right]^T\right) = \operatorname{tr}\left(\mathcal{X}^{(2)}(\mathcal{Y}^{(2)})^T\right).\end{aligned}$$

Furthermore, since $\operatorname{tr}\left(X^T Y\right) = \operatorname{vec}(X)^T \operatorname{vec}(Y)$, this allows us to write

$$\operatorname{tr}\left(\mathcal{X}^{(1)}(\mathcal{Y}^{(1)})^T\right) = \sum_{i=1}^{n} \operatorname{tr}\left(X_i^T Y_i\right) = \sum_{i=1}^{n} \operatorname{vec}(X_i)^T \operatorname{vec}(Y_i).$$

Since the ith rows of $\mathcal{X}^{(3)}$ and $\mathcal{Y}^{(3)}$ are given by $\operatorname{vec}(X_i)^T$ and $\operatorname{vec}(Y_i)^T$, respectively, it holds that $\sum_{i=1}^{n} \operatorname{vec}(X_i)^T \operatorname{vec}(Y_i) = \operatorname{tr}\left(\mathcal{X}^{(3)}(\mathcal{Y}^{(3)})^T\right)$. This concludes the proof. □

Next, we consider a tensor $\mathcal{H} \in \mathbb{R}^{n\times n\times n}$. Also, we assume $\mathcal{H}$ to be symmetric. This means that for given vectors u and v,

$$\mathcal{H}^{(1)}(u \otimes v) = \mathcal{H}^{(1)}(v \otimes u), \tag{2.32}$$

where $\mathcal{H}^{(1)}$ is the mode-1 matricization of $\mathcal{H}$. This condition provides the additional information that the other two matricization modes of $\mathcal{H}$ are the same, i.e.,

$$\mathcal{H}^{(2)} = \mathcal{H}^{(3)}. \tag{2.33}$$

The additional property that the Hessian is symmetric will allow us to derive some new relationships between matricizations and matrices, that will prove to be crucial ingredients in simplifying the expressions arising in the derivation of optimality conditions in Chapter 5.

Lemma 2.28:
Let $\mathcal{H} \in \mathbb{R}^{n\times n\times n}$ be a 3-order tensor, satisfying (2.32) and (2.33), and consider matrices $\mathcal{A}, \mathcal{B}, \mathcal{C} \in \mathbb{R}^{n\times n}$. Then,

$$\mathcal{H}^{(1)}(\mathcal{B} \otimes \mathcal{C})\left(\mathcal{H}^{(1)}\right)^T = \mathcal{H}^{(1)}(\mathcal{C} \otimes \mathcal{B})\left(\mathcal{H}^{(1)}\right)^T, \quad \text{and} \tag{2.34}$$

$$\begin{aligned}(\operatorname{vec}(\mathcal{B}))^T \operatorname{vec}\left(\mathcal{H}^{(2)}(\mathcal{C} \otimes \mathcal{A})(\mathcal{H}^{(2)})^T\right) &= (\operatorname{vec}(\mathcal{C}))^T \operatorname{vec}\left(\mathcal{H}^{(2)}(\mathcal{B} \otimes \mathcal{A})(\mathcal{H}^{(2)})^T\right)\\ &= (\operatorname{vec}(\mathcal{A}))^T \operatorname{vec}\left(\mathcal{H}^{(1)}(\mathcal{C} \otimes \mathcal{B})(\mathcal{H}^{(1)})^T\right). \qquad \Diamond\end{aligned}$$

Proof. We begin by proving the relation in (2.34). The order in the Kronecker product can be changed via pre- and post-multiplication of appropriate permutation matrices; see [86, Sec. 3]. Thus,

$$\mathcal{B} \otimes \mathcal{C} = S(\mathcal{C} \otimes \mathcal{B})S^T,$$

where S is the permutation matrix $S = \sum_{i=1}^{n}((e_i^n)^T \otimes I_n \otimes e_i^n)$. We can then write

$$\mathcal{H}^{(1)}(\mathcal{B} \otimes \mathcal{C}) \left(\mathcal{H}^{(1)}\right)^T = \mathcal{H}^{(1)} S(\mathcal{C} \otimes \mathcal{B}) \left(\mathcal{H}^{(1)} S\right)^T. \tag{2.35}$$

We now manipulate the term $\mathcal{H}^{(1)}S$:

$$\mathcal{H}^{(1)} S = \sum_{i=1}^{n} \mathcal{H}^{(1)}((e_i^n)^T \otimes I_n \otimes e_i^n). \tag{2.36}$$

Furthermore, we can write I_n as the Kronecker product

$$I_n = \sum_{j=1}^{n} (e_j^n)^T \otimes e_j^n, \tag{2.37}$$

and since we know that for vectors $f, g \in \mathbb{R}^q$, $f^T \otimes g = g f^T$, we can write (2.37) in another form as

$$I_n = \sum_{j=1}^{n} e_j^n (e_j^n)^T. \tag{2.38}$$

Substituting these relations in (2.36) leads to

$$\begin{aligned}
\mathcal{H}^{(1)} S &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{H}^{(1)}((e_i^n)^T \otimes (e_j^n)^T \otimes e_j^n \otimes e_i^n) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{H}^{(1)} \left(e_j^n \otimes e_i^n\right) \left((e_i^n)^T \otimes (e_j^n)^T\right) && (\because \text{for } f \in \mathbb{R}^q, f^T \otimes f = f f^T) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{H}^{(1)}(e_i^n \otimes e_j^n)((e_i^n)^T \otimes (e_j^n)^T). && (\because \text{the relation } (2.32))
\end{aligned} \tag{2.39}$$

Next, we use the Kronecker multiplication property in (2.39). Thus, we obtain

$$\begin{aligned}
\mathcal{H}^{(1)} S &= \mathcal{H}^{(1)} \left(\sum_{i=1}^{n} e_i^n (e_i^n)^T \otimes \sum_{j=1}^{n} e_j^n (e_j^n)^T \right) \\
&= \mathcal{H}^{(1)}(I_n \otimes I_n) = \mathcal{H}^{(1)}. && \text{(from (2.38))}
\end{aligned}$$

Substituting the above relation in (2.35) proves (2.34). For the second part, we utilize the trace property (2.29a) to obtain

$$(\text{vec}\,(\mathcal{B}))^T \,\text{vec}\left(\mathcal{H}^{(2)}(\mathcal{C}\otimes\mathcal{A})(\mathcal{H}^{(2)})^T\right) = \text{tr}\left(\underbrace{\mathcal{B}^T\mathcal{H}^{(2)}(\mathcal{C}\otimes\mathcal{A})}_{=\mathcal{L}^{(2)}}(\mathcal{H}^{(2)})^T\right),$$

where $\mathcal{L}^{(2)} \in \mathbb{R}^{n\times n^2}$ can be considered as a mode-2 matricization of a tensor $\mathcal{L}^{n\times n\times n}$. Using Lemma 2.27 and the relations (2.31), we obtain

$$\begin{aligned}\text{tr}\left(\mathcal{L}^{(2)}(\mathcal{H}^{(2)})^T\right) &= \text{tr}\left(\mathcal{L}^{(3)}(\mathcal{H}^{(3)})^T\right) = \text{tr}\left(\mathcal{C}^T\mathcal{H}^{(3)}(\mathcal{B}\otimes\mathcal{A})(\mathcal{H}^{(3)})^T\right)\\ &= \text{tr}\left(\mathcal{C}^T\mathcal{H}^{(2)}(\mathcal{B}\otimes\mathcal{A})(\mathcal{H}^{(2)})^T\right) \qquad \text{(using (2.33))}\\ &= (\text{vec}\,(\mathcal{C}))^T\,\text{vec}\left(\mathcal{H}^{(2)}(\mathcal{B}\otimes\mathcal{A})(\mathcal{H}^{(2)})^T\right).\end{aligned}$$

Furthermore, we also have

$$\begin{aligned}\text{tr}\left(\mathcal{L}^{(2)}(\mathcal{H}^{(2)})^T\right) &= \text{tr}\left(\mathcal{L}^{(1)}(\mathcal{H}^{(1)})^T\right) = \text{tr}\left(\mathcal{A}^T\mathcal{H}^{(1)}(\mathcal{C}\otimes\mathcal{B})(\mathcal{H}^{(1)})^T\right)\\ &= (\text{vec}\,(\mathcal{A}))^T\,\text{vec}\left(\mathcal{H}^{(1)}(\mathcal{C}\otimes\mathcal{B})(\mathcal{H}^{(1)})^T\right),\end{aligned}$$

which completes the proof. □

Next, we prove the connection of a certain permutation matrix to the Kronecker product.

Lemma 2.29:
Consider matrices $X, Y \in \mathbb{R}^{n\times m}$. Define the permutation matrix $T_{(n,m)} \in \{0,1\}^{n^2m^2\times n^2m^2}$ as

$$T_{(n,m)} = I_m \otimes \left[I_m\otimes e_1^n, \ldots, I_m\otimes e_n^n\right]\otimes I_n. \tag{2.40}$$

Then,

$$\text{vec}\,(X\otimes Y) = T_{(n,m)}\,(\text{vec}\,(X)\otimes\text{vec}\,(Y)). \qquad \Diamond$$

Proof. Let us denote the ith columns of X and Y by x_i and y_i, respectively. We can then write

$$\text{vec}\,(X\otimes Y) = \begin{bmatrix}\text{vec}\,(x_1\otimes Y)\\ \vdots\\ \text{vec}\,(x_m\otimes Y)\end{bmatrix}. \tag{2.41}$$

Now we concentrate on the ith block row of $\text{vec}\,(X\otimes Y)$, which, using (2.29b) and (2.29c), can be written as

$$\begin{aligned}\text{vec}\,(x_i\otimes Y) &= \text{vec}\,((x_i\otimes I_n)Y) = (I_m\otimes x_i\otimes I_n)\,\text{vec}\,(Y)\\ &= \left(I_m\otimes\left[x_i^{(1)}e_1^n+\cdots+x_i^{(n)}e_n^n\right]\otimes I_n\right)\text{vec}\,(Y),\end{aligned} \tag{2.42}$$

where $x_i^{(j)}$ is the (j,i)th entry of the matrix X. An alternative way to write (2.42) is

$$\begin{aligned}\operatorname{vec}(x_i \otimes Y) &= [I_m \otimes e_1^n \otimes I_n, \dots, I_m \otimes e_n^n \otimes I_n]\,(x_i \otimes I_{nm})\operatorname{vec}(Y)\\ &= ([I_m \otimes e_1^n, \dots, I_m \otimes e_n^n] \otimes I_n)\,(x_i \otimes \operatorname{vec}(Y)).\end{aligned}$$

This yields

$$\begin{aligned}\operatorname{vec}(X \otimes Y) &= \begin{bmatrix}([I_m \otimes e_1^n, \dots, I_m \otimes e_n^n] \otimes I_n)\,(x_1 \otimes \operatorname{vec}(Y))\\ \vdots \\ ([I_m \otimes e_1^n, \dots, I_m \otimes e_n^n] \otimes I_n)\,(x_m \otimes \operatorname{vec}(Y))\end{bmatrix}\\ &= (I_m \otimes [I_m \otimes e_1^n, \dots, I_m \otimes e_n^n] \otimes I_n)\begin{bmatrix}x_1 \otimes \operatorname{vec}(Y)\\ \vdots \\ x_m \otimes \operatorname{vec}(Y)\end{bmatrix}\\ &= (I_m \otimes [I_m \otimes e_1^n, \dots, I_m \otimes e_n^n] \otimes I_n)\left(\begin{bmatrix}x_1\\ \vdots\\ x_m\end{bmatrix} \otimes \operatorname{vec}(Y)\right)\\ &= (I_m \otimes [I_m \otimes e_1^n, \dots, I_m \otimes e_n^n] \otimes I_n)\,(\operatorname{vec}(X) \otimes \operatorname{vec}(Y)),\end{aligned}$$

which proves the assertion. □

CHAPTER 3

BALANCING-TYPE MODEL REDUCTION METHODS FOR BILINEAR CONTROL SYSTEMS

Contents

3.1. Introduction

In this chapter, we study balancing-type model order reduction (MOR) of a special class of nonlinear systems. In Chapter 2, we have briefly outlined the model reduction via balanced truncation for systems that are linear in the state and in the control input, leading to linear time-invariant (LTI) systems of the form (2.1). As a first step in the direction of nonlinear systems, we discuss the class of *bilinear systems*, which are independently linear in the state and in the input, but not jointly. These systems are considered to be a potential bridge between fully nonlinear systems and linear systems. More precisely, these systems are of the form

$$\Sigma_B : \begin{cases} \dot{x}(t) = Ax(t) + \displaystyle\sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \\ y(t) = Cx(t), \quad x(0) = x_0, \end{cases} \tag{3.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the state, input and output vectors of the system at time t, respectively, and $u_k(t)$ is the kth component of the input $u(t)$. The numbers m and p represent the quantities of inputs and outputs. All system matrices are of appropriate dimensions. One can also think of a mass matrix E in front of $\dot{x}(t)$, but to keep the discussion simple, we consider the matrix E to be the identity. Also, we consider a fixed initial condition x_0 of the system. However, without loss of generality, we assume a zero initial condition, i.e., $x_0 = 0$. In case $x_0 \neq 0$, one can transform the system by defining new appropriate state variables as $\widetilde{x}(t) = x(t) - x_0$, ensuring the zero initial condition of the transformed system, e.g., see [15].

Applications of bilinear systems can be found in various fields such as nuclear fusion, mechanical brakes or biological species [41, 102, 111]. Further, the applicability of the systems (3.1) in MOR for stochastic control problems is studied in [26, 83, 107] and for MOR of a certain class of linear parametric systems in [20]. Moreover, nonlinear systems can be approximated by bilinear systems via Carleman bilinearization [70, 111]. Our goal is to construct another low-dimensional bilinear system, having the same structure as (3.1):

$$\widehat{\Sigma}_B : \begin{cases} \dot{\widehat{x}}(t) &= \widehat{A}\widehat{x}(t) + \sum\limits_{k=1}^{m} \widehat{N}_k \widehat{x}(t) u_k(t) + \widehat{B}u(t), \\ \widehat{y}(t) &= \widehat{C}\widehat{x}(t), \quad \widehat{x}(0) = 0, \end{cases} \tag{3.2}$$

where $\widehat{A}, \widehat{N}_k \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$, $\widehat{B} \in \mathbb{R}^{\widehat{n}\times m}$ and $\widehat{C} \in \mathbb{R}^{p\times\widehat{n}}$ with $\widehat{n} \ll n$, ensuring $y \approx \widehat{y}$ for all admissible inputs $u \in L_2^m[0,\infty)$.

Several methods for linear systems have been extended to bilinear systems such as balanced truncation [5, 26] and interpolation-based MOR [12, 21, 40, 59, 49, 57]. In this chapter, we only focus on balancing-based MOR for bilinear systems and we will discuss interpolation-based MOR for bilinear systems in detail later in Chapter 6. As noted in Chapter 2, the concept of balanced truncation relies on energy functionals, namely controllability and observability energy functionals, of the system. These energy functionals allow us to find those states that are hard to control as well as hard to observe; thus, they are the less important for the input-output behavior of the system. Truncating such states leads to reduced-order systems. This problem for bilinear systems was first discussed in [91], and later on, it was taken up in [5, 6, 26, 49]. For this, algebraic Gramians for bilinear systems were proposed in [5], which allow us to identify less important states with respect to the input/output mapping, although the physical interpretation of these Gramians is not as clear as in the linear case. The connections between these Gramians and the energy functionals of bilinear systems have been studied in [26, 73] for state vectors that are multiples of the canonical unit vectors. Furthermore, the relations between the Gramians and the controllability/observability of bilinear systems have also been studied in [26].

It is worth mentioning that the balancing concept for general nonlinear systems has been studied in a series of papers, see, e.g., [63, 75, 114], where a new notion of controllability and observability energy functionals has been introduced. Although theoretically, the balancing concept for nonlinear systems is appealing, it is seldom applicable in the large-scale setting from the computational perspective. This is due to the fact that the energy functionals are solutions of nonlinear Hamilton-Jacobi equations, which are extremely expensive to solve for large-scale systems, see Section 2.3.

Coming back to the algebraic Gramians of bilinear systems, the main bottleneck in using these Gramians in the MOR context is their computational cost as we are required to solve a couple of generalized Lyapunov equations. Though recently there have been many advances in methods to determine the low-rank solutions of these generalized Lyapunov equations; see [24, 118]. This motivates us to investigate an alternative pair of Gramians for bilinear systems, which we call *truncated* Gramians (TGrams) that are computationally cheaper to compute.

The structure of the chapter is as follows. In the subsequent section, we provide the necessary control theoretic concepts of bilinear systems. This includes the Volterra series representation of bilinear systems, possible stability criteria and derivation of Gramians of bilinear systems. In Section 3.3, we study the connection between these Gramians and energy functionals for an arbitrary state vector, in contrast to [26], where the connection is shown only for the canonical unit vectors and their multiples. Next, we discuss how to construct reduced-order systems using the Gramians for bilinear systems. In Section 3.4, we propose TGrams for bilinear systems and investigate their connections with the controllability and observability of the bilinear systems. Moreover, we reveal the relation between the TGrams and energy functionals of the bilinear systems. Then, we discuss the advantages of considering the TGrams in the MOR context. Subsequently, in Section 3.5, we provide a couple of numerical examples to illustrate the applicability of the TGrams for MOR of bilinear systems and advantages of them over standard Gramians of bilinear systems.

3.2. Control Theoretic Concepts

Here, we gather some basic control theoretic concepts of bilinear systems, which will be useful in the rest of this chapter and in Chapter 6. These concepts and ideas can be found in several textbooks on bilinear system theory, e.g., see [55, 93, 102, 111]. We begin with the Volterra series for bilinear systems, relating the output $y(t)$ of Σ_B to the input of the system. For this, we first derive the expression for the state $x(t)$ in (3.1) in terms of the input and system matrices. Assuming bounded and continuous inputs $u_k(t)$ on a time interval $[0, T]$ and utilizing the Picard-Lindelöf theorem [111, Section

3.1], one can show that there exists $y(t)$ on $[0,T]$, satisfying

$$\begin{aligned} y(t) = \sum_{k=1}^{\infty} \int_0^t \int_0^{\sigma_1} \cdots \int_0^{\sigma_{k-1}} g_k(t,\sigma_1,\ldots,\sigma_k)\Big(u(t-\sigma_1-\cdots-\sigma_k) \otimes \cdots \\ \otimes u(t-\sigma_1) d\sigma_k \cdots d\sigma_1\Big), \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} g_k(t,\sigma_1,\ldots,\sigma_k) = Ce^{At_k}\bar{N}\left(I_m \otimes e^{At_{k-1}}\bar{N}\right) \cdots \\ \times \left(I_{m^{k-2}} \otimes e^{At_2}\bar{N}\right)\left(I_{m^{k-1}} \otimes e^{At_1}B\right). \end{aligned} \tag{3.4}$$

in which $\bar{N} = [N_1,\ldots,N_m]$. The quantities $g_k(t,\sigma_1,\ldots,\sigma_k)$ are also referred to as the *kernels* of the Volterra series.

The above form of the Volterra series has kernels in the *regular form*, e.g., [111]. This, in particular, helps us in deriving the regular transfer functions of a bilinear system. Note that there also exist the kernels of the Volterra series, e.g., in triangular form, see, e.g., [111]. However, the advantage of considering the regular form of the kernels is that all variables are separable. As we know from the linear case that the impulse response or kernel has a one-to-one relation to the transfer function of the system, we thus perform the multi-variate Laplace transform of the degree-k kernels [111], leading to the kth multi-variate transfer function of a bilinear system in the regular form as

$$G_k(s_1,\ldots,s_k) = C\left(\prod_{i=0}^{k-2} I_{m^i} \otimes (s_{k-i}I - A)^{-1}\bar{N}\right)\left(I_{m^{k-1}} \otimes (s_1 I - A)^{-1}B\right). \tag{3.5}$$

Consequently, in an abstract way, the output of a bilinear system in the frequency domain can be characterized by means of these multi-variate transfer functions, extending the transfer function concept of linear systems. However, the real meaning of the variables $s_1,\ldots,s_k$ in the multi-variate transfer functions of a bilinear system is still not very well understood.

Remark 3.1:
We note that the expression of the kth order transfer function for a single-input single-output (SISO) bilinear system is much simpler and is given by

$$G_k(s_1,\ldots,s_k) = C(s_i I - A)^{-1}N \cdots (s_2 I - A)^{-1}N(s_1 I - A)^{-1}B, \tag{3.6}$$

where $N := N_1$. ◊

Next, we review bounded-input-bounded-output stability of a bilinear system. This concept has been investigated in [122] in detail, and we outline the main result in the following theorem.

Theorem 3.2 ([122]):
Consider a bilinear system Σ_B and assume that the matrix A is Hurwitz. This means that there exist scalars $\beta > 0$ and $0 < \alpha \leq -\max_i(\mathrm{Re}\,(\lambda_i(A)))$, satisfying

$$\|e^{At}\| \leq \beta e^{-\alpha t} \quad \forall t \geq 0. \tag{3.7}$$

Furthermore, let M be such that $\|u\| = \sqrt{\sum_{k=1}^{m} |u_k(t)|^2} \leq M$ for all $t \geq 0$, and Γ be defined as $\Gamma := \sum_{k=1}^{m} \|N_k\|$. Then, the output of Σ_B is bounded on $[0, \infty)$ for inputs u if $\Gamma < \dfrac{\alpha}{M\beta}$. ◊

As a next step, we introduce the system matrices, the so-called Gramians for bilinear systems. This concept was first discussed in [50], and later on, in [5] the concept of controllability and observability are generalized based on the kernels of the Volterra series. For this, let us first define

$$\begin{aligned} P_1(t_1) &= e^{At_1}B, \\ P_k(t_1, \ldots, t_k) &= e^{At_k}\left[N_1 P_{k-1}, \ldots, N_m P_{k-1}\right], \quad k = 2, 3, \ldots, \end{aligned} \tag{3.8}$$

which are nothing but the kernels of the Volterra series (3.3). Then, based on (3.8), we define the reachability Gramian P as

$$P = \sum_{k=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} P_k(t_1, \ldots, t_k) P_k(t_1, \ldots, t_k)^T dt_1 \cdots dt_k. \tag{3.9}$$

Similarly, let us define

$$\begin{aligned} Q_1(t_1) &= e^{A^T t_1}C, \\ Q_k(t_1, \ldots, t_k) &= e^{A^T t_k}\left[N_1^T Q_{k-1}, \ldots N_m^T Q_{k-1}\right], \quad k = 2, 3, \ldots, \end{aligned} \tag{3.10}$$

and the observability Gramian Q as

$$Q = \sum_{k=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} Q_k(t_1, \ldots, t_k) Q_k(t_1, \ldots, t_k)^T dt_1 \cdots dt_k. \tag{3.11}$$

Clearly, P and Q are symmetric and positive semi-definite. However, the expressions for P and Q as shown in (3.9) and (3.11), respectively, involve infinite terms. Hence, they may diverge; thus, P and Q may not exist. Therefore, the following theorem, we provide sufficient conditions under which the Gramians for bilinear systems exist, or in other words, the infinite series converge.

Theorem 3.3 ([133]):
Assuming the series (3.9) and (3.11) converge, the reachability Gramian P and the observability Gramian Q of a bilinear system are given by the series (3.9) and (3.11), respectively. These series converge if

(a) the matrix A is Hurwitz, and

(b) $\Gamma_N < \frac{\sqrt{2\alpha}}{\beta}$, where $\Gamma_N = \sqrt{\| \sum_{k=1}^{m} N_k N_k^T \|}$, and α, β are defined in (3.7). ◊

Under these assumptions, in the following, we show the connection between the Gramians P and Q of a bilinear system and generalized linear matrix equations, extending the well-known Lyapunov equations for linear systems.

Theorem 3.4 ([5, 133]):
Assuming the Gramians P and Q of the bilinear system (3.1) exist, they, respectively, solve the following generalized Lyapunov equations

$$AP + PA^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0, \tag{3.12a}$$

$$A^TQ + QA + \sum_{k=1}^{m} N_k^T Q N_k + C^TC = 0. \tag{3.12b}$$

◊

Remark 3.5:
In Theorem 3.3, we have noted the conditions under which the series defining the Gramians of a bilinear system converge, and thus, they solve the generalized Lyapunov equations as indicated in Theorem 3.4. However, as shown in [133], the solutions of the generalized Lyapunov equations may exist even though their corresponding series may diverge. ◊

Before we proceed further, we define the balanced realization of a bilinear system.

Definition 3.6:
Analog to the linear case, a bilinear system Σ_B is said to be balanced if its Gramians $P = Q = \Sigma$ solve

$$A\Sigma + \Sigma A^T + \sum_{k=1}^{m} N_k \Sigma N_k^T + BB^T = 0,$$

$$A^T\Sigma + \Sigma A + \sum_{k=1}^{m} N_k^T \Sigma N_k + C^TC = 0,$$

where $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ and $\sigma_i \geq \sigma_{i+1} \geq 0$. ◊

Remark 3.7:
In general, a bilinear system may not be balanced. However, there exists a balanced transformation $x \mapsto T^{-1}x$, leading to a transformed bilinear system, whose reachability and observability Gramians are equal and diagonal, i.e.,

$$T^{-1}PT^{-T} = T^TQT = \Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_n). \tag{3.13}$$

Analogous to the linear case (see, e.g., [7]), having the factorizations of $P = LL^T$ and $L^TQL = U\Sigma^2U^T$, one finds the corresponding transformation matrix $T = LU\Sigma^{-\frac{1}{2}}$.◊

Now, we turn our attention towards defining a system norm for a bilinear system Σ_B, which help us to measure the quality of a reduced-order system. In this thesis, we are also particularly interested in the $\mathcal{H}_2$-optimal reduced-order system. Thus, we define the $\mathcal{H}_2$-norm for bilinear system. Recall from linear systems that the $\mathcal{H}_2$-norm of a linear system can be determined in terms of the impulse response of the system or kernel of the time-evolution equation. Extending this idea to bilinear systems, one can define the $\mathcal{H}_2$-norm of a bilinear system Σ_B as follows.

Definition 3.8 ([60]):

Let Σ_B be a bounded-input bounded-output (BIBO) stable bilinear system. Then, the $\mathcal{H}_2$-norm of Σ_B is defined by

$$\|\Sigma_B\|_{\mathcal{H}_2} = \sqrt{\operatorname{tr}\left(\sum_{k=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty}\|g_k(t_1,\ldots,t_k)\|_2^2 dt_1\cdots dt_k\right)}, \tag{3.14}$$

where $g_k(t_1,\ldots,t_k)$ is as given in (3.4). ◊

Note that Definition 3.8 makes sense in the case when the series in (3.14) convergence, or equivalently, when the generalized reachability and observability Gramians exist. Furthermore, the connection between the system Gramians, namely reachability and observability Gramians, and the $\mathcal{H}_2$-norm of Σ_B is derived in [133], which is given by

$$\|\Sigma_B\|_{\mathcal{H}_2} = \sqrt{\operatorname{tr}(CPC^T)} = \sqrt{\operatorname{tr}(B^TQB)}, \tag{3.15}$$

where P and Q are the reachability and observability Gramians and are solutions of the generalized Lyapunov equations as shown in Theorem 3.4.

Before we conclude the section, we look at an alternative characterization of the kth order transfer function (3.5) in the pole-residue formulation. This will be very useful while dealing with bilinear descriptor systems in Chapter 6. Analogous to the linear case, the kth order multi-variate transfer function can be written in the aforementioned formulation as follows.

Proposition 3.9 ([60]):

Consider the multi-variate transfer function

$$G_k(s_1,s_2,\ldots,s_k) = C(s_kI-A)^{-1}N\cdots(s_2I-A)^{-1}N(s_1I-A)^{-1}B$$

of a SISO bilinear system and let $\{\lambda_1,\lambda_2,\ldots,\lambda_n\} \subset \mathbb{C}$ be the n distinct eigenvalues of the matrix A. Then, the multi-variate transfer function can also be written in the pole-residue formulation as follows:

$$G_k(s_1,s_2,\ldots,s_k) = \sum_{l_1=1}^{n}\sum_{l_2=1}^{n}\cdots\sum_{l_k=1}^{n}\frac{\phi_{l_1,\ldots,l_k}}{\prod_{i=1}^{k}(s_i-\lambda_{l_i})},$$

where

$$\phi_{l_1,\ldots,l_k} = \lim_{s_k \to \lambda_{l_k}} (s_k - \lambda_{l_k}) \lim_{s_{k-1} \to \lambda_{l_{k-1}}} (s_{k-1} - \lambda_{l_{k-1}}) \cdots \lim_{s_1 \to \lambda_{l_1}} (s_1 - \lambda_{l_1}) G_k(s_1, \ldots, s_k). \tag{3.16}$$

◊

This pole-residue decomposition of the transfer functions can also be utilized to derive another expression for the $\mathcal{H}_2$-norm of a bilinear system. We summarize in the following theorem.

Theorem 3.10 ([22, 60]):
Let Σ_B be a SISO BIBO-stable bilinear system. Assuming the $\mathcal{H}_2$-norm of Σ_B exists, then it can be given in the pole-residue formulation as follows:

$$\|\Sigma_B\|_{\mathcal{H}_2} = \sum_{l_1=1}^{n} \sum_{l_2=1}^{n} \cdots \sum_{l_k=1}^{n} \phi_{l_1,\ldots,l_k} G_k(-\lambda_{l_1}, \ldots, -\lambda_{l_k}), \tag{3.17}$$

where $\phi_{l_1,\ldots,l_k}$ are as defined in (3.16). ◊

The pole-residue formulation of multi-variate transfer functions and the $\mathcal{H}_2$-norm expression as shown in Theorem 3.10 will play an important role while extending the existing model reduction techniques for bilinear ODEs to bilinear descriptor systems. These are discussed in detail in Chapter 6.

3.3. Standard Balanced Truncation Technique for Bilinear Systems

In this section, we discuss how the Gramians of a bilinear system are related to energy functionals and then show how to remove states less important for the input-output behavior, leading to a reduced-order system. We begin by comparing the controllability and observability energy functionals, denoted by E_c and E_o, respectively, with certain quadratic forms, given in terms of the Gramians P and Q. As we have seen in the linear case, the energy functionals E_c and E_o can be exactly determined in terms of the Gramians of a linear system. This is no longer possible when it comes to nonlinear systems including bilinear systems. As a remedy, it is desirable to ensure bounds on the energy functionals in terms of algebraic Gramians of the systems, allowing us still to find the states that are hard to control and hard to observe, at least locally.

This problem initially was investigated in [48, 73], where the expressions for the gradients of E_c and E_o, in the neighborhood of the origin have been derived. These are:

$$\nabla E_c(x) = \widetilde{P}(x)^{-1} x \qquad \text{and} \qquad \nabla E_o(x) = \widetilde{Q}(x) x, \tag{3.18}$$

where $\widetilde{P}(x)$ and $\widetilde{Q}(x)$ solve

$$
\begin{aligned}
A\widetilde{P}(x) + \widetilde{P}(x)A^T &= -\sum_{k=1}^{m}(N_k x + b_j)(N_k x + b_j)^T,\\
A^T\widetilde{Q}(x) + \widetilde{Q}(x)A &= -\sum_{k=1}^{m}\widetilde{Q}(x)N_k x x^T N_k^T \widetilde{Q}(x) - C^T C.
\end{aligned}
$$

Based on it, the following bounds for energy functions are derived, assuming $P > 0$ and one of the N_k is of full rank:

$$
E_c(x_0) > x_0^T P^{-1} x_0 \qquad \text{and} \qquad E_o(x_0) < x_0^T Q x_0. \tag{3.19}
$$

However, going from (3.18) to (3.19) requires an additional integrability condition, which may not hold in general. This issue is discussed in detail in [26, 127], showing that this is a key obstacle and where examples explaining the problem are provided.

Coming back to the relation between energy functionals and Gramians, the authors in [26] have provided another interpretation of these Gramians, which is outlined in the following theorem.

Theorem 3.11 ([26]):
Let a bilinear system Σ_B be a balanced system, i.e., $P = Q = \Sigma > 0$. Then, there exists $\epsilon > 0$ such that the following bounds hold for all canonical unit vectors e_i:

$$
E_c(\epsilon e_i) > \epsilon^2 e_i^T P^{-1} e_i = \epsilon^2/\sigma_i \tag{3.20}
$$

and

$$
E_o(\epsilon e_i) < \epsilon^2 e_i^T Q e_i = \epsilon^2 \sigma_i, \tag{3.21}
$$

where σ_i is the ith diagonal entry of Σ. ◊

Clearly, it can be seen that bounds for the energy functionals bounds shown in Theorem 3.11 are very restrictive since they only hold for canonical unit vectors or its multiples. Therefore, in the following, we provide a result, showing under what conditions bounds for the energy functionals of bilinear systems hold for an arbitrary given state. Before we state the corresponding theorem, we introduce a *homogeneous* bilinear system, which is used to characterize the observability energy in the system:

$$
\begin{aligned}
\dot{x}(t) &= Ax(t) + \sum_{k=1}^{m} N_k x(t) u_k(t),\\
y(t) &= Cx(t), \quad x(0) = x_0.
\end{aligned} \tag{3.22}
$$

Theorem 3.12:
Consider the bilinear system (3.1), with a stable A, and assume that the system is asymptotically any state x_0 reachable from 0. Let P, $Q > 0$ be the Gramians of the

system, respectively, and $E_c(x)$ and $E_o(x)$, respectively, be the controllability and observability energy functionals of the bilinear system. Then, there exists a small neighborhood W of 0, where the following relation holds:

$$E_c(x_0) \geq \frac{1}{2} x_0^T P^{-1} x_0 \quad \text{if} \quad x_0 \in W(0). \tag{3.23}$$

Furthermore, for a sufficiently small input $u(t)$ in the homogeneous bilinear system (3.22), the following relation holds:

$$E_o(x_0, u) \leq \frac{1}{2} x_0^T Q x_0. \tag{3.24}$$

$\diamondsuit$

Proof. To prove (3.23), we follow the lines of reasoning in [26]. Let us assume that $x_0 \in \mathbb{R}^n$ is controlled by the input $u = u_{x_0} \in L_2^m[0, \infty)$, minimizing the input cost functional in the definition of $E_c(x_0)$. Using this input, we consider the homogeneous linear differential equation given by

$$\dot{\phi} = \left(A + \sum_{k=1}^{m} N_k u_k(t) \right) \phi =: A_u(t)\phi(t),$$

whose fundamental solution is denoted by Φ_u. Thus, if we consider the time-varying system

$$\dot{x}(t) = A_u(t)x + Bu(t), \tag{3.25}$$

then its reachability Gramian [120, 128] is given by

$$P_u = \int_{-\infty}^{0} \Phi_u(0, \tau) B B^T \Phi_u(0, \tau)^T d\tau.$$

Obviously, the input u also steers the time-varying system (3.25) from 0 to x_0. Moreover, to steer the system (3.25), the minimum required input energy is equal to $x_0^T P_u^{\#} x_0$, where $P_u^{\#}$ denotes the Moore-Penrose pseudo inverse of P_u. This implies that $\|u\|_{L_2}^2$ is larger than the $x_0^T P_u^{\#} x_0$, i.e.,

$$\|u\|_{L_2}^2 \geq x_0^T P_u^{\#} x_0,$$

Alternatively, one can also determine P_u as an observability Gramian:

$$P_u = \int_0^{\infty} \Psi_u(t, 0) B B^T \Psi_u(t, 0)^T dt,$$

where Ψ_u is the fundamental solution of the dual system, satisfying

$$\dot{\Psi}_u = \left(A^T + \sum_{k=1}^{m} N_k^T u_k(-t) \right) \Psi_u, \quad \Psi_u(t, t) = I.$$

Now, we define $\widetilde{x}(t) = \Psi_u(t,0)x_0$. Further, since the system is assumed to be reachable, there exists an input $u \in L_2^m[0,\infty)$, which can steer the system from 0 to x_0. This implies that $\widetilde{x}(t) \to 0$ as $t \to \infty$. Thus, we have

$$\begin{aligned}
x_0^T P x_0 &= -\int_0^\infty \frac{d}{dt}\left(\widetilde{x}(t)^T P \widetilde{x}(t)\right) dt \\
&= -\int_0^\infty \widetilde{x}(t)^T \left(AP + \sum_{k=1}^m N_k P u_k(-t) + PA^T + \sum_{k=1}^m P N_k^T u_k(-t)\right)\widetilde{x}(t)dt \\
&= -\int_0^\infty \widetilde{x}(t)^T \left(AP + PA^T + \sum_{k=1}^m N_k P N_k^T\right)\widetilde{x}(t)dt \\
&\quad + \sum_{k=1}^m \int_0^\infty \widetilde{x}(t)^T \left(N_k P N_k^T - N_k P u_k(-t) - P N_k^T u_k(-t)\right)\widetilde{x}(t)dt.
\end{aligned}$$

Thus, we get

$$\begin{aligned}
-\int_0^\infty \widetilde{x}(t)^T \left(AP + PA^T + \sum_{k=1}^m N_k P N_k^T\right)\widetilde{x}(t)dt &= \int_0^\infty \widetilde{x}(t)^T B B^T \widetilde{x}(t)dt \\
&= x_0^T P_u x_0.
\end{aligned}$$

Moreover, if

$$\int_0^\infty \widetilde{x}(t)^T \sum_{k=1}^m \left(N_k P N_k^T - N_k P u_k(-t) - P N_k^T u_k(-t)\right)\widetilde{x}(t)dt \geq 0, \tag{3.26}$$

then we have $x_0^T P x_0 \geq x_0^T P_u x_0$. Furthermore, if we assume that the reachable state x_0 lies in a sufficiently small ball W in the neighborhood of 0, then x_0 is reached with a sufficiently small input u, guaranteeing that the condition (3.26) is satisfied for all states $x_0 \in W(0)$. Hence, we obtain

$$x_0^T P^{-1} x_0 \leq x_0^T P_u^{-1} x_0, \quad \text{where} \quad x_0 \in W(0).$$

Next, we prove the relation (3.24), We begin with the definition of the output energy functional (Definition 2.21). That is,

$$E_o(x_0,u) = \frac{1}{2}\int_0^\infty \|y(t)\|^2 = \frac{1}{2}\int_0^\infty \|Cx(t)\|^2 = \frac{1}{2}\int_0^\infty x(t)^T C^T C x(t).$$

Substituting for $C^T C$ from (3.12b), we obtain

$$E_o(x_0,u) = \frac{1}{2}\int_0^\infty \left(-2x(t)^T Q A x(t) - \sum_{k=1}^m x(t)^T N_k^T Q N_k x(t)\right) dt.$$

Next, we substitute for $Ax(t)$ from (3.22) to get

$$\begin{aligned}E_o(x_0,u) &= \frac{1}{2}\int\limits_0^{\infty}\Big(-2x(t)^TQ\dot{x}(t)+2\sum_{k=1}^{m}x(t)^TQN_kx(t)u_k(t)-\sum_{k=1}^{m}x(t)^TN_k^TQN_kx(t)\Big)dt\\ &= \frac{1}{2}\int_0^{\infty}-\frac{d}{dt}(x(t)^TQx(t))dt+\frac{1}{2}\int_0^{\infty}x(t)^T\Big(\sum_{k=1}^{m}[QN_k+N_k^TQ]u_k(t)\\ &\quad-\sum_{k=1}^{m}N_k^TQN_k\Big)x(t)dt.\end{aligned}$$

This gives

$$E_o(x_0,u)-\frac{1}{2}x_0^TQx_0=\int_0^{\infty}x(t)^T\mathcal{R}(u(t))x(t)dt,$$

where $\mathcal{R}(u(t))=\frac{1}{2}\sum\limits_{k=1}^{m}\big(QN_ku_k(t)+N_k^TQu_k(t)-N_k^TQN_k\big)$. Hence, we get

$$E_o(x_0)-\frac{1}{2}x_0^TQx_0=\max_{\substack{u\in\mathcal{B}_{(\alpha,\beta)}\\ \\ x(0)=x_0,x(\infty)=0}}\int_0^{\infty}x(t)^T\mathcal{R}(u(t))x(t),$$

where $B_{(\alpha,\beta)}\overset{\text{def}}{=}\{u\in L_2^m[0,\infty),\|u\|_{L_2}\leq\alpha,\|u\|_{L_\infty}\leq\beta\}$. First, note that if the term $x(t)^TN_k^TQN_kx(t)$ in the expression of $x(t)^T\mathcal{R}(u(t))x(t)$ is equal to zero for some t, then $x(t)^TQN_ku_k(t)$ is also equal to zero. Thus, if $\|u\|_\infty$ is sufficiently small, meaning β is small, then $\mathcal{R}(u(t))$ is always a negative semi-definite matrix. Therefore, we have

$$E_o(x_0)-\frac{1}{2}x_0^TQx_0\leq 0,$$

concluding the proof. □

Theorem 3.12 reveals that energy functionals for bilinear systems can be bounded by the quadratic form given in terms of Gramians in the neighbourhood of the origin and for small inputs. However, in Theorem 3.12, it is assumed that the Gramians are positive definite matrices, which might not be the case in general. This issue has been addressed in [26], and therein, the relations between energy functionals and the state, lying in $\ker(P)$ or in $\ker(Q)$ are provided. It says that if the desired state $x_0\notin\operatorname{range}(P)$, then $E_c(x_0)=\infty$, and similarly, if an initial state x_0 belongs to $\ker(Q)$, then $E_o(x_0)=0$. This shows that the states x_0 with $x_0\in\ker(Q)$ or $x_0\notin\operatorname{range}(P)$ do not play any role in the dynamics of the system; hence, they can be removed. The main idea of balanced truncation lies in furthermore neglecting the almost uncontrollable and almost unobservable states (hard to control and hard to observe states). In order to guarantee that hard to control and hard to observe states are truncated simultaneously,

Algorithm 3.1: Balanced truncation for bilinear systems.

```
1 Input: System matrices $A, N_k, B$ and $C$, and the order of the reduced-order system $\widehat{n}$.
2 Output: The reduced-order system's matrices $\widehat{A}, \widehat{N}_k, \widehat{B}$ and $\widehat{C}$.
3 Determine low-rank approximations of the truncated Gramians $P \approx RR^T$ and $Q \approx SS^T$;
4 Compute SVD of $S^T R$:
      $S^T R = U\Sigma V = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \operatorname{diag}(\Sigma_1, \Sigma_2) \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T$,
      where $\Sigma_1$ contains the $\widehat{n}$ largest singular values of $S^T R$;
5 Construct the projection matrices $\mathcal{V}$ and $\mathcal{W}$:
      $\mathcal{V} = SU_1\Sigma_1^{-\frac{1}{2}}$ and $\mathcal{W} = RV_1\Sigma_1^{-\frac{1}{2}}$;
6 Determine the reduced-order system's realization:
      $\widehat{A} = \mathcal{W}^T A\mathcal{V}, \widehat{N}_k = \mathcal{W}^T N_k \mathcal{V}, \quad \widehat{B} = \mathcal{W}^T B, \quad \widehat{C} = C\mathcal{V}.$
```

we need to find a transformation of the bilinear into a balanced bilinear system via an appropriate transformation as stated in Remark 3.7. Without loss of generality, we consider the following bilinear system being a balanced bilinear system:

$$\begin{aligned}\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \sum_{k=1}^{m} \begin{bmatrix} N_k^{(11)} & N_k^{(12)} \\ N_k^{(21)} & N_k^{(22)} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} u_k(t) + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} x_1^T(t) & x_2^T(t) \end{bmatrix}^T, \quad \begin{bmatrix} x_1^T(0), x_2^T(0) \end{bmatrix} = [0, 0]\end{aligned}$$

with the reachability and observability Gramians equal to Σ :

$$\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

$\sigma_i > \sigma_{i+1}$ and $x_1(t) \in \mathbb{R}^{\widehat{n}}$ and $x_2(t) \in \mathbb{R}^{n-\widehat{n}}$. Fixing $\widehat{n}$ such that $\sigma_{\widehat{n}} > \sigma_{\widehat{n}+1}$, we determine a reduced-order system of order $\widehat{n}$ by setting $x_2 = 0$ as follows:

$$\begin{aligned}\dot{x}_1(t) &= A_{11}x_1(t) + \sum_{k=1}^{m} N_k^{(11)} x_1(t) u_k(t) + B_1 u(t), \\ \widehat{y}(t) &= C_1 x_1(t), \quad x_1(0) = 0.\end{aligned} \tag{3.27}$$

Finally, we again summarize all steps to determine a reduced-order system for the bilinear system via balancing in Algorithm 3.1.

This provides a good local reduced-order system, but unlike in the case of linear systems, it is not clear how to quantify the error, occurring due to x_2 being removed. Moreover, it is worth noting that in the linear case, the reduced-order system obtained via balanced truncation is a balanced one, which is not the case for bilinear systems, i.e., the Gramians of (3.27) are neither diagonal nor equal in general.

The main challenge in applying balanced truncation for bilinear systems is that it requires the solution of two generalized Lyapunov equations. In this direction, there have many advancements to determine low-factors of these generalized Lyapunov equations in recent times. In [24], the authors have extended the Alternating Direction Implicit (ADI) iteration method to the generalized Lyapunov, and recently, an efficient method using the rational Krylov subspace is also proposed in [118]. But methods are still quite expensive in a large-scale setting. This motivates us to look at an alternative pair of Gramians for bilinear systems, which are cheaper to compute and still provide us some energy functional interpretations.

3.4. Alternative Balancing Approach for Bilinear Systems

In this section, we seek to investigate an alternative pair of Gramians, the so-called truncated Gramians (TGrams) for bilinear systems, and discuss their advantages in balancing-type model reduction. We define TGrams for bilinear systems by considering only the first two terms in the series in (3.9) and (3.11), which depend on the first two *kernels* of the Volterra series of the bilinear system and its dual, as follows:

$$P_T = \int_0^\infty P_1(t_1)P_1(t_1)^T dt + \int_0^\infty \int_0^\infty P_2(t_1,t_2)P_2(t_1,t_2)^T dt_1 dt_2, \tag{3.28a}$$

$$Q_T = \int_0^\infty Q_1(t_1)Q_1(t_1)^T dt_1 + \int_0^\infty \int_0^\infty Q_2(t_1,t_2)Q_2(t_1,t_2)^T dt_1 dt_2, \tag{3.28b}$$

where P_i and Q_i are defined in (3.8) and (3.10), respectively. Next, we establish the relations between these TGrams and the solutions of Lyapunov equations.

Lemma 3.13:
Consider the bilinear system (3.1) with the matrix A being Hurwitz. Let P_T and Q_T be the truncated reachability and observability Gramians of the system as defined in (3.28). Then, P_T and Q_T satisfy the following Lyapunov equations:

$$AP_T + P_T A^T + \sum_{k=1}^m N_k P_l N_k^T + BB^T = 0, \tag{3.29a}$$

$$A^T Q_T + Q_T A + \sum_{k=1}^m N_k^T Q_l N_k + C^T C = 0, \tag{3.29b}$$

respectively, where P_l and Q_l are the Gramians of the corresponding linear part, which solve

$$AP_l + P_l A^T + BB^T = 0, \tag{3.30}$$

$$A^TQ_l + Q_lA + C^TC = 0. \tag{3.31}$$

◊

Proof. The lemma can be proven in a similar way as done in [5, Thm. 1]. However, for the sake of completeness, we sketch the proof here. We begin with the truncated reachability Gramian for the bilinear system. The first term in (3.28a) is

$$P_l = \int_0^\infty P_1(t_1)P(t_1)^T dt_1,$$

where $P_1(t)$ is defined in (5.7), solving

$$AP_l + P_lA^T + BB^T = 0, \tag{3.32}$$

provided A is a Hurwitz matrix, e.g., see [7]. Next, using the second term in (3.28a), we define

$$\begin{aligned}
\mathcal{P}_2 &:= \int_0^\infty \int_0^\infty \sum_{k=1}^m e^{At_2}N_k e^{At_1}BB^Te^{A^Tt_1}N_k^Te^{A^Tt_2}dt_1dt_2\\
&= \sum_{k=1}^m \int_0^\infty e^{At_2}N_k\left(\int_0^\infty e^{At_1}BB^Te^{A^Tt_1}dt_1\right)N_k^Te^{A^Tt_2}dt_2\\
&= \sum_{k=1}^m \int_0^\infty e^{At_2}N_kP_lN_k^Te^{A^Tt_2}dt_1dt_2.
\end{aligned}$$

Again using the integral representation of the Lyapunov equation, e.g. [7], we know that $\mathcal{P}_2$ solves the following (provided A is stable):

$$A\mathcal{P}_2 + \mathcal{P}_2A^T + \sum_{k=1}^m N_kP_lN_k^T = 0. \tag{3.33}$$

Since P_T is defined as the sum of P_l and $\mathcal{P}_2$, it satisfies

$$AP_T + P_TA^T + \sum_{k=1}^m N_kP_lN_k^T + BB^T = 0.$$

Similarly, we can show the result for the truncated observability Gramian for the bilinear system. This completes the proof. □

The main aim of introducing these TGrams is to use them in the balancing model reduction technique for bilinear systems; therefore, it is necessary to investigate relations of these TGrams with the energy functionals of bilinear systems. As a next step, we compare the energy functionals of the bilinear system and the quadratic forms of the TGrams.

Lemma 3.14:
Consider the bilinear system (3.1), with a stable matrix A and let the system be asymptotically any state x_0 reachable from 0. Let P, $Q > 0$ and P_T, $Q_T > 0$ be the Gramians and TGrams of the system, respectively, and $E_c(x)$ and $E_o(x)$, respectively, be the controllability and observability energy functionals of the bilinear system. Furthermore, assume that one of the N_k in (3.1) is of full rank. Then, there exists a small neighborhood W of 0, where the following relations hold:

$$E_c(x) \geq \frac{1}{2} x^T P_T^{-1} x \geq \frac{1}{2} x^T P^{-1} x \quad \text{and} \quad x \in W(0). \tag{3.34}$$

Furthermore, if a bilinear system is locally controllable, then, for sufficiently small inputs, the following relation also holds:

$$E_o(x) \leq \frac{1}{2} x^T Q_T x \leq \frac{1}{2} x^T Q x. \tag{3.35}$$

$\Diamond$

Proof. Let us assume that $x_0 \in \mathbb{R}^n$ is controlled by the input $u = u_{x_0} \in L_2^m(-\infty, 0]$, minimizing the input cost functional in the definition of $E_c(x_0)$. Let Ψ_u be the fundamental solution of the dual system satisfying

$$\dot{\Psi}_u = \left(A^T + \sum_{k=1}^{m} N_k^T u_k(-t) \right) \Psi_u, \quad \Psi_u(t,t) = I. \tag{3.36}$$

From the proof of Theorem 3.12, we know that we can define $\widetilde{x}(t) = \Psi_u(t,0)x_0$, satisfying $\widetilde{x}(t) \to 0$ as $t \to \infty$. Now, we consider

$$\begin{aligned}
x_0^T P_T x_0 &= -\int_0^\infty \frac{d}{dt} \left(\widetilde{x}(t)^T P_T \widetilde{x}(t) \right) dt \\
&= -\int_0^\infty \widetilde{x}(t)^T \left(A P_T + \sum_{k=1}^{m} N_k P_T u_k(-t) \right. \\
&\qquad \left. + P_T A^T + \sum_{k=1}^{m} P_T N_k^T u_k(-t) \right) \widetilde{x}(t) dt \\
&= -\int_0^\infty \widetilde{x}(t)^T \left(A P_T + P_T A^T + \sum_{k=1}^{m} N_k P_l N_k^T \right) \widetilde{x}(t) dt \\
&\quad + \sum_{k=1}^{m} \int_0^\infty \widetilde{x}(t)^T \left(N_k P_l N_k^T - N_k P_T u_k(-t) - P_T N_k^T u_k(-t) \right) \widetilde{x}(t) dt.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}-\int_0^\infty \widetilde{x}(t)^T\left(AP_T+P_TA^T+\sum_{k=1}^m N_kP_lN_k^T\right)\widetilde{x}(t)dt &= \int_0^\infty \widetilde{x}(t)^TBB^T\widetilde{x}(t)dt\\ &= x_0^TP_ux_0.\end{aligned}$$

Moreover, if

$$\int_0^\infty \widetilde{x}(t)^T\sum_{k=1}^m\left(N_kP_lN_k^T-N_kP_Tu_k(-t)-P_TN_k^Tu_k(-t)\right)\widetilde{x}(t)dt\geq 0, \qquad (3.37)$$

then we have $x_0^TP_Tx_0 \geq x_0^TP_ux_0$.

Furthermore, if we assume that the reachable state x_0 lies in a sufficiently small ball W in the neighborhood of 0, then x_0 is reached with a sufficiently small input u, guaranteeing that the condition (3.37) is satisfied for all states $x_0 \in W(0)$. Hence, we obtain

$$x_0^TP_T^{-1}x_0 \leq x_0^TP_u^{-1}x_0, \quad \text{where} \quad x_0\in W(0).$$

Furthermore, if the reachability Gramian P, which is the solution of (3.12a), is given by the series (3.8), and P_T is determined by the sum of the first two terms of this series, then it is easy to conclude that $P \geq P_T > 0$. That means, $x_0^TP^{-1}x_0 \leq x_0^TP_T^{-1}x_0$. Thus, we have $x_0^TP^{-1}x_0 \leq x_0^TP_T^{-1}x_0 \leq x_0^TP_u^{-1}x_0$, where $x_0 \in W(0)$.

Furthermore, along the lines of Theorem 3.12 (the second part), we can show that

$$E_o(x_0,u)-\frac{1}{2}x_0^TQ_Tx_0 = \int_0^\infty x(t)^T\mathcal{R}(u(t))x(t)dt,$$

where $\mathcal{R}(u(t)) = \frac{1}{2}\sum_{k=1}^m\left(Q_TN_ku_k(t)+N_k^TQ_Tu_k(t)-N_k^TQ_lN_k\right)$. As argued in Theorem 3.12, if $\|u\|_{L_\infty}$ is sufficiently small, then it can be seen that $\mathcal{R}(u(t))$ is a negative semidefinite matrix. Hence, we get

$$E_o(x_0)-\frac{1}{2}x_0^TQ_Tx_0 \leq \max_{\substack{u\in\mathcal{B}_{(\alpha,\beta)}\\ x(0)=x_0,x(\infty)=0}}\int_0^\infty x(t)^T\mathcal{R}(u(t))x(t)dt \quad\Rightarrow\quad E_o(x_0)\leq\frac{1}{2}x_0^TQ_Tx_0.$$

Moreover, if the observability Gramian is determined as a series with positive semidefinite summands, then it can also be seen that $Q \geq Q_T$; hence

$$E_o(x_0)\leq\frac{1}{2}x_0^TQ_Tx_0\leq\frac{1}{2}x_0^TQx_0.$$

This concludes the proof. □

To illustrate the relation between energy functionals, Gramians and TGrams of bilinear systems, we consider the same scalar example as considered in [73].

Example 3.1:
Consider a scalar example (a, b, c, η). We assume $a < 0$, $\eta^2 + 2a < 0$ and $bc \neq 0$ to ensure the existence of $P, Q > 0$. The energy functionals of the system can be determined by solving the corresponding nonlinear PDEs [73], yielding

$$E_c(x) = \frac{2a}{\eta^2}\left[\frac{\eta x}{\eta x + b} + \log\left(\frac{b}{\eta x + b}\right)\right] \quad \text{and} \quad E_o(x) = -\frac{1}{2}\left(\frac{c^2}{2a}\right)x^2.$$

The approximations of the energy functionals using the Gramians are:

$$E_c^{(G)}(x) = \frac{1}{2}\left(\frac{\eta^2 + 2a}{-b^2}\right)x^2 \quad \text{and} \quad E_o^{(G)}(x) = \frac{1}{2}\left(\frac{-c^2}{\eta^2 + 2a}\right)x^2.$$

The approximations of the energy functionals using TGrams are:

$$E_c^{(T)}(x) = a\left(-b^2 + \frac{\eta^2 b^2}{2a}\right)^{-1} x^2 \quad \text{and} \quad E_o^{(T)}(x) = \frac{1}{4a}\left(-c^2 + \frac{\eta^2 c^2}{2a}\right)x^2. \qquad \diamondsuit$$

The comparison of these quantities by setting the parameters to $-a = b = c = \eta = 1$ is illustrated in Figure 3.1.

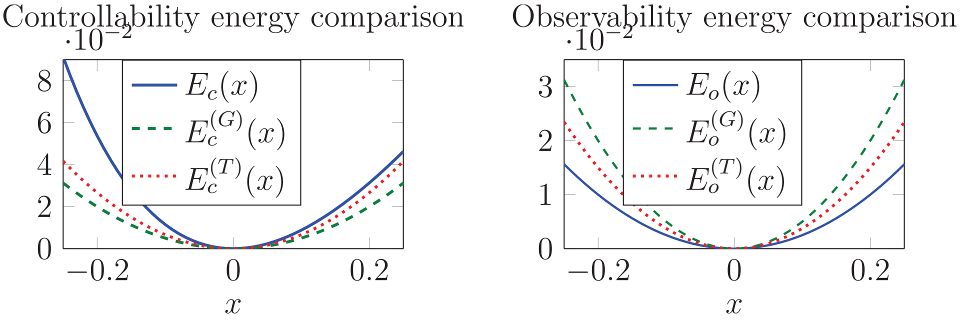


Figure 3.1.: Comparison of the energy functionals of the bilinear system and their approximations via Gramians and TGrams.

From Lemma 3.14, it is clear that the TGrams for bilinear systems can also be used to determine the states that absorb a lot of energy, and still produce very little output energy, at least for a small neighborhood of the origin. However, there are at least a couple of advantages of considering the TGrams over the Gramians for bilinear systems in the model reduction framework. Firstly, TGrams can approximate the energy functionals of the bilinear systems more accurately (at least locally) as proven in Lemma 3.14 and also illustrated in Example 3.1. Secondly, in order to compute TGrams, we require the solutions of four conventional Lyapunov equations, whereas the Gramians require the solutions of the generalized Lyapunov equations (3.12), which are indeed much more computationally cumbersome.

Finally, our conclusion is here that if a bilinear system is excited by small external inputs, then it is worth to consider TGrams for reducing such bilinear systems, although it is difficult to say how small an input can be, which probably depends on the system matrices. A similar kind of observation has also been noticed in the literature while investing the truncated $\mathcal{H}_2$-optimal interpolation model reduction for bilinear systems [60]; therein, it has been shown by means of numerical examples that sometimes a reduced-order system, which locally minimizes the $\mathcal{H}_2$-norm based on the TGrams, performs better in the time-domain simulations than the $\mathcal{H}_2$-optimal reduced-order system. This phenomenon can be explained by TGrams as truncated $\mathcal{H}_2$-optimal model reduction (with truncation index 2) tries to minimize the norm of the error system (difference between the original and reduced-order systems) based on the proposed TGrams.

Furthermore, bilinear control systems are strongly connected to parametric linear systems, see [20]. If the variations of the parameters are small, then it is also worth applying TGrams over the Gramians. The phenomenon has been also observed while investigating the truncation $\mathcal{H}_2$-model reduction problem [60] for bilinear systems.

3.5. Numerical Experiments

In this section, we illustrate the efficiency of the reduced-order systems obtained via the proposed TGrams for the bilinear system and compare it with that of the actual Gramians as discussed in Section 3.3. We denote the Gramians for the bilinear system by SGrams (*standard* Gramians) from now on. In order to determine the low-rank factors of the Gramians for bilinear systems, we employ the most recently proposed algorithm in [118], which utilizes many of the properties of inexact solutions and uses the extended Krylov subspace method (EKSM) to solve the conventional Lyapunov equation up to a desired accuracy. To determine the low-rank factors of the linear Lyapunov equation, we also utilize EKSM to be in the same line. All the simulations were carried out in MATLAB version 8.0.0.783(R2012b) on a board with 4 Intel® Xeon® E7-8837 CPUs with a 2.67-GHz clock speed, 8 Cores each and 1TB of total RAM.

3.5.1. Burgers' equation

We consider a viscous Burgers' equation, which is one of the standard test examples for bilinear systems; see, e.g., [40]. The dynamics of the system are governed by

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = \mu\frac{\partial^2 v}{\partial x^2}, \qquad (x,t) \in (0,1) \times (0,T)$$

with boundary conditions:

$$v(x,0) = 0, \quad x \in (0,1), \quad v(0,t) = u(t), \quad v(1,t) = 0, \quad t \geq 0.$$

As shown in [40], a spatial semi-discretization of the governing equation using k equidistant nodes leads to an ODE system having quadratic nonlinearity. However, a quadratic nonlinear system can be approximated as a bilinear system using Carleman bilinearization; see, e.g., [111]. The dimension of the approximated bilinearized system is $n = k + k^2$. We set the viscosity $\mu = 0.1$ and $k = 40$, and choose the observation vector C such that it yields an average value for the variable v in the spatial domain. Note that the bilinearized system is not an $\mathcal{H}_2$ system, which can be checked by looking at the eigenvalues of the matrix $\mathcal{X} := (I \otimes A + A \otimes I + N \otimes N)$. If $\sigma(\mathcal{X}) \not\subset \mathbb{C}^-$, then the series determining its controllability Gramians diverges. To overcome this issue, we choose a scaling factor γ for the matrices B and N_k, and the input $u(t)$ is scaled by $\frac{1}{\gamma}$. Such an idea was first proposed for bilinear systems in [26]. For this example, we set $\gamma = 0.1$, ensuring $\sigma(\mathcal{X}) \subset \mathbb{C}^-$.

We determine reduced-order systems of orders $r = 5$ and $r = 10$ using SGrams and TGrams, and compare the quality of the reduced-order systems by using two arbitrary control inputs $u^{(1)}(t) = te^{-t}\sin(\pi t)$ and $u^{(2)}(t) = te^{-t} + 1$ as shown in Figure 3.2. More importantly, we also show the CPU-time to determine the low-rank factors of SGrams and TGrams in the same figure.

Figure 3.2 also shows that computing TGrams is much cheaper than SGrams. Moreover, we observe that the reduced-order systems based on TGrams are quite competitive to those computed by SGrams for both control inputs and both orders in this example.

3.5.2. Electricity cable impacted by wind

Below, we discuss an example of a damped wave equation with Lévy noise [108], whose governing equation is given by

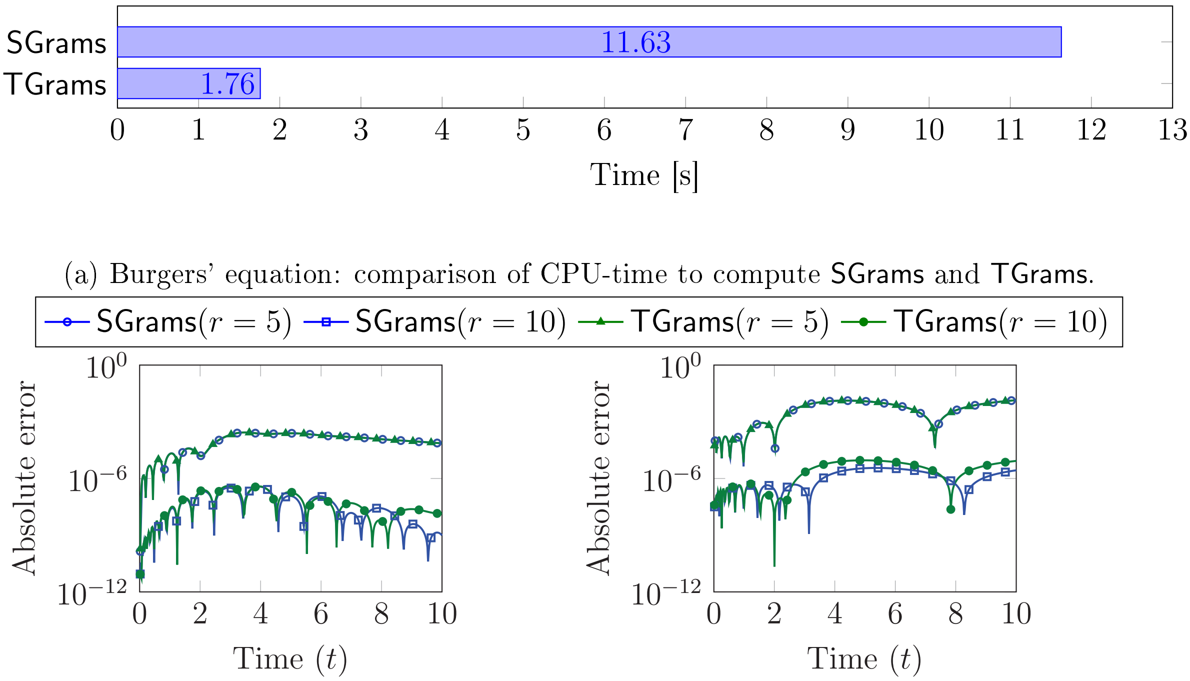
$$\frac{\partial^2}{\partial t^2}X(t,z) + 2\frac{\partial}{\partial t}X(t,z) = \frac{\partial^2}{\partial z^2}X(t,z) + \mathrm{e}^{-(z-\frac{\pi}{2})^2}\,u(t) + 2\,\mathrm{e}^{-(z-\frac{\pi}{2})^2}\,X(t-,z)\frac{\partial M(t)}{\partial t}$$

for $t, z \in [0,\pi]$, where M is a scalar, square integrable Lévy process with mean zero. The boundary and initial conditions are:

$$X(t,0) = X(t,\pi) = 0 \quad \text{and} \quad X(0,z) = 0, \; \frac{\partial}{\partial t}X(t,z)\bigg|_{t=0} \equiv 0.$$

An approximation for the position of the middle of the cable yields the output

$$Y(t) = \frac{1}{2\epsilon}\int_{\frac{\pi}{2}-\epsilon}^{\frac{\pi}{2}+\epsilon} X(t,z)dz, \quad \epsilon > 0.$$



(a) Burgers' equation: comparison of CPU-time to compute SGrams and TGrams.

(b) Input $u^{(1)}(t) = te^{-t}\sin(\pi t)$.

(c) Input $u^{(2)}(t) = te^{-t} + 1$.

Figure 3.2.: Burgers' equation: comparisons of CPU-time and transient responses of the original and reduced-order systems for two different orders and for two inputs.

As can be seen, the governing equation is second order stochastic PDE (SPDE) but is first transformed into a first order SPDE followed by the discretization in space. Following [108], a semi-discretized version of the governing SPDE leads to a linear stochastic system with $x(0) = 0$ and $t \in [0, \pi]$:

$$dx(t) = [Ax(t) + Bu(t)]\,dt + Nx(s-)dM(s), \quad y(t) = Cx(t). \tag{3.38}$$

Here, A, $N \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $x(t-) := \lim_{s\uparrow t} x(s)$ and y is the corresponding output. Moreover, we assume that the adapted control satisfies $\|u\|^2_{\mathcal{L}^2_T} := \mathbb{E}\int_0^T \|u(t)\|^2_{\mathbb{R}^m}\,dt < \infty$. For a detailed discussion, we refer to [108].

In contrast to [108], we fix a different noise process, which allows the wind to come from two directions instead of just one. The noise term we choose is represented by a compound Poisson process $M(t) = \sum_{i=1}^{N(t)} Z_i$ with $(N(t))_{t\in[0,\pi]}$ being a Poisson process with parameter 1. Furthermore, $Z_1, Z_2, \ldots$ are independent uniformly distributed random variables with $Z_i \sim \mathcal{U}\left(-\sqrt{3}, \sqrt{3}\right)$, which are also independent of $(N(t))_{t\in[0,\pi]}$. This choice implies $\mathbb{E}\left[M(t)\right] = 0$ and $\mathbb{E}\left[M^2(1)\right] = 1$. Balanced truncation for such an Ito type SDE (3.38) with the particular choice of M is also based on Gramians, which fulfill

(3.12) with $m = 1$ and $N := N_1$. Note that although the Gramians of linear stochastic systems and bilinear systems coincide for this case, these Gramians have different interpretations in the case of a SDE. Nonetheless, to compute a reduced-order system, we can blindly use SGrams and TGrams for balancing. We fix the dimension of (3.38) to $n = 1000$ and set $u(t) = e^{w(t)}\sin(t)$ and then run several numerical experiments.

In Figure 3.3, we present three trajectories of the output y. This output shows the position of the middle of the string. The purely positive input function $u(t) = e^{w(t)}\sin(t)$, $t \in [0, \pi]$, pushes the cable up, where w is a Wiener process that is independent of the compound Poisson process $M(t) = \sum_{i=1}^{N(t)} Z_i$. Due to the randomness represented by the Wiener process, the input can have completely different intensities. This randomness, of course, leads to completely different outputs in Figure 3.3. At the same time the wind can randomly effect the cable which can either move the cable up or down. If the wind is strong enough, the appearance of the wind is marked by peaks, where we also have little jumps that are not visible in the graphs in Figure 3.3 due to their small size.

We apply balanced truncation based on SGrams as described in [35] and compute the reduced-order systems of order $r = 3$ and $r = 6$. We blindly repeat the procedure for a linear stochastic system but we now replace the SGrams by the TGrams. We again mention that although the Gramians for linear stochastic systems are the same as for bilinear systems, they, however, have some different energy interpretations which have extensively been studied in [107]. We again compute the reduced-order systems of dimensions $r = 3, 6$ based on the truncated Gramians.

Next, we discuss the quality of these derived the reduced-order systems and computational cost to determine the low-rank factors of SGrams and TGrams. In Figure 3.4a, we see that the TGrams are computationally much cheaper as compared to the SGrams. In order to compare the quality of the reduced-order systems, we determine the point-wise deviation and the mean error of the large-scale output with the reduced output based on the SGrams and the TGrams in Figure 3.4. For the $r = 3$ case, clearly, the reduced-order system based on the TGrams outperforms the one based on the SGrams for all three trajectories (see Figure 3.4b). This also applies for the mean deviation as shown in Figure 3.4d (left). For the $r = 6$ case, it is not that obvious anymore. The reduced-order system obtained by SGrams seems to be marginally more accurate, but still, both methods result in very competitive reduced-order systems, see Figures 3.4c and 3.4d (right).

3.6. Conclusions

In this chapter, we have discussed balanced truncation for bilinear systems. Firstly, we studied the relation between energy functionals and the quadratic form of the Gramians for an arbitrary state vector. Then, we have introduced a concept of truncated

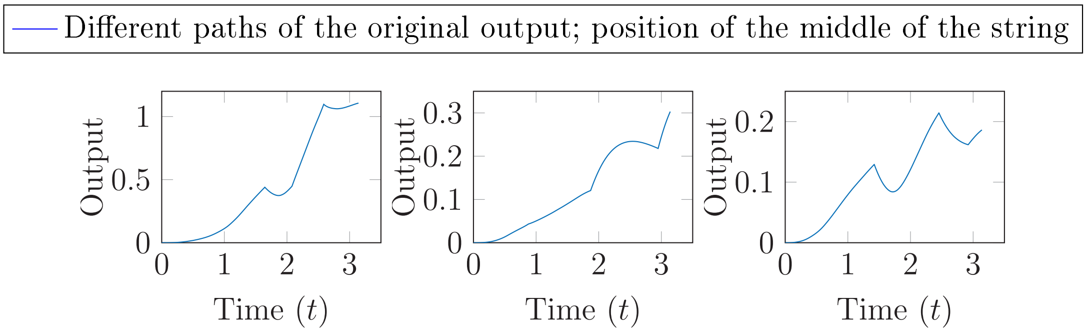
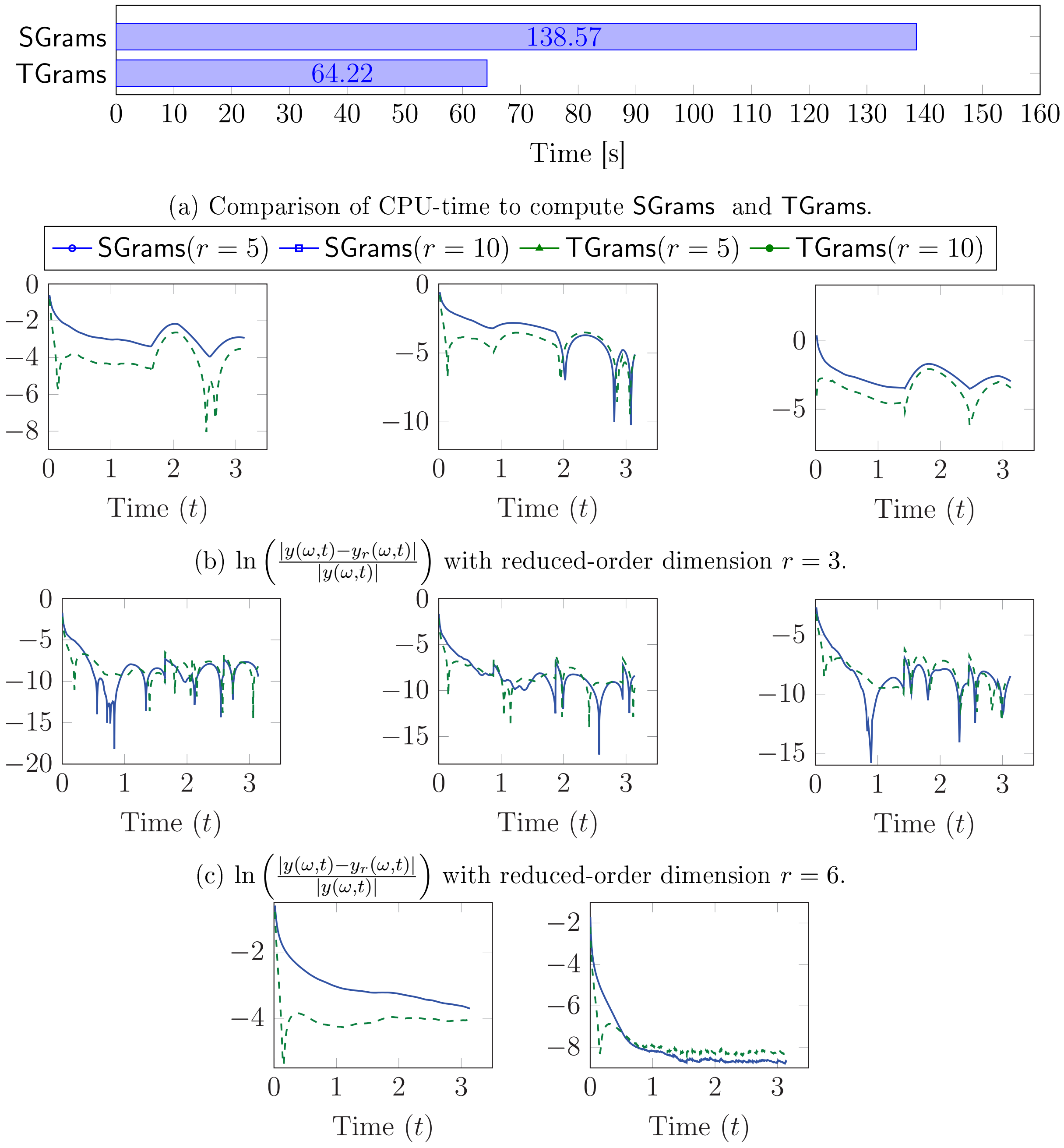


Figure 3.3.: Electricity cable: trajectories for input $u(t) = e^{w(t)} \sin(t)$.

Gramians for bilinear systems. These also allow us to find the states, which are both hard to control and hard to observe, like the Gramians for bilinear systems, in the neighborhood of the origin. We have also shown that the truncated Gramians approximate the energy functionals of bilinear systems better (at least locally) as compared to the Gramians of the latter systems. Moreover, we have discussed advantages of the truncated Gramians in the model reduction context. In the end, we have demonstrated the efficiency of the proposed truncated Gramians in the model reduction framework by means of two numerical examples.

As we have seen, these TGrams also provide qualitatively good reduced-order system for linear stochastic systems as well; however, the energy interpretations for linear stochastic systems in terms of TGrams is still not clear. So, as a further research topic, it would be interesting to study the connection. Moreover, it is yet to be studied how to bound the error between the original and reduced-order systems due to the truncation of the singular values. It would also be compelling to investigate and derive an error bound, ensuring an error in the output with a desired tolerance.



(a) Comparison of CPU-time to compute SGrams and TGrams.

(b) $\ln\left(\frac{|y(\omega,t)-y_r(\omega,t)|}{|y(\omega,t)|}\right)$ with reduced-order dimension $r = 3$.

(c) $\ln\left(\frac{|y(\omega,t)-y_r(\omega,t)|}{|y(\omega,t)|}\right)$ with reduced-order dimension $r = 6$.

(d) $\ln\left(\frac{\mathbb{E}|y(t)-y_r(t)|}{\mathbb{E}|y(t)|}\right)$, where $r = 3$ (left), $r = 6$ (right).

Figure 3.4.: Electricity cable: comparison of reduced-order systems for $u(t) = e^{w(t)}\sin(t)$.

CHAPTER 4

BALANCING-BASED MODEL REDUCTION FOR QUADRATIC-BILINEAR CONTROL SYSTEMS

Contents

4.1. Introduction

In the previous chapter, we have discussed balancing-type model reduction method for a class of nonlinear systems, the so-called bilinear systems, acting as a bridge between fully nonlinear and linear systems. Moving one more step towards nonlinear systems, in this chapter, we address another vital class of nonlinear systems, the so-called *quadratic-bilinear* (QB) systems. These are of the form

$$\Sigma_{QB}: \begin{cases} \dot{x}(t) = Ax(t) + H\left(x(t)\otimes x(t)\right) + \displaystyle\sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \\ y(t) = Cx(t), \quad x(0) = 0, \end{cases} \tag{4.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the states, inputs, and outputs of the systems at time t, respectively; n is the state dimension, and u_k is the kth entry of u. Furthermore, $A, N_k \in \mathbb{R}^{n\times n}$ for $k \in \{1,\ldots,m\}$, $H \in \mathbb{R}^{n\times n^2}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$.

There is a variety of applications where the system inherently contains a quadratic nonlinearity, which can be modeled in the QB form (4.1); e.g., spatial discretizations of the Burgers' equation, the Allen-Cahn or Chafee-Infante equation, and many other models from engineering and physics. Moreover, a large class of smooth nonlinear systems, involving combinations of elementary functions like exponential, trigonometric, and polynomial functions, etc., can be equivalently rewritten as QB systems (4.1) as shown in [25, 78]. We discuss this in a great detail in the subsequent section.

For a given QB system Σ_{QB} of order n as shown in (4.1), our aim is to construct a reduced-order system

$$\widehat{\Sigma}_{QB}: \begin{cases} \dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t) + \widehat{H}\left(\widehat{x}(t)\otimes \widehat{x}(t)\right) + \displaystyle\sum_{k=1}^{m} \widehat{N}_k \widehat{x}(t) u_k(t) + \widehat{B}u(t), \\ \widehat{y}(t) = \widehat{C}\widehat{x}(t), \quad \widehat{x}(0) = 0, \end{cases} \tag{4.2}$$

where $\widehat{A}$, $\widehat{N}_k \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ for $k \in \{1,\ldots,m\}$, $\widehat{H} \in \mathbb{R}^{\widehat{n}\times\widehat{n}^2}$, $\widehat{B} \in \mathbb{R}^{\widehat{n}\times m}$, and $\widehat{C} \in \mathbb{R}^{p\times\widehat{n}}$ with $\widehat{n} \ll n$ such that the output of the reduced-order system $\widehat{y}$ approximates very well the output of the original system y in a proper norm for all admissible inputs $u \in L_2^m[0,\infty)$.

Similar to the linear and bilinear cases, we construct a reduced-order system (4.2) via projections. Towards this goal, we construct the two basis matrices V, $W \in \mathbb{R}^{n\times\widehat{n}}$ such that W^TV is invertible. Then, the reduced matrices in (4.2) are computed as:

$$\begin{aligned} &\widehat{A} = (W^TV)^{-1}W^TAV, && \widehat{N}_k = (W^TV)^{-1}W^TN_kV, \quad \text{for} \quad k \in \{1,\ldots,m\}, \\ &\widehat{H} = (W^TV)^{-1}W^TH(V\otimes V), && \widehat{B} = (W^TV)^{-1}W^TB, \quad \text{and} \quad \widehat{C} = CV. \end{aligned}$$

It can be easily seen that the quality of the reduced-order system depends on the choice of the reduction subspaces spanned by the columns of V and W, respectively. There exist various MOR approaches in the literature to determine these subspaces. One of the earlier and popular methods for nonlinear systems is proper orthogonal decomposition (POD); see, e.g., [10, 47, 88, 96]. POD relies on the Galerkin projection

$\mathcal{P} = \mathcal{V}\mathcal{V}^T$, where $\mathcal{V}$ is determined based on extracting the dominant modes of the system dynamics from a selection of snapshots of the solution trajectories computed using some training input. A Petrov-Galerkin-type projection can be obtained using the dual/adjoining system in either time or frequency domain [110, 130]. Morover, in this method, one can perform the computation related to $\widehat{H}(\widehat{x}(t) \otimes \widehat{x}(t))$ in the reduced-order system (4.2) even more cheaply and quite accurately. For this, there are some advanced methodologies such as the empirical interpolation method (EIM), the discrete empirical interpolation method (DEIM), the best point interpolation method (BPIM), etc. For detail, we refer to [13, 47, 53, 76].

Another popular trajectory-based MOR technique is trajectory piecewise linear (TPWL) method [109], where nonlinear functions are replaced by a weighted combination of linear systems. These linear systems can then be reduced by applying well-established MOR techniques for linear systems such as balanced truncation or the interpolation-based iterative method (IRKA); see, e.g., [7, 79]. In recent years, reduced basis methods have been successfully applied to nonlinear systems to obtain reduced-order systems [13, 76]. In spite of all these, the trajectory-based MOR techniques have the drawback of being input dependent. This makes the obtained reduced-order systems inadequate to control applications, where the variation of the input is inherent to the problem.

Some other ideas, based on interpolation or moment-matching, have been extended from linear systems to QB systems, with the aim of capturing the input-output behavior of the underlying system independent of a training input. One-sided interpolatory projection for QB systems is studied in [11, 78, 105, 106]. Recently, one-sided projection method has been extended to two-sided interpolatory projection in [23, 25], ensuring more moments to be matched, for a given order of a reduced-order system. These methods result in reduced-order systems which does not rely on the training data or the solution trajectories of specific inputs; see also the survey [15] for some related approaches. Thus, the determined reduced-order systems can be used in input-varying applications. Although these methods have evolved as an effective MOR technique for nonlinear systems in recent times, shortcomings of these methods are: how to choose an appropriate order of a reduced-order system and how to select good interpolation points. Furthermore, the two-sided interpolatory projection method [25] is only applicable to single-input single-output systems, which is very restrictive, and additionally the stability of the resulting reduced-order systems also remains another major issue. We note here that the method proposed in this thesis does not resolve this issue. It remains an open problem, even in the case of linear systems, to give general conditions for stability preservation of two-sided projection methods. Moreover, we mention that there exist methods which relies on the data of the multi-variate transfer functions of quadratic-bilinear systems, see, e.g., [69, 92]. However, collecting the required data is a challenging task.

In this chapter, our focus rather lies on a balancing-type MOR technique for QB systems. As mentioned earlier, this technique mainly depends on controllability and observability energy functionals, or in other words, Gramians of the system. In the previous chapter, we have studied this methodology for bilinear control system, and in this chapter, we aim at extending this methodology to the QB control systems.

The structure of the chapter is as follows. In the subsequent section, we first review a quadratic-bilinearization process for a nonlinear system and discuss the symmetric structure of the Hessian H in (4.1). In Section 4.3, we propose the reachability Gramian and its truncated version for QB systems based on the underlying Volterra series of the system. Additionally, we determine the observability Gramian and its truncated version based on the dual system associate to the QB system. Furthermore, we establish relations between the solutions of a certain type of quadratic Lyapunov equations and the proposed Gramians. Later on, we develop the connection between the proposed Gramians and the energy functionals of the QB systems, and also reveal their relations to controllability and observability of the system. Consequently, we utilize these Gramians for balancing of QB systems, which allows us to determine those states that are hard to control as well as hard to observe. Truncation of such states leads to reduced-order systems. In Section 4.3.4, we discuss the related computational issues and advantages of the truncated version of Gramians in the MOR framework. We further discuss the stability of these reduced-order systems. In Section 4.4, we test the efficiency of the proposed balanced truncation MOR technique for various semi-discretized nonlinear PDEs and compare it with the existing moment-matching techniques for the QB systems. We finally conclude with a short summary and future research topics.

4.2. Quadratic-Bilinearization and Hessian Properties

As noted in the introduction that a large class of nonlinear systems, containing smooth monovariate variables can be rewritten into the quadratic-bilinear form, we begin by outlining the process of rewriting a nonlinear system into the QB form (4.1).

4.2.1. Quadratic-bilinearization of nonlinear systems

Let us consider a nonlinear system of the form:

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{4.3}$$

where $f(x) \in \mathbb{R}^n \to \mathbb{R}^n$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{p\times n}$, and $x(t)$, $y(t)$ and $u(t)$ are the state, output and input vectors at time t. Rewriting a nonlinear system into the QB form is called *quadratic-bilinearization*. This process involves essentially two steps; the first step is to polynomialization of the nonlinear system, which is followed by the quadratic-bilinearization of the obtained polynomialized system. We first define the nonlinear system (4.3), which have the polynomialization form.

Definition 4.1:
A nonlinear system is said to be in the polynomialization form if its all nonlinearities are in polynomial with respect to its state vector and linear with respect to the control input. ♢

Polynomialization of the nonlinear system (4.3) can be done in two different ways. The first step is to introduce appropriate new variables, followed by either taking the Lie derivatives of these newly introduced variables or defining some algebraic constraints in terms of new variables. In the following, we summarize the important steps in order to polynomialize a nonlinear system:

1. Introduce some appropriate new variables.
2. Replace the nonlinear terms by using new variables, thus resulting in a polynomialized nonlinear system.
3. Derive differential equations or define algebraic constraints.

Example 4.1:
To illustrate this process, we consider a nonlinear ODE as follows:

$$\dot{x}_1(t) = -x_1(t) + x_2^3(t) + e^{-x_2(t)}, \tag{4.4a}$$
$$\dot{x}_2(t) = -x_1 + u(t). \tag{4.4b}$$

The system (4.4) has cubic and exponential nonlinearities. To polynomialize it, we first introduce a new variable:

$$z_1(t) := e^{-x_2(t)}.$$

Now, if we substitute the variable $z_1(t)$ in (4.4a), then we obtain

$$\dot{x}_1(t) = -x_1(t) + x_2^3(t) + z_1(t), \tag{4.5a}$$
$$\dot{x}_2(t) = -x_1 + u(t). \tag{4.5b}$$

Next, we derive the differential equation for the variable $z_1(t)$, which is:

$$\dot{z}_1(t) = -e^{-x_2(t)}\dot{x}_2(t) = z_1(t)x_1(t) - z_1(t)u(t).$$

This illustrates how a nonlinear system can be polynomialized. ♢

The next step is to rewrite a polynomialized nonlinear system into the quadratic-bilinear form. Once we polynomialize nonlinear systems, we introduce higher-order terms as new variables and derive the corresponding differential equations. For illustration purpose, we again consider the same Example 4.1. After polynomialization of (4.4), we obtain the following set of ODEs:

$$\dot{x}_1(t) = -x_1(t) + x_2^3(t) + z_1(t), \tag{4.6a}$$

$$\dot{x}_2(t) = -x_1(t) + u(t), \tag{4.6b}$$

$$\dot{z}_1(t) = z_1(t)x_1(t) - z_1(t)u(t). \tag{4.6c}$$

Since Eq. (4.6a) has a cubic order term, we introduce a new variable as $z_2(t) := x_2^2(t)$ and derive the corresponding differential equation:

$$\dot{z}_2(t) = 2x_2(t)\dot{x}_2(t) = -2x_2(t)x_1(t) + 2x_2(t)u(t). \tag{4.7}$$

Thus, the cubic term $x_2^3(t)$ in (4.6a) can be replaced by $x_2(t)z_2(t)$. Hence, we have completely transformed the nonlinear system (4.4) into a quadratic-bilinear system by introducing new variables and by deriving their derivatives. However, there are two major disadvantages of the quadratic-bilinearization. Firstly, the quadratic-bilinearization of a nonlinear system destroys the structure of the original nonlinearities; however, this transformation is exact, i.e., it requires no approximation and does not introduce any error. Secondly, the dimension of the resulting QB system becomes higher than the original nonlinear system, whereas our aim is to reduce the state dimensions, but it eases the model reduction process.

Note that the transformation of a nonlinear system to the QB form is not unique, and the *minimal*, or an optimal transformation of a nonlinear system is yet an open problem. Furthermore, we briefly like to mention that the similar ideas have been in the literature for a long time, and it is known as McCormick-Relaxation, see [99]; but, this idea was used in the model reduction framework for the first time in [78].

4.2.2. Symmetrization of the Hessian

Having collected basic properties of tensor algebra in Section 2.4, we here discuss a connection between the Hessian H of QB systems (4.1) and tensor matricization. Since H lies in $\mathbb{R}^{n\times n^2}$, one can interpret it as an unfolding of a tensor $\mathcal{H}^{n\times n\times n}$. Without loss of generality, we, in this thesis, assume the Hessian H to be the mode-1 matricization of $\mathcal{H}$, i.e., $H = \mathcal{H}^{(1)}$. Recall from Section 2.4 that a symmetric tensor $\mathcal{H}$ satisfies the following for given vectors $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$:

$$H(v\otimes w) = \mathcal{H}^{(1)}(v\otimes w) = \mathcal{H}^{(1)}(w\otimes v) = H(w\otimes v). \tag{4.8}$$

Moreover, if the tensor $\mathcal{H}$ is a symmetric one, then $\mathcal{H}^{(1)}$ or the Hessian H has a symmetric structure. However, the Hessian H of a QB system, obtained via semi-discretization

```
Algorithm 4.1: Symmetrize the Hessian.
  Input: The Hessian $H$.
  Output: The symmetric Hessian $\widetilde{H}$.
  Determine the tensor $\mathcal{H}$ such that its mode-1 matricization is $H$.
  Compute mode-2 and mode-3 matricization of $\mathcal{H}$, denoting respectively by $\mathcal{H}(2)$
    and $\mathcal{H}^{(3)}$.
  Determine another tensor $\widetilde{\mathcal{H}}$ such that its mode-2 matricization $\widetilde{\mathcal{H}}^{(2)}$ is given by
    $\widetilde{\mathcal{H}}^{(2)} = \frac{1}{2}\left(\mathcal{H}^{(2)} + \mathcal{H}^{(3)}\right)$.
  Then, the symmetric Hessian $\widetilde{H} = \widetilde{\mathcal{H}}^{(1)}$.
```

of the governing equation or after the quadratic-bilinearization of a nonlinear system, might not have a symmetric structure. But as shown in [25, 39], the Hessian H of a QB system can be modified in such a way that it has a symmetric structure, without any change in the dynamics of the systems, i.e., $H(x(t) \otimes x(t)) = \widetilde{H}(x(t) \otimes x(t))$, where $\widetilde{H}$ has a symmetric structure. In Algorithm 4.1, we summarize the steps to ensure the symmetric structure of the Hessian, without changing the system dynamics.

Example 4.2 ([39]):
To illustrate this process, we consider a two-dimensional purely quadratic system as follows:

$$\dot{x}(t) = H(x(t) \otimes x(t)),$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \qquad H = \begin{bmatrix} a & b & c & d \\ e & f & g & h \end{bmatrix}. \tag{4.9}$$

Clearly, the Hessian H does not have the symmetric structure, i.e., $H(v \otimes w) \neq H(w \otimes v)$ for given arbitrary vectors v and w. However, the Hessian H in (4.9) can be changed to a matrix $\widetilde{H}$ such that $H(x(t) \otimes x(t)) = \widetilde{H}(x(t) \otimes x(t))$ but $\widetilde{H}$ has a symmetric structure. To construct such $\widetilde{H}$, we follow the steps from Algorithm 4.1:

Step 1. We determine a tensor $\mathcal{H} \in \mathbb{R}^{2\times 2\times 2}$, whose the first and second layers, respectively, are:

$$\mathcal{H}(:,:,1) = \begin{bmatrix} a & b \\ e & f \end{bmatrix}, \qquad \mathcal{H}(:,:,2) \begin{bmatrix} c & d \\ g & h \end{bmatrix},$$

ensuring $\mathcal{H}^{(1)} = H$.

Step 2. Determine mode-2 and mode-3 matricizations of $\mathcal{H}$:

$$\mathcal{H}^{(2)} = \begin{bmatrix} a & e & c & g \\ b & f & d & h \end{bmatrix}, \qquad \mathcal{H}^{(3)} = \begin{bmatrix} a & e & b & f \\ c & g & d & h \end{bmatrix}. \tag{4.10}$$

Step 3. Compute $\widetilde{\mathcal{H}}^{(2)} = \dfrac{1}{2}\left(\mathcal{H}^{(2)} + \mathcal{H}^{(3)}\right)$:

$$\widetilde{\mathcal{H}}^{(2)} = \begin{bmatrix} a & e & \frac{1}{2}(b+c) & \frac{1}{2}(f+g) \\ \frac{1}{2}(b+c) & \frac{1}{2}(f+g) & d & h \end{bmatrix}. \tag{4.11}$$

Step 4. Then, the symmetric Hessian $\widetilde{H} =: \widetilde{\mathcal{H}}^{(1)}$ is given by

$$\widetilde{H} = \begin{bmatrix} a & \frac{1}{2}(b+c) & \frac{1}{2}(b+c) & d \\ e & \frac{1}{2}(f+g) & \frac{1}{2}(f+g) & h \end{bmatrix}. \tag{4.12}$$

Using simple algebra, it can be verified that $\widetilde{H}(x(t)\otimes x(t)) = H(x(t)\otimes x(t))$. Moreover, $\widetilde{H}(v\otimes w) = \widetilde{H}(w\otimes v)$ for all vectors v and w. This process of symmetrizing the Hessian can be efficiently applied to large state-space, sparse dynamical systems. We also mention that when a tensor is symmetric, then its mode-2 and mode-3 matricizations coincide, i.e., $\mathcal{H}^{(2)} = \mathcal{H}^{(3)}$, and it also allows us to established many important relations among tensors and matrices as noted in Section 2.4. Thus, in this thesis, without loss of generality, we assume that the Hessian of a QB system has a symmetric structure. ♢

Before we move on deriving algebraic Gramains for QB systems, we point out that the quadratic-bilinear control systems (4.1) is somehow a little different than the problem considered in [78], where the following quadratic-type system has been considered:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + H\left(x(t)\otimes x(t)\right) + \sum_{k=1}^{m} N_k x(t)u_k(t) + \sum_{k=1}^{m} L_k(x(t)\otimes x(t))u_k(t) + Bu(t), \\ y(t) &= Cx(t), \quad x(0) = 0. \end{aligned} \tag{4.13}$$

where $L_k \in \mathbb{R}^{n\times n^2}$ and the dimensions of all other matrices are the same as in (4.1). The quadratic-type system (4.13) contains extra terms which is coupling between quadratic term and inputs. However, in this thesis, we focus on the quadratic-bilinear systems of the form as in (4.1), which can be interpreted as a combination of a purely quadratic system and a purely bilinear system.

4.3. Algebraic Gramians for Quadratic-Bilinear systems and Model Order Reduction

The main purpose of this section is to determine algebraic Gramians for QB systems and study their usage in the model order reduction context. Let us consider QB systems of the form

$$\dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \tag{4.14a}$$

$$y(t) = Cx(t), \quad x(0) = 0, \tag{4.14b}$$

where $A, N_k \in \mathbb{R}^{n \times n}$, $H \in \mathbb{R}^{n \times n^2}$, $B \in \mathbb{R}^{n \times m}$ and $C \in \mathbb{R}^{p \times n}$. Furthermore, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ denote the state, input and output vectors of the system, respectively. We begin by deriving the reachability Gramian of the QB system and its connection with a certain type of quadratic Lyapunov equation.

4.3.1. Reachability Gramian for QB system

In order to derive the reachability Gramian, we first formulate the Volterra series for the QB system (4.14). Before we proceed further, for ease we define the following short-hand notation:

$$u^{(k)}_{\sigma_1,\ldots,\sigma_l}(t) := u_k(t - \sigma_1 \cdots - \sigma_l) \quad \text{and} \quad x_{\sigma_1,\ldots,\sigma_l}(t) := x(t - \sigma_1 \cdots - \sigma_l).$$

We integrate both sides of the differential equation (4.14a) in the state variables with respect to time to obtain

$$\begin{aligned} x(t) = \int_0^t e^{A\sigma_1} B u_{\sigma_1}(t) d\sigma_1 + \sum_{k=1}^{m} \int_0^t e^{A\sigma_1} N_k x_{\sigma_1}(t) u^{(k)}_{\sigma_1}(t) d\sigma_1 \\ + \int_0^t e^{A\sigma_1} H \left(x_{\sigma_1}(t) \otimes x_{\sigma_1}(t) \right) d\sigma_1. \end{aligned} \tag{4.15}$$

Based on the above equation, we obtain an expression for $x_{\sigma_1}(t)$ as follows:

$$\begin{aligned} x_{\sigma_1}(t) = \int_0^{t-\sigma_1} e^{A\sigma_2} B u_{\sigma_1,\sigma_2}(t) d\sigma_2 + \sum_{k=1}^{m} \int_0^{t-\sigma_1} e^{A\sigma_2} N_k x_{\sigma_1,\sigma_2}(t) u^{(k)}_{\sigma_1,\sigma_2}(t) d\sigma_2 \\ + \int_0^{t-\sigma_1} e^{A\sigma_2} H \left(x_{\sigma_1,\sigma_2}(t) \otimes x_{\sigma_1,\sigma_2}(t) \right) d\sigma_2 \end{aligned}$$

and substitute it in (4.15) to have

$$\begin{aligned} x(t) = \int_0^t e^{A\sigma_1} B u_{\sigma_1}(t) d\sigma_1 + \sum_{k=1}^{m} \int_0^t \int_0^{t-\sigma_1} e^{A\sigma_1} N_k e^{A\sigma_2} B u^{(k)}_{\sigma_1}(t) u_{\sigma_1,\sigma_2}(t) d\sigma_1 d\sigma_2 \\ + \int_0^t \int_0^{t-\sigma_1} \int_0^{t-\sigma_1} e^{A\sigma_1} H (e^{A\sigma_2} B \otimes e^{A\sigma_3} B) \left(u_{\sigma_1,\sigma_2}(t) \otimes u_{\sigma_1,\sigma_3}(t) \right) d\sigma_1 d\sigma_2 d\sigma_3 + \cdots . \end{aligned}$$

Repeating this process by repeatedly substituting for the state yields the Volterra series for the QB system. Having carefully analyzed the *kernels* of the Volterra series for the system, we define the reachability mapping $\bar{P}$ as follows:

$$\bar{P} = [\bar{P}_1,\ \bar{P}_2,\ \bar{P}_3, \ldots], \tag{4.16}$$

where the $\bar{P}_i$'s are:

$$\begin{aligned}
\bar{P}_1(t_1) &= e^{At_1}B, && (4.17a)\\
\bar{P}_2(t_1,t_2) &= e^{At_2}\left[N_1,\ldots,N_m\right]\left(I_m \otimes \bar{P}_1(t_1)\right), && (4.17b)\\
\vdots \qquad & \qquad \vdots \\
\bar{P}_i(t_1,\ldots,t_i) &= e^{At_i}\Big[H\big[\bar{P}_1(t_1)\otimes \bar{P}_{i-2}(t_2,\ldots,t_{i-1}), \bar{P}_2(t_1,t_2)\otimes \bar{P}_{i-3}(t_3,\ldots,t_{i-1}),\\
&\qquad \ldots, \bar{P}_{i-2}(t_1,\ldots,t_{i-2})\otimes \bar{P}_1(t_{i-1})\big],\\
&\qquad \left[N_1,\ldots,N_m\right]\left(I_m\otimes \bar{P}_{i-1}(t_1,\ldots,t_{i-1})\right)\Big], \forall\ i \geq 3. && (4.17c)
\end{aligned}$$

Using the mapping $\bar{P}$ (4.16), we define the reachability Gramian P as

$$P = \sum_{i=1}^{\infty} P_i \qquad \text{with} \qquad P_i = \int_0^{\infty}\cdots\int_0^{\infty} \bar{P}_i(t_1,\ldots,t_i)\bar{P}_i^T(t_1,\ldots,t_i)dt_1\cdots dt_i. \tag{4.18}$$

In what follows, we show the equivalence between the above proposed reachability Gramian and the solution of a certain type of quadratic Lyapunov equation.

Theorem 4.2:
Consider the QB system (4.14) with a stable matrix A. If the reachability Gramian P of the system defined as in (4.18) exists, then the Gramian P satisfies the generalized quadratic Lyapunov equation, given by

$$AP + PA^T + H(P\otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0. \tag{4.19}$$

◊

Proof. We begin by considering the first term in the summation (4.18). This is,

$$P_1 = \int_0^{\infty} \bar{P}_1\bar{P}_1^T dt_1 = \int_0^{\infty} e^{At_1}BB^Te^{A^Tt_1}dt_1.$$

As shown, e.g., in [7], P_1 satisfies the following Lyapunov equation (provided A is stable):

$$AP_1 + P_1A^T + BB^T = 0. \tag{4.20}$$

Next, we consider the second term in the summation (4.18):

$$\begin{aligned}
P_2 &= \int_0^\infty \int_0^\infty \bar{P}_2 \bar{P}_2^T dt_1 dt_2 \\
&= \int\limits_0^\infty \int\limits_0^\infty e^{At_2} \left[N_1, \ldots, N_m\right] \left(I_m \otimes \left(e^{At_1} BB^T e^{A^T t_1}\right)\right) \left[N_1, \ldots N_m\right]^T e^{A^T t_2} dt_1 dt_2 \\
&= \sum_{k=1}^m \int_0^\infty e^{At_2} N_k \Big(\int_0^\infty e^{At_1} BB^T e^{A^T t_1} dt_1 \Big) N_k^T e^{A^T t_2} dt_1 dt_2 \\
&= \sum_{k=1}^m \int_0^\infty e^{At_2} N_k P_1 N_k^T e^{A^T t_2} dt_2.
\end{aligned}$$

Again using the integral representation of the solution to Lyapunov equations [7], we see that P_2 is the solution of the following Lyapunov equation:

$$AP_2 + P_2 A^T + \sum_{k=1}^m N_k P_1 N_k^T = 0. \tag{4.21}$$

Finally, we consider the ith term, for $i \geq 3$, which is

$$\begin{aligned}
P_i &= \int_0^\infty \cdots \int_0^\infty \bar{P}_i \bar{P}_i^T dt_1 \cdots dt_i \\
&= \int\limits_0^\infty e^{At_i} \Bigg[H \Bigg[\int\limits_0^\infty \mathcal{F}\left(\bar{P}_1(t_1)\right) dt_1 \otimes \int\limits_0^\infty \cdots \int\limits_0^\infty \mathcal{F}\left(\bar{P}_{i-2}(t_2, \ldots, t_{i-1})\right) dt_2 \cdots dt_{i-1} \\
&\quad + \cdots + \int\limits_0^\infty \cdots \int\limits_0^\infty \mathcal{F}\left(\bar{P}_{i-2}(t_1, \ldots, t_{i-2})\right) dt_1 \cdots dt_{i-2} \otimes \int\limits_0^\infty \mathcal{F}\left(\bar{P}_1(t_{i-1})\right) dt_{i-1} \Bigg] H^T \\
&\quad + \sum_{k=1}^m N_k \left(\int_0^\infty \cdots \int_0^\infty \mathcal{F}\left(\bar{P}_{i-1}(t_1, \ldots, t_{i-1})\right) \right) N_k^T \Bigg] e^{A^T t_i} dt_i,
\end{aligned}$$

where we use the shorthand $\mathcal{F}(\mathcal{A}) := \mathcal{A}\mathcal{A}^T$. Thus, we have

$$P_i = \int_0^\infty e^{At_i} \Big[H(P_1 \otimes P_{i-2} + \cdots + P_{i-2} \otimes P_1) H^T + \sum_{k=1}^m N_k P_{i-1} N_k^T \Big] e^{A^T t_i} dt_i.$$

Similar to P_1 and P_2, we can show that P_i satisfies the following Lyapunov equation, given in terms of the preceding P_k, for $k \in \{1, \ldots, i-1\}$:

$$AP_i + P_i A^T + H(P_1 \otimes P_{i-2} + \cdots + P_{i-2} \otimes P_1) H^T + \sum_{k=1}^m N_k P_{i-1} N_k^T = 0. \tag{4.22}$$

To the end, adding (4.20), (4.21) and (4.22) yields

$$A\sum_{i=1}^{\infty}P_i+\sum_{i=1}^{\infty}P_i\,A^T+H\left(\sum_{i=1}^{\infty}P_i\otimes\sum_{i=1}^{\infty}P_i\right)H^T+\sum_{k=1}^{m}N_k\left(\sum_{i=1}^{\infty}P_i\right)N_k^T+BB^T=0.$$

This implies that $P=\sum_{i=1}^{\infty}P_i$ solves the generalized quadratic Lyapunov equation given by (4.19). □

4.3.2. Dual system and observability Gramian for QB system

We first derive the dual system for the QB system; the dual system plays an important role in determining the observability Gramian for the QB system (4.14). We aim at determining the observability Gramian in a similar fashion as done for the reachability Gramian in the preceding subsection. From linear and bilinear systems, we know that the observability Gramian of the dual system is the same as the reachability Gramian; here, we also consider the same analogy. If we compare the system (4.14) with the general nonlinear system as shown in (2.25), it turns out that for the system (4.14)

$$\mathcal{A}(x,u,t)=A+H(x\otimes I)+\sum_{k=1}^{m}N_ku_k,\quad \mathcal{B}(x,u,t)=B\ \text{ and }\ \mathcal{C}(x,u,t)=C.$$

Using Lemma 2.22, we can write down the state-space realization of the nonlinear Hilbert adjoint operator of the QB system as follows:

$$\dot{x}(t)=Ax(t)+H(x(t)\otimes x(t))+\sum_{k=1}^{m}N_kx(t)u_k(t)+Bu(t),\qquad x(0)=0,\tag{4.23a}$$

$$\dot{z}(t)=-A^Tz(t)-(x(t)^T\otimes I)H^Tz(t)-\sum_{k=1}^{m}N_k^Tz(t)u_k(t)-C^Tu_d(t),\ z(\infty)=0,\tag{4.23b}$$

$$y_d(t)=B^Tz(t),\tag{4.23c}$$

where $z(t)\in\mathbb{R}^n, u_d(t)\in\mathbb{R}$ and $y_d(t)\in\mathbb{R}$ can be interpreted as the dual state, dual input and dual output vectors of the system at time t, respectively. Next, we attempt to utilize the knowledge for the tensor multiplications and matricization (see Section 2.4) to simplify the term $(x(t)^T\otimes I)H^Tz(t)$ in the system (4.23) and to write it in the form of $x(t)\otimes z(t)$.

Note that the matrix $H\in\mathbb{R}^{n\times n^2}$ in the system denotes a Hessian, which can be seen as an unfolding of a 3-dimensional tensor $\mathcal{H}\in\mathbb{R}^{n\times n\times n}$. Here, we choose the tensor $\mathcal{H}\in\mathbb{R}^{n\times n\times n}$ such that its mode-1 matricization is the same as the Hessian H, i.e.,

$H = \mathcal{H}^{(1)}$. Next, let us consider a tensor $\mathcal{T} \in \mathbb{R}^{1\times n\times 1}$, whose mode-1 matricization $\mathcal{T}^{(1)}$ is given by

$$\mathcal{T}^{(1)} = z(t)^T H(x(t) \otimes I) = z(t)^T \mathcal{H}^{(1)}(x(t) \otimes I).$$

We then observe that the mode-1 matricization of the tensor $\mathcal{T}$ is a transpose of the mode-2 matricization, i.e., $\mathcal{T}^{(1)} = \left(\mathcal{T}^{(2)}\right)^T$, leading to

$$\mathcal{T}^{(1)} = \left(\mathcal{T}^{(2)}\right)^T = (x(t) \otimes z(t))^T \, (\mathcal{H}^{(2)})^T.$$

Therefore, we can rewrite the system (4.23) as:

$$\dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \qquad x(0) = 0, \quad (4.24a)$$

$$\dot{z}(t) = -A^T z(t) - \mathcal{H}^{(2)}\,(x(t) \otimes z(t)) - \sum_{k=1}^{m} N_k^T u_k(t) z(t) - C^T u_d(t), \;\; z(\infty) = 0, \quad (4.24b)$$

$$y_d(t) = B^T z(t). \quad (4.24c)$$

In the meantime, we like to point out that there are two possibilities to define $\mathcal{A}(x,u,t)$ in the case of the QB system (4.1). One is $\mathcal{A}(x,u,t) = A + H(x \otimes I) + \sum_{k=1}^{m} N_k u_k$, which we have used in the above discussion; however, there is another possibility to define $\mathcal{A}(x,u,t)$ as $\widetilde{\mathcal{A}}(x,u,t) = A + H(I \otimes x) + \sum_{k=1}^{m} N_k u_k$, leading to a nonlinear Hilbert adjoint operator whose state-space realization is given by:

$$\dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \qquad x(0) = 0, \quad (4.25a)$$

$$\dot{z}(t) = -A^T z(t) - \mathcal{H}^{(3)}(x(t) \otimes z(t)) - \sum_{k=1}^{m} N_k^T u_k(t) z(t) - C^T u_d(t), \;\; z(\infty) = 0, \quad (4.25b)$$

$$y_d(t) = B^T z(t). \quad (4.25c)$$

It can be noticed that the realizations (4.24) and (4.25) are the same, except the appearance of $\mathcal{H}^{(2)}$ in (4.24) instead of $\mathcal{H}^{(3)}$ in (4.25). Nonetheless, if one assumes that the Hessian H is symmetric, i.e., $H(u \otimes v) = H(v \otimes u)$ for $u, v \in \mathbb{R}^n$, then the mode-2 and mode-3 matricizations coincide, i.e., $\mathcal{H}^{(2)} = \mathcal{H}^{(3)}$. We have discussed in Subsection 4.2.2 that, without loss of generality, we can assume that the Hessian H is symmetric.

Now, we turn our attention towards determining the observability Gramian for the QB system by utilizing the state-space realization of the Hilbert adjoint operator (dual system). For this, we follow the same steps as used for determining the reachability

Gramian. Using the dual system (4.24), one can write the dual state $z(t)$ of the dual system at time t as follows:

$$\begin{aligned} z(t) &= \int_{\infty}^{t} e^{-A^T(t-\sigma_1)} C^T u_d(\sigma_1) d\sigma_1 + \sum_{k=1}^{m} \int_{\infty}^{t} e^{-A^T(t-\sigma_1)} N_k^T z(\sigma_1) u_k(\sigma_1) d\sigma_1, \\ &+ \int_{\infty}^{t} e^{-A^T(t-\sigma_1)} \mathcal{H}^{(2)} \left(x(\sigma_1) \otimes z(\sigma_1)\right) d\sigma_1, \end{aligned}$$

which after an appropriate change of variable leads to

$$\begin{aligned} z(t) &= \int_{\infty}^{0} e^{A^T\sigma_1} C^T u^{(d)}(t+\sigma_1) d\sigma_1 + \sum_{k=1}^{m} \int_{\infty}^{0} e^{A^T\sigma_1} N_k^T z(t+\sigma_1) u_k(t+\sigma_1) d\sigma_1 \\ &+ \int_{\infty}^{0} e^{A^T\sigma_1} \mathcal{H}^{(2)} \big(x(t+\sigma_1) \otimes z(t+\sigma_1)\big) d\sigma_1. \end{aligned} \tag{4.26}$$

Eq. (4.25a) gives the expression for $x(t+\sigma_1)$. This is

$$\begin{aligned} x(t+\sigma_1) &= \int_{0}^{t+\sigma_1} e^{A\sigma_2} Bu(t+\sigma_1-\sigma_2) d\sigma_2 + \sum_{k=1}^{m} \int_{0}^{t+\sigma_1} \Big(e^{A\sigma_2} N_k x(t+\sigma_1-\sigma_2) \\ &\times u_k(t+\sigma_1-\sigma_2)\Big) d\sigma_2 + \int\limits_{0}^{t+\sigma_1} e^{A\sigma_2} H(x(t+\sigma_1-\sigma_2) \otimes x(t+\sigma_1-\sigma_2)) d\sigma_2. \end{aligned}$$

We substitute for $x(t+\sigma_1)$ using the above equation, and $z(t+\sigma_1)$ using (4.26), which gives rise to the following expression:

$$\begin{aligned} z(t) &= \int_{\infty}^{0} e^{A^T\sigma_1} C^T u_d(t+\sigma_1) d\sigma_1 + \sum_{k=1}^{m} \int_{\infty}^{0} \int_{\infty}^{0} e^{A^T\sigma_1} N_k^T \\ &\times e^{A^T\sigma_2} C^T u_d(t+\sigma_1+\sigma_2) u_k(t+\sigma_1) d\sigma_1 d\sigma_2 + \int_{\infty}^{0} \int_{0}^{t+\sigma_1} \int_{\infty}^{0} e^{A^T\sigma_1} \\ &\times \mathcal{H}^{(2)} \Big(e^{A\sigma_2} B \otimes e^{A^T\sigma_3} C^T\Big) u(t+\sigma_1-\sigma_2) u_d(t+\sigma_1+\sigma_3) d\sigma_1 d\sigma_2 d\sigma_3 + \cdots . \end{aligned} \tag{4.27}$$

By repeatedly substituting for the state x and the dual state z, we derive the Volterra series for the dual system, although the notation becomes much more complicated. Carefully inspecting the kernels of the Volterra series of the dual system, we define the observability mapping $\bar{Q}$, similar to the reachability mapping, as follows:

$$\bar{Q} = [\bar{Q}_1,\ \bar{Q}_2,\ \bar{Q}_3, \ldots], \tag{4.28}$$

in which

$$\begin{aligned}
\bar{Q}_1(t_1) &= e^{A^Tt_1}C^T,\\
\bar{Q}_2(t_1,t_2) &= e^{A^Tt_2}\left[N_1^T,\dots,N_m^T\right]\left(I_m\otimes\bar{Q}_1(t_1)\right),\\
\vdots\quad &\qquad \vdots\\
\bar{Q}_i(t_1,\dots,t_i) &= e^{A^Tt_i}\Big[\mathcal{H}^{(2)}\left[\bar{P}_1(t_1)\otimes\bar{Q}_{i-2}(t_2,\dots,t_{i-1}),\dots,\bar{P}_{i-2}(t_1,\dots,t_{i-2})\otimes\bar{Q}_1(t_{i-1})\right],\\
&\qquad\qquad \left[N_1^T,\dots,N_m^T\right]\left(I_m\otimes\bar{Q}_{i-1}(t_1,\dots,t_{i-1})\right)\Big],\quad \forall\, i\geq 3.
\end{aligned}$$

where $\bar{P}_i(t_1,\dots,t_i)$ are defined in (4.17). Based on the above observability mapping, we define the observability Gramian Q of the QB system as

$$Q=\sum_{i=1}^{\infty}Q_i \quad\text{with}\quad Q_i=\int_0^{\infty}\cdots\int_0^{\infty}\bar{Q}_i\bar{Q}_i^T\,dt_1\cdots dt_i. \tag{4.29}$$

Analogous to the reachability Gramian, we next show a relation between the observability Gramian and the solution of a generalized quadratic Lyapunov equation.

Theorem 4.3:
Consider the QB system (4.14) with a stable matrix A, and let Q, defined in (4.29), be the observability Gramian of the system and assume it exists. Then, the Gramian Q satisfies the following Lyapunov equation:

$$A^TQ+QA+\mathcal{H}^{(2)}(P\otimes Q)(\mathcal{H}^{(2)})^T+\sum_{k=1}^{m}N_k^TQN_k+C^TC=0, \tag{4.30}$$

where P is the reachability Gramian of the system, i.e., the solution of the generalized quadratic Lyapunov equation (4.19). ♢

Proof. The proof of the above theorem is analogous to the proof of Theorem 4.2; therefore, for the brevity, we skip it. □

Remark 4.4:
As one would expect, the Gramians for QB systems reduce to the Gramians for bilinear systems (see, e.g., [26] or Chapter 3 of this thesis) if the quadratic term is zero, i.e., $H=0$. ♢

Furthermore, it will also be interesting to look at a truncated versions of the Gramians of the QB system based on the leading kernels of the Volterra series. We call a truncated version of the Gramians *truncated* Gramians of QB systems. For this, let us consider approximate reachability and observability mappings as follows:

$$\widetilde{P}_{\mathcal{T}}=\left[\widetilde{P}_1,\widetilde{P}_2,\widetilde{P}_3\right],\qquad \widetilde{Q}_{\mathcal{T}}=\left[\widetilde{Q}_1,\widetilde{Q}_2,\widetilde{Q}_3\right],$$

where

$$\begin{aligned}
\widetilde{P}_1(t_1) &= e^{At_1}B, & \widetilde{Q}_1(t_1) &= e^{A^Tt_1}C^T,\\
\widetilde{P}_2(t_1,t_2) &= e^{At_2}\left[N_1,\ldots,N_m\right]\left(I_m\otimes\widetilde{P}_1(t_1)\right),\\
\widetilde{Q}_2(t_1,t_2) &= e^{A^Tt_2}\left[N_1^T,\ldots,N_m^T\right]\left(I_m\otimes\widetilde{Q}_1(t_1)\right),\\
\widetilde{P}_3(t_1,t_2,t_3) &= e^{At_3}H(\widetilde{P}_1(t_1)\otimes\widetilde{P}_1(t_2)), & \widetilde{Q}_3(t_1,t_2,t_3) &= e^{A^Tt_3}\mathcal{H}^{(2)}(\widetilde{P}_1(t_1)\otimes\widetilde{Q}_1(t_2)).
\end{aligned}$$

Then, one can define the truncated reachability and observability Gramians in a similar fashion as the Gramians of the system:

$$P_{\mathcal{T}} = \sum_{i=1}^{3}\widehat{P}_i, \quad \text{where} \quad \widehat{P}_i = \int_0^\infty \widetilde{P}_i(t_1,\ldots,t_i)\widetilde{P}_i^T(t_1,\ldots,t_i)dt_1\cdots dt_i, \tag{4.31a}$$

$$Q_{\mathcal{T}} = \sum_{i=1}^{3}\widehat{Q}_i, \quad \text{where} \quad \widehat{Q}_i = \int_0^\infty \widetilde{Q}_i(t_1,\ldots,t_i)\widetilde{Q}_i^T(t_1,\ldots,t_i)dt_1\cdots dt_i, \tag{4.31b}$$

respectively. Similar to the Gramians P and Q, in the following we derive the relation between these truncated Gramians and the solutions of the Lyapunov equations.

Corollary 4.5:
Let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be the truncated Gramians of the QB system as defined in (4.31), and assume that the matrix A in the QB system (4.1) is stable. Then, $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ satisfy the following Lyapunov equations:

$$AP_{\mathcal{T}} + P_{\mathcal{T}}A^T + H(\widehat{P}_1\otimes\widehat{P}_1)H^T + \sum_{k=1}^{m}N_k\widehat{P}_1N_k^T + BB^T = 0, \quad \text{and} \tag{4.32a}$$

$$A^TQ_{\mathcal{T}} + Q_{\mathcal{T}}A + \mathcal{H}^{(2)}(\widehat{P}_1\otimes\widehat{Q}_1)(\mathcal{H}^{(2)})^T + \sum_{k=1}^{m}N_k^T\widehat{Q}_1N_k + C^TC = 0, \tag{4.32b}$$

respectively, where P_1 and Q_1 are solutions to the following Lyapunov equations:

$$A\widehat{P}_1 + \widehat{P}_1A^T + BB^T = 0, \quad \text{and} \tag{4.33}$$

$$A^T\widehat{Q}_1 + \widehat{Q}_1A + C^TC = 0, \quad \text{respectively.} \tag{4.34}$$

◊

Proof. We begin by showing the relation between the truncated reachability Gramian $P_{\mathcal{T}}$ and the solutions of Lyapunov equation. First, note that the first two terms of the reachability Gramian P (4.31a) and the truncated reachability Gramian $P_{\mathcal{T}}$ (4.18) are the same, i.e., $\widehat{P}_1 = P_1$ and $\widehat{P}_2 = P_2$. Hence, $\widehat{P}_1$ and $\widehat{P}_2$ are the unique solutions of the following Lyapunov equations for a stable matrix A:

$$A\widehat{P}_1 + \widehat{P}_1 A^T + BB^T = 0, \quad \text{and} \tag{4.35a}$$

$$A\widehat{P}_2 + \widehat{P}_2 A^T + \sum_{k=1}^{m} N_k \widehat{P}_1 N_k^T = 0. \tag{4.35b}$$

Now, we consider the third term in the summation (4.31a). This is

$$\begin{aligned}
P_3 &= \int_0^\infty \int_0^\infty \int_0^\infty \widetilde{P}_3(t_1,t_2,t_3)\widetilde{P}_3^T(t_1,t_2,t_3)dt_1dt_2dt_3 \\
&= \int_0^\infty \int_0^\infty \int_0^\infty e^{At_3} H(\widetilde{P}_1(t_1)\widetilde{P}^T(t_1) \otimes \widetilde{P}_1(t_2)\widetilde{P}^T(t_2))H^T e^{A^Tt_3}dt_1dt_2dt_3 \\
&= \int_0^\infty e^{At_3} H \left(\left(\int_0^\infty \widetilde{P}_1(t_1)\widetilde{P}^T(t_1)dt_1 \right) \otimes \left(\int_0^\infty \widetilde{P}_1(t_2)\widetilde{P}^T(t_2)dt_2 \right) \right) H^T e^{A^Tt_3} dt_3 \\
&= \int_0^\infty e^{At_3} H \left(\widehat{P}_1 \otimes \widehat{P}_1 \right) H^T e^{A^Tt_3} dt_3.
\end{aligned}$$

Furthermore, we use the relation between the above integral representation and the solution of Lyapunov equation to show that $\widehat{P}_3$ solves:

$$A\widehat{P}_3 + \widehat{P}_3 A^T + H(\widehat{P}_1 \otimes \widehat{P}_1)H^T = 0. \tag{4.36}$$

Summing (4.35a), (4.35b) and (4.36) yields

$$AP_{\mathcal{T}} + \mathcal{P}_{\mathcal{T}} A^T + H(\widehat{P}_1 \otimes \widehat{P}_1) + \sum_{k=1}^{m} N_k \widehat{P}_1 N_k + BB^T = 0. \tag{4.37}$$

Analogously, we can show that $Q_{\mathcal{T}}$ solves (4.32b). This concludes the proof. □

We will investigate the advantages of these truncated Gramians in the model reduction framework in the later part of this chapter in detail. Since our primary aim of introducing the Gramians and its truncated version is to use them in the balancing-type model reduction framework, it is important to investigate connections between the proposed Gramians and energy functionals, namely, controllability and observability energy functionals. Also, we show how the definiteness of the Gramians are related to reachability and observability of the QB systems. We start by establishing the conditions under which the Gramians approximate the energy functionals of the QB system, in certain quadratic forms.

4.3.3. Comparison of energy functionals with Gramians

By using Theorem 2.20, we obtain the following nonlinear partial differential equation, whose solution gives the controllability energy functional for the QB system:

$$\begin{aligned}&\frac{\partial E_c}{\partial x}(Ax + H(x\otimes x)) + (Ax + H(x\otimes x))^T\frac{\partial E_c}{\partial x}^T\\ &\quad+\frac{\partial E_c}{\partial x}\big([N_1,\dots,N_m](I_m\otimes x) + B\big)\big([N_1,\dots,N_m](I_m\otimes x) + B\big)^T\frac{\partial E_c}{\partial x}^T = 0,\end{aligned} \tag{4.38}$$

with $E_c(0) = 0$. For nonlinear systems, the energy functionals are rather complicated nonlinear functions. Thus, we aim at providing some bounds between the certain quadratic form of the proposed Gramians for QB systems and energy functionals. For the controllability energy functional, we extend the reasoning given in Theorem 3.12 for bilinear systems.

Theorem 4.6:
Consider a controllable QB system (4.14) with a stable matrix A. Let $P > 0$ be its reachability Gramian, which exists and is the unique definite solution of the quadratic Lyapunov equation (4.19), and $E_c(x)$ denote the controllability energy functional of the QB system, solving (4.38). Then, there exists a neighborhood W of 0 such that

$$E_c(x) \geq \frac{1}{2}x^TP^{-1}x, \text{ where } x\in W(0). \qquad \lozenge$$

Proof. Consider a state x_0 and let a control input $u = u_0 : (-\infty, 0] \to \mathbb{R}^m$, which minimizes the input energy in the definition of $E_o(x_0)$ and steers the system from 0 to x_0. Now, we consider the time-varying homogeneous nonlinear differential equation

$$\dot{\phi} = \left(A + H(\phi\otimes I) + \sum_{k=1}^{m} N_k u_k(t)\right)\phi =: A_u\phi(t), \tag{4.39}$$

and its fundamental solution $\Phi_u(t,\tau)$. The system (4.39) can thus be interpreted as a time-varying system. The reachability Gramian of the time-varying control system [120, 128] $\dot{x} = A_u x(t) + Bu(t)$ is given by

$$P_u = \int_{-\infty}^{0}\Phi(0,\tau)BB^T\Phi(0,\tau)^T d\tau.$$

The input u also steers the time-varying system from 0 to x_0. Moreover, the minimum input energy required to steer the time-varying system is equal to $x_0^TP_u^{\#}x_0$, where $P_u^{\#}$ denotes the Moore-Penrose pseudo inverse of P_u. Thus, we have

$$\|u\|_{L_2}^2 \geq \frac{1}{2}x_0^TP_u^{\#}x_0.$$

An alternative way to determine P_u can be given by

$$P_u = \int_0^\infty \widetilde{\Phi}(t,0)^T BB^T \widetilde{\Phi}(t,0)dt,$$

where $\widetilde{\Phi}$ is the fundamental solution of the following differential equation

$$\dot{\widetilde{\Phi}} = \left(A^T + \mathcal{H}^{(2)}(x(-t)\otimes I) + \sum_{k=1}^{m} N_k^T u_k(-t)\right)\widetilde{\Phi} \quad \text{with} \quad \Phi(t,t) = I, \tag{4.40}$$

and $x(t)$ is the solution of

$$\dot{x}(t) = Ax(t) + H(x\otimes x) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t).$$

Then, we define $\eta(t)$, satisfying $\eta(t) = \widetilde{\Phi}(t,0)x_0$. Since it is assumed that the QB system is reachable, the state x_0 can be reached by using a finite input energy, i.e., $\|u\|_{L_2} < \infty$. Hence, the input $u(t)$ is a square-integrable function over $t \in (-\infty, 0]$ and so is $x(t)$. This implies that $\lim\limits_{t\to\infty} \eta(t) \to 0$, provided A is stable. Thus, we have

$$\begin{aligned}
x_0^T P x_0 &= -\int_0^\infty \frac{d}{dt}\left(\eta(t)^T P \eta(t)\right) dt \\
&= -\int_0^\infty \eta(t)^T \left(\left(A + H(x(-t)\otimes I) + \sum_{k=1}^{m} N_k u_k(-t)\right) P\right. \\
&\qquad \left. + P\left(A^T + \mathcal{H}^{(2)}(x(-t)\otimes I) + \sum_{k=1}^{m} N_k^T u_k(-t)\right)\right)\eta(t)dt \\
&= -\int_0^\infty \eta(t)^T \left(AP + PA^T + H(P\otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T\right)\eta(t) \\
&\qquad + \left(H(P\otimes P) - H(x(-t)\otimes I)P - P\mathcal{H}^{(2)}(x(-t)\otimes I)\right. \\
&\qquad \left. + \sum_{k=1}^{m}\left(N_k P N_k - PN_k^T u_k(-t) - N_k^T P u_k(-t)\right)\right)\eta(t)dt.
\end{aligned}$$

Now, we have

$$\begin{aligned}
-\int_0^\infty \eta(t)^T \left(AP + PA^T + H(P\otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T\right)\eta(t) \\
= \int_0^\infty \eta(t)^T BB^T \eta(t) = x_0^T P_u x_0.
\end{aligned}$$

Hence, if

$$\int_0^\infty \eta(t)^T\Big(H(P\otimes P)H^T - H(x(-t)\otimes I)P - P\mathcal{H}^{(2)}(x(-t)\otimes I) + \sum_{k=1}^{m}\left(N_kPN_k - PN_k^Tu_k(-t) - N_k^TPu_k(-t)\right)\Big)\eta(t)dt \geq 0, \tag{4.41}$$

then $x_0^TPx_0 \geq x_0^TP_ux_0$. Further, if x_0 lies in the neighborhood of the origin, i.e., $x_0 \in W(0)$, then a small input u is sufficient to steer the system from 0 to x_0 and $x(t) \in W(0)$ for $t \in (-\infty, 0]$, which ensures that the relation (4.41) holds for all $x_0 \in W(0)$. Therefore, we have $x_0^TP^{-1}x_0 \leq x_0^TP_u^{-1}x_0$ if $x_0 \in W(0)$. □

Similarly, we next show an upper bound for the observability energy functional for the QB system in terms of the observability Gramian (in the quadratic form).

Theorem 4.7:
Consider a QB system (4.14) with $B \equiv 0$ and an initial condition x_0, and let E_o be the observability energy functional. Let $P > 0$ and $Q \geq 0$ be solutions to the generalized Lyapunov equations (4.19) and (4.30), respectively. Then, for small inputs, there exists a neighborhood $\widetilde{W}$ of the origin such that

$$E_o(x_0) \leq \frac{1}{2}x^TQx, \quad \text{where} \quad x \in \widetilde{W}(0). \qquad \lozenge$$

Proof. Using the observability energy functional definition (see Definition 2.21), we have

$$E_o(x_0) = \max_{\substack{u\in\mathcal{B}_{(\alpha,\beta)} \\ x(0)=x_0, x(\infty)=0}} \frac{1}{2}\int_0^\infty \widetilde{E}_o(x_0,u)dt, \tag{4.42}$$

where $\mathcal{B}_{(\alpha,\beta)} \overset{\text{def}}{=} \{u \in L_2^m[0,\infty), \|u\|_{L_2} \leq \alpha, \|u\|_{L_\infty} \leq \beta\}$ and $\widetilde{E}_o(x_0,u) := \|y(t)\|_2$. Thus, we have

$$\widetilde{E}_o(x_0,u) = \|y(t)\|_2 = \|Cx(t)\|_2 = x(t)^TC^TCx(t).$$

Substituting for C^TC from (4.30), we obtain

$$\widetilde{E}_o(x_0,u) = -2x(t)^TQAx(t) - x(t)^T\mathcal{H}^{(2)}(P\otimes Q)\left(\mathcal{H}^{(2)}\right)^Tx(t) - \sum_{k=1}^{m}x(t)^TN_k^TQN_kx(t).$$

Next, we substitute for Ax from (4.14) (with $B = 0$) to have

$$\begin{aligned}
\widetilde{E}_o(x_0, u) &= -2x(t)^TQ\dot{x}(t) + 2x(t)^TQHx(t) \otimes x(t) + 2\sum_{k=1}^{m} x(t)^TQN_kx(t)u_k(t) \\
&\qquad - x(t)^T\mathcal{H}^{(2)}\left(P \otimes Q\right)\left(\mathcal{H}^{(2)}\right)^T x(t) - \sum_{k=1}^{m} x(t)^TN_k^TQN_kx(t) \\
&= -\frac{d}{dt}\left(x(t)^TQx(t)\right) + x(t)^T\Big(QH(I \otimes x(t)) + (I \otimes x(t)^T)H^TQ \\
&\quad + \sum_{k=1}^{m}(QN_k + N_k^TQ)u_k(t) - \mathcal{H}^{(2)}(P \otimes Q)\big(\mathcal{H}^{(2)}\big)^T - \sum_{k=1}^{m} N_k^TQN_k\Big)x(t).
\end{aligned}$$

This gives

$$\begin{aligned}
E_o(x_0) &= \max_{\substack{u\in\mathcal{B}_{(\alpha,\beta)} \\ x(0)=x_0,x(\infty)=0}} \frac{1}{2}\int_0^\infty \widetilde{E}_o(x_0, u)dt \\
&= \frac{1}{2}x_0^TQx_0 + \max_{\substack{u\in\mathcal{B}_{(\alpha,\beta)} \\ x(0)=x_0,x(\infty)=0}} \frac{1}{2}\int_0^\infty x(t)^T\left(R_H(x,u) + \sum_{k=1}^{m} R_{N_k}(x,u)\right)x(t)dt,
\end{aligned}$$

where

$$\begin{aligned}
R_H(x,u) &:= QH(I \otimes x) + (I \otimes x)H^TQ - \mathcal{H}^{(2)}(P \otimes Q)\big(\mathcal{H}^2\big)^T, \\
R_{N_k}(x,u) &:= \left(QN_ku_k + N_k^TQu_k - N_k^TQN_k\right).
\end{aligned}$$

First, note that if for a vector v, $v^TN_k^TQN_kv = 0$, then $QN_kv = 0$. Therefore, there exist inputs u for which $\|u\|_{L_\infty}$ is small, ensuring $R_{N_k}(x,u)$ is a negative semidefinite. Similarly, if for a vector w, $w^T\mathcal{H}^{(2)}(P \otimes Q)\big(\mathcal{H}^2\big)^T w = 0$ and $P > 0$, then $(I \otimes Q)\big(\mathcal{H}^2\big)^T w = 0$. Using tensor-matrix product properties (2.31), it can be shown that $QH(w \otimes I) = QH(I \otimes w) = 0$, when H is a symmetric structure. Now, we consider an initial condition x_0 lies in the small neighborhood of the origin and $u \in \mathcal{B}_{(\alpha,\beta)}$, ensuring that the resulting trajectory $x(t)$ for all time t is such that $R_H(x,u)$ is a negative semidefinite. Finally, we get

$$E_o(x_0) - \frac{1}{2}x_0^TQx_0 \leq 0,$$

for x_0 lies in the neighborhood of the origin and for the inputs u, which have small L_2-norm as well as L_∞ norm. This concludes the proof. □

Until this point, we have proven that in the neighborhood of the origin, the energy functionals of the QB system can be approximated by the Gramians in the quadratic form. However, one can also prove similar bounds for the energy functionals using the truncated Gramians for QB systems (defined in Corollary 4.5). We summarize this in the following corollary.

Corollary 4.8:
Consider the system (4.14), having a stable matrix A, to be locally reachable and observable. Let $E_c(x)$ and $E_o(x)$ be controllability and observability energy functionals of the system, respectively, and the truncated Gramians $P_{\mathcal{T}} > 0$ and $Q_{\mathcal{T}} > 0$ be solutions to the Lyapunov equations as shown in Corollary 4.5. Furthermore, assume that at least one of matrices H or N_k is of full rank. Then,

(i) there exists a neighborhood $W_{\mathcal{T}}$ of the origin such that

$$E_c(x) \geq \frac{1}{2} x^T P_{\mathcal{T}}^{-1} x, \text{ where } x \in W_{\mathcal{T}}(0).$$

(ii) Moreover, there also exists a neighborhood $\widetilde{W}_{\mathcal{T}}$ of the origin, where

$$E_o(x) \leq \frac{1}{2} x^T Q_{\mathcal{T}} x, \text{ where } x \in \widetilde{W}_{\mathcal{T}}(0). \qquad \diamondsuit$$

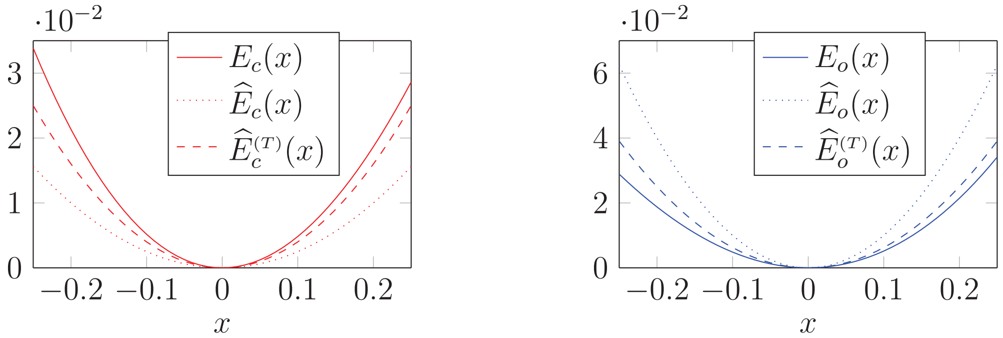
In what follows, we illustrate the above bounds using Gramians and truncated Gramians by considering a scalar dynamical system, where A, H, N, B, C are scalars, and are denoted by a, h, n, b, c, respectively.

Example 4.3:
Consider a scalar system (a, h, n, b, c), where $a < 0$ (stability) and nonzero h, b, c. For simplicity, we take $n = 0$ so that we can easily obtain analytic expressions for the controllability and observability energy functionals, denoted by $E_c(x)$ and $E_o(x)$, respectively. Assume that the system is reachable on $\mathbb{R}$. Then, $E_c(x)$ and $E_o(x)$ can be determined via solving partial differential equations (2.23) and (2.24) (with $g(x) = 0$), respectively. These are:

$$E_c(x) = -\left(ax^2 + \tfrac{2}{3}hx^3\right)\frac{1}{b^2}, \qquad E_o(x) = -\frac{c^2}{2h}\left(x - \frac{a}{h}\log\left(\frac{a + hx}{a}\right)\right).$$

The quadratic approximations of these energy functionals by using the Gramians, are:

$$\begin{aligned} \widehat{E}_c(x) &= \frac{x^2}{2P} \quad \text{with} \quad P = -\frac{a + \sqrt{a^2 - h^2 b^2}}{h^2}, \\ \widehat{E}_o(x) &= \frac{Qx^2}{2} \quad \text{with} \quad Q = -\frac{c^2}{2a + h^2 P}, \end{aligned}$$



(a) Comparison of the controllability energy functional and its approximations.

(b) Comparison of the observability energy functional and its approximations.

Figure 4.1.: Comparison of exact energy functionals of a QB system with approximated energy functionals via the Gramians and truncated Gramians.

and the approximations in terms of the truncated Gramians are:

$$\widehat{E}_c^{(\mathcal{T})}(x) = \frac{x^2}{2P_{\mathcal{T}}} \quad \text{with} \quad P_{\mathcal{T}} = -\frac{h^2b^4 + 4a^2b^2}{8a^3},$$
$$\widehat{E}_o^{(\mathcal{T})}(x) = \frac{Q_{\mathcal{T}}x^2}{2} \quad \text{with} \quad Q_{\mathcal{T}} = -\frac{h^2b^2c^2 + 4a^2c^2}{8a^3}.$$

In order to compare these functionals, we set $a = -2, b = c = 2$ and $h = 1$ and plot the resulting energy functionals in Figure 4.1.

Clearly, Figure 4.1 illustrates the lower and upper bounds for the controllability and observability energy functionals, respectively, at least locally. Moreover, we observe that the bounds for the energy functionals, given in terms of truncated Gramians are closer to the actual energy functionals of the system in the small neighborhood of the origin. ◊

So far, we have shown the bounds for the energy functionals in terms of the Gramians of the QB system. In order to prove those bounds, it is assumed that P is a positive definite. However, this assumption might not be fulfilled for many QB systems, especially arising from semi-discretization of PDEs. Therefore, our next objective is to provide another interpretation of the proposed Gramians, that is, the connection of Gramians with reachability and observability of the system. For the observability energy functional, we consider the output y of the following *homogeneous* QB system:

$$\begin{aligned}\dot{x}(t) &= Ax + Hx(t) \otimes x(t) + \sum_{k=1}^{m} N_k x(t) u_k(t),\\ y(t) &= Cx(t), \qquad x(0) = x_0,\end{aligned} \tag{4.43}$$

as considered for bilinear systems in [26, 73]. However, it might also be possible to consider an *inhomogeneous* system by setting the control input u completely zero, as shown in [114]. We first investigate how the proposed Gramians are related to reachability and observability of the QB systems, analogues to derivation for bilinear systems in [26].

Theorem 4.9:

(a) Consider the QB system (4.14), and assume the reachability Gramian P to be the solution of (4.19). If the system is steered from 0 to x_0, where $x_0 \notin \operatorname{range}(P)$, then $E_c(x_0) = \infty$ for all input functions u.

(b) Furthermore, consider the *homogeneous* QB system (4.43) and assume $P > 0$ and Q to be the reachability and observability Gramians of the QB system, which are solutions of (4.19) and (4.30), respectively. If the initial state satisfies $x_0 \in \ker(Q)$, then $E_o(x_0) = 0$. ♢

Proof. (a) By assumption, P satisfies

$$AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0. \tag{4.44}$$

Next, we consider a vector $v \in \ker(P)$ and multiply the above equation from the left and right with v^T and v, respectively, to obtain

$$\begin{aligned}0 &= v^TAPv + v^TPA^Tv + v^TH(P \otimes P)H^Tv + \sum_{k=1}^{m} v^TN_kPN_k^Tv + v^TBB^Tv\\ &= v^TH(P \otimes P)H^Tv + \sum_{k=1}^{m} v^TN_kPN_k^Tv + v^TBB^Tv.\end{aligned}$$

This implies $B^Tv = 0$, $PN_k^Tv = 0$ and $(P \otimes P)H^Tv = 0$. From (4.44), we thus obtain $PA^Tv = 0$. Now, we consider an arbitrary state vector $x(t)$, which is the solution of (4.14) at time t for any given input function u. If $x(t) \in \operatorname{range}(P)$ for some t, then we have

$$\dot{x}(t)^Tv = x(t)^TA^Tv + (x(t) \otimes x(t))^T H^Tv + \sum_{k=1}^{m} u_k(t)x(t)^TN_k^Tv + u(t)B^Tv = 0.$$

The above relation indicates that $\dot{x}(t) \perp v$ if $v \in \ker(P)$ and $x(t) \in \operatorname{range}(P)$. It shows that $\operatorname{range}(P)$ is invariant under the dynamics of the system. Since the initial condition 0 lies in $\operatorname{range}(P)$, $x(t) \in \operatorname{range}(P)$ for all $t \geq 0$. This reveals that if the final state $x_0 \notin \operatorname{range}(P)$, then it cannot be reached from 0; hence, $E_c(x_0) = \infty$.

(b) Following the above discussion, we can show that $(I \otimes Q)\left(\mathcal{H}^{(2)}\right)^T \ker(Q) = 0$, $QN_k \ker(Q) = 0$, $QA \ker(Q) = 0$, and $C \ker(Q) = 0$. Let $x(t)$ denote the solution of the homogeneous system at time t. If $x(t) \in \ker(Q)$ and a vector $\widetilde{v} \in \operatorname{range}(Q)$, then we have

$$\begin{aligned}\widetilde{v}^T \dot{x}(t) &= \underbrace{\widetilde{v}Ax(t)}_{=0} + \widetilde{v}^T H(x(t) \otimes x(t))) + \sum_{k=1}^{m} \underbrace{\widetilde{v}^T N_k x(t) u_k(t)}_{=0} \\ &= x(t)^T \mathcal{H}^{(2)}(x(t) \otimes \widetilde{v}) = \underbrace{x(t)^T \mathcal{H}^{(2)}(I \otimes \widetilde{v})}_{=0} x(t) = 0.\end{aligned}$$

This implies that if $x(t) \in \ker(Q)$, then $\dot{x}(t) \in \ker(Q)$. Therefore, if the initial condition $x_0 \in \ker(Q)$, then $x(t) \in \ker(Q)$ for all $t \geq 0$, resulting in $y(t) = C \underbrace{x(t)}_{\in \ker(Q)} = 0$; hence, $E_o(x_0) = 0$. □

The above theorem suggests that the state components, belonging to $\ker(P)$ or $\ker(Q)$, do not play a major role as far as the system dynamics are concerned. This shows that the states, which belong to $\ker(P)$, are uncontrollable, and similarly, the states, lying in $\ker(Q)$ are unobservable once the uncontrollable states are removed. Furthermore, we have shown in Theorems 4.6 and 4.7 the lower and upper bounds for the controllability and observability energy functions in the quadratic form of the Gramians P and Q of QB systems (at least in the neighborhood of the origin). This coincides with the concept of balanced truncation model reduction which aims at eliminating weakly controllable and weakly observable state components. Such states are corresponding to zero or small singular values of P and Q. In order to find these states simultaneously, we utilize the balancing tools similar to the linear case; see, e.g., [6, 7]. For this, one needs to determine the Cholesky factors of the Gramians as $P =: S^T S$ and $Q =: R^T R$, and compute the SVD of $SR^T =: U\Sigma V^T$, resulting in a transformation matrix $T = S^T U \Sigma^{-\frac{1}{2}}$. Using the matrix T, we obtain an equivalent QB system

$$\begin{aligned}\dot{\widetilde{x}}(t) &= \widetilde{A}\widetilde{x}(t) + \widetilde{H}\widetilde{x}(t) \otimes \widetilde{x}(t) + \sum_{k=1}^{m} \widetilde{N}_k \widetilde{x}(t) u_k(t) + \widetilde{B}u(t), \\ y(t) &= \widetilde{C}\widetilde{x}(t), \quad \widetilde{x}(0) = 0,\end{aligned} \tag{4.45}$$

where $\widetilde{A} = T^{-1}AT$, $\widetilde{H} = T^{-1}H(T \otimes T)$, $\widetilde{N}_k = T^{-1}N_k T$, $\widetilde{B} = T^{-1}B$, $\widetilde{C} = CT$. Then, the above transformed system (4.45) is a balanced system, as the Gramians $\widetilde{P}$ and $\widetilde{Q}$ of the system (4.45) are equal and diagonal, i.e., $\widetilde{P} = \widetilde{Q} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$. The attractiveness of the balanced system is that it allows us to find state components corresponding to small singular values of both $\widetilde{P}$ and $\widetilde{Q}$. If $\sigma_{\widehat{n}} > \sigma_{\widehat{n}+1}$, for some $\widehat{n} \in \mathbb{N}$, then it is easy to see that states related to $\{\sigma_{\widehat{n}+1}, \ldots, \sigma_n\}$ are not only hard to control but also hard to observe; hence, they can be eliminated. In order to determine a reduced-order system of order $\widehat{n}$, we partition $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}$ and $T^{-1} = \begin{bmatrix} S_1^T & S_2^T \end{bmatrix}^T$, where $T_1, S_1^T \in \mathbb{R}^{n \times \widehat{n}}$, and define the reduced-order system's realization as follows:

$$\widehat{A} = S_1 A T_1, \quad \widehat{H} = S_1 H (T_1 \otimes T_1), \quad \widehat{N}_k = S_1 N_k T_1, \quad \widehat{B} = S_1 B, \quad \widehat{C} = C T_1, \tag{4.46}$$

which is generally a locally good approximate of the original system, though it is not a straightforward task to estimate the error occurring due to the truncation singular values unlike in the case of linear systems.

Up to now, we have proposed the Gramians for the QB systems and have showed their relations to energy functionals of the system which allows us to determine the reduced-order systems. Next, we discuss computational issues which one might face while utilizing the proposed Gramians in the MOR framework.

4.3.4. Computational issues

One of the major concerns in applying balanced truncation MOR is that it requires the solutions of two Lyapunov equations (4.19) and (4.30). These equations are quadratic in nature, which are not trivial to solve, and they appear to be computationally expensive. So far, it is not clear how to solve these generalized quadratic Lyapunov equation efficiently; however, under some assumptions, a fix point iteration scheme can be employed, which is based on the theory of convergent splitting presented in [52, 116]. This has been studied for solving generalized Lyapunov equation for bilinear systems in [51], wherein the proposed stationary method is as follows:

$$\mathcal{L}(X_i) = \mathcal{N}(X_{i-1}) - BB^T, \qquad i = 1, 2, \dots, \tag{4.47}$$

with $\mathcal{L}(X) = AX + XA^T$ and $\mathcal{N}(X_i) = -\sum_{k=1}^{m} N_k X_i N_k^T$. To perform this stationary iteration, we require the solution of a conventional Lyapunov equation at each iteration. Assuming $\sigma(A) \subset \mathbb{C}^-$ and spectral radius of $\mathcal{L}^{-1}\mathcal{N} < 1$, the iteration (4.47) linearly converges to a positive semidefinite solution X of the generalized Lyapunov equation for bilinear systems, which is

$$AX + XA^T + \sum_{k=1}^{m} N_k X N_k^T + BB^T = 0.$$

Later on, the efficiency of this iterative method was improved in [118] by utilizing tools for inexact solution of $Ax = b$. The main idea was to determine a low-rank factor of $\mathcal{N}(X_{i-1}) - BB^T$ by truncating the columns, corresponding to small singular values and to increase the accuracy of the low-rank solution of the linear Lyapunov equation (4.47) with each iteration. In total, this results in an efficient method to determine a low-rank solution of the generalized Lyapunov equation for bilinear systems with the desired tolerance. For detailed insights, we refer to [118].

One can utilize the same tools to determine the solutions of generalized quadratic-type Lyapunov equations. We begin with the inexact form equation, which on convergence gives the reachability Gramian, this is,

$$\mathcal{L}(X_i) = \Pi(X_{i-1}) - BB^T, \quad i = 1, 2, \ldots \tag{4.48}$$

where $\mathcal{L}(X) = AX + XA^T$ and $\Pi(X) = -H(X \otimes X)H^T - \sum_{k=1}^{m} N_k X N_k^T$. Similar to the bilinear case, if $\sigma(A) \subset \mathbb{C}^-$ and the spectral radius of $\mathcal{L}^{-1}\Pi < 1$, then the iteration (4.48) converges to a positive semidefinite solution of the generalized quadratic Lyapunov equation. Next, we determine a low-rank approximation of $\Pi(X) = -H(X \otimes X)H^T - \sum_{k=1}^{m} N_k X N_k^T$. For this, let us assume a low-rank product $X := FDF^T$, where $F \in \mathbb{R}^{n\times k}$ and a QR decomposition of $F := Q_F R_F$. We then perform an eigenvalue decomposition of the relatively small matrix $R_F D R_F^T := U\Sigma U^T$, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$ with $\sigma_j \geq \sigma_{j+1}$. Assuming there exists a scalar β such that

$$\sqrt{\sigma_{\beta+1}^2 + \cdots + \sigma_k^2} \leq \tau\sqrt{\sigma_1^2 + \cdots + \sigma_k^2},$$

where τ is a given tolerance, this gives us a low-rank approximation of X as:

$$X \approx \widetilde{F}\widetilde{D}\widetilde{F}^T,$$

where $\widetilde{F} = Q_F\widetilde{U}$ and $\widetilde{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_\beta)$. Following the short-hand notation, we denote $\widetilde{Z} = \mathcal{T}_\tau(Z)$, which gives the low-rank approximation of ZZ^T with the tolerance τ, i.e., $ZZ^T \approx \widetilde{Z}\widetilde{Z}^T$. Considering a low-rank factor of $X_{k-1} \approx Z_{k-1}Z_{k-1}^T$, the right-hand side of (4.48)

$$\begin{aligned}\Pi(X_{k-1}) - BB^T \approx -[H(Z_{k-1} \otimes Z_{k-1}), [N_1, \ldots, N_m]\, Z_{k-1}, B]& \\ \times [H(Z_{k-1} \otimes Z_{k-1}), [N_1, \ldots, N_m]\, Z_{k-1}, B]^T&\end{aligned}$$

can be replaced with its truncated version $\mathcal{T}_\tau(\Pi(X_{k-1}) - BB^T) =: -\mathbb{F}_k\mathbb{F}_k^T$ with the desired tolerance. This indicates that we now need to solve the following linear Lyapunov equation at each step:

$$AX_k + X_kA = -\mathbb{F}_k\mathbb{F}_k^T, \tag{4.49}$$

which can be solved very efficiently by using any of the recently developed low-rank solvers for Lyapunov equations; see, e.g., [36, 121].

Remark 4.10:
At step 7 of Algorithm 4.2, one can check the accuracy of solutions by measuring the relative changes in the solutions, i.e., $\dfrac{\|P_k - P_{k-1}\|}{\|P_k\|}$ and $\dfrac{\|Q_k - Q_{k-1}\|}{\|Q_k\|}$. When these relative changes are smaller than a *tolerance* level, e.g., the square root of the machine precision, then one can stop the iterations to have sufficiently accurate solutions of the quadratic Lyapunov equations. $\Diamond$

Algorithm 4.2: Iterative scheme to determine Gramians for QB systems.

```
  Input: System matrices $A, H, N_1, \ldots, N_m, B, C$ and tolerance $\tau$.
  Output: Low-rank factors of the Gramians: $Z_k$ $(P \approx Z_k Z_k^T)$ and
          $X_k$ $(Q \approx X_k X_k^T)$
1 Solve approximately $AM + MA^T + BB^T = 0$ for $P_1 \approx Z_1 Z_1^T$.
2 Solve approximately $A^T G + GA + C^T C = 0$ for $Q_1 \approx X_1 X_1^T$.
3 for $k = 2, 3, \ldots$ do
4     Determine low-rank factors:
          $\mathbb{B}_k = \mathcal{T}_\tau([H(Z_{k-1} \otimes Z_{k-1}), N_1 Z_{k-1}, \ldots, N_m Z_{k-1}, B])$,
          $\mathbb{C}_k = \mathcal{T}_\tau([\mathcal{H}^{(2)}(Z_{k-1} \otimes X_{k-1}), N_1^T X_{k-1}, \ldots, N_m^T X_{k-1}, C^T])$.
5     Solve approximately $AM + MA^T + \mathbb{B}_k \mathbb{B}_k^T = 0$ for $P_k \approx Z_k Z_k^T$.
6     Solve approximately $A^T G + GA + \mathbb{C}_k \mathbb{C}_k^T = 0$ for $Q_k \approx X_k X_k^T$.
7     if solutions are sufficiently accurate then
8         stop
```

Algorithm 4.3: An efficient way to perform Kronecker product.

```
  Input: $H$, $Z$.
  Output: $H_z := H(Z_i \otimes Z_i)$
1 Determine $\mathcal{Y} \in \mathbb{R}^{n_z \times n \times n}$ such that $\mathcal{Y}^{(2)} = Z_i^T \mathcal{H}^{(2)}$.
2 Determine $\mathcal{K} \in \mathbb{R}^{n \times n_z \times n_z}$ such that $\mathcal{K}^{(3)} = Z_i^T \mathcal{Y}^{(3)}$.
3 Then, $H_z = \mathcal{K}^{(1)}$.
```

Remark 4.11:
In Algorithm 4.2, we propose to determine the low-rank solutions of the Lyapunov equation at each intermediate step with the same tolerance. However, one can also consider to increase the tolerance adaptively for computing the low-rank solution of the Lyapunov equation with each iteration which probably can speed up even more, see [118] for the generalized Lyapunov equations for bilinear systems. $\Diamond$

In order to employ Algorithm 4.2, the right-hand side of the conventional Lyapunov equation (see step 3) requires the computation of $H(Z_i \otimes Z_i) =: \Gamma$ at each step, which is also computationally and memory-wise expensive. If $Z_i \in \mathbb{R}^{n \times n_z}$, then the direct multiplication of $Z_i \otimes Z_i$ would have complexity of $\mathcal{O}(n^2 \cdot n_z^2)$, leading to an unmanageable task for large-scale systems, even on modern computer architectures. However, it is shown in [25] that Γ can be determined efficiently by making use of the tensor multiplication properties, which are reported in Section 2.4. In Algorithm 4.3, we provide the procedure to compute Γ efficiently. This way, we can avoid determining the full matrix $Z_i \otimes Z_i$. Analogously, we can also compute the term $\mathcal{H}^{(2)}(Z_i \otimes X_i)$. Note that Algorithm 4.3 does not rely on any particular structure of the Hessian H. However, a QB system resulting from semi-discretization of PDEs usually leads to a Hessian, which has a special structure related to that particular PDE and the choice of the discretization method.

Therefore, we propose another efficient way to compute $H(Z_i \otimes Z_i)$ that utilizes a particular the sparsity structure of the Hessian, arising from the governing PDEs or ODEs. Generally, the term $H(x \otimes x)$ in the QB system (4.1) can be written as

$$H(x \otimes x) = \sum_{j=1}^{p} (\mathcal{A}^{(j)}x) \circ (\mathcal{B}^{(j)}x),$$

where $\circ$ denotes the Hadamard product, and $\mathcal{A}^{(j)}$ and $\mathcal{B}^{(j)}$ are sparse matrices, depending on the nonlinear operators in the underlying PDE and the discretization scheme, and p is generally a very small integer; for example, it is equal to 1 in case of Burgers' equations. Furthermore, using the ith rows of $\mathcal{A}^{(j)}$ and $\mathcal{B}^{(j)}$, we can construct the ith row of the Hessian:

$$H(i,:) = \sum_{j=1}^{p} \mathcal{A}^{(j)}(i,:) \otimes \mathcal{B}^{(j)}(i,:),$$

where $H(i,:)$, $\mathcal{A}^{(j)}(i,:)$ and $\mathcal{B}^{(j)}(i,:)$ represent the ith rows of the matrices H, $\mathcal{A}^{(j)}$ and $\mathcal{B}^{(j)}$, respectively. This clearly shows that there is a particular Kronecker product structure of the Hessian H, which can be used in order to determine $H(Z_i \otimes Z_i)$.

Example 4.4:
Here, we consider the Chafee-Infante equation, which is discretized over the spatial domain via a finite difference scheme. The MOR problem for this example will be considered in Section 4.4.2, where one can also find the governing equations and boundary conditions. For this particular example, the Hessian (after having rewritten the system into the QB form) is given by

$$\begin{aligned} H(i,:) &= -\frac{1}{2}e_i^n \otimes e_{k+i}^n - \frac{1}{2}e_{k+i}^n \otimes e_i^n, \quad i \in \{1,\dots,k\}, \\ H(i,:) &= -2(e_i^n \otimes e_i^n) + e_{i-k}^n \otimes \begin{bmatrix} X(i-k,:) & 0 \end{bmatrix} + \begin{bmatrix} X(i-k,:) & 0 \end{bmatrix} \otimes e_{i-k}^n, \\ &\qquad\qquad i \in \{k+1,\dots,n\}, \end{aligned}$$

where k is the number of grid points, $n = 2k$, and $H(i,:)$ is the ith row vector of the matrix H; $X(i,:)$ also denotes the ith row vector of the matrix $X \in \mathbb{R}^{k\times k}$ defined as

$$X = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix}. \tag{4.50}$$

◊

Algorithm 4.4: Chafee-Infante equation: an illustration of computing Kronecker product, utilizing Hessian structure of PDEs.

```
Input: $Z_i \in \mathbb{R}^{2k \times n_z}, X \in \mathbb{R}^{k \times k}$ (as defined in (4.50)).
Output: $H_z := H(Z_i \otimes Z_i)$
Compute $Z_x := XZ_i(1:k,:)$.
for $i = 1:k$ do
    $H_z(i,:) = -\frac{1}{2}Z_i(i,:) \otimes Z_i(k+i,:) - \frac{1}{2}Z)i(i,:) \otimes Z_i(k+i,:)$,
    $H_z(k+i,:) = -2\left(Z_i(i,:) \otimes Z_i(i,:)\right) + Z_i(i,:) \otimes Z_x(i,:) + Z_x(i,:) \otimes Z(i,:)$.
```

The Kronecker product representation of each row of the matrix H allows us to compute the rows of $H_v := H(V \otimes V)$ by selecting only the required rows of V. This way, we can determine H_v efficiently in large-scale, sparse settings, and then multiply with W^T to obtain the desired reduced Hessian. We describe this procedure i in Algorithm 4.4 that shows how one can determine the reduced Hessian for the Chafee-Infante example.

Furthermore, computation of a reduced Hessian $\widehat{H} := W^T H(V \otimes V)$, where V and W are the projection matrices, involves the Kronecker products. In this case as well, one should avoid the explicit computation of $V \otimes V$. To determine $\widehat{H}$, one can make use of either Algorithm 4.3 or can exploit the structure of the Hessian as illustrated by the Chafee-Infante equation. We first outline the steps in Algorithm 4.5 which utilizes the properties of tensor products.

If one makes use of the Kronecker product structure of the Hessian H, then one first needs to determine $H_v = H(V \otimes V)$, followed by simply multiplication of W^T, yielding $\widehat{H} = W^T H_v$. In order to show the effectiveness of the proposed methodology that uses the special Kronecker product structure of the Hessian H, we compute $\widehat{H} = W^T H(V \otimes V)$ for different orders of original and reduced-order systems and show the required CPU-time to compute it in Figure 4.2. The simulations were performed on a board with 4 Intel® Xeon® E7-8837 CPUs with a 2.67-GHz clock speed using

Algorithm 4.5: Computation of the Hessian of the reduced QB system [25].

```
Input: $H, V \in \mathbb{R}^{n \times \widehat{n}}, W \in \mathbb{R}^{n \times \widehat{n}}$.
Output: $\widehat{H} := W^T H(V \otimes V)$
Determine $\mathcal{Y} \in \mathbb{R}^{\widehat{n} \times n \times n}$, such that $\mathcal{Y}^{(1)} = W^T H$.
Determine $\mathcal{Z} \in \mathbb{R}^{\widehat{n} \times \widehat{n} \times n}$, such that $\mathcal{Z}^{(2)} = V^T \mathcal{Y}^{(2)}$.
Determine $\mathcal{X} \in \mathbb{R}^{\widehat{n} \times \widehat{n} \times \widehat{n}}$, such that $\mathcal{X}^{(3)} = V^T \mathcal{Z}^{(3)}$.
Then, the reduced Hessian is $\widehat{H} = \mathcal{X}^{(1)}$.
```

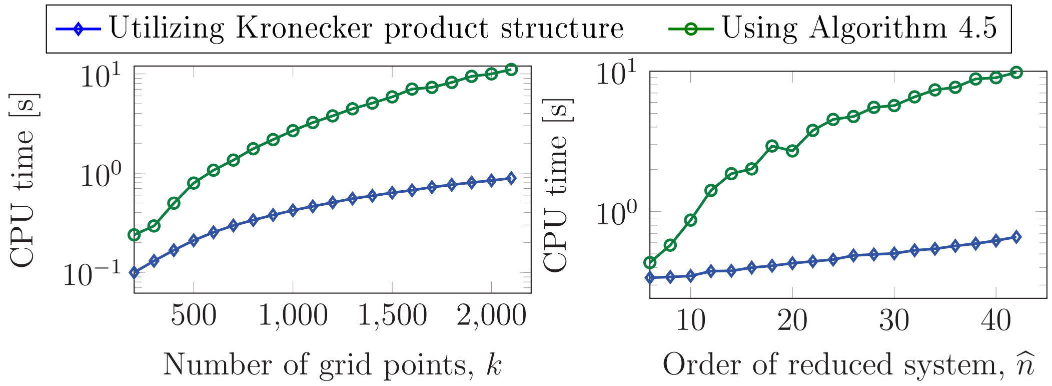


Figure 4.2.: The left figure shows the computational time for $\widehat{H} := W^T H(V \otimes V)$ by varying the number of grid points in the spatial domain by fixing the order of the reduced-order system to $\widehat{n} = 20$. In the right figure, we show the computational time for different orders of the reduced-order system using a fix number of grid points, $k = 1000$.

MATLAB 8.0.0.783 (R2012b).

Figure 4.2 illustrates that the computational cost for constructing the reduced Hessian by using the proposed method, which exploits the Kronecker product structure of the Hessian H, grows much slower than the cost in Algorithm 4.5. Therefore, we conclude here that it is worth exploiting the Kronecker product structure of the Hessian of the system for an efficient computation of $\widehat{H}$ in large-scale settings.

Next, we discuss the existence of the solutions of quadratic type generalized Lyapunov equations. As noted in Algorithm 4.2, one can determine the solution of these Lyapunov equations using fixed point iterations. In the following, we discuss sufficient conditions under which these iterations converge to finite solutions.

Theorem 4.12:
Consider a QB system as defined in (4.14) and let P and Q be its reachability and observability Gramians, respectively, and assume they exist and solve (4.19) and (4.30), respectively. Assume that the Gramians P and Q are determined using fixed point iterations as shown in Algorithm 4.2. Then, the Gramian P converges to a positive semidefinite solution if

(i) A is stable, i.e., there exist $0 < \alpha \leq -\max(\lambda_i(A))$ and $\beta > 0$ such that $\|e^{At}\| \leq \beta e^{-\alpha t}$.

(ii) $\dfrac{\beta^2 \Gamma_N}{2\alpha} < 1$, where $\Gamma_N := \sum_{k=1}^{m} \|N_k\|^2$.

(iii) $1 > \mathcal{D}^2 - \dfrac{\beta^2 \Gamma_H}{\alpha} \dfrac{\beta^2 \Gamma_B}{\alpha} > 0$, *where* $\mathcal{D} := 1 - \dfrac{\beta^2 \Gamma_N}{2\alpha}$, where $\Gamma_B := \|BB^T\|$,

$$\Gamma_H := \|H\|^2.$$

and $\|P\|$ is bounded by

$$\|P\| \leq \frac{2\alpha}{\beta^2\Gamma_H}\left(\mathcal{D} - \sqrt{\mathcal{D}^2 - 4\frac{\beta^2\Gamma_H}{2\alpha}\frac{\beta^2\Gamma_B}{2\alpha}}\right) =: \mathcal{P}_\infty. \tag{4.51}$$

Furthermore, the Gramian Q also converges to a positive semidefinite solution if in addition to the above conditions (i)–(iii), the following condition satisfies

$$\frac{\beta^2}{2\alpha}\left(\Gamma_N + \widetilde{\Gamma}_H\mathcal{P}_\infty\right) < 1,$$

where $\widetilde{\Gamma}_H := \|\mathcal{H}^{(2)}\|^2$. Moreover, $\|Q\|$ is bounded by

$$\|Q\| \leq \frac{\beta^2}{2\alpha}\Gamma_C\left(1 - \frac{\beta^2}{2\alpha}\left(\Gamma_N + \widetilde{\Gamma}_H\mathcal{P}_\infty\right)\right)^{-1},$$

where $\Gamma_C := \|C^TC\|$. ◊

Proof. Let us first consider the equation corresponding to P_1:

$$AP_1 + AP_1 + BB^T = 0.$$

Alternatively, if A is stable, we can write P_1 in the integral form as

$$P_1 = \int_0^\infty e^{At}BB^Te^{A^Tt}dt,$$

implying

$$\|P_1\| \leq \beta^2\|BB^T\|\int_0^\infty e^{-2\alpha t}dt = \frac{\beta^2\Gamma_B}{2\alpha},$$

where $\Gamma_B := \|BB^T\|$. Next, we look at the equation corresponding to P_k, $\forall k \geq 2$, which is given in terms of P_{k-1}:

$$AP_k + P_kA^T + H(P_{k-1}\otimes P_{k-1})H^T + \sum_{k=1}^m N_kP_{k-1}N_k + BB^T = 0.$$

We can also write P_k in the integral form, provided A is stable:

$$\begin{aligned}P_k &= \int_0^\infty e^{At}\left(H(P_{k-1}\otimes P_{k-1})H^T + \sum_{k=1}^m N_kP_{k-1}N_k + BB^T\right)e^{A^Tt}dt\\ &\leq \beta^2\left(\Gamma_H\|P_{k-1}\|^2 + \Gamma_N\|P_{k-1}\| + \Gamma_B\right)\int_0^\infty e^{-2\alpha t}dt\\ &\leq \beta^2\frac{(\Gamma_H\|P_{k-1}\|^2 + \Gamma_N\|P_{k-1}\| + \Gamma_B)}{2\alpha},\end{aligned}$$

where $\Gamma_H := \|H\|^2$ and $\Gamma_N := \sum_{k=1}^m \|N_k\|^2$. If we consider an upper bound for the norm of P_{k-1} in order to provide an upper bound for P_k and apply Lemma A.1, then we know that $\lim_{k\to\infty} \|P_k\|$ is bounded if

$$1 > \mathcal{D}^2 - 4\frac{\beta^2\Gamma_H}{2\alpha}\frac{\beta^2\Gamma_B}{2\alpha} > 0, \quad \textit{where} \;\; \mathcal{D} := 1 - \frac{\beta^2\Gamma_N}{2\alpha} \quad \text{and} \quad \frac{\beta^2\Gamma_N}{2\alpha} < 1.$$

Moreover, a bound for $\lim_{k\to\infty} \|P_k\|$ can be given by

$$\lim_{k\to\infty} \|P_k\| \le \frac{2\alpha}{\beta^2\Gamma_H}\left(\mathcal{D} - \sqrt{\mathcal{D}^2 - 4\frac{\beta^2\Gamma_H}{2\alpha}\frac{\beta^2\Gamma_B}{2\alpha}}\right) =: \mathcal{P}_\infty.$$

Now, we consider the equation corresponding to Q_1:

$$A^TQ_1 + A^TQ_1 + C^TC = 0,$$

which can be rewritten as:

$$Q_1 = \int_0^\infty e^{A^Tt}C^TCe^{At}dt$$

if A is stable. This implies

$$\|Q_1\| \le \beta^2\Gamma_C \int_0^\infty e^{-2\alpha t}dt = \beta^2\frac{\Gamma_C}{2\alpha},$$

where $\Gamma_c := \|C^TC\|$. Next, we look at the equation corresponding to Q_k, that is,

$$A^TQ_k + Q_kA + \mathcal{H}^{(2)}(P_{k-1}\otimes Q_{k-1})\left(\mathcal{H}^{(2)}\right)^T + \sum_{k=1}^m N_k^TQ_{k-1}N_k + C^TC = 0.$$

This yields

$$\|Q_k\| \le \frac{\beta^2}{2\alpha}\left(\left(\Gamma_N + \widetilde{\Gamma}_H\|P_{k-1}\|\right)Q_{k-1} + \Gamma_C\right),$$

where $\widetilde{\Gamma}_H := \|\mathcal{H}^{(2)}\|$. Since $\|P_{k-1}\| \le \mathcal{P}_\infty$ for all $k \ge 1$, we further have

$$\|Q_k\| \le \frac{\beta^2}{2\alpha}\left(\left(\Gamma_N + \widetilde{\Gamma}_H\mathcal{P}_\infty\right)\|Q_{k-1}\| + \Gamma_C\right).$$

A sufficient condition under which the above recurrence formula in $\|Q_k\|$ converges is:

$$\frac{\beta^2}{2\alpha}\left(\Gamma_N + \widetilde{\Gamma}_H\mathcal{P}_\infty\right) < 1,$$

and $\lim_{k\to\infty}\|Q_k\|$ is then bounded by

$$\lim_{k\to\infty}\|Q_k\| \le \frac{\beta^2}{2\alpha}\Gamma_C\left(1 - \frac{\beta^2}{2\alpha}\left(\Gamma_N + \widetilde{\Gamma}_H\mathcal{P}_\infty\right)\right)^{-1}.$$

This concludes the proof. □

Algorithm 4.6: Balanced truncation for QB systems (truncated version).

```
Input: System matrices $A, H, N_k, B$, $C$, and $\widehat{n}$.
Output: The reduced-order system's matrices $\widehat{A}, \widehat{H}, \widehat{N}_k, \widehat{B}, \widehat{C}$.
Determine low-rank approximations of the truncated Gramians $P_{\mathcal{T}} \approx RR^T$ and
  $Q_{\mathcal{T}} \approx SS^T$.
Compute SVD of $S^TR$:
    $S^TR = U\Sigma V = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \operatorname{diag}(\Sigma_1, \Sigma_2) \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T$,
    where $\Sigma_1$ contains the $\widehat{n}$ largest singular values of $S^TR$.
Construct the projection matrices $\mathcal{V}$ and $\mathcal{W}$:
    $\mathcal{V} = SU_1\Sigma_1^{-\frac{1}{2}}$ and $\mathcal{W} = RV_1\Sigma_1^{-\frac{1}{2}}$.
Determine the reduced-order system's realization:
    $\widehat{A} = \mathcal{W}^TA\mathcal{V}, \quad \widehat{H} = \mathcal{W}^TH(\mathcal{V} \otimes \mathcal{V}), \quad \widehat{N}_k = \mathcal{W}^TN_k\mathcal{V}, \quad \widehat{B} = \mathcal{W}^TB, \quad \widehat{C} = C\mathcal{V}$.
```

4.3.5. Advantages of truncated Gramians in the model reduction context

As noted in Section 4.3.3, the quadratic forms of both actual Gramians and its truncated versions (truncated Gramians) impose bounds for the energy functionals of QB systems, at least in the neighborhood of the origin, and we also provide the interpretation of reachability and observability of the system in terms of Gramians. We have seen in the previous subsection that determining Gramians P and Q is very challenging task for large-scale settings. Moreover, the convergence of Algorithm 4.2 highly depends on the spectral radius condition $\mathcal{L}^{-1}\Pi$, which should be less than 1. This condition might not be satisfied for large H and N_k; thus, Algorithm 4.2 may not convergence. On the other hand, in order to compute the truncated Gramians, there is no such convergence issue. Furthermore, it can also be observed that the bounds for energy functionals using the truncated Gramains can be much better (in the neighborhood of the origin), see, for example Figure 4.1.

This motivates us to use the truncated Gramians to determine the reduced-order models, and we present the square-root balanced truncation for QB systems based on these truncated Gramians in Algorithm 4.6. Furthermore, we will see in Section 4.4 as well that these truncated Gramians also yield very good qualitative reduced-order systems for QB systems.

4.3.6. Stability of the reduced-order systems

We now discuss the stability of the reduced-order systems, obtained by using Algorithm 4.6. For this, we consider only the autonomous part of the QB system as follows:

$$\dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)), \tag{4.52}$$

where $x_{eq} = 0$ is a stable equilibrium. In the following, we discuss the Lyapunov stability of x_{eq}. For this, we first note the definition of the latter stability.

Proposition 4.13:
Consider a QB system with $u \equiv 0$ (4.52). If there exists a Lyapunov function $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$\mathcal{F}(x) > 0 \quad \text{and} \quad \frac{d}{dt}\mathcal{F}(x) < 0, \qquad \forall x \in \mathcal{B}_{0,r} \backslash \{0\},$$

where $\mathcal{B}_{0,r}$ is a ball of radius r centered around 0, then $x_{eq} = 0$ is a locally asymptotically stable. ♢

However, many other notions of the stability of nonlinear systems are available in the literature, for instance based on a certain dissipation inequality [38], which might be difficult to apply in the large-scale setting. In this thesis, we stick to the notion of the Lyapunov-based stability for the reduced-order systems.

Theorem 4.14:
Consider the QB system (4.14) with a stable matrix A. Let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be its truncated reachability and observability Gramians, defined in Corollary 4.5, respectively. If the reduced-order system is determined as shown in Algorithm 4.6, then for a Lyapunov function $\mathcal{F}(\widehat{x}) = \widehat{x}^T\Sigma_1\widehat{x}$, we have

$$\mathcal{F}(\widehat{x}) > 0, \qquad \frac{d}{dt}(\mathcal{F}(\widehat{x})) < 0 \qquad \forall\, \widehat{x} \in \mathcal{B}_{0,r} \backslash \{0\},$$

where $r = \dfrac{\sigma_{\min}(\mathcal{V}^T\mathcal{G}\mathcal{V})}{2\|\Sigma_1\|\|\widehat{H}\|}$ and $\mathcal{G} = \mathcal{H}^{(2)}(P_1 \otimes Q_1)\left(\mathcal{H}^{(2)}\right)^T + \sum_{k=1}^m N_k^T Q_1 N_k + C^TC$ with P_1 and Q_1 being the solutions of (4.33) and (4.34), respectively. ♢

Proof. First, we establish the relation between $\mathcal{V}$, $\mathcal{W}$, $Q_{\mathcal{T}}$ and Σ_1. For this, we consider

$$\begin{aligned}\mathcal{W}\Sigma_1 &= RV_1\Sigma_1^{\frac{1}{2}} = RV_1\begin{bmatrix}\Sigma_1 & 0\end{bmatrix}^T U^TU_1\Sigma_1^{-\frac{1}{2}} = RV\Sigma U^TU_1\Sigma_1^{-\frac{1}{2}}\\ &= RR^TS^TU_1\Sigma_1^{-\frac{1}{2}} = Q_{\mathcal{T}}\mathcal{V}.\end{aligned}$$

Keeping in mind the above relation, we get

$$\begin{aligned}\widehat{A}^T\Sigma_1 + \Sigma_1\widehat{A} + \mathcal{V}^T\mathcal{G}\mathcal{V} &= \mathcal{V}^TA^T\mathcal{W}\Sigma_1 + \Sigma_1\mathcal{W}^TA\mathcal{V} + \mathcal{V}^T\mathcal{G}\mathcal{V}\\ &= \mathcal{V}^TA^TQ_{\mathcal{T}}\mathcal{V} + \mathcal{V}^TQ_{\mathcal{T}}A\mathcal{V} + \mathcal{V}^T\mathcal{G}\mathcal{V} = \mathcal{V}^T(A^TQ_{\mathcal{T}} + Q_{\mathcal{T}}A + \mathcal{G})\mathcal{V} = 0.\end{aligned} \tag{4.53}$$

Since $\mathcal{G}$ is a positive semidefinite matrix and $\mathcal{V}$ has full column rank, $\mathcal{V}^T\mathcal{G}\mathcal{V}$ is also positive semidefinite. This implies that $\eta(\widehat{A}) \leq 0$, where $\eta(\cdot)$ denotes the spectral abscissa of a matrix. Coming back to the Lyapunov function $\mathcal{F}(\widehat{x}) = \widehat{x}^T\Sigma_1\widehat{x}$, which is always greater than 0 for all $\widehat{x} \neq 0$ due to Σ_1 being a positive definite matrix, we compute the derivative of the Lyapunov function as

$$\begin{aligned}\frac{d}{dt}\mathcal{F}(\widehat{x}) &= \dot{\widehat{x}}^T\Sigma_1\widehat{x} + \widehat{x}^T\Sigma_1\dot{\widehat{x}}\\ &= \widehat{x}^T\widehat{A}^T\Sigma_1\widehat{x} + (\widehat{x}^T\otimes\widehat{x}^T)\widehat{H}^T\Sigma_1\widehat{x} + \widehat{x}^T\Sigma_1\widehat{A}\widehat{x} + \widehat{x}^T\Sigma_1\widehat{H}(\widehat{x}\otimes\widehat{x})\\ &= \widehat{x}^T(\widehat{A}^T\Sigma_1 + \Sigma_1\widehat{A})\widehat{x} + (\widehat{x}^T\otimes\widehat{x}^T)\widehat{H}^T\Sigma_1\widehat{x} + \widehat{x}^T\Sigma_1\widehat{H}(\widehat{x}\otimes\widehat{x}).\end{aligned}$$

Substituting $\widehat{A}^T\Sigma_1 + \Sigma_1\widehat{A} = -\mathcal{V}^T\mathcal{G}\mathcal{V}$ from (4.53) in the above equation yields

$$\frac{d}{dt}\mathcal{F}(\widehat{x}) = -\widehat{x}^T\mathcal{V}^T\mathcal{G}\mathcal{V}\widehat{x} + 2\widehat{x}^T\Sigma_1\widehat{H}(\widehat{x}\otimes\widehat{x}). \tag{4.54}$$

As

$$\widehat{x}^T\mathcal{V}^T\mathcal{G}\mathcal{V}\widehat{x} \geq \sigma_{\min}(\mathcal{V}^T\mathcal{G}\mathcal{V})\|\widehat{x}\|^2,$$

implying

$$-\widehat{x}^T\mathcal{V}^T\mathcal{G}\mathcal{V}x \leq -\sigma_{\min}(\mathcal{V}^T\mathcal{G}\mathcal{V})\|\widehat{x}\|^2,$$

inserting the above inequality in (4.54) leads to

$$\tfrac{d}{dt}\mathcal{F}(\widehat{x}) \leq -\sigma_{\min}(\mathcal{V}^T\mathcal{G}\mathcal{V})\|\widehat{x}\|^2 + 2\|\widehat{x}\|^3\|\Sigma_1\|\|\widehat{H}\|.$$

For locally asymptotic stability of the reduced-order system, we require

$$\tfrac{d}{dt}\mathcal{F}(\widehat{x}) \leq -\sigma_{\min}(\mathcal{V}^T\mathcal{G}\mathcal{V})\|\widehat{x}\|^2 + 2\|\widehat{x}\|^3\|\Sigma_1\|\|\widehat{H}\| < 0,$$

which gives rise to the following bound on $\|\widehat{x}\|$:

$$\|\widehat{x}\| < \tfrac{\sigma_{\min}(\mathcal{V}^T\mathcal{G}\mathcal{V})}{2\|\Sigma_1\|\|\widehat{H}\|}.$$

This concludes the proof. □

4.4. Numerical Experiments

In this section, we consider MOR of several QB control systems and evaluate the efficiency of the proposed balanced truncation (BT) technique (Algorithm 4.6). For this, we need to solve a number of conventional Lyapunov equations. In our numerical experiments, we determine low-rank factors of these Lyapunov equations by using the ADI method as proposed in [32]. We compare the proposed methodology with the existing

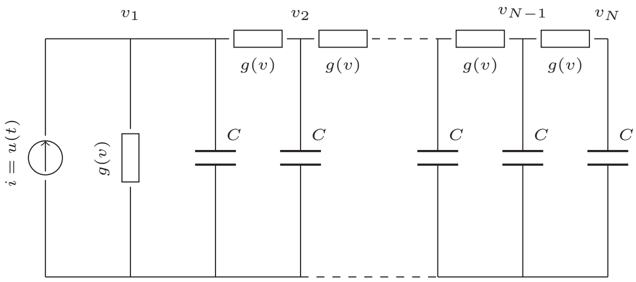


Figure 4.3.: Nonlinear RC ladder diagram.

moment-matching techniques for QB systems, namely one-sided moment-matching [78] and its recent extension to two-sided moment-matching [25]. These moment-matching methods aim at approximating the underlying generalized transfer functions of the system. Moreover, we need interpolation points to apply the moment-matching methods; thus, we choose l linear $\mathcal{H}_2$-optimal interpolation points, determined by applying *IRKA* [79] to the corresponding linear part. This leads to a reduced QB system of order $\widehat{n} = 2l$. All the simulations were done on on a board with 4 Intel® Xeon® E7-8837 CPUs with a 2.67-GHz clock speed using MATLAB 8.0.0.783 (R2012b).

4.4.1. Nonlinear RC ladder

The first example, we discuss, is a nonlinear RC ladder as shown in Section 4.4.1. It is a well-known example and is used as one of the benchmark problems in the community of nonlinear model reduction; see, e.g., [11, 40, 78, 98, 106]. The ladder consists of nonlinear resistors g and capacitors $\widetilde{C}$. Let $v_i(t)$ and $u(t)$ denote the voltage between the ith node and the ground and the input signal to the independent current source, respectively. Applying the Kirchoff's law at each node leads to the following set of equations:

$$\widetilde{C}\dot{v}_1(t) + g(v_1(t)) + g(v_1(t) - v_2(t)) = u(t), \tag{4.55a}$$

$$\widetilde{C}\dot{v}_k(t) + g(v_k(t) - v_{k-1}(t)) = g(v_{k-1}(t) - v_k(t)), \quad k = 2, \dots, N-1 \tag{4.55b}$$

$$\widetilde{C}\dot{v}_N(t) = g(v_{N-1}(t) - v_N(t)). \tag{4.55c}$$

Furthermore, the current-voltage relation of the resistor g is given as $g(v) = e^{40v} - v - 1$, and for simplicity, we set all capacitors $\widetilde{C} = 1$, leading to a nonlinear control system as

$$\begin{aligned}\dot{v}(t) &= f(v(t)) + Bu(t),\\ y(t) &= Cv(t),\end{aligned}$$

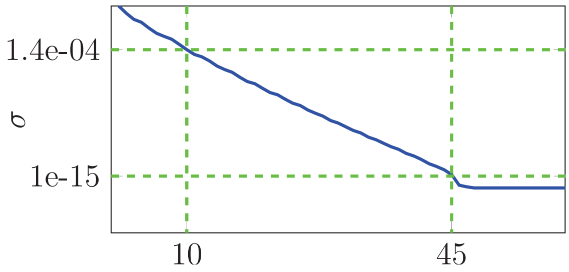


Figure 4.4.: RC ladder: decay of the normalized singular values based the truncated Gramians, and the dotted lines show the normalized singular value for $\widehat{n} = 10$ and the order of the reduced-order system corresponding to the normalized singular value $1e{-}15$.

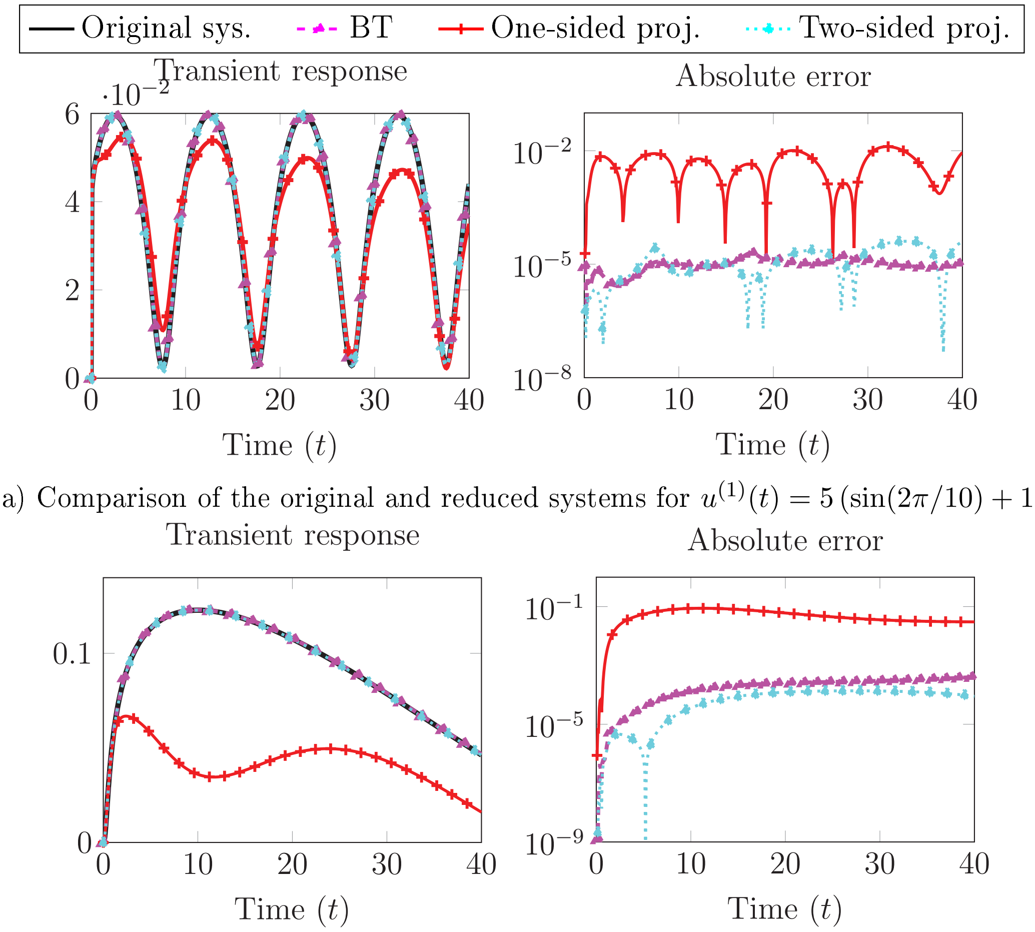
where

$$v(t) = \begin{bmatrix} v_1(t) \\ v_(t)2 \\ \vdots \\ v_k(t) \\ \vdots \\ v_N(t) \end{bmatrix}, \quad f(v(t)) = \begin{bmatrix} -g(v_1) - g(v_1 - v_2) \\ g(v_1 - v_2) - g(v_2 - v_3) \\ \vdots \\ g(v_{k-1} - v_k) - g(v_k - v_{k+1}) \\ \vdots \\ g(v_{N-1} - v_N) \end{bmatrix}, \quad B = C^T = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

As shown in [78], introducing some appropriate new variables allows us to write the system dynamics in the QB form. For this example, if one defines new state variables as $x_1(t) = v_1(t)$ and $x_i(t) = v_i(t) - v_{i+1}(t)$ and define new state variables $z_1 = e^{40v_1 - 1}$ and $z_i = e^{40x_i}$, the system (4.55) can be written in the QB form of dimension $2N$. A more detailed discussion can be found in [78].

We consider 500 capacitors in the ladder, resulting in a QB system of order $n = 1000$. For this particular example, the matrix A is a semi-stable matrix, i.e., $0 \subset \sigma(A)$. As a result, the truncated Gramians of the system may not exist; therefore, we replace the matrix A by $A_s := A{-}0.05I_n$, where I_n is the identity matrix, to determine these Gramians. Note that we project the original system with the matrix A to compute a reduced-order system but the projection matrices are computed using the Gramians obtained via the shifted matrix A_s. In Figure 4.4, we show the decay of the singular values, determined by the truncated Gramians (with the shifted A). We then compute the reduced-order system of order $\widehat{n} = 10$ by using balanced truncation. Also, we determine five $\mathcal{H}_2$-optimal linear interpolation points and compute reduced-order systems of order $\widehat{n} = 10$ via one-sided and two-sided projection methods.

To compare the quality of these approximations, we simulate these systems for the input signals $u^{(1)}(t) = 5\,(\sin(2\pi/10) + 1)$ and $u^{(2)}(t) = 10\,(t^2 \exp(-t/5))$. Figure 4.5 presents the transient responses and relative errors of the output for these input signals,



(a) Comparison of the original and reduced systems for $u^{(1)}(t) = 5\left(\sin(2\pi/10) + 1\right)$.

(b) Comparison of the original and reduced systems for $u^{(2)}(t) = 10\left(t^2 \exp(-t/5)\right)$.

Figure 4.5.: RC ladder: comparison of reduced-order systems obtained by BT and moment-matching methods for two arbitrary control inputs.

which shows that balanced truncation outperforms the one-sided interpolatory method; on the other hand, we see that balanced truncation is competitive to the two-sided interpolatory projection for this example for both considered inputs.

4.4.2. One-dimensional Chafee-Infante equation

As a second example, we consider the one-dimensional Chafee-Infante (Allen-Cahn) equation. This nonlinear system has been widely studied in the literature; see, e.g., [45, 82], and its model reduction related problem was recently considered in [25]. The governing equation, subject to initial conditions and boundary control, have a cubic

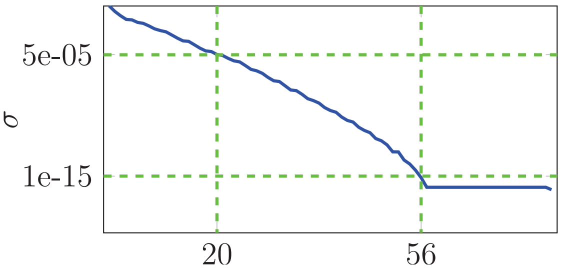


Figure 4.6.: Chafee-Infante equation: decay of the normalized singular values based the truncated Gramians, and dotted line shows the normalized singular value for $\widehat{n} = 20$ and the order of the reduced-order system corresponding to the normalized singular value $1e{-}15$.

nonlinearity:

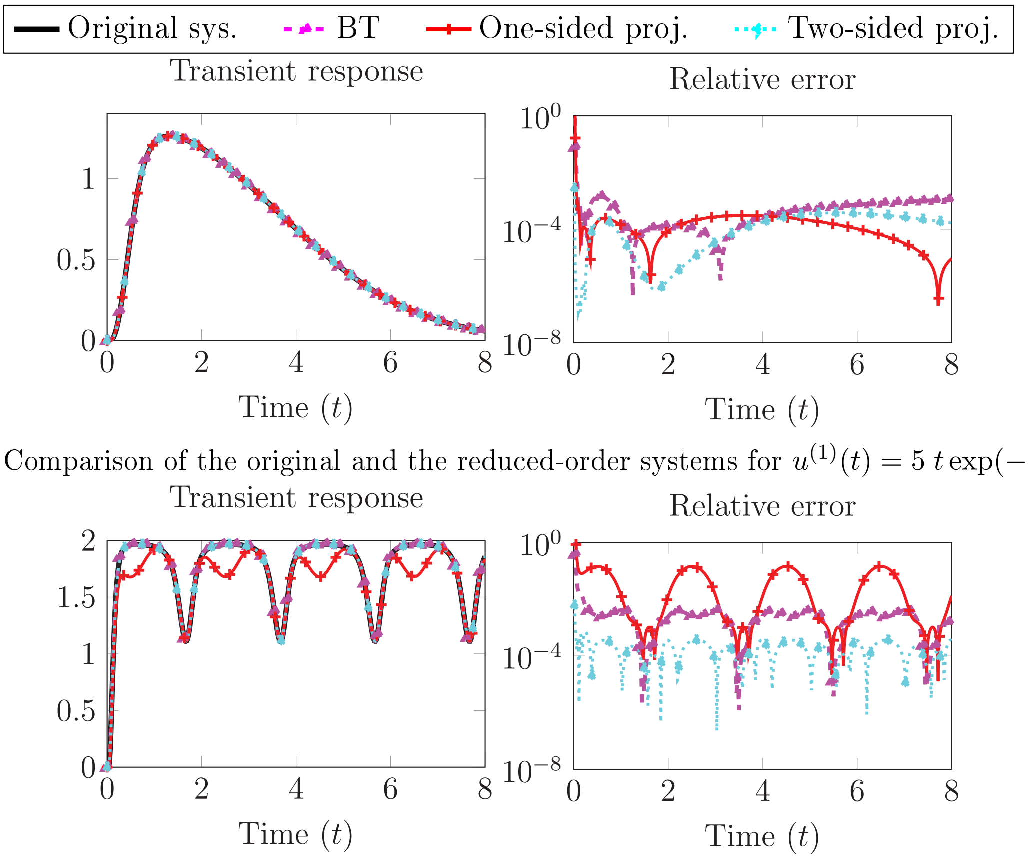
$$\begin{aligned} &\dot{v} + v^3 = v_{xx} + v, && (0, L) \times (0, T), && v(0, \cdot) = u(t), && (0, T), \\ &v_x(L, \cdot) = 0, && (0, T), && v(x, 0) = 0, && (0, L). \end{aligned} \tag{4.56}$$

Here, we make use of a finite difference scheme and consider k grid points in the spatial domain, leading to a semi-discretized nonlinear ODE. However, the system (4.56) with the cubic nonlinearity can be rewritten in the QB form by defining new variables $w_i = v_i^2$ with derivate $\dot{w}_i = 2v_i\dot{v}_i$. We observe the response at the right boundary at $x = L$. We use the number of grid points $k = 500$, which results in a QB system of dimension $n = 2 \cdot 500 = 1000$ and set the length $L = 1$. In Figure 4.6, we show the decay of the normalized singular values based on the truncated Gramians of the system.

We determine reduced-order systems of order $\widehat{n} = 20$ by using balanced truncation, and one-sided and two-sided interpolatory methods. To compare the quality of these reduced-order systems, we observe the outputs of the original and reduced-order systems for two arbitrary control inputs $u(t) = 5t\exp(-t)$ and $u(t) = 30(\sin(\pi t) + 1)$ in Figure 4.7.

Figure 5.2 shows that the reduced-order systems obtained via balanced truncation and one-sided and two-sided interpolatory projection methods are almost of the same quality for input $u^{(1)}$. But for the input $u^{(2)}$, the reduced-order system obtained via the one-sided interpolatory projection method completely fails to capture the dynamics of the system, while balanced truncation and two-sided interpolatory projection can reproduce the system dynamics with a slight advantage over the two-sided projection regarding accuracy.

However, it is worthwhile to mention that as we increase the order of the reduced-order system, the two-sided interpolatory projection method tends to produce unstable reduced-order systems. On the other hand, the accuracies of the reduced-order systems obtained by balanced truncation and one-sided moment-matching increase with the order of the reduced-order systems.



(a) Comparison of the original and the reduced-order systems for $u^{(1)}(t) = 5\,t\exp(-t)$.

(b) Comparison of the original and the reduced-order systems for $u^{(2)}(t) = 30\,(\sin(\pi t) + 1)$.

Figure 4.7.: Chafee-Infante equation: comparison of the reduced-order systems obtained via balanced truncation and moment-matching methods for the inputs $u^{(1)}(t) = 5\,(t\exp(-t))$ and $u^{(2)}(t) = 30\,(\sin(\pi t) + 1)$.

4.4.3. The FitzHugh-Nagumo (F-N) system

Lastly, we consider the F-N system, a simplified neuron model of the Hodgkin-Huxley model, describing activation and deactivation dynamics of a spiking neuron. This model has been considered in the framework of POD-based [47] and moment-matching model reduction techniques [23]. The dynamics of the system are governed by the following nonlinear coupled differential equations:

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

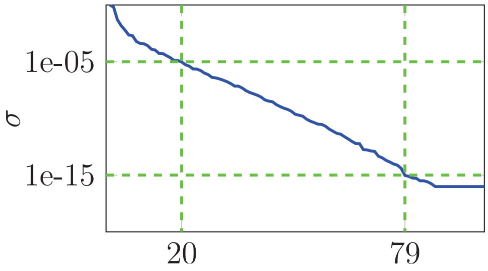


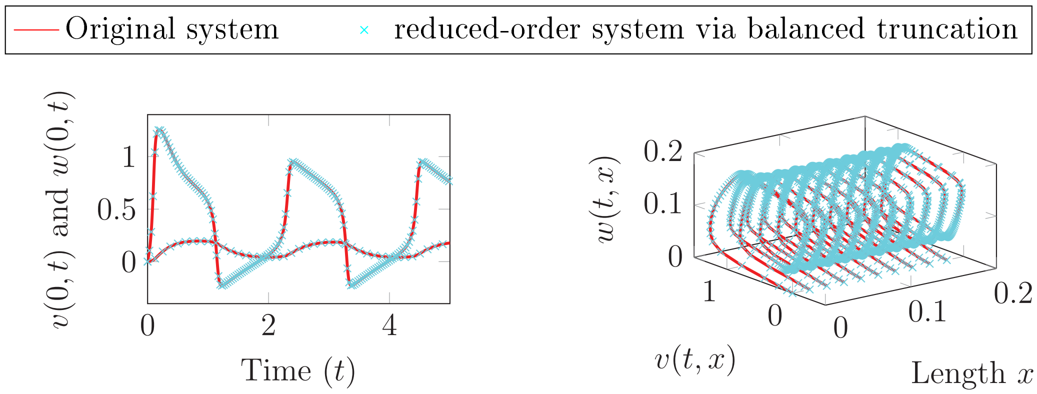
Figure 4.8.: FitzHugh-Nagumo system: decay of the normalized singular values based the truncated Gramians of the system, and the dotted lines show the normalized singular value for $\widehat{n} = 20$ and the order of the reduced-order system corresponding to the normalized singular value $1e{-}15$.

with the nonlinear function $f(v(x,t)) = v(v-0.1)(1-v)$ and the initial and boundary conditions:

$$\begin{aligned} v(x,0) &= 0, & w(x,0) &= 0, & x &\in [0,L] \\ v_x(0,t) &= i_0(t), & v_x(1,t) &= 0, & t &\geq 0, \end{aligned}$$

where $\epsilon = 0.015$, $h = 0.5$, $\gamma = 2$, $q = 0.05$. We set the length $L = 0.2$. The stimulus i_0 acts as an actuator, taking the values $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and the variables v and w denote the voltage and recovery voltage, respectively. We also assume the same outputs of interest as considered in [23], which are $v(0,t)$ and $w(0,t)$. These outputs describe nothing but the limit cyclic at the left boundary. Using a finite difference discretization scheme, one can obtain a system with two inputs and two outputs of dimension $2k$ with cubic nonlinearities, where k is the number of degrees of freedom. Similar to the previous example, the F-H system can also be transformed into a QB system of dimension $n = 3k$ by introducing a new state variable $z_i = v_i^2$. We set $k = 500$, resulting in a QB system of order $n = 1500$. Figure 4.8 shows the decay of the singular values based on the truncated Gramians for the QB system.

Next, we determine reduced-order systems of order $\widehat{n} = 20$ by using balanced truncation. In order to apply one-sided moment-matching, we take four linear $\mathcal{H}_2$-optimal points to construct the projection matrix, and then take first the most 20 dominant modes to construct a reduced-order system of order 20. Moreover, theoretically, two-sided projection method [25] applies only to SISO QB system, and this particular example is a multi-input multi-output (MIMO) QB system. However, for sake of comparison, we blindly apply two-sided projection using block Krylov system as done in the case of one-sided projection. We observe that the reduced-order systems, obtained via the moment-matching methods with linear $\mathcal{H}_2$-optimal interpolations, both one-sided and two-sided, fail to capture the dynamics and limit cycles. We made several attempts to adjust the order of the reduced-order systems; but we were unable to determine a stable reduced-order system via these methods with linear $\mathcal{H}_2$-optimal points which could replicate the dynamics. Contrary to these methods, the balanced truncation



(a) The response $v(t)$ and $w(t)$ at the left boundary.

(b) Limit-cycles.

Figure 4.9.: FitzHugh-Nagumo system: comparison of the response at the left boundary and the limit cycle behavior of the original system and the reduced-order (balanced truncation) system. The reduced-order systems determined by moment-matching methods were unable to produce these limit cycles.

replicates the dynamics of the system faithfully as can be seen in Figure 4.9a. As the dynamics of the system produces limit cycles for each spatial variable x, we, therefore, plot the solutions v and w over the spatial domain x, which is also captured by the reduced-order system very well.

Note that the reduced-order system reported in [23] was obtained using higher-order moments in a trial-and-error fashion but cannot be reproduced by an automated algorithm.

4.5. Conclusions and Outlook

In this chapter, we have investigated balanced truncation model reduction for the QB control systems by extending the ideas for bilinear control systems. We have proposed the Gramians, namely reachability and observability Gramians, for QB systems based on the kernels of their underlying Volterra series. Additionally, we have also introduced a truncated version of the Gramians. We have further compared the controllability and observability energy functionals of QB system with certain quadratic forms of the proposed Gramians for the system and have investigated the connection between the Gramians and reachability/observability of the QB system. Also, we have discussed the advantages of the truncated version of Gramians in the MOR framework and studied the Lyapunov stability of the reduced-order systems, obtained via the square-root balanced truncation. By means of various semi-discretized nonlinear PDEs, we have demonstrated the efficiency the proposed balanced methods for QB systems and compared it with the existing moment-matching techniques.

As a future work, it would be important to investigate low-rank solvers for quadratic Lyapunov equations which solve the Gramians, and use them in the model reduction. Furthermore, it is important to study the error between the original and reduced-order systems, occurring due to the truncation, which allows us to bound an error in the outputs of the original and reduced-order systems. In various applications, the transient response for a limited time is of much interest; therefore, an extension of time-limited balanced truncation from linear systems, see, e.g., [33] to quadratic-bilinear systems will be very useful. And an extension to quadratic-bilinear descriptor systems will be promising, especially due to applications in flow problems (Navier-Stokes equations).

CHAPTER 5

INTERPOLATORY-BASED $\mathcal{H}_2$-QUASI-OPTIMAL MODEL REDUCTION FOR QUADRATIC-BILINEAR CONTROL SYSTEMS

Contents

5.1. Introduction

In the previous chapter, we have extended balanced truncation model reduction from linear/bilinear systems to quadratic-bilinear (QB) systems. We have discussed how to determine the states which are *less important* for the system dynamics; hence, removing such states leads to reduced-order systems. However, in order to apply balanced truncation to QB systems using the truncated Gramians, we require the solutions of four conventional Lyapunov equations, which could be computationally cumbersome in large-scale settings, though there have been many advancements in recent times related to computing the low-rank solutions of the Lyapunov equations, see, e.g., [36, 121].

Another popular input-independent MOR approach for linear and bilinear systems is based on computing models that satisfy optimality conditions for the best approximation in the $\mathcal{H}_2$ system norm. Therefore, in this chapter, we study the $\mathcal{H}_2$-optimal approximation problem for QB systems. As noted in the previous chapter, interpolation-based ideas from linear or bilinear systems have been extended to QB systems, see, e.g., [13, 25, 76]. However, an important question which still remains is how to choose these interpolation points, which leads to optimal reduced-order systems in a system norm. we show how to choose the model reduction bases in a two-sided projection framework for QB systems so that the reduced-order system approximately minimizes the cost encoding the approximation error in the $\mathcal{H}_2$-norm. The structure of the chapter is as follows. In Section 5.2, we first define the $\mathcal{H}_2$-norm of the QB system (4.1) based on the *kernels* of its Volterra series (input/output mapping), and also derive an expression for a truncated $\mathcal{H}_2$-norm for QB systems. Subsequently, based on the truncated $\mathcal{H}_2$-norm of the error system, we derive first-order necessary conditions for optimal model reduction of QB systems. We then propose an iterative algorithm to construct reduced models that *approximately* satisfy the newly derived optimality conditions. In Section 5.3, we illustrate the efficiency of the proposed method for various semi-discretized nonlinear PDEs and compare it with existing methods such as balanced truncation (proposed in the previous chapter) as well as the one-sided and two-sided interpolatory methods for QB systems [25, 78]. We conclude the paper with a short summary and potential future directions in Section 5.4.

5.2. $\mathcal{H}_2$-Norm for QB Systems and Optimality Conditions

In this section, we first define the $\mathcal{H}_2$-norm for the QB systems (4.1) and its truncated version. Then, based on the truncated $\mathcal{H}_2$ measure, we derive first-order necessary conditions for optimal model reduction. These optimality conditions will naturally lead to a numerical algorithm to construct quasi-optimal reduced models for QB systems that are independent of training data. The proposed model reduction framework extends the optimal $\mathcal{H}_2$ methodology from linear [79] and bilinear systems [21, 59] to QB nonlinear systems.

5.2.1. $\mathcal{H}_2$-norm of QB systems

In order to define $\mathcal{H}_2$-norm for QB systems and its truncated version, we first require the input/output representation for QB systems. In other words, we aim at obtaining the solution of QB systems with the help of *Volterra* series. We have derived the

Volterra series for QB systems in Subsection 4.3.1 while deriving reachability Gramians by aiming at directly writing an expression for the state $x(t)$ in the convolution form. However, we here utilize another approach to derive the same Volterra series for QB system by using the *variational analysis* [111, Section 3.4]. Since the QB system comes falls under the class of linear-analytic systems, for a scalar α, we can write the solution $x(t)$ for an input $\alpha u(t)$ as

$$x(t) = \sum_{s=1}^{\infty} \alpha^s x_s(t),$$

where $x_s(t) \in \mathbb{R}^n$. Thus, we obtain

$$\begin{aligned}\left(\sum_{s=1}^{\infty} \alpha^s \dot{x}_s(t)\right) &= A(\sum_{s=1}^{\infty} \alpha^s x_s(t)) + H\left(\left(\sum_{s=1}^{\infty} \alpha^s x_s(t)\right) \otimes \left(\sum_{s=1}^{\infty} \alpha^s x_s(t)\right)\right) \\ &\quad + \sum_{k=1}^{m} \alpha N_k \sum_{s=1}^{\infty} \alpha^s x_s(t) u_k(t) + \alpha B u(t).\end{aligned} \tag{5.1}$$

Since the expression (5.1) holds for arbitrary α, the coefficients of α^i, $i = \{1, 2, \ldots\}$ can be equated in both sides of (5.1), leading to

$$\begin{aligned}\dot{x}_1(t) &= Ax_1(t) + Bu(t), \\ \dot{x}_2(t) &= Ax_2(t) + H\left(x_1(t) \otimes x_1(t)\right) + \sum_{k=1}^{m} N_k x_1 u_k(t), \\ \dot{x}_s(t) &= Ax_s(t) + \sum_{\substack{i,j \geq 1 \\ i+j=s}} H\left(x_i(t) \otimes x_j(t)\right) + \sum_{k=1}^{m} N_k x_{s-1}(t) u(t), \quad s \geq 3.\end{aligned} \tag{5.2}$$

Then, let $\alpha = 1$ so that $x(t) = \sum_{s=1}^{\infty} x_s(t)$, where $x_s(t)$ solves the coupled linear differential equation (5.2). The equation for $x_1(t)$ corresponds to a linear system, thus allowing us to write the expression for $x_1(t)$ as a convolution:

$$x_1(t) = \int_0^t e^{At_1} B u(t - t_1) dt_1. \tag{5.3a}$$

Using the expression for $x_1(t)$, we can obtain an explicit expression for $x_2(t)$:

$$\begin{aligned}x_2(t) = &\int\limits_0^t \int\limits_0^{t-t_3} \int\limits_0^{t-t_3} e^{At_3} H\left(e^{At_2} B \otimes e^{At_1} B\right) u(t - t_2 - t_3) \otimes u(t - t_1 - t_3) dt_1 dt_2 dt_3 \\ &+ \sum_{k=1}^{m} \int\limits_0^t \int\limits_0^{t-t_2} e^{At_2} N_k e^{At_1} B u(t - t_1 - t_2) u_k(t - t_2) dt_1 dt_2.\end{aligned}$$

Similarly, one can write down explicit expressions for $x_s(t), s \geq 3$, as well, but the notation and expression become tedious, and we skip them for brevity. Then, we can write the output $y(t)$ of the QB system as $y(t) = \sum_{s=1}^{\infty} Cx_s(t)$, leading to the input/output relation of the QB system (4.1)

$$\begin{aligned}
y(t) = &\int_0^t Ce^{At_1}Bu(t-t_1)dt_1 + \\
&\int_0^t \int_0^{t-t_3} \int_0^{t-t_3} e^{At_3} H\left(e^{At_2}B \otimes e^{At_1}B\right) u(t-t_2-t_3) \otimes u(t-t_1-t_3)dt_1dt_2dt_3 + \\
&\int_0^t \int_0^{t-t_2} e^{At_2}\left[N_1, \ldots, N_m\right]\left(I_m \otimes e^{At_1}B\right)\left(u(t-t_2) \otimes u(t-t_1-t_2)\right)dt_1dt_2 + \cdots .
\end{aligned} \tag{5.5}$$

Examining the structure of (5.5) reveals that the *kernels* $f_i(t_1, \ldots, t_i)$ of (5.5) are given by the recurrence formula

$$f_i(t_1, \ldots, t_i) = Cg_i(t_1, \ldots, t_i), \tag{5.6}$$

where

$$\begin{aligned}
g_1(t_1) &= e^{At_1}B, \\
g_2(t_1, t_2) &= e^{At_2}\left[N_1, \ldots, N_m\right]\left(I_m \otimes e^{At_1}B\right), \\
g_i(t_1, \ldots, t_i) &= e^{At_i}\left[H\left[g_1(t_1) \otimes g_{i-2}(t_2, \ldots, t_{i-1}), \ldots, g_{i-2}(t_1, \ldots, t_{i-2}) \otimes g_1(t_{i-1})\right],\right. \\
&\qquad\qquad \left.\left[N_1, \ldots, N_m\right]\left(I_m \otimes g_{i-1}\right)\right], \quad i \geq 3.
\end{aligned} \tag{5.7}$$

As shown in [133], the $\mathcal{H}_2$-norm of a bilinear system can be defined in terms of a series of kernels, corresponding to its input/output mapping. Inspired by this definition, next we introduce the $\mathcal{H}_2$-norm of a QB system based on these kernels.

Definition 5.1:
Consider the QB system (4.1) with its Volterra kernels, defined in (5.6). Then, we define the $\mathcal{H}_2$-norm of the QB system by

$$\|\Sigma_{QB}\|_{\mathcal{H}_2} := \sqrt{\operatorname{tr}\left(\sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} f_i(t_1, \ldots, t_i) f_i^T(t_1, \ldots, t_i) dt_1 \ldots dt_i\right)}. \tag{5.8}$$

$\Diamond$

Fortunately, we can find an alternative way to compute the norm in a numerically efficient way using matrix equations. We know from the cases of linear and bilinear systems that the $\mathcal{H}_2$-norms of these systems can be computed in terms of the certain system Gramians. We next show that this is also the case for QB systems. The algebraic Gramians for QB systems were studied in the previous chapter. So, in the following, we extend such relations between the $\mathcal{H}_2$-norm, see Definition 5.1, and the systems Gramians to QB systems.

Lemma 5.2:
Consider a QB system with a stable matrix A, and let P and Q, respectively, be the controllability and observability Gramians of the system, which are the unique positive semidefinite solutions of the following quadratic-type Lyapunov equations:

$$AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0, \quad \text{and} \tag{5.9}$$

$$A^T Q + QA + \mathcal{H}^{(2)}(P \otimes Q)\left(\mathcal{H}^{(2)}\right)^T + \sum_{k=1}^{m} N_k^T Q N_k + C^T C = 0. \tag{5.10}$$

Assuming the $\mathcal{H}_2$-norm of the QB system exists, i.e., the series in (5.8) converges, then the $\mathcal{H}_2$-norm of the QB system can be computed as

$$\|\Sigma_{QB}\|_{\mathcal{H}_2} := \sqrt{\operatorname{tr}\left(CPC^T\right)} = \sqrt{\operatorname{tr}\left(B^T QB\right)}. \tag{5.11}$$

◊

Proof. We begin with the definition of the $\mathcal{H}_2$-norm of a QB system, that is,

$$\begin{aligned}\|\Sigma_{QB}\|_{\mathcal{H}_2} &= \sqrt{\operatorname{tr}\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} f_i(t_1,\ldots,t_i)f_i^T(t_1,\ldots,t_i)dt_1\ldots dt_i\right)}\\ &= \sqrt{\operatorname{tr}\left(C\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} g_i(t_1,\ldots,t_i)g_i^T(t_1,\ldots,t_i)dt_1\ldots dt_i\right)C^T\right)},\end{aligned}$$

where $f_i(t_1,\ldots,t_i)$ and $g_i(t_1,\ldots,t_i)$ are defined in (5.6) and(5.7), respectively. It is shown in [28] that

$$\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} g_i(t_1,\ldots,t_i)g_i^T(t_1,\ldots,t_i)dt_1\ldots dt_i\right) = P, \tag{5.12}$$

where P solves (5.9) if the series in (5.12) converges. Thus,

$$\|\Sigma_{QB}\|_{\mathcal{H}_2} = \sqrt{\operatorname{tr}\left(CPC^T\right)}.$$

Next, we prove that $\operatorname{tr}\left(CPC^T\right) = \operatorname{tr}\left(B^T QB\right)$, where Q solves (5.10). Making use of the Kronecker product properties (2.29), we can write $\operatorname{tr}\left(CPC^T\right)$ as

$$\operatorname{tr}\left(CPC^T\right) = \mathcal{I}_p^T(C \otimes C)\operatorname{vec}(P).$$

Vectorizing both sides of (5.9) yields

$$\left(A \otimes I_n + I_n \otimes A + \sum_{k=1}^{m} N_k \otimes N_k\right) \operatorname{vec}(P) + (H \otimes H) \operatorname{vec}(P \otimes P) + (B \otimes B)\mathcal{I}_m = 0. \tag{5.13}$$

Using Lemma 2.29 in the above equation and performing some simple manipulations yields an expression for $\operatorname{vec}(P)$ as

$$\operatorname{vec}(P) = \mathcal{G}^{-1}(B \otimes B)\mathcal{I}_m =: P_v,$$

where

$$\mathcal{G} = -\left(A \otimes I_n + I_n \otimes A + \sum_{k=1}^{m} N_k \otimes N_k + (H \otimes H)T_{(n,n)}(I_{n^2} \otimes \operatorname{vec}(P))\right).$$

Thus,

$$\operatorname{tr}\left(CPC^T\right) = \mathcal{I}_p^T(C \otimes C)\mathcal{G}^{-1}(B \otimes B)\mathcal{I}_m = \mathcal{I}_m^T(B^T \otimes B^T)\mathcal{G}^{-T}(C^T \otimes C^T)\mathcal{I}_p. \tag{5.14}$$

Now, let $Q_v = \mathcal{G}^{-T}(C^T \otimes C^T)\mathcal{I}_p^T$. As a result, we obtain

$$\begin{aligned}(C^T \otimes C^T)\mathcal{I}_p^T &= \operatorname{vec}\left(C^T C\right) = \mathcal{G}^T Q_v \\ &= -\left(A^T \otimes I_n + I_n \otimes A^T + \sum_{k=1}^{m} N_k^T \otimes N_k^T\right) Q_v \\ &\qquad + \left((H \otimes H)\, T_{(n,n)}\, (I_{n^2} \otimes P_v)\right)^T Q_v.\end{aligned}$$

Next, we consider a matrix $\widetilde{Q}$ such that $\operatorname{vec}\left(\widetilde{Q}\right) = Q_v$, which further simplifies the above equation as

$$\operatorname{vec}\left(C^T C\right) = -\operatorname{vec}\left(A^T\widetilde{Q} + \widetilde{Q}A + \sum_{k=1}^{m} N_k^T\widetilde{Q}N_k\right) - \left((H \otimes H)T_{(n,n)}(I_{n^2} \otimes P_v)\right)^T Q_v. \tag{5.15}$$

Now, we focus on the transpose of the second part of (5.15), that is,

$$\begin{aligned}&Q_v^T(H \otimes H)T_{(n,n)}\left(I_{n^2} \otimes \operatorname{vec}(P)\right) \\ &\quad = Q_v^T(H \otimes H)T_{(n,n)}\left[e_1^{n^2} \otimes \operatorname{vec}(P), \ldots, e_{n^2}^{n2} \otimes \operatorname{vec}(P)\right] \\ &\quad = Q_v^T(H \otimes H)\left[\operatorname{vec}(\Psi_1 \otimes P), \ldots, \operatorname{vec}(\Psi_{n^2} \otimes P)\right] =: \Xi, \qquad \text{(using Lemma 2.29)}\end{aligned}$$

where $\Psi_i \in \mathbb{R}^{n\times n}$ is such that $e_i^{n^2} = \operatorname{vec}(\Psi_i)$. Using (2.29) and Lemma 2.28, we further analyze the above equation:

$$\begin{aligned}
\Xi &= \operatorname{vec}\left(\widetilde{Q}\right)^T \left[\operatorname{vec}\left(H\left(\Psi_1 \otimes P\right)H^T\right), \ldots, \operatorname{vec}\left(H\left(\Psi_{n^2} \otimes P\right)H^T\right)\right] \\
&= \operatorname{vec}\left(\widetilde{Q}\right)^T \left[\operatorname{vec}\left(H\left(P \otimes \Psi_1\right)H^T\right), \ldots, \operatorname{vec}\left(H\left(P \otimes \Psi_{n^2}\right)H^T\right)\right] \\
&= \Bigg[\operatorname{vec}\left(\Psi_1\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P\otimes\widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right), \ldots, \\
&\qquad\qquad \operatorname{vec}\left(\Psi_{n^2}\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P\otimes\widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right)\Bigg] \\
&= \left[\left(e_1^{n^2}\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P\otimes\widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right), \ldots, \left(e_{n^2}^{n^2}\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P\otimes\widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right)\right] \\
&= \left(\operatorname{vec}\left(\mathcal{H}^{(2)}\left(P\otimes\widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right)\right)^T.
\end{aligned}$$

Substituting this relation into (5.15) yields

$$\operatorname{vec}\left(C^TC\right) = -\operatorname{vec}\left(A^T\widetilde{Q} + \widetilde{Q}A + \sum_{k=1}^{m} N_k^T\widetilde{Q}N_k + \mathcal{H}^{(2)}\left(P\otimes\widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right),$$

which shows that $\widetilde{Q}$ solves (5.10) as well. Since it is assumed that Eq. Eq. (5.10) has a unique solution, we get $\widetilde{Q} = Q$. Replacing $\mathcal{G}^{-T}(C^T\otimes C^T)\mathcal{I}_p^T$ by $\operatorname{vec}(Q)$ in (5.14) and using (2.29) results in

$$\operatorname{tr}\left(CPC^T\right) = \mathcal{I}_m^T(B^T\otimes B^T)\operatorname{vec}(Q) = \operatorname{tr}\left(B^TQB\right).$$

This concludes the proof. □

It can be seen that if H is zero, the expression (5.11) boils down to the $\mathcal{H}_2$-norm of bilinear systems, and if all N_k are also set to zero then it provides us the $\mathcal{H}_2$-norm of stable linear systems as one would expect.

Remark 5.3:

In Lemma 5.2, we have assumed that the solutions of (5.9) and (5.10) exist, and that they are unique and positive semidefinite. Equivalently, the series appearing in the definition of the $\mathcal{H}_2$-norm is finite (see Definition 5.1); hence, the $\mathcal{H}_2$-norm exists. Naturally, the stability of the matrix A is necessary for the existence of Gramians, and a detailed study of the solutions of (5.9) and (5.10) has been carried out in the previous chapter. However, as for bilinear systems, these Gramians may not have the desired properties such as uniqueness and positive semi-definiteness when $\|N_k\|$ and $\|H\|$ are large.

Nonetheless, from a MOR point of view, a solution of these problems can be obtained via rescaling of the system as has been done in the bilinear case [48]. For this, we need to rescale the input variable $u(t)$ as well as the state vector $x(t)$. More precisely, we can replace $x(t)$ and $u(t)$ by $x(t) =: \gamma\widetilde{x}(t)$ and $u(t) =: \gamma\widetilde{u}(t)$ in (4.1). This leads to

$$\begin{aligned}\gamma\dot{\widetilde{x}}(t) &= \gamma A\widetilde{x}(t) + \gamma^2 H\left(\widetilde{x}(t)\otimes\widetilde{x}(t)\right) + \gamma^2\sum_{k=1}^{m} N_k\widetilde{x}(t)\widetilde{u}_k(t) + \gamma B\widetilde{u}(t),\\ y(t) &= \gamma C\widetilde{x}(t), \quad \widetilde{x}(0) = 0.\end{aligned} \tag{5.16}$$

For $\gamma \neq 0$, we get a scaled system as follows:

$$\begin{aligned}\dot{\widetilde{x}}(t) &= A\widetilde{x}(t) + (\gamma H)\left(\widetilde{x}(t)\otimes\widetilde{x}(t)\right) + \sum_{k=1}^{m}(\gamma N_k)\widetilde{x}(t)\widetilde{u}_k(t) + B\widetilde{u}(t),\\ \widetilde{y}(t) &= C\widetilde{x}(t), \quad \widetilde{x}(0) = 0,\end{aligned} \tag{5.17}$$

where $\widetilde{y}(t) = y(t)/\gamma$. Comparing the systems (4.1) and (5.17) shows that the input/output mappings differ by the scaling factor γ. Hence, we can use (5.17) as an auxiliary system during the MOR process; more precisely, to compute the model reduction basis. However, note that the reduced-order system is constructed by applying Petrov-Galerkin projection applied to the original, unscaled matrices in (4.1)
. ◊

Our primary aim is to determine a reduced-order system that minimizes the $\mathcal{H}_2$-norm of the error system. From the derived $\mathcal{H}_2$-norm expression for the QB system, it is clear that the true $\mathcal{H}_2$-norm has a complicated structure as defined in (5.8) and does not lend itself well to deriving necessary conditions for optimality. Therefore, to simplify the problem, we focus only on the three leading terms of the series (5.5). The main reason for considering the first three terms is that it is the minimum number of terms containing contributions from all the system matrices (A, H, N_k, B, C); in other words, linear, bilinear and quadratic terms are already contained in these first three terms. Our approach is also inspired by [59], where a truncated $\mathcal{H}_2$ norm is defined for bilinear systems and used to construct high-fidelity reduced-order models minimizing corresponding error measures. Therefore, based on these three leading terms, we define a truncated $\mathcal{H}_2$-norm for QB systems, denoted by $\|\Sigma_{QB}\|_{\mathcal{H}_2^{(\mathcal{T})}}$. Precisely, the truncated norm can be defined as follows:

$$\|\Sigma_{QB}\|_{\mathcal{H}_2^{(\mathcal{T})}} := \sqrt{\operatorname{tr}\left(\sum_{i=1}^{3}\int_0^\infty\cdots\int_0^\infty \widetilde{f}_i(t_1,\ldots,t_i)\left(\widetilde{f}_i(t_1,\ldots,t_i)\right)^T dt_1\cdots dt_i\right)}, \tag{5.18}$$

where

$$\widetilde{f}_i(t_1,\ldots,t_i) = C\widetilde{g}_i(t_1,\ldots,t_i), \qquad i\in\{1,2,3\}, \tag{5.19}$$

and

$$\begin{aligned}\widetilde{g}_1(t_1) &= e^{At_1}B, \qquad \widetilde{g}_2(t_1,t_2) = e^{At_2}\left[N_1,\ldots,N_m\right]\left(I_m\otimes e^{At_1}B\right),\\ \widetilde{g}_3(t_1,t_2,t_3) &= e^{At_3}H(e^{At_2}B\otimes e^{At_1}B).\end{aligned}$$

Analogous to the $\mathcal{H}_2$-norm of the QB system, a truncated $\mathcal{H}_2$-norm of QB systems can be determined by truncated controllability and observability Gramians associated with the QB system, denoted by $P_\mathcal{T}$ and $Q_\mathcal{T}$, respectively, see Corollary 4.5. If the matrix A is stable, these truncated Gramians (in the integral form) exist, and are the unique and positive semidefinite solutions of the following Lyapunov equations

$$AP_\mathcal{T} + P_\mathcal{T}A^T + \sum_{k=1}^{m} N_k P_l N_k^T + H(P_l \otimes P_l)H^T + BB^T = 0, \tag{5.20a}$$

$$A^TQ_\mathcal{T} + Q_\mathcal{T}A + \sum_{k=1}^{m} N_k^T Q_l N_k + \mathcal{H}^{(2)}(P_l \otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T + C^TC = 0, \tag{5.20b}$$

where $\mathcal{H}^{(2)}$ is the mode-2 matricization of the QB Hessian, and P_l and Q_l are the unique solutions of the following Lyapunov equations:

$$AP_l + P_lA^T + BB^T = 0, \tag{5.21a}$$

$$A^TQ_l + Q_lA + C^TC = 0. \tag{5.21b}$$

In what follows, we show the connection between the truncated $\mathcal{H}_2$-norm and the defined truncated Gramians for QB systems.

Lemma 5.4:
Let Σ_{QB} be the QB system (4.1) with a stable A matrix. Then the truncated $\mathcal{H}_2$-norm based on the first three terms of the Volterra series is given by

$$\|\Sigma_{QB}\|_{\mathcal{H}_2^{(\mathcal{T})}} = \sqrt{\operatorname{tr}\left(CP_\mathcal{T}C^T\right)} = \sqrt{\operatorname{tr}\left(B^TQ_\mathcal{T}B\right)},$$

where $P_\mathcal{T}$ and $Q_\mathcal{T}$ are truncated controllability and observability Gramians of the system, satisfying (5.20). ♢

Proof. First, we note that (5.20) and (5.21) are standard Lyapunov equations. As A is assumed to be stable, these equations have unique solutions [14]. Next, let $\mathcal{R}_i$ be

$$\mathcal{R}_i = \int_0^\infty \cdots \int_0^\infty \widetilde{f}_i(t_1,\ldots,t_i)\left(\widetilde{f}_i(t_1,\ldots,t_i)\right)^T dt_1\cdots dt_i,$$

where $\widetilde{f}_i(t_1,\ldots,t_i)$ are as defined in (5.19). Thus, $\|\Sigma_{QB}\|^2_{\mathcal{H}_2^{(\mathcal{T})}} = \operatorname{tr}\left(C\left(\sum_{i=1}^3 \mathcal{R}_i\right)C^T\right)$. It is shown in Corollary 4.5 that $\sum_{i=1}^3 \mathcal{R}_i = P_\mathcal{T}$ solves the Lyapunov equation (5.20a). Hence,

$$\|\Sigma\|^2_{\mathcal{H}_2^{(\mathcal{T})}} = \operatorname{tr}\left(CP_\mathcal{T}C^T\right).$$

Next, we show that $\operatorname{tr}\left(CP_\mathcal{T}C^T\right) = \operatorname{tr}\left(B^TQ_\mathcal{T}B\right)$. For this, we use the trace property (2.29b) to obtain:

$$\operatorname{tr}\left(CP_\mathcal{T}C^T\right) = (\mathcal{I}_p)^T(C\otimes C)\operatorname{vec}(P_\mathcal{T}) \text{ and } \operatorname{tr}\left(B^TQ_\mathcal{T}B\right) = (\operatorname{vec}(Q_\mathcal{T}))^T(B\otimes B)\mathcal{I}_m.$$

Applying vec $(\cdot)$ to both sides of (5.20) results in

$$\begin{aligned}\operatorname{vec}(P_{\mathcal{T}}) &= \mathcal{L}^{-1}\left((B\otimes B)\,\mathcal{I}_m + \sum_{k=1}^{m}(N_k\otimes N_k)\,\mathcal{L}^{-1}(B\otimes B)\,\mathcal{I}_m\right.\\ &\qquad\qquad\left. + \operatorname{vec}\left(H\,(P_l\otimes P_l)\,H^T\right)\right), \text{ and}\\ \operatorname{vec}(Q_{\mathcal{T}}) &= \mathcal{L}^{-T}\left((C\otimes C)^T\mathcal{I}_p + \sum_{k=1}^{m}(N_k\otimes N_k)^T\mathcal{L}^{-T}(C\otimes C)^T\mathcal{I}_p\right.\\ &\qquad\qquad\left. + \operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\right),\end{aligned}$$

where $\mathcal{L} = -(A\otimes I_n + I_n\otimes A)$, and P_l and Q_l solve (5.21). Thus,

$$\begin{aligned}\operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right) = &\left((\mathcal{I}_p)^T(C\otimes C) + (\mathcal{I}_p)^T(C\otimes C)\mathcal{L}^{-1}\sum_{k=1}^{m}(N_k\otimes N_k)\right.\\ &\left. + \left(\operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\right)^T\right)\mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m.\end{aligned} \tag{5.22}$$

Since P_l and Q_l are the unique solutions of (5.21a) and (5.21b), this gives $\operatorname{vec}(P_l) = \mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m$ and $\operatorname{vec}(Q_l) = \mathcal{L}^{-T}(C\otimes C)^T\mathcal{I}_p$. This implies that

$$\begin{aligned}&\left(\operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\right)^T\operatorname{vec}(P_l)\\ &\qquad = \operatorname{vec}(P_l)^T\operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\\ &\qquad = \operatorname{vec}(Q_l)^T\operatorname{vec}\left(H(P_l\otimes P_l)H^T\right) && \text{(using Lemma 2.27)}\\ &\qquad = (\mathcal{I}_p)^T(C\otimes C)\mathcal{L}^{-1}\operatorname{vec}\left(H(P_l\otimes P_l)H^T\right).\end{aligned}$$

Substituting the above relation in (5.22) yields

$$\begin{aligned}\operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right) &= (\mathcal{I}_p)^T(C\otimes C)\mathcal{L}^{-1}\Big((B\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(N_k\otimes N_k)\mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m\\ &\qquad + \operatorname{vec}\left(H(P_l\otimes P_l)H^T\right)\Big)\\ &= (\mathcal{I}_p)^T(C\otimes C)\operatorname{vec}(P_{\mathcal{T}}) = \operatorname{tr}\left(CP_{\mathcal{T}}C^T\right).\end{aligned}$$

This concludes the proof. □

Remark 5.5:
One can consider the first $\mathcal{M}$ terms of the corresponding Volterra series and, based on these $\mathcal{M}$ kernels, another truncated $\mathcal{H}_2$-norm can be defined. However, this significantly increases the complexity of the problem. In this paper, we stick to the truncated $\mathcal{H}_2$-norm for the QB system that depends on the first three terms of the input/output mapping. We intend to construct reduced-order systems (4.2) such that this truncated $\mathcal{H}_2$-norm of the error system is minimized. Another motivation for the derived truncated $\mathcal{H}_2$-norm for QB systems is that for bilinear systems, the authors in [59] showed that the $\mathcal{H}_2$-optimal model reduction based on a truncated $\mathcal{H}_2$-norm (with only two terms of the Volterra series of a bilinear system) also mimics the accuracy of the true $\mathcal{H}_2$-optimal approximation very closely. $\Diamond$

5.2.2. Optimality conditions based on the truncated $\mathcal{H}_2$-norm

We now derive necessary conditions for optimal model reduction based on the truncated $\mathcal{H}_2$-norm of the error system. First, we define the QB error system. For the full QB model Σ_{QB} in (4.1) and the reduced QB model $\widehat{\Sigma}_{QB}$ in (4.2), we can write the error system as

$$\begin{bmatrix}\dot{x}(t)\\ \dot{\widehat{x}}(t)\end{bmatrix} = \underbrace{\begin{bmatrix}A & \mathbf{0}\\ \mathbf{0} & \widehat{A}\end{bmatrix}}_{A^e}\underbrace{\begin{bmatrix}x(t)\\ \widehat{x}(t)\end{bmatrix}}_{x^e(t)} + \begin{bmatrix}H\left(x(t)\otimes x(t)\right)\\ \widehat{H}\left(\widehat{x}(t)\otimes \widehat{x}(t)\right)\end{bmatrix} + \sum_{k=1}^{m}\underbrace{\begin{bmatrix}N_k & \mathbf{0}\\ \mathbf{0} & \widehat{N}_k\end{bmatrix}}_{N_k^e}\begin{bmatrix}x(t)\\ \widehat{x}(t)\end{bmatrix}u_k(t) + \underbrace{\begin{bmatrix}B\\ \widehat{B}\end{bmatrix}}_{B^e}u(t),$$
$$y^e(t) = y(t) - \widehat{y}(t) = \underbrace{\begin{bmatrix}C & -\widehat{C}\end{bmatrix}}_{C^e}\begin{bmatrix}x^T(t) & \widehat{x}^T(t)\end{bmatrix}^T, \quad x^e(0) = 0. \tag{5.23}$$

It can be seen that the error system (5.23) is not in the conventional QB form due to the absence of the quadratic term $x^e(t)\otimes x^e(t)$. However, we can rewrite the system (5.23) into a regular QB form by using an appropriate Hessian of the error system (5.23) as follows:

$$\Sigma^e := \begin{cases}\dot{x}^e(t) = A^e x^e(t) + H^e\left(x^e(t)\otimes x^e(t)\right) + \displaystyle\sum_{k=1}^{m} N_k^e x^e(t)u_k(t) + B^e u(t),\\ y^e(t) = C^e x^e(t), \quad x^e(0) = 0,\end{cases} \tag{5.24}$$

where $H^e = \begin{bmatrix}H\mathcal{F}\\ \widehat{H}\widehat{\mathcal{F}}\end{bmatrix}$ with $\mathcal{F} = \begin{bmatrix}I_n & \mathbf{0}\end{bmatrix}\otimes\begin{bmatrix}I_n & \mathbf{0}\end{bmatrix}$ and $\widehat{\mathcal{F}} = \begin{bmatrix}\mathbf{0} & I_{\widehat{n}}\end{bmatrix}\otimes\begin{bmatrix}\mathbf{0} & I_{\widehat{n}}\end{bmatrix}$. Next, we consider the truncated $\mathcal{H}_2$-norm, as defined in Lemma 5.4, for the error system (5.24). For the existence of this norm for the system (5.24), it is necessary to assume that the matrix A^e is stable, i.e., the matrices A and $\widehat{A}$ are stable. Further, we assume that the matrix $\widehat{A}$ is diagonalizable. Then, by performing basic algebraic manipulations and making use of Lemma 2.29, we obtain the expression for the error functional $\mathcal{E}$ based on the truncated $\mathcal{H}_2$-norm of the error system (5.24) as shown next.

Corollary 5.6:
Let Σ_{QB} be the original system, having a stable matrix A, and let $\widehat{\Sigma}_{QB}$ be the reduced-order system, having a stable and diagonalizable matrix $\widehat{A}$. Then,

$$\mathcal{E}^2 := \|\Sigma^e\|^2_{\mathcal{H}_2^{(\mathcal{T})}} = (\mathcal{I}_p)^T(C^e \otimes C^e)(-A^e \otimes I_{n+\widehat{n}} - I_{n+\widehat{n}} \otimes A^e)^{-1}\Big((B^e \otimes B^e)\mathcal{I}_m \\ + \sum_{k=1}^{m}(N_k^e \otimes N_k^e)\,\mathrm{vec}\,(P_l^e) + \mathrm{vec}\left(H^e(P_l^e \otimes P_l^e)\,(H^e)^T\right)\Big), \quad (5.25)$$

where P_l^e solves

$$A^eP_l^e + P_l^e(A^e)^T + B^e(B^e)^T = 0.$$

Furthermore, let $\widehat{A} = \widehat{R}\widehat{\Lambda}\widehat{R}^{-1}$ be the spectral decomposition of $\widehat{A}$, and define $\widetilde{B} = \widehat{R}^{-1}\widehat{B}$, $\widetilde{C} = \widehat{C}\widehat{R}$, $\widetilde{N}_k = \widehat{R}^{-1}\widehat{N}_k\widehat{R}$ and $\widetilde{H} = \widehat{R}^{-1}\widehat{H}(\widehat{R} \otimes \widehat{R})$. Then, the error can be rewritten as

$$\mathcal{E}^2 = (\mathcal{I}_p)^T\left(\widetilde{C}^e \otimes \widetilde{C}^e\right)\left(-\widetilde{A}^e \otimes I_{n+\widehat{n}} - I_{n+\widehat{n}} \otimes \widetilde{A}^e\right)^{-1}\Bigg(\left(\widetilde{B}^e \otimes \widetilde{B}^e\right)\mathcal{I}_m \\ + \sum_{k=1}^{m}\left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e\right)\mathcal{P}_l + \left(\widetilde{H}^e \otimes \widetilde{H}^e\right)T_{(n+\widehat{n},n+\widehat{n})}(\mathcal{P}_l \otimes \mathcal{P}_l)\Bigg), \quad (5.26)$$

where

$$\widetilde{A}^e = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & \Lambda \end{bmatrix}, \widetilde{N}_k^e = \begin{bmatrix} N_k & \mathbf{0} \\ \mathbf{0} & \widetilde{N}_k \end{bmatrix}, \widetilde{H}^e = \begin{bmatrix} H\mathcal{F} \\ \widetilde{H}\widehat{\mathcal{F}} \end{bmatrix}, \widetilde{B}^e = \begin{bmatrix} B \\ \widetilde{B} \end{bmatrix}, \widetilde{C}^e = \begin{bmatrix} C^T \\ -\widetilde{C}^T \end{bmatrix}^T,$$

$$\mathcal{P}_l = \begin{bmatrix} \mathcal{P}_l^{(1)} \\ \mathcal{P}_l^{(2)} \end{bmatrix} = \begin{bmatrix} \left(-A \otimes I_{n+\widehat{n}} - I_n \otimes \widetilde{A}^e\right)^{-1}\left(B \otimes \widetilde{B}^e\right)\mathcal{I}_m \\ \left(-\Lambda \otimes I_{n+\widehat{n}} - I_{\widehat{n}} \otimes \widetilde{A}^e\right)^{-1}\left(\widetilde{B} \otimes \widetilde{B}^e\right)\mathcal{I}_m \end{bmatrix}, \text{and} \quad (5.27)$$

$$T_{(n+\widehat{n},n+\widehat{n})} = I_{n+\widehat{n}} \otimes \left[I_{n+\widehat{n}} \otimes e_1^{n+\widehat{n}}, \ldots, I_{n+\widehat{n}} \otimes e_{n+\widehat{n}}^{n+\widehat{n}}\right] \otimes I_{n+\widehat{n}}.$$

◊

The above spectral decomposition for $\widehat{A}$ is computationally useful in simplifying the expressions as we will see later. It reduces the number of optimization variables by $r(r-1)$ since Λ becomes a diagonal matrix without changing the value of the cost function (this is a state-space transformation of the reduced model, which does not change the input-output mapping). Even though it limits the reduced-order systems to those only having diagonalizable $\widehat{A}$, as observed in the linear [79] and bilinear cases [21, 59], it is extremely rare in practice that the optimal $\mathcal{H}_2$ models will have a non-diagonalizable

$\widehat{A}$; therefore, this diagonalizability assumption does not incur any restriction from a practical perspective.

Our aim is to choose the optimization variables Λ, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{N}_k$ and $\widetilde{H}$ such that the $\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2^{(\mathcal{T})}}$, i.e., equivalently the error expression (5.26), is minimized. Before we proceed further, we introduce a particular permutation matrix

$$M_{pqr} = \left[I_p \otimes \begin{bmatrix} I_q \\ \mathbf{0} \end{bmatrix} \quad I_p \otimes \begin{bmatrix} \mathbf{0} \\ I_{\widehat{n}} \end{bmatrix} \right], \tag{5.28}$$

which will prove helpful in simplifying the expressions related to the Kronecker product of block matrices. For example, consider matrices $\mathcal{A} \in \mathbb{R}^{p\times p}$, $\mathcal{B} \in \mathbb{R}^{q\times q}$ and $\mathcal{C} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$. Then, the following relation holds:

$$M_{pqr}^T \left(\mathcal{A} \otimes \begin{bmatrix} \mathcal{B} & \mathbf{0} \\ \mathbf{0} & \mathcal{C} \end{bmatrix} \right) M_{pqr} = \begin{bmatrix} \mathcal{A} \otimes \mathcal{B} & \mathbf{0} \\ \mathbf{0} & \mathcal{A} \otimes \mathcal{C} \end{bmatrix}.$$

Similar block structures can be found in the error expression $\mathcal{E}$ in Corollary 5.6, which can be simplified analogously. Moreover, due to the presence of many Kronecker products, it will be convenient to derive necessary conditions for optimality in the Kronecker product formulation itself. Furthermore, these conditions can be easily translated into a theoretically equivalent framework of Sylvester equations, which are more concise, are more easily interpretable, and, more importantly, automatically lead to an effective numerical algorithm for model reduction. To this end, let $V_i \in \mathbb{R}^{n\times\widehat{n}}$ and $W_i \in \mathbb{R}^{n\times\widehat{n}}$, $i \in \{1, 2\}$, be the solutions of the following standard Sylvester equations:

$$V_1(-\Lambda) - AV_1 = B\widetilde{B}^T, \tag{5.29a}$$

$$W_1(-\Lambda) - A^TW_1 = C^T\widetilde{C}, \tag{5.29b}$$

$$V_2(-\Lambda) - AV_2 = \sum_{k=1}^{m} N_kV_1\widetilde{N}_k^T + H(V_1 \otimes V_1)\widetilde{H}^T, \text{ and,} \tag{5.29c}$$

$$W_2(-\Lambda) - A^TW_2 = \sum_{k=1}^{m} N_k^TW_1\widetilde{N}_k + 2 \cdot \mathcal{H}^{(2)}(V_1 \otimes W_1)(\widetilde{\mathcal{H}}^{(2)})^T, \tag{5.29d}$$

where Λ, $\widetilde{N}_k$, $\widetilde{B}$ and $\widetilde{C}$ are as defined in Corollary 5.6. Furthermore, we define trial and test basis matrices $V \in \mathbb{R}^{n\times\widehat{n}}$ and $W \in \mathbb{R}^{n\times\widehat{n}}$ as

$$V = V_1 + V_2 \quad \text{and} \quad W = W_1 + W_2. \tag{5.30}$$

We also define $\widehat{V} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ and $\widehat{W} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ (which will appear in the optimality conditions as we see later) as follows:

$$\widehat{V} = \widehat{V}_1 + \widehat{V}_2 \quad \text{and} \quad \widehat{W} = \widehat{W}_1 + \widehat{W}_2, \tag{5.31}$$

where $\widehat{V}_i \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$, $\widehat{W}_i \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$, $i \in \{1,2\}$ are the solutions of the set of equations in (5.29) but with the original system's state-space matrices being replaced with the reduced-order system ones; for example, A with $\widehat{A}$ and B with $\widehat{B}$, etc. Next, we present first order necessary conditions for optimality, which aim at minimizing the error expression (5.26). The following theorem extends the truncated $\mathcal{H}_2$ optimal conditions from the bilinear case to the much more general quadratic-bilinear nonlinearities.

Theorem 5.7:
Let Σ_{QB} and $\widehat{\Sigma}_{QB}$ be the original and reduced-order systems as defined in (4.1) and (4.2), respectively. Let $\widehat{\Lambda} = \widehat{R}^{-1}\widehat{A}\widehat{R}$ be the spectral decomposition of $\widehat{A}$, and define $\widetilde{H} = \widehat{R}^{-1}\widehat{H}(\widehat{R}\otimes\widehat{R})$, $\widetilde{N}_k = \widehat{R}^{-1}\widehat{N}_k\widehat{R}$, $\widetilde{C} = \widehat{C}R$, $\widetilde{B} = \widehat{R}^{-1}\widehat{B}$. If $\widehat{\Sigma}_{QB}$ is a reduced-order system that minimizes the truncated $\mathcal{H}_2$-norm of the error system (5.24) subject to $\widehat{A}$ being diagonalizable, then $\widehat{\Sigma}_{QB}$ satisfies the following conditions:

$$\operatorname{tr}\left(CVe_i^{\widehat{n}}\left(e_j^p\right)^T\right) = \operatorname{tr}\left(\widehat{C}\widehat{V}e_i^{\widehat{n}}\left(e_j^p\right)^T\right), \qquad i \in \{1,\dots,\widehat{n}\},\quad j \in \{1,\dots,p\}, \tag{5.32a}$$

$$\operatorname{tr}\left(B^TWe_i^{\widehat{n}}\left(e_j^m\right)^T\right) = \operatorname{tr}\left(\widehat{B}^T\widehat{W}e_i^{\widehat{n}}\left(e_j^m\right)^T\right), \qquad i \in \{1,\dots,\widehat{n}\},\quad j \in \{1,\dots,m\}, \tag{5.32b}$$

$$(W_1(:,i))^TN_kV_1(:,j) = (\widehat{W}_1(:,i))^T\widehat{N}_k\widehat{V}_1(:,j), \qquad i,j \in \{1,\dots,\widehat{n}\},\quad k \in \{1,\dots,m\}, \tag{5.32c}$$

$$(W_1(:,i))^TH(V_1(:,j)\otimes V_1(:,l)) = (\widehat{W}_1(:,i))^T\widehat{H}(\widehat{V}_1(:,j)\otimes\widehat{V}_1(:,l)), \qquad i,j,l \in \{1,\dots,\widehat{n}\}, \tag{5.32d}$$

$$(W_1(:,i))^TV(:,i) + (W_2(:,i))^T V_1(:,i) = (\widehat{W}_1(:,i))^T\widehat{V}(:,i) + \left(\widehat{W}_2(:,i)\right)^T \widehat{V}_1(:,i), \qquad i \in \{1,\dots,\widehat{n}\}. \tag{5.32e}$$

◊

Proof. We start with deriving the optimality conditions by taking the derivative of the error functional $\mathcal{E}$ (5.26) with respect to $\widetilde{C}$. By using Lemma B.1, we obtain

$$\frac{\partial\mathcal{E}^2}{\partial\widetilde{C}_{ij}} = 2(\mathcal{I}_p)^T\left(\left[0 \ \ -e_i^p(e_j^{\widehat{n}})^T\right]\otimes\widetilde{C}^e\right)\left(-\widetilde{A}^e\otimes I_{n+\widehat{n}} - I_{n+\widehat{n}}\otimes\widetilde{A}^e\right)^{-1}$$
$$\left(\left(\widetilde{B}^e\otimes\widetilde{B}^e\right)\mathcal{I}_m + \sum_{k=1}^m\left(\widetilde{N}_k^e\otimes\widetilde{N}_k^e\right)\mathcal{P}_l + \left(\widetilde{H}^e\otimes\widetilde{H}^e\right)T_{(n+\widehat{n},n+\widehat{n})}(\mathcal{P}_l\otimes\mathcal{P}_l)\right),$$

where $\mathcal{P}_l$ is defined in (5.27). Simplifying this expression, we get

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T \left(-e_i^p(e_j^{\widehat{n}})^T \otimes \widetilde{C}^e\right) \left(-\Lambda \otimes I_{n+\widehat{n}} - I_{\widehat{n}} \otimes \widetilde{A}^e\right)^{-1} \left(\left(\widetilde{B} \otimes \widetilde{B}^e\right) \mathcal{I}_m \right.\\ &\quad \left. + \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{P}_l^{(2)} + \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})}(\mathcal{P}_l \otimes \mathcal{P}_l)\right),\\ &= 2(\mathcal{I}_p)^T \left(-e_i^p(e_j^{\widehat{n}})^T \otimes \widetilde{C}^e\right) \left(M_{\widehat{n}n\widehat{n}}\left(-\mathcal{J}_\Lambda - \mathcal{J}_A\right) M_{\widehat{n}n\widehat{n}}^T\right)^{-1} \left(\left(\widetilde{B} \otimes \widetilde{B}^e\right) \mathcal{I}_m \right.\\ &\quad \left. + \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{P}_l^{(2)} + \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})}(\mathcal{P}_l \otimes \mathcal{P}_l)\right),\end{aligned} \tag{5.33}$$

where

$$\mathcal{J}_\Lambda = \begin{bmatrix} \Lambda \otimes I_n & 0 \\ 0 & \Lambda \otimes I_{\widehat{n}} \end{bmatrix}, \qquad \mathcal{J}_A = \begin{bmatrix} I_{\widehat{n}} \otimes A & 0 \\ 0 & I_{\widehat{n}} \otimes \Lambda \end{bmatrix}, \quad \text{and}$$

$\mathcal{P}_l^{(2)}$ is the lower block row of $\mathcal{P}_l$ as shown in (5.27). Furthermore, since $M_{\widehat{n}n\widehat{n}}$ is a permutation matrix, this implies $M_{\widehat{n}n\widehat{n}}M_{\widehat{n}n\widehat{n}}^T = I$. Using this relation in (5.33), we obtain

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T \left(-e_i^p(e_j^{\widehat{n}})^T \otimes \begin{bmatrix} C & -\widetilde{C}\end{bmatrix}\right) M_{\widehat{n}n\widehat{n}}\left(-\mathcal{J}_\Lambda - \mathcal{J}_A\right)^{-1} \left(M_{\widehat{n}n\widehat{n}}^T \left(\widetilde{B} \otimes \widetilde{B}^e\right) \mathcal{I}_m \right.\\ &\quad \left. + M_{\widehat{n}n\widehat{n}}^T \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{P}_1^{(2)} + M_{\widehat{n}n\widehat{n}}^T \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})}(\mathcal{P}_l \otimes \mathcal{P}_l)\right)\\ &= 2(\mathcal{I}_p)^T \left(\begin{bmatrix} -e_i^p(e_j^{\widehat{n}})^T \otimes C & e_i e_j^T \otimes \widetilde{C}\end{bmatrix}\right) \left(-\mathcal{J}_\Lambda - \mathcal{J}_A\right)^{-1} \left(\begin{bmatrix} \widetilde{B} \otimes B \\ \widetilde{B} \otimes \widetilde{B}\end{bmatrix} \mathcal{I}_m \right.\\ &\quad + \sum_{k=1}^{m} \begin{bmatrix} \widetilde{N}_k \otimes N_k & 0 \\ 0 & \widetilde{N}_k \otimes \widetilde{N}_k \end{bmatrix} M_{\widehat{n}n\widehat{n}}^T \mathcal{P}_1^{(2)}\\ &\quad \left. + \begin{bmatrix} \left(\widetilde{H}\widehat{\mathcal{F}} \otimes H\mathcal{F}\right) T_{(n+\widehat{n},n+\widehat{n})}(M \otimes M)(M^T \otimes M^T)(\mathcal{P}_l \otimes \mathcal{P}_l) \\ \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}\widehat{\mathcal{F}}\right) T_{(n+\widehat{n},n+\widehat{n})}(M \otimes M)(M^T \otimes M^T)(\mathcal{P}_l \otimes \mathcal{P}_l)\end{bmatrix} \right),\end{aligned} \tag{5.34}$$

where M is the permutation matrix defined in (B.1). The multiplication of M^T and $\mathcal{P}_l$ yields

$$M^T\mathcal{P}_l = \begin{bmatrix} M_{nn\widehat{n}}\mathcal{P}_l^{(1)} \\ M_{\widehat{n}n\widehat{n}}\mathcal{P}_1^{(2)} \end{bmatrix} = \begin{bmatrix} p_1^T & p_2^T & p_3^T & p_4^T \end{bmatrix}^T =: \widetilde{\mathcal{P}}_l,$$

where

$$\begin{aligned} p_1 &= (-A\otimes I_n - I_n\otimes A)^{-1}(B\otimes B)\,\mathcal{I}_m, \quad p_2 = (-A\otimes I_{\widehat{n}} - I_n\otimes \Lambda)^{-1}\left(B\otimes \widetilde{B}\right)\mathcal{I}_m,\\ p_3 &= (-\Lambda\otimes I_n - I_{\widehat{n}}\otimes A)^{-1}\left(\widetilde{B}\otimes B\right)\mathcal{I}_m, \quad p_4 = (-\Lambda\otimes I_{\widehat{n}} - I_{\widehat{n}}\otimes \Lambda)^{-1}\left(\widetilde{B}\otimes \widetilde{B}\right)\mathcal{I}_m. \end{aligned} \tag{5.35}$$

Moreover, note that $p_3 = \operatorname{vec}(V_1)$, where V_1 solves (5.29a). Applying the result of Lemma B.2 in (5.34) yields

$$\begin{aligned} \frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes C\right)(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes A)^{-1}\Bigg((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes N_k)p_3\\ &\qquad + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Bigg) - 2(\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes \widetilde{C}\right)\left(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes \widetilde{A}\right)^{-1}\\ &\qquad \times\Bigg((\widetilde{B}\otimes \widetilde{B})\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes \widetilde{N}_k)p_4 + (\widetilde{H}\otimes \widetilde{H})T_{(\widehat{n},\widehat{n})}(p_4\otimes p_4)\Bigg)\\ &= 2(\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes C\right)(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes A)^{-1}\Bigg((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes N_k)p_3\\ &\qquad + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Bigg) - 2(\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes \widehat{C}\right)\left(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes \widehat{A}\right)^{-1}\\ &\qquad \times\Bigg((\widetilde{B}\otimes \widehat{B})\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes \widehat{N}_k)\widehat{p}_4 + (\widetilde{H}\otimes \widehat{H})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes \widehat{p}_4)\Bigg), \end{aligned} \tag{5.36}$$

where $\widehat{p}_4 = \left(-\Lambda\otimes I_{\widehat{n}} - I_{\widehat{n}}\otimes \widehat{A}\right)^{-1}\left(\widetilde{B}\otimes \widehat{B}\right)\mathcal{I}_m = \operatorname{vec}\left(\widehat{V}_1\right)$, where $\widehat{V}_1$ is as defined in (5.31). Setting (5.36) equal to zero results in a necessary condition with respect to $\widetilde{C}$ as follows:

$$\begin{aligned} &(\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes C\right)(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes A)^{-1}\Bigg((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes N_k)p_3\\ &\qquad + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Bigg)\\ &= (\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes \widehat{C}\right)\left(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes \widehat{A}\right)^{-1}\\ &\qquad \times\Bigg((\widetilde{B}\otimes \widehat{B})\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes \widehat{N}_k)\widehat{p}_4 + (\widetilde{H}\otimes \widehat{H})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes \widehat{p}_4)\Bigg). \end{aligned} \tag{5.37}$$

Now, we first manipulate the left-hand side of (5.37). Using Lemma 2.29 and (2.29), we get

$$(\mathcal{I}_p)^T\left(e_i^p(e_j^{\widehat{n}})^T\otimes C\right)(-\Lambda\otimes I_n - I_{\widehat{n}}\otimes A)^{-1}\Bigg((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^m(\widetilde{N}_k\otimes N_k)p_3$$

$$
\begin{aligned}
&\quad + (\widetilde{H} \otimes H) T_{(n,\widehat{n})}(p_3 \otimes p_3)\Big) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^{\widehat{n}})^T \otimes C\right) (-\Lambda \otimes I_n - I_{\widehat{n}} \otimes A)^{-1} \left(\operatorname{vec}\left(B\widetilde{B}^T\right) + \sum_{k=1}^{m} \operatorname{vec}\left(N_k V_1 \widetilde{N}_k^T\right) \right. \\
&\quad \left. + (\widetilde{H} \otimes H) \operatorname{vec}(V_1 \otimes V_1) \right) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^{\widehat{n}})^T \otimes C\right) (-\Lambda \otimes I_n - I_{\widehat{n}} \otimes A)^{-1} \left(\operatorname{vec}\left(B\widetilde{B}^T + \sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T\right) \right. \\
&\quad \left. + \operatorname{vec}\left(H(V_1 \otimes V_1)\widetilde{H}^T\right) \right) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^{\widehat{n}})^T \otimes C\right) (\operatorname{vec}(V_1) + \operatorname{vec}(V_2)) = \operatorname{tr}\left(C(V_1 + V_2) e_j^{\widehat{n}} (e_i^p)^T\right) \\
&= \operatorname{tr}\left(C V e_j^{\widehat{n}} (e_i^p)^T\right),
\end{aligned}
$$

where V_2 solves (5.29c) and $V = V_1 + V_2$. Using the similar steps, we can show that the righthand side of (5.37) is equal to $\operatorname{tr}\left(\widehat{C}\widehat{V} e_j^{\widehat{n}}(e_i^p)^T\right)$, where $\widehat{V}$ is defined in (5.31). Therefore, Eq. (5.37) is the same as (5.32a).

Necessary conditions with respect to Λ

By utilizing Lemma B.1, we aim at deriving the necessary condition with respect to the ith diagonal entry of Λ. We differentiate $\mathcal{E}$ with respect to λ_i to obtain

$$
\begin{aligned}
\frac{\partial \mathcal{E}^2}{\partial \lambda_i} &= 2(\mathcal{I}_p)^T \left(\widetilde{C}^e \otimes \widetilde{C}^e\right) \mathcal{L}_e^{-1} \mathbb{E} \mathcal{L}_e^{-1} \Big(\left(\widetilde{B}^e \otimes \widetilde{B}^e\right) \mathcal{I}_m + \sum_{k=1}^{m} \left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e\right) \mathcal{P}_l && (5.38) \\
&\quad + \left(\widetilde{H}^e \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})} (\mathcal{P}_l \otimes \mathcal{P}_l) \Big) + (\mathcal{I}_p)^T \left(\widetilde{C}^e \otimes \widetilde{C}^e\right) \mathcal{L}_e^{-1} && (5.39) \\
&\quad \times \left(2 \sum_{k=1}^{m} \left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e\right) \mathcal{L}_e^{-1} \mathbb{E} \mathcal{P}_l + 4 \left(\widetilde{H}^e \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})} \left(\left(\mathcal{L}_e^{-1} \mathbb{E} \mathcal{P}_l\right) \otimes \mathcal{P}_l \right) \right), \\
& && (5.40)
\end{aligned}
$$

where

$$
\mathcal{L}_e = -\left(\widetilde{A}^e \otimes I_{n+\widehat{n}} + I_{n+\widehat{n}} \otimes \widetilde{A}^e\right) \quad \text{and} \quad \mathbb{E} = \begin{bmatrix} 0 & 0 \\ 0 & e_i^{\widehat{n}} (e_i^{\widehat{n}})^T \end{bmatrix} \otimes I_{n+\widehat{n}}.
$$

Performing some algebraic calculations gives rise to the following expression:

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \lambda_i} &= 2(\mathcal{I}_p)^T \left(-\widetilde{C} \otimes \widetilde{C}^e\right) \mathcal{Z}_e^{-1} \Xi_{n+r} \mathcal{Z}_e^{-1} \Big(\left(\widetilde{B} \otimes \widetilde{B}^e\right) \mathcal{I}_m + \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{P}_1^{(2)} \\ &\quad + \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})} (\mathcal{P}_l \otimes \mathcal{P}_l)\Big) + 2(\mathcal{I}_p)^T \left(-\widetilde{C} \otimes \widetilde{C}^e\right) \mathcal{Z}_e^{-1} \\ &\quad \times \left(\sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{Z}_e^{-1} \Xi_{n+r} \mathcal{P}_1^{(2)} + 2 \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+\widehat{n},n+\widehat{n})} (\mathcal{L}_e^{-1} \mathbb{E} \mathcal{P}_l \otimes \mathcal{P}_l)\right),\end{aligned}$$

where $\mathcal{Z}_e := -\left(\Lambda \otimes I_{n+\widehat{n}} + I_{\widehat{n}} \otimes A^e\right)$ and $\Xi_m := (e_i^{\widehat{n}}(e_i^{\widehat{n}})^T \otimes I_m)$. Next, we utilize Lemma B.2 and use the permutation matrix M (as done while deriving the necessary conditions with respect to $\widetilde{C}$) to obtain

$$\begin{aligned}&\frac{\partial \mathcal{E}^2}{\partial \lambda_i} \\ &\quad = 2(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B} \otimes B)\mathcal{I}_m + \textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) p_3 + (\widetilde{H} \otimes H) T_{(n,\widehat{n})} (p_3 \otimes p_3)\Big) \\ &\quad - 2(\mathcal{I}_p)^T \widetilde{\mathcal{S}}\Big((\widetilde{B} \otimes \widetilde{B})\mathcal{I}_m + \textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes \widetilde{N}_k) p_4 + (\widetilde{H} \otimes \widetilde{H}) T_{(\widehat{n},\widehat{n})} (p_4 \otimes p_4)\Big) \\ &\quad + 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes C\right) L^{-1}\Big(\textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) L^{-1} \Xi_n p_3 + 2(\widetilde{H} \otimes H) T_{(n,\widehat{n})} (L^{-1} \Xi_n (p_3 \otimes p_3))\Big) \\ &\quad - 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes \widetilde{C}\right) \widetilde{L}^{-1}\Big(\textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes \widetilde{N}_k) \widetilde{L}^{-1} \Xi_{\widehat{n}} p_4 + 2(\widetilde{H} \otimes \widetilde{H}) T_{(\widehat{n},\widehat{n})} (\widetilde{L}^{-1} \Xi_{\widehat{n}} (p_4 \otimes p_4))\Big) \\ &\quad = 2(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B} \otimes B)\mathcal{I}_m + \textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) p_3 + (\widetilde{H} \otimes H) T_{(n,\widehat{n})} (p_3 \otimes p_3)\Big) \\ &\quad - 2(\mathcal{I}_p)^T \widehat{\mathcal{S}}\Big((\widetilde{B} \otimes \widehat{B})\mathcal{I}_m + \textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes \widehat{N}_k) \widehat{p}_4 + (\widetilde{H} \otimes \widehat{H}) T_{(\widehat{n},\widehat{n})} (\widehat{p}_4 \otimes \widehat{p}_4)\Big) \\ &\quad + 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes C\right) L^{-1}\Big(\textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) L^{-1} \Xi_n p_3 + 2(\widetilde{H} \otimes H) T_{(n,\widehat{n})} (L^{-1} \Xi_n (p_3 \otimes p_3))\Big) \\ &\quad - 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{L}^{-1}\Big(\textstyle\sum_{k=1}^{m} (\widetilde{N}_k \otimes \widehat{N}_k) \widehat{L}^{-1} \Xi_{\widehat{n}} \widehat{p}_4 + 2(\widetilde{H} \otimes \widehat{H}) T_{(\widehat{n},\widehat{n})} (\widehat{L}^{-1} \Xi_{\widehat{n}} (\widehat{p}_4 \otimes \widehat{p}_4)\Big),\end{aligned}$$

where p_3 and p_4 are the same as defined in (5.35), and

$$\begin{aligned}\mathcal{S} &:= \left(\widetilde{C} \otimes C\right) L^{-1}(e_i^{\widehat{n}}(e_i^{\widehat{n}})^T \otimes I_n) L^{-1}, & \widetilde{\mathcal{S}} &:= \left(\widetilde{C} \otimes \widetilde{C}\right) \widetilde{L}^{-1}(e_i^{\widehat{n}}(e_i^{\widehat{n}})^T \otimes I_{\widehat{n}}) \widetilde{L}^{-1}, \\ \widehat{\mathcal{S}} &:= \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{L}^{-1}(e_i^{\widehat{n}}(e_i^{\widehat{n}})^T \otimes I_{\widehat{n}}) \widehat{L}^{-1}, & L &:= -\left(\Lambda \otimes I_n + I_{\widehat{n}} \otimes A\right), \\ \widetilde{L} &:= -\left(\Lambda \otimes I_{\widehat{n}} + I_{\widehat{n}} \otimes \Lambda\right), & \widehat{L} &:= -\left(\Lambda \otimes I_{\widehat{n}} + I_{\widehat{n}} \otimes \widehat{A}\right).\end{aligned}$$

By using the properties derived in Lemma 2.27, we can simplify the above equation:

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \lambda_i} &= 2(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \textstyle\sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Big)\\
&\quad - 2(\mathcal{I}_p)^T \widehat{\mathcal{S}}\Big((\widetilde{B}\otimes \widehat{B})\mathcal{I}_m + \textstyle\sum_{k=1}^{m}(\widetilde{N}_k \otimes \widehat{N}_k)p_4 + (\widetilde{H}\otimes \widehat{H})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes \widehat{p}_4)\Big)\\
&\quad + 2(\mathcal{I}_m)^T\left(\widetilde{B}\otimes B\right)L^{-T}\Xi_n L^{-T}\Big(\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \mathcal{H}^{(2)})T_{(n,\widehat{n})}(p_3\otimes q_3)\Big)\\
&\quad - 2(\mathcal{I}_m)^T\left(\widetilde{B}\otimes \widehat{B}\right)\widehat{L}^{-T}\Xi_{\widehat{n}}\widehat{L}^{-T}\Big(\sum_{k=1}^{m}(\widetilde{N}_k\otimes \widehat{N}_k)^T \widehat{q}_4 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \widehat{\mathcal{H}}^{(2)})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes \widehat{q}_4)\Big),\end{aligned}$$

where

$$q_3 = (-\Lambda\otimes I_n - I_{\widehat{n}}\otimes A)^{-T}\left(\widetilde{C}\otimes C\right)\mathcal{I}_p \quad\text{and}\quad \widehat{q}_4 = (-\Lambda\otimes I_{\widehat{n}} - I_{\widehat{n}}\otimes A_r)^{-T}\left(\widetilde{C}\otimes \widehat{C}\right)\mathcal{I}_p.$$

Once again, we determine an interpolation-based necessary condition with respect to Λ_i by setting the last equation equal to zero:

$$\begin{aligned}&(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Big)\\
&\quad + (\mathcal{I}_m)^T\left(\widetilde{B}\otimes B\right)L^{-T}\Xi_n L^{-T}\Big(\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \mathcal{H}^{(2)})T_{(n,\widehat{n})}(p_3\otimes q_3)\Big)\\
&\quad = (\mathcal{I}_p)^T \widehat{\mathcal{S}}\Big((\widetilde{B}\otimes \widehat{B})\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes \widehat{N}_k)p_4 + (\widetilde{H}\otimes \widehat{H})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes \widehat{p}_4)\Big),\\
&\quad + (\mathcal{I}_m)^T\left(\widetilde{B}\otimes \widehat{B}\right)\widehat{L}^{-T}\Xi_{\widehat{n}}\widehat{L}^{-T}\Big(\sum_{k=1}^{m}(\widetilde{N}_k\otimes \widehat{N}_k)^T \widehat{q}_4 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \widehat{\mathcal{H}}^{(2)})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes \widehat{q}_4)\Big).\end{aligned} \tag{5.41}$$

Now, we first simplify the left-hand side of the above equation using Lemma 2.29 and (2.29). We first focus of the first part of the left-hand side of (5.41). This yields

$$\begin{aligned}&(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Big)\\
&\qquad = (\mathcal{I}_p)^T\left(\widetilde{C}\otimes C\right)L^{-1}(e_i^{\widehat{n}}(e_i^{\widehat{n}})^T\otimes I_n)\\
&\qquad\quad \times L^{-1}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\Big)\\
&\qquad = \underbrace{(\mathcal{I}_p)^T\left(\widetilde{C}\otimes C\right)L^{-1}}_{(\mathrm{vec}(W_1))^T}(e_i^{\widehat{n}}(e_i^{\widehat{n}})^T\otimes I_n)\,\mathrm{vec}\,(V) = \mathrm{tr}\left(V e_i^{\widehat{n}}(e_i^{\widehat{n}})^T W_1^T\right)\\
&\qquad = \left(V_1(:,i)\right)^T W(:,i) = \left(W_1(:,i)\right)^T V(:,i),\end{aligned}$$

where W_1 solves (5.29b). Analogously, we can show that

$$\begin{aligned}(\mathcal{I}_m)^T\left(\widetilde{B}\otimes B\right)L^{-T}\Xi_n L^{-T}\Big(\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3+2(\widetilde{\mathcal{H}}^{(2)}\otimes\mathcal{H}^{(2)})T_{(n,\widehat{n})}(p_3\otimes q_3)\Big)\\ =(W(:,i))^T V_1(:,i).\end{aligned} \tag{5.42}$$

Thus, the left-hand side of (5.41) is equal to $(W(:,i))^T V_1(:,i)+(W_1(:,i))^T V(:,i)$. Using similar steps, we can also show that the right-hand side of (5.41) is equal to $\left(\widehat{W}(:,i)\right)^T\widehat{V}_1(:,i)+\left(\widehat{W}_1(:,i)\right)^T\widehat{V}(:,i)$. Thus, we obtain the optimality conditions with respect to Λ given in (5.32e).

The necessary conditions with respect to $\widetilde{B}$, $\widetilde{N}$ and $\widetilde{H}$ can also be determined in a similar manner as for $\widetilde{C}$ and λ_i. For brevity of the paper, we skip detailed derivations; however, we state the final optimality conditions. A necessary condition for optimality with respect to the (i,j) entry of $\widetilde{N}_k$ is

$$(\mathcal{I}_p)^T\left(\widetilde{C}\otimes C\right)L^{-1}\left((e_i^{\widehat{n}}(e_j^{\widehat{n}})^T\otimes N_k)p_3\right)=(\mathcal{I}_p)^T\left(\widetilde{C}\otimes\widehat{C}\right)\widehat{L}^{-1}\left((e_i^{\widehat{n}}(e_j^{\widehat{n}})^T\otimes\widehat{N}_k)\widehat{p}_4\right),$$

which then yields (5.32c) in the Sylvester equation form. A similar optimality condition with respect to the (i,j) entry of $\widetilde{H}$ is given by

$$\begin{aligned}&(\mathcal{I}_p)^T\left(\widetilde{C}\otimes C\right)L^{-1}\left((e_i^{\widehat{n}}(e_j^{\widehat{n}^2})^T\otimes H)T_{(n,\widehat{n})}(p_3\otimes p_3)\right)\\ &\qquad=(\mathcal{I}_p)^T\left(\widetilde{C}\otimes\widehat{C}\right)\widehat{L}^{-1}\left((e_i^{\widehat{n}}(e_j^{\widehat{n}^2})^T\otimes\widehat{H})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes\widehat{p}_4)\right),\end{aligned}$$

which can be equivalently described as (5.32d). Finally, the necessary condition appearing with respect to the (i,j) entry of $\widetilde{B}$ is

$$\begin{aligned}&(\mathcal{I}_m)^T\left(e_i^{\widehat{n}}(e_j^{m})^T\otimes B\right)L^{-T}\Big((\widetilde{C}\otimes C)\mathcal{I}_p+\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3\\ &\qquad+2(\widetilde{\mathcal{H}}^{(2)}\otimes\mathcal{H}^{(2)})T_{(n,\widehat{n})}(p_3\otimes q_3)\Big),\\ &=(\mathcal{I}_m)^T\left(e_i^{\widehat{n}}(e_j^{m})^T\otimes\widehat{B}\right)\widehat{L}^{-T}\Big((\widetilde{C}\otimes\widehat{C})\mathcal{I}_p+\sum_{k=1}^{m}(\widetilde{N}_k\otimes\widehat{N}_k)^T\widehat{q}_4\\ &\qquad+2(\widetilde{\mathcal{H}}^{(2)}\otimes\widehat{\mathcal{H}}^{(2)})T_{(\widehat{n},\widehat{n})}(\widehat{p}_4\otimes\widehat{q}_4)\Big),\end{aligned}$$ □

which gives rise to (5.32b).

5.2.3. Truncated quadratic-bilinear iterative rational Krylov algorithm

The remaining challenge is now to develop a numerically efficient model reduction algorithm to construct a reduced QB system satisfying the first-order optimality conditions in Theorem 5.7. However, as in the linear [79] and bilinear [21, 59] cases, since the optimality conditions involve the matrices $V, W, \widehat{V}, \widehat{W}$, which depend on the reduced-order system matrices we are trying to construct, it is not a straightforward task to determine a reduced-order system directly that satisfies all the necessary conditions for optimality, i.e., (5.32a)–(5.32e). We propose Algorithm 5.1, which upon convergence leads to reduced-order systems that *approximately* satisfy the first-order necessary conditions for optimality given in Theorem 5.7. Throughout the paper, we denote the algorithm by truncated QB-IRKA, or TQB-IRKA.

Remark 5.8:
Ideally, *upon convergence* implies that the reduced-order quantities $\widehat{A}$, $\widehat{H}$, $\widehat{N}_k$, $\widehat{B}$, $\widehat{C}$ in Algorithm 5.1 stagnate. In a numerical implementation, one can check the stagnation based on the change of eigenvalues of the reduced matrix $\widehat{A}$ and terminate the algorithm once the relative change in the eigenvalues of $\widehat{A}$ is of the order of the machine precision. However, in all of our numerical experiments, we run TQB-IRKA until the relative change in the eigenvalues of $\widehat{A}$ is less than 10^{-5}. We observe that the quality of reduced-order systems does not change significantly thereafter, as in the cases of IRKA, B-IRKA, and TB-IRKA.

Our next goal is to show how the reduced-order system resulting from TQB-IRKA upon convergence relates to first-order optimality conditions (5.32). As a first step, we provide explicit expressions showing how far away the resulting reduced-order system is from satisfying the optimality conditions. Later, based on these expressions, we discuss how far the reduced-order systems, obtained from TQB-IRKA for weakly nonlinear QB systems, satisfy the optimality condition with small perturbations. We also illustrate using our numerical examples in Section 5.3 that in practice, the reduced-order system seemingly often satisfies the optimality conditions quite accurately.

Theorem 5.9:
Let Σ_{QB} be a QB system (4.1) and let $\widehat{\Sigma}_{QB}$ be the reduced-order QB system (4.2), computed by TQB-IRKA upon convergence. Let V_i, W_i, for $i \in \{1, 2\}$, be the matrices that solve (5.29), and let V and W be the matrices defining the projection used for model order reduction, as defined in (5.30). Similarly, let $\widehat{V}_i, \widehat{W}_i$, for $i \in \{1, 2\}$, be the matrices that solve (5.29), where the original system's state-space matrices are being replaced with their reduced-order counterparts. Moreover, let $\widehat{V}$ and $\widehat{W}$ be the matrices defined in (5.31). Assume that $\sigma(\widehat{A}) \cap \sigma(-\Pi A) = \emptyset$ and $\sigma(\widehat{A}) \cap \sigma(-\Pi^T A^T) = \emptyset$, where $\Pi = V(W^TV)^{-1}W^T$ and $\sigma(\cdot)$ denotes the eigenvalue spectrum of a matrix.

Algorithm 5.1: TQB-IRKA for QB systems.

```
Input: The system matrices: $A, H, N_1, \ldots, N_m, B, C$.
Symmetrize the Hessian $H$ and determine its mode-2 matricization $\mathcal{H}^{(2)}$.
Make an initial guess for the reduced matrices $\widehat{A}, \widehat{H}, \widehat{N}_1, \ldots, N_m, \widehat{B}, \widehat{C}$ with $\widehat{A}$
  being diagonalizable.
while relative change in $\{\lambda_i\} >$ tol convergence do
    Perform the spectral decomposition of $\widehat{A}$ and define:
    $\widehat{\Lambda} = \widehat{R}^{-1}\widehat{A}\widehat{R},\ \widetilde{N}_k = \widehat{R}^{-1}\widehat{N}_k\widehat{R},\ \widetilde{H} = \widehat{R}^{-1}\widehat{H}\left(\widehat{R}\otimes\widehat{R}\right),\ \widetilde{B} = \widehat{R}^{-1}\widehat{B},\ \widetilde{C} = \widehat{C}\widehat{R}.$
    Compute mode-2 matricization $\widetilde{\mathcal{H}}^{(2)}$.
    Solve for $V_1$ and $V_2$:
        $-V_1\Lambda - AV_1 = B\widetilde{B}^T,$
        $-V_2\Lambda - AV_2 = H(V_1\otimes V_1)\widetilde{H}^T + \sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T.$
    Solve for $W_1$ and $W_2$:
        $-W_1\Lambda - A^TW_1 = C^T\widetilde{C},$
        $-W_2\Lambda - A^TW_2 = 2\cdot\mathcal{H}^{(2)}(V_1\otimes W_1)(\widetilde{\mathcal{H}}^{(2)})^T + \sum_{k=1}^{m} N_k^T W_1 \widetilde{N}_k.$
    Compute $V$ and $W$:
        $V := V_1 + V_2, \qquad W := W_1 + W_2.$
    $V = \mathrm{orth}\,(V)$, $W = \mathrm{orth}\,(W)$.
    Determine the reduced matrices:
        $\widehat{A} = (W^TV)^{-1}W^TAV, \qquad \widehat{H} = (W^TV)^{-1}W^TH(V\otimes V),$
        $\widehat{N}_k = (W^TV)^{-1}W^TN_kV, \qquad \widehat{B} = (W^TV)^{-1}W^TB, \qquad \widehat{C} = CV.$
Output: $\widehat{A}, \widehat{H}, \widehat{N}_1, \ldots, \widehat{N}_m, \widehat{B}, \widehat{C}$.
```

Furthermore, assume $\Pi_v = V_1(W^TV_1)^{-1}W^T$ and $\Pi_w = W_1(V^TW_1)^{-1}V^T$ exist. Then, the reduced-order system $\widehat{\Sigma}_{QB}$ satisfies the following relations: [a]

$$\operatorname{tr}\left(CVe_i^{\widehat{n}}\left(e_j^p\right)^T\right) = \operatorname{tr}\left(\widehat{C}\widehat{V}e_i^{\widehat{n}}\left(e_j^p\right)^T\right) + \epsilon_C^{(i,j)},$$
$$i \in \{1,\ldots,\widehat{n}\},\quad j \in \{1,\ldots,p\}, \tag{5.43a}$$

$$\operatorname{tr}\left(B^TWe_i^{\widehat{n}}\left(e_j^m\right)^T\right) = \operatorname{tr}\left(\widehat{B}^T\widehat{W}e_i^{\widehat{n}}\left(e_j^m\right)^T\right) + \epsilon_B^{(i,j)},$$
$$i \in \{1,\ldots,\widehat{n}\},\quad j \in \{1,\ldots,m\}, \tag{5.43b}$$

$$(W_1(:,i))^TN_kV_1(:,j) = (\widehat{W}_1(:,i))^T\widehat{N}_k\widehat{V}_1(:,j) + \epsilon_N^{(i,j,k)},$$
$$i,j \in \{1,\ldots,\widehat{n}\},\quad k \in \{1,\ldots,m\}, \tag{5.43c}$$

$$(W_1(:,i))^T H(V_1(:,j) \otimes V_1(:,l)) = (\widehat{W}_1(:,i))^T \widehat{H}(\widehat{V}_1(:,j) \otimes \widehat{V}_1(:,l)) + \epsilon_H^{(i,j,l)}, \quad i,j,l \in \{1,\dots,\widehat{n}\}, \tag{5.43d}$$

$$(W_1(:,i))^T V(:,i) + (W_2(:,i))^T V_1(:,i) = (\widehat{W}_1(:,i))^T \widehat{V}(:,i) + \left(\widehat{W}_2(:,i)\right)^T \widehat{V}_1(:,i) + \epsilon_\lambda^{(i)}, \quad i \in \{1,\dots,\widehat{n}\}. \tag{5.43e}$$

where

$$\begin{aligned}
\epsilon_C^{(i,j)} &= -\operatorname{tr}\left(CV\Gamma_v e_i^{\widehat{n}} \left(e_j^p\right)^T\right), \\
\epsilon_B^{(i,j)} &= -\operatorname{tr}\left(B^T W (W^T V)^{-T} \Gamma_w e_i^{\widehat{n}} \left(e_j^m\right)^T\right), \\
\epsilon_N^{(i,j,k)} &= (\epsilon_w(:,i))^T N_k (V_1(:,j) - \epsilon_v(:,j)) + (W_1(:,i))^T N_k(\epsilon_v(:,j)), \\
\epsilon_H^{(i,j,l)} &= (W_1(:,i) - \epsilon_w(:,i))^T H(\epsilon_v(:,j) \otimes (V_1(:,l) - \epsilon_v(:,l)) + V_1(:,j) \otimes \epsilon_v(:,l)) \\
&\quad + (\epsilon_w(:,i))^T H((V(:,j) - \epsilon_v(:,j)) \otimes (V_1(:,l) - \epsilon_v(:,l))), \text{ and} \\
\epsilon_\lambda^{(i)} &= -\left(\widehat{W}(:,i)\right)^T \Gamma_v(:,i) - (\Gamma_w(: .i))^T \left(\widehat{V}(:,i) - \Gamma_v(:,i)\right) \\
&\quad - (W_2(:,i))^T V_2(:,i) + (\widehat{W}_2(:,i))^T \widehat{V}_2(:,i),
\end{aligned}$$

in which ϵ_v, ϵ_w, Γ_v and Γ_w, respectively, solve

$$\epsilon_v \Lambda + \Pi A \epsilon_w = (\Pi - \Pi_v)(AV_1 + B\widetilde{B}^T), \tag{5.44a}$$

$$\epsilon_w \Lambda + (A\Pi)^T \epsilon_w = (\Pi^T - \Pi_w)(A^T W_1 + C^T \widetilde{C}), \tag{5.44b}$$

$$\Gamma_v \Lambda + \widehat{A}\Gamma_v = -(W^T V)^{-1} W^T \left(\sum_{k=1}^m N_k \epsilon_v \widetilde{N}_k^T + H(\epsilon_v \otimes (V_1 + \epsilon_v) + V_1 \otimes \epsilon_v)\widetilde{H}^T\right) \tag{5.44c}$$

$$\Gamma_w \Lambda + \widehat{A}^T \Gamma_w = V^T \left(\sum_{k=1}^m N_k^T \epsilon_w \widetilde{N}_k + \mathcal{H}^{(2)}(\epsilon_v \otimes (W_1 + \epsilon_w) + V_1 \otimes \epsilon_w) \left(\mathcal{H}^{(2)}\right)^T\right). \tag{5.44d}$$

◊

Proof. We begin by establishing a relationship between $V_1 \in \mathbb{R}^{n \times \widehat{n}}$, $\widehat{V}_1 \in \mathbb{R}^{\widehat{n} \times \widehat{n}}$ and $V \in \mathbb{R}^{n \times \widehat{n}}$. For this, consider the Sylvester equation related to V_1

$$-V_1\Lambda - AV_1 = B\widetilde{B}^T, \tag{5.45}$$

and the oblique projector $\Pi_v := V_1(W^T V_1)^{-1} W^T$. Then, we apply the projector Π_v to the Sylvester equation (5.45) from the left to obtain

$$-V_1\Lambda - \Pi_v AV_1 = \Pi_v B\widetilde{B}^T, \text{ and}$$

$$-V_1\Lambda - \Pi AV_1 * = (\Pi_v - \Pi)AV_1 + \Pi_v B\widetilde{B}^T, \tag{5.46}$$

where $\Pi := V(W^TV)^{-1}W^T$. Now, we recall that $\widehat{V}_1$ satisfies the Sylvester equation

$$-\widehat{V}_1\Lambda - \widehat{A}\widehat{V}_1 = \widehat{B}\widetilde{B}^T.$$

We next multiply it by V from the left and substitute for $\widehat{A}$ and $\widehat{B}$ to obtain

$$-V\widehat{V}_1\Lambda - \Pi AV\widehat{V}_1 = \Pi B\widetilde{B}^T. \tag{5.47}$$

Subtracting (5.46) from (5.47) yields

$$(V_1 - V\widehat{V}_1)\Lambda + \Pi A(V_1 - V\widehat{V}_1) = (\Pi - \Pi_v)\left(AV_1 + B\widetilde{B}^T\right).$$

Since it is assumed that $\sigma(\widehat{A}) \cap \sigma(-\Pi A) = \emptyset$, this implies that $\Lambda \otimes I_n + I_{\widehat{n}} \otimes (\Pi A)$ is invertible. Therefore, we can write

$$V_1 = V\widehat{V}_1 + \epsilon_v, \tag{5.48}$$

where ϵ_v solves the Sylvester equation

$$\epsilon_v\Lambda + \Pi_v A\epsilon_v = (\Pi - \Pi_v)\left(AV_1 + B\widetilde{B}^T\right). \tag{5.49}$$

Similarly, one can show that

$$W_1 = W(W^TV)^{-T}\widehat{W}_1 + \epsilon_w, \tag{5.50}$$

where ϵ_w solves

$$\epsilon_w\Lambda + \Pi^TA^T\epsilon_w = (\Pi^T - \Pi_w)(A^TW_1 + C^T\widetilde{C}),$$

in which $\Pi_w := W_1(V^TW)V^T$. Using (5.48) and (5.50), we obtain

$$\begin{aligned}\widehat{W}_1(:,i)^T\widehat{N}_k\widehat{V}_1(:,j) &= \widehat{W}_1(:,i)^T(W^TV)^{-1}W^TN_kV\widehat{V}_1(:,j)\\ &= \left(W_1(:,i) - \epsilon_w(:,i)\right)^T N_k\left(V_1(:,j) - \epsilon_v(:,j)\right)\\ &= W_1(:,i)^TN_kV_1(:,j) - \left(\epsilon_w(:,i)\right)^T N_k(V_1(:,j) - \epsilon_v(:,j))\\ &\quad - \left(W_1(:,i)\right)^T N_k(\epsilon(:,j)),\end{aligned}$$

which is (5.43c) in Theorem 5.7. Similarly, one can prove (5.43d). To prove (5.43a), we consider the following Sylvester equation for V:

$$V(-\Lambda) - AV = B\widetilde{B}^T + \sum_{k=1}^{m} N_kV_1\widetilde{N}_k^T + H(V_1 \otimes V_1)\widetilde{H}^T. \tag{5.51}$$

Applying Π to both sides of the above Sylvester equation yields

$$V\left(I_{\widehat{n}}(-\Lambda) - \widehat{A}I_{\widehat{n}}\right) = V\left(\widehat{B}\widetilde{B}^T + \mathcal{Y}\right), \tag{5.52}$$

where $\mathcal{Y} = (W^TV)^{-1}W^T\left(\sum_{k=1}^m N_kV_1\widetilde{N}_k^T + H(V_1\otimes V_1)\widetilde{H}^T\right)$. This implies that

$$I_{\widehat{n}}(-\Lambda) - \widehat{A}I_{\widehat{n}} = \widehat{B}\widetilde{B}^T + \mathcal{Y}. \tag{5.53}$$

Next, we consider the Sylvester equation for $\widehat{V}$,

$$\widehat{V}(-\Lambda) - \widehat{A}\widehat{V} = \widehat{B}\widetilde{B}^T + \sum_{k=1}^m \widehat{N}_k\widehat{V}_1\widetilde{N}_k^T + \widehat{H}(\widehat{V}_1\otimes\widehat{V}_1)\widetilde{H}^T. \tag{5.54}$$

We then subtract (5.54) and (5.53) to obtain

$$\begin{aligned}(I_{\widehat{n}} - \widehat{V})(-\Lambda) - \widehat{A}(I_{\widehat{n}} - \widehat{V}) = &\sum_{k=1}^m (W^TV)^{-1}W^TN_k\left(V_1 - V\widehat{V}_1\right)\widetilde{N}_k^T\\ &+ (W^TV)^{-1}W^TH\left(V_1\otimes V_1 - (V\widehat{V}_1\otimes V\widehat{V}_1)\right)\widetilde{H}^T.\end{aligned}$$

Substituting $V\widehat{V}_1$ from (5.48) gives

$$\begin{aligned}(I_{\widehat{n}} - \widehat{V})(-\Lambda) - \widehat{A}(I_{\widehat{n}} - \widehat{V}) = &\sum_{k=1}^m (W^TV)^{-1}W^TN_k\epsilon_v\widetilde{N}_k^T\\ &+ (W^TV)^{-1}W^TH\left(\epsilon_v\otimes V_1 + V_1\otimes\epsilon_v + \epsilon_v\otimes\epsilon_v\right)\widetilde{H}^T.\end{aligned}$$

Since Λ contains the eigenvalues of $\widehat{A}$ and $\widehat{A}$ is stable, Λ and $-\widehat{A}$ cannot have any common eigenvalues. Hence, the matrix $\Lambda\otimes I_{\widehat{n}} + I_{\widehat{n}}\otimes\widehat{A}$ is invertible. Therefore the above Sylvester equations for $\Gamma := \widehat{V} - I_{\widehat{n}}$ have a unique solution and can be written as

$$\begin{aligned}\Gamma_v\Lambda + \widehat{A}\Gamma_v = &\sum_{k=1}^m (W^TV)^{-1}W^TN_k\epsilon_v\widetilde{N}_k^T\\ &+ (W^TV)^{-1}W^TH\left(\epsilon_v\otimes V_1 + V_1\otimes\epsilon_v + \epsilon_v\otimes\epsilon_v\right)\widetilde{H}^T.\end{aligned}$$

To prove (5.43a), we observe that

$$\begin{aligned}\operatorname{tr}\left(\widehat{C}\widehat{V}e_i^{\widehat{n}}\left(e_j^p\right)^T\right) &= \operatorname{tr}\left(CV(I_{\widehat{n}} + \Gamma_v)e_i^{\widehat{n}}\left(e_j^p\right)^T\right)\\ &= \operatorname{tr}\left(CVe_i^{\widehat{n}}\left(e_j^p\right)^T\right) + \operatorname{tr}\left(CV\Gamma_ve_i^{\widehat{n}}\left(e_j^p\right)^T\right).\end{aligned}$$

Thus,

$$\operatorname{tr}\left(CVe_i^{\widehat{n}}\left(e_j^p\right)^T\right) = \operatorname{tr}\left(\widehat{C}\widehat{V}e_i^{\widehat{n}}\left(e_j^p\right)^T\right) + \epsilon_C^{(i,j)}.$$

Analogously, we can prove that there exists Γ_w such that $\widehat{W} = (W^TV)^T + \Gamma_w$ and that it satisfies

$$\Gamma_w\Lambda + \widehat{A}^T\Gamma_w = V^T\left(\sum_{k=1}^{m} N_k^T\epsilon_w\widetilde{N}_k + \mathcal{H}^{(2)}(\epsilon_v \otimes (W_1+\epsilon_w) + V_1\otimes\epsilon_w)\left(\mathcal{H}^{(2)}\right)^T\right).$$

To prove (5.43b), we observe that

$$\operatorname{tr}\left(\widehat{B}^T\widehat{W}e_i^{\widehat{n}}\left(e_j^m\right)^T\right) = \operatorname{tr}\left(B^TW(W^TV)^{-T}((W^TV)^T+\Gamma_v)e_i^{\widehat{n}}\left(e_j^p\right)^T\right).$$

Thus,

$$\operatorname{tr}\left(\widehat{B}^T\widehat{W}e_i^{\widehat{n}}\left(e_j^m\right)^T\right) = \operatorname{tr}\left(B^TW^T + B^TW(W^TV)^{-T}\Gamma_w)e_i^{\widehat{n}}\left(e_j^p\right)^T\right).$$

Since we now know that $\widehat{V} = I_{\widehat{n}} + \Gamma_v$ and $\widehat{W} = (W^TV)^T + \Gamma_w$, we get

$$V\widehat{V} = V + V\Gamma_v \quad \text{and} \quad W(W^TV)^{-T}\widehat{W} = W + W(W^TV)^{-T}\Gamma_w. \tag{5.55}$$

We make use of (5.55) to prove (5.43e) in the following:

$$\begin{aligned}
&(W_1(:,i))^TV(:,i) + (W_2(:,i))^T V_1(:,i)\\
&\quad= (W(:,i))^TV(:,i) - (W_2(:,i))^TV_2(:,i)\\
&\quad= \left(W(W^TV)^{-T}\left(\widehat{W}(:,i) - \Gamma_w(:.i)\right)\right)^T V\left(\widehat{V}(:,i) - \Gamma_v(:,i)\right)\\
&\qquad - (W_2(:,i))^TV_2(:,i)\\
&\quad= \left(\widehat{W}(:,i) - \Gamma_w(:.i)\right)^T\left(\widehat{V}(:,i) - \Gamma_v(:,i)\right) - (W_2(:,i))^TV_2(:,i)\\
&\quad= \left(\widehat{W}(:,i)\right)^T\widehat{V}(:,i) - \left(\widehat{W}(:,i)\right)^T\Gamma_v(:,i) - (\Gamma_w(:.i))^T\left(\widehat{V}(:,i) - \Gamma_v(:,i)\right)\\
&\qquad - (W_2(:,i))^TV_2(:,i)\\
&\quad= (\widehat{W}_1(:,i))^T\widehat{V}(:,i) + \left(\widehat{W}_2(:,i)\right)^T\widehat{V}_1(:,i) + \epsilon_\lambda^{(i)},
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_\lambda^{(i)} = &-\left(\widehat{W}(:,i)\right)^T\Gamma_v(:,i) - (\Gamma_w(:.i))^T\left(\widehat{V}(:,i) - \Gamma_v(:,i)\right)\\
&- (W_2(:,i))^TV_2(:,i) + (\widehat{W}_2(:,i))^T\widehat{V}_2(:,i).
\end{aligned}$$

This completes the proof. . □

Remark 5.10:
In Theorem 5.9, we have presented measures, e.g., the distance between $\operatorname{tr}\left(CVe_i^r(e_j^p)^T\right)$ and $\operatorname{tr}\left(\widehat{C}\widehat{V}e_i^r(e_j^p)^T\right)$, denoted by $\epsilon_C^{(i,j)}$, with which the reduced-order system via TQB-IRKA satisfies the optimality conditions (5.32). But Theorem 5.9 in general does

not provide a guarantee for the smallness of these distances. However, we provide an intuition for the weakly nonlinear QB systems, i.e., QB systems for which $\|H\|$ and $\|N_k\|$ are small with respect to $\|B\|$ and $\|C\|$. Recall that V_1 and V_2 solve the Sylvester equations (5.29a) and (5.29c), respectively, and the right-hand side for V_2 is quadratic in H and N_k. Therefore, for a weakly nonlinear QB system, $\|V_2\|$ will be relatively small compared to $\|V_1\|$. Hence, V is expected to be close to V_1. Thus, one could anticipate that the projectors $\Pi = V(W^TV)W^T$ and $\Pi_v = V_1(W^TV_1)W^T$ will be close to each other. As a result, the right-hand side of the Sylvester equation (5.44a) will be small, and hence so is ϵ_v. In a similar way, one can argue that ϵ_w in (5.44b) will be small. Therefore, it can be shown that in the case of weakly nonlinear QB systems (4.1), all ϵ's in (5.43) such as $\epsilon_C^{(i,j)}$ should be small.

Indeed, the situation in practice proves much better. We observe in our numerical results (see Section 5.3) that even for strongly nonlinear QB systems, i.e., $\|H\|$ and $\|N_k\|$ are comparable or even much larger than $\|B\|$ and $\|C\|$, Algorithm 5.1 still yields reduced-order systems which satisfy the optimality conditions (5.32) almost exactly with negligible perturbations. $\diamondsuit$

Remark 5.11:
Algorithm 5.1 can be seen as an extension of the *truncated* B-IRKA with truncation index 2 [59, Algo. 2] from bilinear systems to QB systems. In [59], the truncation index $\mathcal{N}$, which denotes the number of terms in the underlying Volterra series for bilinear systems, is free, and as $\mathcal{N} \to \infty$, all the perturbations go to zero. However, it is shown in [59] that in most cases, a small $\mathcal{N}$, for example 2 or 3, is enough to satisfy all optimality conditions closely. In our case, a similar convergence will occur if we let the number of terms in the underlying Volterra series of the QB system grow; however, this is not numerically feasible since the subsystems in the QB case become rather complicated after the first three terms. Indeed, because of this, [78], [25] and [4] have considered the interpolation of multivariate transfer functions corresponding to only the first two subsystems. Moreover, even in the case of balanced truncation for QB systems [28], it is shown by means of numerical examples that the truncated Gramians for QB systems based on the first three terms of the underlying Volterra series produce quantitatively accurate reduced-order systems. Our numerical examples show that this is the case here as well. $\diamondsuit$

Remark 5.12:
So far in all of our discussions, we have assumed that the reduced matrix $\widehat{A}$ is diagonalizable. This is a reasonable assumption since non-diagonalizable matrices lie in a set of Lebesgue measure zero. The probability of entering this set by any numerical algorithm including TQB-IRKA is zero with respect to Lebesgue measure. Thus, TQB-IRKA can be considered safe in this regard.

Furthermore, throughout the analysis, it has been assumed that the reduced matrix $\widehat{A}$ is Hurwitz. However, in case $\widehat{A}$ is not Hurwitz, then the truncated $\mathcal{H}_2$-norm of the error system will be unbounded; thus the reduced-order systems indeed cannot be (locally) optimal. Nonetheless, a mathematical study to ensure the stability from $\mathcal{H}_2$ iterative schemes are still under investigation even for linear systems. However, a simple fix to this problem is to reflect the unstable eigenvalues of $\widehat{A}$ in every step back to the left-half plane. Also, see [94] for a more involved approach to stabilize a reduced order system.

Theorem 5.9 assumes that TQB-IRKA has converged. As stated in Remark 5.11, TQB-IRKA extends IRKA, B-IRKA and TB-IRKA to the kind of QB systems we consider. Even for the linear case, i.e., for IRKA, convergence cannot be theoretically guaranteed despite overwhelming numerical evidence that IRKA (and (T)B-IRKA), in most cases, converge rapidly to a local minimum. Convergence of IRKA can be guaranteed theoretically only for the symmetric case [58]. Moreover, in [16] and [58], variants of IRKA with guaranteed global convergence have been introduced; however due to the success of regular IRKA and its simple implementation, these modifications have not been as widely used. Therefore, guaranteed theoretical convergence in this iterative setting is an open issue even for the linear and bilinear cases, and naturally for the QB case as well. ♢

Remark 5.13:
As mentioned above, so far the analysis is based on the assumption that the reduced matrix $\widehat{A}$ is diagonalizable. For a reduced matrix $\widehat{A}$ with Jordan blocks, one would need to extend the derivation of the Sylvester-equation based $\mathcal{H}_2$ optimality conditions in [131], where Wilson [131] differentiates the $\mathcal{H}_2$ error with respect to the reduced matrix $\widehat{A}$ as opposed to individual eigenvalues $\{\lambda_i\}$ as we do here. An interpolation interpretation of the Jordan blocks in the linear case has also been established; see [126]. However, since the Jordan blocks in the optimal reduced models so far have never been observed in practice, extensions of the $\mathcal{H}_2$ theory to the bilinear case have focused on the diagonalizabilty assumption; thus, we keep the same assumption here. However, based on how the Sylvester-equation based conditions for the linear case appear, for QB systems with non-diagonalizable $\widehat{A}$, one can reasonably expect an algorithm similar to Algorithm 5.1, where the steps 6–7 are replaced by solving consecutively for V_1, V_2, W_1, W_2 in the following Sylvester equations

$$\begin{aligned}
-V_1\widehat{A} - AV_1 &= B\widehat{B}^T,\\
-V_2\widehat{A} - AV_2 &= H(V_1 \otimes V_1)\widehat{H}^T + \sum_{k=1}^{m} N_k V_1 \widehat{N}_k^T,\\
-W_1\widehat{A}^T - A^TW_1 &= C^T\widehat{C},\\
-W_2\widehat{A}^T - A^TW_2 &= 2\cdot\mathcal{H}^{(2)}(V_1 \otimes W_1)(\widehat{H}^{(2)})^T + \sum_{k=1}^{m} N_k^T W_1 \widehat{N}_k.
\end{aligned}$$

Note that with this formulation, $\widehat{A}$ enters into the algorithm directly without diagonalization. However, due to the reasons listed before, we leave this theoretical development for future work. ◊

Remark 5.14:
In oder to employ Algorithm 5.1, we need to perform computations such as $H(V_1 \otimes V_1)\widetilde{H}^T$, $\mathcal{H}^{(2)}(V_1 \otimes W_1)\left(\widetilde{H}^{(2)}\right)^T$. In Subsection 4.3.4, we have discussed in detail how these terms can be computed efficiently in large-scale settings. ◊

5.2.4. Generalized state-space QB systems

Thus far, we have used $E = I$ in front of $\dot{x}(t)$ in the QB system (4.1). However, it is also common that the spatial discretization of a nonlinear PDE results in a mass matrix $E \neq I$. Thus, in the following, we consider a generalized state-space QB of the form:

$$\Sigma_{QB} : \begin{cases} E\dot{x}(t) = Ax(t) + H\left(x(t) \otimes x(t)\right) + \displaystyle\sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \\ y(t) = Cx(t), \quad x(0) = 0, \end{cases} \tag{5.56}$$

with dimension as in (4.1), where the matrix $E \in \mathbb{R}^{n\times n}$ is considered to be non-singular. In general, we aim at constructing a reduced generalized state-space QB as follows:

$$\Sigma_{QB} : \begin{cases} \widehat{E}\dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t) + \widehat{H}\left(\widehat{x}(t) \otimes \widehat{x}(t)\right) + \displaystyle\sum_{k=1}^{m} \widehat{N}_k \widehat{x}(t) u_k(t) + \widehat{B}u(t), \\ \widehat{y}(t) = \widehat{C}\widehat{x}(t), \quad \widehat{x}(0) = 0, \end{cases} \tag{5.57}$$

where $\widehat{x}(t) \in \mathbb{R}^{\widehat{n}}$, $u(t) \in \mathbb{R}^m$ and $\widehat{y}(t) \in \mathbb{R}^p$ are the reduced state, input and output of the reduced-order system at time (t). To simplify the $\mathcal{H}_2$-optimal model reduction problem for the generalized QB systems, we assume that $\widehat{E} = I_{\widehat{n}}$ and $\widehat{A}$ is diagonalizable. Under these assumptions, one can apply the result derived for the mass matrix $E = I_n$. One obvious way is to invert E, but this is inadmissible in a large-scale setting. Moreover, the resulting matrices may be dense, making the algorithm computationally expensive. Nevertheless, Algorithm 5.1 can be employed without inverting E. For this, we need to modify steps 6 and 7 in Algorithm 5.1 as follows:

$$\begin{aligned}
-EV_1\Lambda - AV_1 &= B\widetilde{B}^T, \\
-EV_2\Lambda - AV_2 &= H(V_1 \otimes V_1)\widetilde{H}^T + \textstyle\sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T, \\
-E^TW_1\Lambda - A^TW_1 &= C^T\widetilde{C}, \\
-E^TW_2\Lambda - A^TW_2 &= 2\cdot\mathcal{H}^{(2)}(V_1 \otimes W_1)(\widetilde{\mathcal{H}}^{(2)})^T + \textstyle\sum_{k=1}^{m} N_k^T W_1 \widetilde{N}_k,
\end{aligned}$$

Algorithm 5.2: TQB-IRKA for generalized QB systems.

```
Input: The system matrices: $E, A, H, N_1, \ldots, N_m, B, C$.
Symmetrize the Hessian $H$ and determine its mode-2 matricization $\mathcal{H}^{(2)}$.
Make an initial guess for the reduced matrices $\widehat{E}, \widehat{A}, \widehat{H}, \widehat{N}_1, \ldots, N_m, \widehat{B}, \widehat{C}$ with $\widehat{A}$
  being diagonalizable.
while relative change in $\Lambda(\widehat{A}, \widehat{E})$ > tol convergence do
    Compute mode-2 matricization $\widetilde{\mathcal{H}}^{(2)}$.
    Solve for $V_1$ and $V_2$:
        $-EV_1\widehat{A} - AV_1\widehat{E}^T = B\widehat{B}^T,$
        $-EV_2\widehat{A} - AV_2\widehat{E}^T = H(V_1 \otimes V_1)\widehat{H}^T + \sum_{k=1}^m N_k V_1 \widehat{N}_k^T.$
    Solve for $W_1$ and $W_2$:
        $-E^T W_1 \widehat{A}^T - A^T W_1 \widehat{E} = C^T \widehat{C},,$
        $-E^T W_2 \widehat{A}^T - A^T W_2 \widehat{E} = 2 \cdot \mathcal{H}^{(2)}(V_1 \otimes W_1)(\widehat{H}^{(2)})^T + \sum_{k=1}^m N_k^T W_1 \widehat{N}_k.$
    Compute $V$ and $W$:
        $V := V_1 + V_2, \qquad W := W_1 + W_2.$
    $V = \operatorname{orth}(V)$, $W = \operatorname{orth}(W)$.
    Determine the reduced matrices:
        $\widehat{E} = W^T E V, \qquad \widehat{A} = W^T A V, \qquad \widehat{H} = W^T H(V \otimes V),$
        $\widehat{N}_k = W^T N_k V, \qquad \widehat{B} = W^T B, \qquad \widehat{C} = CV.$
Output: $\widehat{E}, \widehat{A}, \widehat{H}, \widehat{N}_1, \ldots, \widehat{N}_m, \widehat{B}, \widehat{C}$.
```

and replace $(W^TV)^{-1}$ with $(W^TEV)^{-1}$, assuming W^TEV is invertible while determining the reduced-order system matrices in step 9 of Algorithm 5.1. Then, the modified iterative algorithm with the matrix E also provides a reduced-order system, *approximately* satisfying optimality conditions under the considered assumptions. We skip the rigorous proof for the $E \neq I$ case, but it can be proven along the lines of $E = I$.

Furthermore, one can aim at deriving the Wilson-type $\mathcal{H}_2$-optimality conditions for the generalized QB system by considering $\widehat{E}$, $\widehat{A}$, $\widehat{H}$, $\widehat{N}_k$ and $\widehat{C}$ as optimization variable which does not require any assumption on the reduced matrices. Then, one can expect an algorithm as sketched in Algorithm 5.2, which, upon convergence, yields a reduced-order system, satisfying the Wilson-type optimality conditions *approximately*. However, a detailed theoretical discussion on Wilson-type optimality conditions, we leave as a potential future work.

5.3. Numerical Experiments

In this section, we illustrate the behavior of the proposed model reduction method TQB-IRKA for QB systems by means of several semi-discretized nonlinear PDEs, and compare it with the existing MOR techniques, such as one-sided and two-sided subsystem-based interpolatory projection methods [25, 78, 106], balanced truncation (BT) for QB systems (proposed in the previous chapter), and POD, e.g., see [88, 96] in terms of the accuracy of the time-domain performance and the truncated $\mathcal{H}_2$-norm. We iterate Algorithm 5.1 until the relative change in the eigenvalues of $\widehat{A}$ becomes smaller than a given tolerance, which we set to 10^{-5}. Moreover, we determine the interpolation points for the one-sided and two-sided interpolatory projection methods applying IRKA [79] to the corresponding linear part, which appear to be a good set of interpolation points as shown in [25]. All the simulations were done on a board with 4 Intel® Xeon® E7-8837 CPUs with a 2.67-GHz clock speed using MATLAB 8.0.0.783 (R2012b). Some more details related to the numerical examples are as follows:

1. For all time domain simulations, the original and reduced-order systems are integrated by the routine `ode15s` in MATLAB with a relative error tolerance of 10^{-8} and an absolute error tolerance of 10^{-10}.

2. We measure the output at 500 equidistant points within the time interval $[0, T]$, where T is defined in each numerical example.

3. In order to employ BT, we need to solve four standard Lyapunov equations. For this, we use `mess_lyap.m` from M.E.S.S.-1.0.1 [113] which is based on one of the latest ADI methods proposed in [32].

4. We initialize TQB-IRKA (Algorithm 5.1) by choosing an arbitrary reduced-order system the `rand` command in MATLAB, while ensuring $\widehat{A}$ is Hurwitz and diagonalizable.

5. Since POD can be applied to a general nonlinear system, we apply POD to the original nonlinear system, without transforming it into a QB system as we observe that this way, POD yields better reduced-order systems.

6. One of the aims of the numerical examples is to determine the residuals in Theorem 5.9. For this, we first define $\Phi_C^e \in \mathbb{R}^{r\times p}$, $\Phi_B^e \in \mathbb{R}^{r\times m}$, $\Phi_N^e \in \mathbb{R}^{r\times r\times m}$, $\Phi_H^e \in \mathbb{R}^{r\times r\times r}$ and $\Phi_\Lambda^e \in \mathbb{R}^{r}$ such that $\epsilon_C^{(i,j)}$ is the (i,j)th entry of Φ_C^e, $\epsilon_B^{(i,j)}$ is the (i,j)th entry of Φ_B^e, $\epsilon_N^{(i,j,k)}$ is the (i,j,k)th entry of Φ_N^e, $\epsilon_H^{(i,j,k)}$ is the (i,j,k)th entry of Φ_H^e, and $\epsilon_\Lambda^{(i)}$is ith entry of Φ_Λ^e.

   Furthermore, we define Φ_C, Φ_B, Φ_N, Φ_H, and Φ_Λ to be the terms on the left hand side of equations (5.43a) – (5.43e) in Theorem 5.9, e.g., the (i,j)th entry of

Φ_C is $\operatorname{tr}\left(CVe_i^r\left(e_j^p\right)^T\right)$. As a result, we define relative perturbation measures as follows:

$$\mathcal{E}_C = \frac{\|\Phi_C^e\|_2}{\|\Phi_C\|_2},\ \mathcal{E}_B = \frac{\|\Phi_B^e\|_2}{\|\Phi_B\|_2},\ \mathcal{E}_N = \frac{\|\Phi_N^{e1}\|_2}{\|\Phi_N^{(1)}\|_2},\ \mathcal{E}_H = \frac{\|\Phi_H^{e1}\|_2}{\|\Phi_H^{(1)}\|_2},\ \mathcal{E}_\Lambda = \frac{\|\Phi_\Lambda^e\|_2}{\|\Phi_\Lambda\|_2}, \quad (5.58)$$

where $\Phi^{(1)}_{\{N,H\}}$ and $\Phi^{e1}_{\{N,H\}}$ are mode-1 matricizations of the tensors $\Phi_{\{N,H\}}$ and $\Phi^{e1}_{\{N,H\}}$, respectively.

7. We also address a numerical issue which one might face while employing Algorithm 5.1. In step 8 of Algorithm 5.1, we need to take the sum of the two matrices V_1 and V_2. If H and N_k are too large, then the norm of V_2 can be much larger than that of V_1. Thus, a direct sum might reduce the effect of V_1. As a remedy we propose to use a scaling factor γ for H and N_k, resulting in matrices V_1 and V_2 such that $\frac{\|V_2\|}{\|V_1\|} \in \mathcal{O}\left(10^0 - 10^2\right)$. We have already noted in Remark 5.3 that this scaling just scales the input-output mapping. Once again we emphasize that we just compute the model reduction basis matrices V and W using the scaled system, but we project the original, unscaled system to construct the reduced-order system.

5.3.1. One dimensional Chafee-Infante equation

As a first example, we consider the same Chafee-Infante example as discussed in Subsection 4.4.2. Note that the original system is of order $n = 1000$. We construct reduced-order systems of order $\widehat{n} = 10$ using TQB-IRKA, BT, one-sided and two-sided interpolatory projection methods, and POD. Having initialized TQB-IRKA randomly, it takes 9 iterations to converge, and for this example, we choose the scaling factor $\gamma = 10^{-3}$. For the POD based approximation, we collect 500 snapshots of the true solution for the training input $u^{(1)}(t) = (1+\sin(\pi t))\exp(-t/5)$ and compute the projection by taking the 10 dominant basis vectors.

In order to compare the quality of these reduced-order systems with respect to the original system, we first simulate them using the same training input used to construct the POD basis, i.e., $u^{(1)}(t) = (1+\sin(\pi t))\exp(-t/5)$. We plot the transient responses and relative output errors for this input in Figure 5.1. As expected, since we are comparing the reduced models for the same forcing term used for POD, Figure 5.1 shows that the POD approximation outperforms the other methods for the input $u^{(1)}$. However, the interpolatory methods also provide adequate reduced-order systems for $u^{(1)}(t)$ even though the reduction is performed without any knowledge of $u^{(1)}(t)$.

To test the robustness of the reduced-order systems, we compare the time-domain simulations of the reduced-order systems with the original one in Figure 5.2 for a

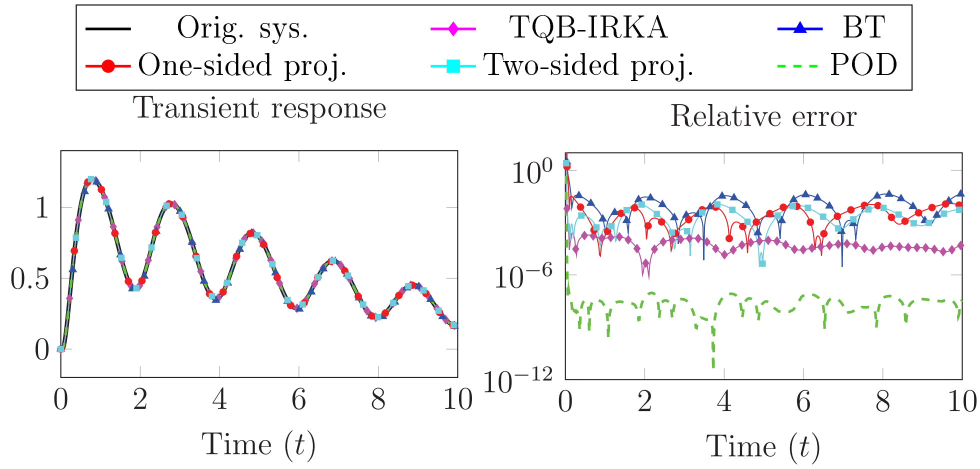


Figure 5.1.: Chafee-Infante: comparison of responses for the boundary control input $u^{(1)}(t) = (1 + \sin(\pi t)) \exp(-t/5)$.

Input	TQB-IRKA	BT	One-sided	Two-sided	POD
$u^{(1)}(t)$	$6.54 \cdot 10^{-5}$	$1.40 \cdot 10^{-2}$	$4.30 \cdot 10^{-3}$	$3.51 \cdot 10^{-3}$	$2.87 \cdot 10^{-8}$
$u^{(2)}(t)$	$1.63 \cdot 10^{-3}$	$1.43 \cdot 10^{-2}$	$4.59 \cdot 10^{-1}$	$6.65 \cdot 10^{-3}$	$6.70 \cdot 10^{-2}$

Table 5.1.: Chafee-Infante: the mean relative errors of the output.

slightly different input, namely $u^{(2)}(t) = 25\,(1 + \sin(\pi t))$. First, observe that the POD approximation fails to reproduce the system's dynamics for the input $u^{(2)}$ accurately as POD is input-dependent. Moreover, the one-sided interpolatory projection method also performs worse for the input $u^{(2)}$. On the other hand, TQB-IRKA, BT, and the two-sided interpolatory projection method, all yield very accurate reduced-order systems of comparable qualities; TQB-IRKA produces marginally better reduced-order systems. Once again it is important to emphasize that neither $u^{(1)}(t)$ nor $u^{(2)}(t)$ have entered the model reduction procedure in TQB-IRKA. To give a quantitative comparison of the reduced-order systems for both inputs, $u^{(1)}$ and $u^{(2)}$, we report the mean relative errors in Section 5.3.1 as well, which also provides us a similar information.

Furthermore, we study the impact of the scaling factor γ, as discussed in Remark 5.3, on the performance reduced-order systems obtained via TQB-IRKA. For the same inputs $u^{(i)}$, $i \in \{1, 2\}$, we plot the relative errors in the time-domain responses for different values of the scaling factor in Figure 5.3. For this example, we observe that for $\gamma = 10^{-3}$, TQB-IRKA produces a slightly better reduced-order system in terms of the accuracy of the time-domain simulations than for all other tested values of γ; however, all scaling factors $\gamma \in \{10^{0}, 10^{-1}, \ldots, 10^{-4}\}$ produce comparable reduced-order systems. For very small values of γ such as $\gamma = \{10^{-5}, 10^{-6}\}$, TQB-IRKA yields very

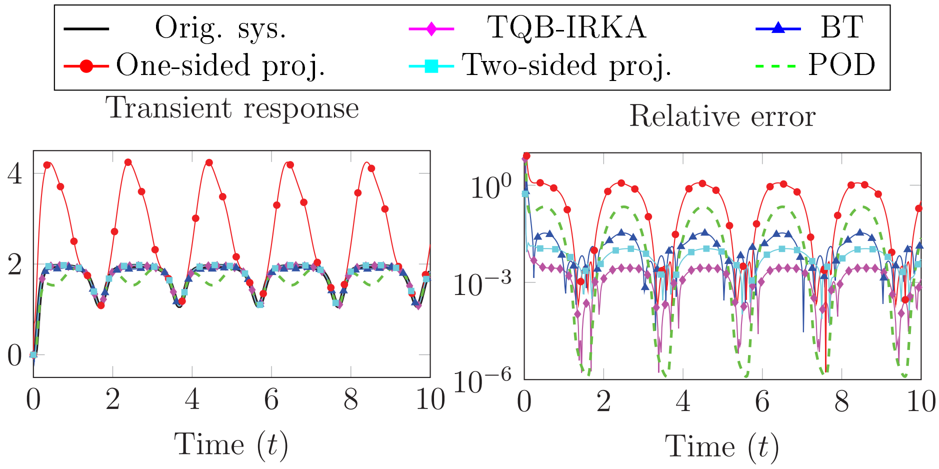


Figure 5.2.: Chafee-Infante: comparison of responses for the boundary control input $u^{(2)}(t) = 25(1 + \sin(\pi t))$.

$\mathcal{E}_C$	$\mathcal{E}_B$	$\mathcal{E}_N$	$\mathcal{E}_H$	$\mathcal{E}_\lambda$
1.35×10^{-8}	8.85×10^{-12}	8.84×10^{-16}	1.77×10^{-13}	1.44×10^{-11}

Table 5.2.: Chafee-Infante: perturbations to the optimality conditions.

poor reduced-order systems. This is expected since by choosing a very small scaling factor, the effect of the quadratic and bilinear terms is reduced significantly and the model reduction basis matrices almost correspond to the linear term only; hence, poor reduced-order systems result. We have observed that if a scaling factor is chosen such that $\dfrac{\|V_2\|}{\|V_1\|} \approx \mathcal{O}\left(10^0\text{–}10^2\right)$, then TQB-IRKA not only provides a better reduced-order system but also converges faster, although we do not have a theoretical justification for this observation yet. Therefore, as future work, it would be interesting to investigate the influence of the scaling factor on the quality of the obtained reduced-order systems also from a theoretical point of view.

In Theorem 5.9, we have presented the quantities, denoted by ϵ_C, ϵ_B, ϵ_λ, ϵ_N, and ϵ_H, which measure how far the reduced-order system of TQB-IRKA is from satisfying the optimality conditions (5.32) upon convergence. These quantities can be computed as shown in (5.58), and are listed in Table 5.2, showing a very small magnitude perturbations. In Remark 5.10, we have argued that for a weakly nonlinear QB system, we expect these quantities to be small. However, even for this example with strong nonlinearity, i.e., $\|H\|$ and $\|N_k\|$ are not small at all, the reduced-order system computed by TQB-IRKA satisfies the optimality conditions (5.32) very accurately. This result also strongly supports the discussion of Remark 5.11 that a small truncation index is

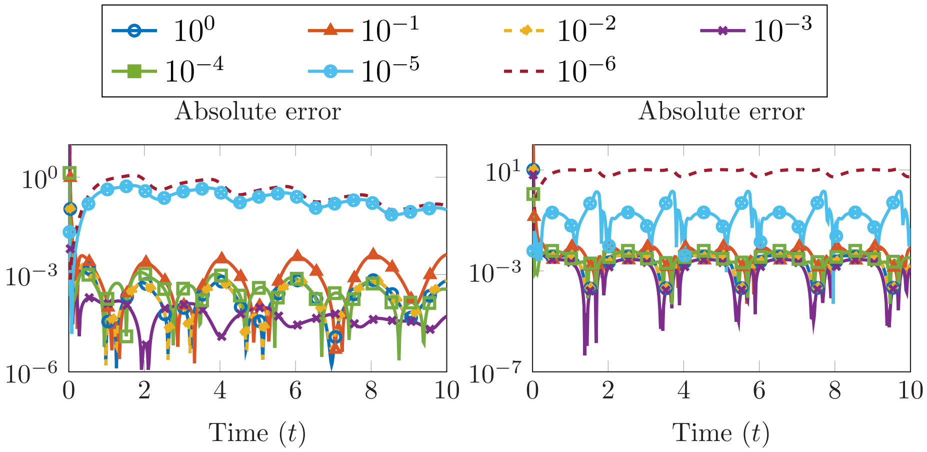


Figure 5.3.: Absolute error between the original and reduced-order systems ($r = 10$) obtained using TQB-IRKA for different scaling factors γ for inputs $u^{(1)}$ and $u^{(2)}$.

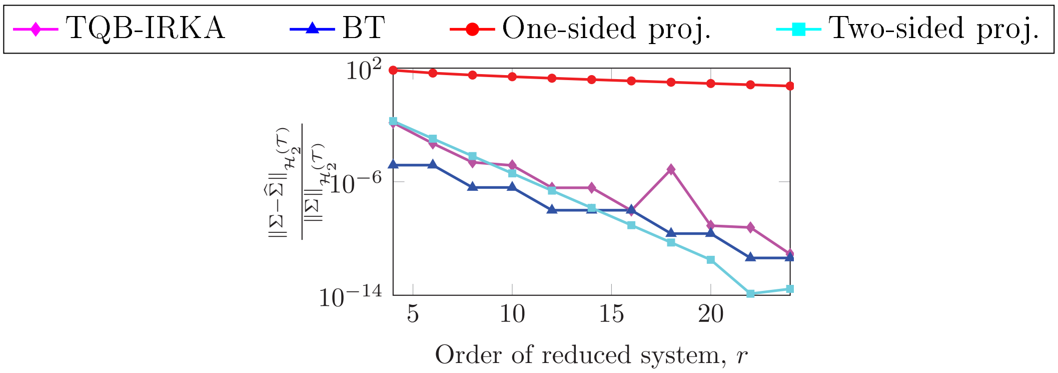


Figure 5.4.: Chafee-Infante: comparison of the truncated $\mathcal{H}_2$-norm of the error system, having obtained reduced-order systems of different orders via different methods.

expected to be enough in many cases.

Furthermore, since TQB-IRKA approximately minimizes the truncated $\mathcal{H}_2$-norm of the error system, i.e., $\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2^{(\mathcal{T})}}$, we also compare the truncated $\mathcal{H}_2$-norm of the error system in Figure 5.4, where the reduced models are constructed by various methods of different orders. As mentioned before, the reduced-order systems obtained via POD preserve the structure of the original nonlinearities; therefore, the truncated $\mathcal{H}_2$-norm definition, given in Lemma 5.4, does not apply.

Figure 5.4 indicates that the reduced-order systems obtained via one-sided interpolatory projection perform worst in the truncated $\mathcal{H}_2$-norm measure. Moreover, while BT performs better as compared to TQB-IRKA and the two-sided projection method for small reduced orders with respect to the truncated $\mathcal{H}_2$-norm, for higher reduced orders, the two-sided interpolatory method yields the best reduced-order systems. However, it is important to emphasize that unlike in the case of linear dynamical systems, the $\mathcal{H}_2$-norm and the L_∞-norm of the output for nonlinear systems, including QB systems, are not as strongly connected as in the linear case. This can be seen in Figure 5.4; for reduced order $r = 10$, even though BT yields the smallest truncated $\mathcal{H}_2$ error, in the time-domain simulations for inputs $u^{(1)}$ and $u^{(2)}$, it is not the best in terms of the L_∞-norm of the output. Nevertheless, the truncated $\mathcal{H}_2$-norm of the error system is still a robust indicator for the quality of the reduced-order system, because this norm is defined by the *kernels*, which define the mapping from the input to the output. Thus, if the kernels are ensured to be close enough, then one can expect an accurate approximation of the output.

5.3.2. Nonlinear RC ladder

Next, we discuss the same nonlinear RC ladder example as in Subsection 4.4.1. We set the number of capacitors in the ladder to $k = 500$, resulting in a QB system of order $n = 1000$. Note that the matrix A of the resulting QB system has eigenvalues at zero; therefore, the truncated $\mathcal{H}_2$-norm may not exist. Moreover, BT also cannot be employed as we need to solve Lyapunov equations that require a stable A matrix. Thus, we shift the matrix A to $A_s := A - 0.01I_n$ to determine the projection matrices for TQB-IRKA and BT, but we project the original system matrices.

We construct reduced-order systems of order $r = 10$ using all five different methods. In this example as well, we initialize TQB-IRKA randomly and it converges after 27 iterations. We choose the scaling factor $\gamma = 0.01$. In order to compute a reduced-order system via POD, we first obtain 500 snapshots of the true solution for the training input $u^{(1)}(t) = e^{-t}$ and then use the 10 dominant modes to determine the projection.

We first compare the accuracy of these reduced-order systems for the same training input $u^{(1)}(t) = e^{-t}$ that is also used to compute the POD basis. Figure 5.5 shows the transient responses and relative errors of the output for the input $u^{(1)}$. As one would expect, POD outperforms all other methods since the control input $u^{(1)}$ is the same as the training input for POD. Nonetheless, TQB-IRKA, BT, and two-sided projection also yield very good reduced-order systems, considering they are obtained without any prior knowledge of the input.

We also test the reduced-order systems for an input different from the training input, precisely, $u^{(2)}(t) = 2.5\,(\sin(\pi t/5) + 1)$. Figure 5.6 shows the transient responses and relative errors of the output for the input $u^{(2)}$. We observe that POD does perform almsot as well as TQB-IRKA, BT and two-sided projection methods even for this input, and the one-sided projection method completely fails to capture the system dynamics for the input $u^{(2)}$. This can also be observed from Section 5.3.2, where the mean relative errors of the outputs are reported.

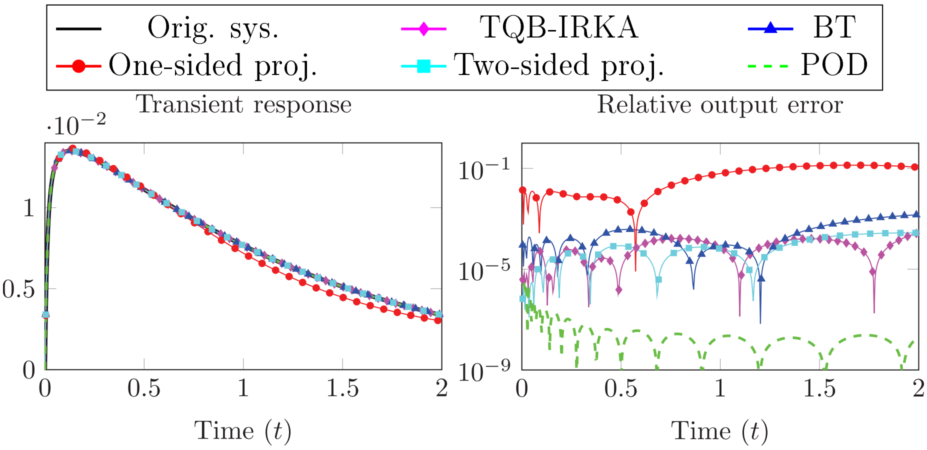


Figure 5.5.: An RC circuit: comparison of responses for the input $u^{(1)}(t) = e^{-t}$.

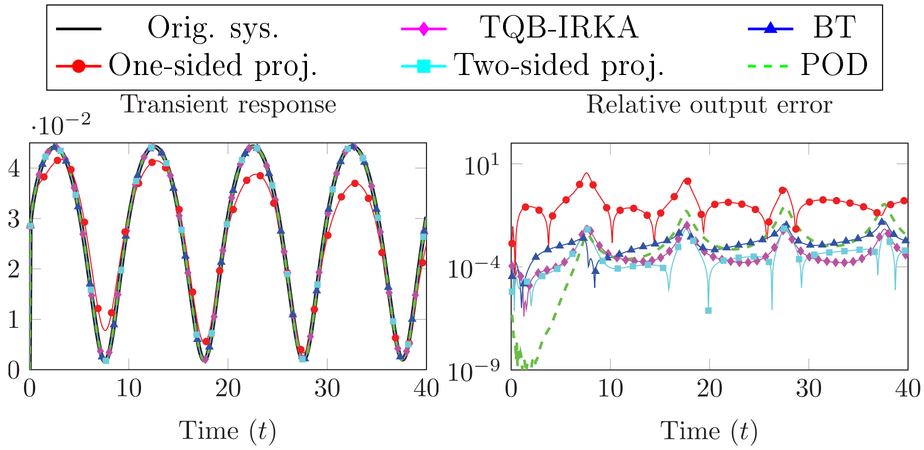


Figure 5.6.: RC circuit: comparison of responses for the input $u^{(2)} = 2.5\,(\sin(\pi t/5) + 1)$.

Input	TQB-IRKA	BT	One-sided	Two-sided	POD
$u^{(1)}(t)$	$8.82 \cdot 10^{-5}$	$3.67 \cdot 10^{-4}$	$6.50 \cdot 10^{-2}$	$1.01 \cdot 10^{-4}$	$7.24 \cdot 10^{-8}$
$u^{(2)}(t)$	$1.12 \cdot 10^{-3}$	$2.15 \cdot 10^{-3}$	$2.32 \cdot 10^{-1}$	$7.80 \cdot 10^{-4}$	$7.8 \cdot 10^{-3}$

Table 5.3.: RC circuit: the mean absolute errors of the output.

Further, we compute the quantities as defined in (5.58) using the reduced-order system of order $r = 10$ obtained upon convergence of TQB-IRKA and list them in Section 5.3.2. This also indicates that the obtained reduced-order system using TQB-IRKA satisfies all the optimality conditions (5.32) very accurately even though the nonlinear part of the system plays a significant role in the system dynamics.

$\mathcal{E}_C$	$\mathcal{E}_B$	$\mathcal{E}_N$	$\mathcal{E}_H$	$\mathcal{E}\lambda$
3.99×10^{-10}	4.68×10^{-8}	3.91×10^{-7}	3.37×10^{-8}	3.91×10^{-8}

Table 5.4.: RC circuit: perturbations to the optimality conditions.

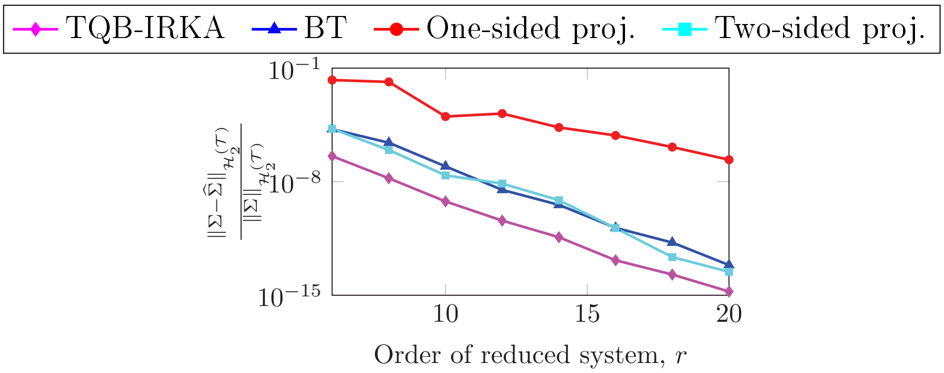


Figure 5.7.: RC circuit: comparison of the truncated $\mathcal{H}_2$-norm of the error system obtained via different methods of various orders.

Next, we also compare the truncated $\mathcal{H}_2$-norm of the error system, i.e., $\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2^{(\mathcal{T})}}$, in Figure 5.7, where the reduced models are constructed by various methods of different orders. The figure shows that TQB-IRKA yields the best reduced-order systems with respect to the truncated $\mathcal{H}_2$-norm among the investigated methods.

Note that we apply POD to the original system with exponential nonlinearities; therefore, we cannot compute the truncated $\mathcal{H}_2$-norm defined in Lemma 5.4. Hence, POD is omitted in Figure 5.7.

5.3.3. The FitzHugh-Nagumo (F-N) system

This example considers the F-N system, describing activation and deactivation dynamics of spiking neurons which is also considered in Subsection 4.4.3. Recall the governing nonlinear coupled PDEs:

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

with the nonlinear function $f(v(x,t)) = v(v-0.1)(1-v)$, and initial and boundary conditions as follows:

$$\begin{aligned}&v(x,0) = 0, && w(x,0) = 0, && x \in (0,L),\\ &v_x(0,t) = i_0(t), && v_x(1,t) = 0, && t \geq 0,\end{aligned}$$

where $\epsilon = 0.015$, $h = 0.5$, $\gamma = 2$, $q = 0.05$, and $i_0(t)$ is an actuator, acting as a control input. We set $L = 0.3$. The voltage and recovery voltage are denoted by v and w, respectively. Furthermore, we also consider the same output as considered in Subsection 4.4.3, which is the limit-cycle at the left boundary, i.e., $x = 0$. The system can be considered as having two inputs, namely q and $i_0(t)$; it has also two outputs, which are $v(0, t)$ and $w(0, t)$. This means that the system is a multi-input multi-output (MIMO) system as opposed to the two previous examples. We discretize the governing equations using a finite difference scheme. This leads to an ODE system, having cubic nonlinearity, which can then be transformed into the QB form. We consider $k = 300$ grid points, resulting in a QB system of order $3k = 900$.

We next determine reduced-order systems of order $\widehat{n} = 35$ using TQB-IRKA, BT, and POD. We choose the scaling factor $\gamma = 1$ in TQB-IRKA and it requires 26 iterations to converge. In order to apply POD, we first collect 500 snapshots of the original system for the time interval $t \in (0, 10]$ using $i_0(t) = 50(\sin(2\pi t) - 1)$ and then determine the projection based on the 35 dominant modes. The one-sided and two-sided subsystem-based interpolatory projection methods have major disadvantages in the MIMO QB case. The one-sided interpolatory projection approach of [78] can be applied to MIMO QB systems, however the dimension of the subspace V, and thus the dimension of the reduced model, increases quadratically due to the $V \otimes V$ term. As we mentioned in Section 5.1, two-sided interpolatory projection is only applicable to single-input single output (SISO) QB systems. When the number of inputs and outputs are the same, which is the case in this example, one can still employ [25, Algo. 1] to construct a reduced-order system. This is exactly what we did here. However, it is important to note that even though the method can be applied numerically, it no longer ensures the theoretical subsystem interpolation property. Despite these drawbacks, for completeness of the comparison, we still construct reduced models using both one-sided and two-sided subsystem-based interpolatory projections.

Since the F-N system has two inputs and two outputs, each interpolation point yields 6 columns of the projection matrices V and W. Thus, in order to apply the two-sided projection, we use 6 linear $\mathcal{H}_2$-optimal points and determine the reduced-order system of order 35 by taking the 35 dominant vectors. We do the same for the one-sided interpolatory projection method to compute the reduced-order system.

Next, we compare the quality of the reduced-order systems and plot the transient responses and the absolute errors of the outputs in Figure 5.8 for the training input $i_0(t) = 50(\sin(2\pi t) - 1)$.

As anticipated, POD provides a very good reduced-order system since the POD basis is constructed by using the same trajectory. Note that despite not reporting CPU times for the offline phases in this paper, due to the very different levels of the implementations used for the various methods, we would like to mention that in this example the construction of the POD basis with the fairly sophisticated MATLAB in-

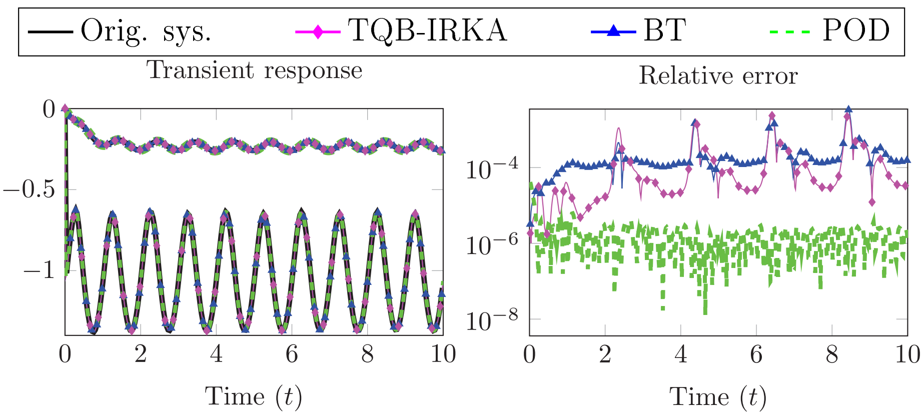


Figure 5.8.: The FitzHugh-Nagumo system: comparison of the limit-cycle at the left boundary, $x = 0$ for $i_0(t) = 50(\sin(2\pi t) - 1)$.

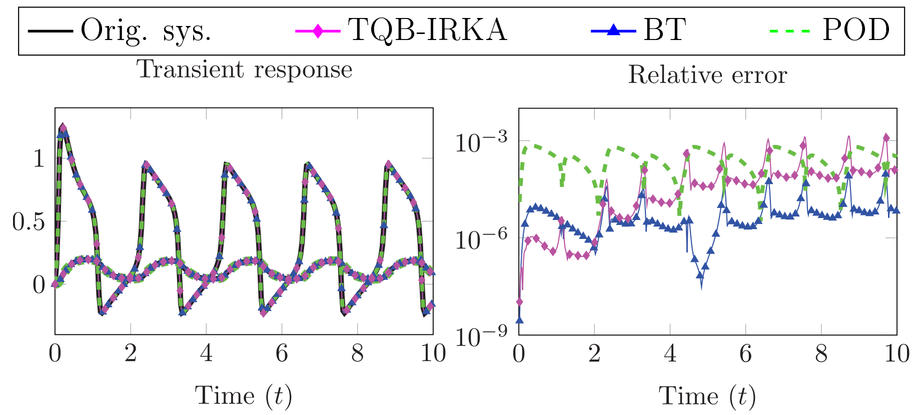


Figure 5.9.: The FitzHugh-Nagumo system: comparison of the limit-cycle at the left boundary, $x = 0$ for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$.

tegrator `ode15s` takes roughly 1.5 more CPU time than constructing the TQB-IRKA reduced-order model with our vanilla implementation.

Comparing TQB-IRKA and BT, TQB-IRKA gives a marginally better reduced-order system as compared to BT for $i_0(t) = 50(\sin(2\pi t) - 1)$, but still both are very competitive. In contrast, the one-sided and two-sided interpolatory projection methods produce unstable reduced-order systems and are therefore omitted from the figures.

To test the robustness of the obtained reduced-order systems, we choose a different control input $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and compare the transient responses in Figure 5.9. In this figure, we observe that BT performs the best among all methods for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and POD and TQB-IRKA produce reduced-order systems of almost the same quality. One-sided and two-sided projection result in unstable reduced-order systems for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$ as well. Furthermore, we also

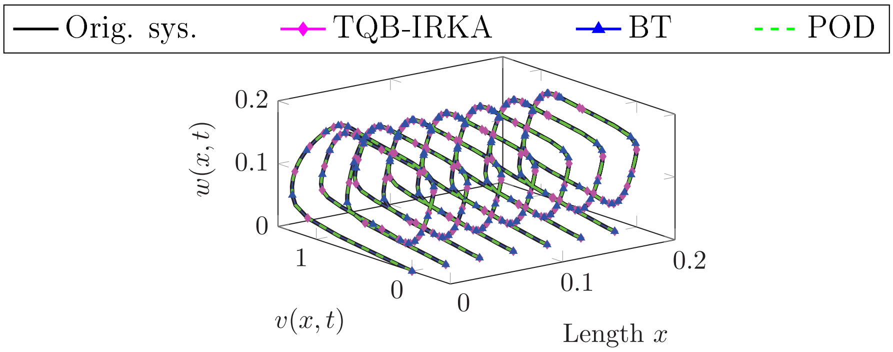


Figure 5.10.: The FitzHugh-Nagumo system: limit-cycle behavior of the original and reduced-order systems in the spatial domain.

$\mathcal{E}_C$	$\mathcal{E}_B$	$\mathcal{E}_N$	$\mathcal{E}_H$	$\mathcal{E}_\lambda$
8.76×10^{-8}	7.35×10^{-9}	1.78×10^{-11}	4.27×10^{-9}	9.14×10^{-10}

Table 5.5.: The FitzHugh-Nagumo system: perturbations to the optimality conditions.

show the limit-cycles on the full space obtained from the original and reduced-order systems in Figure 5.10 for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and observe that the reduced-order systems obtained using POD, TQB-IRKA, and BT, enable us to reproduce the limit-cycles, which is a typical neuronal dynamics as shown in Figures 5.8 and 5.10

Stressing here again that for particular interpolation points and higher-order moments, it might be possible to construct reduced-order systems via one-sided and two-sided interpolatory projection methods, which can reconstruct the limit-cycles as shown in [23]. But as discussed in [23], stability of the reduced-order systems is highly sensitive to these specific choices and even slight modifications may lead to unstable systems. For the $\mathcal{H}_2$ linear optimal interpolation points selection we made here, the one-sided and two-sided approaches were not able to reproduce the limit-cycles; thus motivating the usage of TQB-IRKA and BT once again, especially for the MIMO case.

Moreover, we report how far the reduced-order system of order $\widehat{n} = 35$ due to TQB-IRKA is from satisfying the optimality conditions (5.32). For this, we compute the perturbations (5.58) and list them in Section 5.3.3. This clearly indicates that the reduced-order system almost satisfies all optimality conditions.

Lastly, we measure the truncated $\mathcal{H}_2$-norm of the error systems, using the reduced-order systems obtained via different methods of various orders. We plot the relative truncated $\mathcal{H}_2$-norm of the error systems in Figure 5.11. We observe that TQB-IRKA produces better reduced-order systems with respect to the truncated $\mathcal{H}_2$-norm as compared to BT and one-sided projection. Furthermore, since we require stability of the matrix $\widehat{A}$ in the reduced QB system (4.2) to be able to compute the truncated $\mathcal{H}_2$-norm

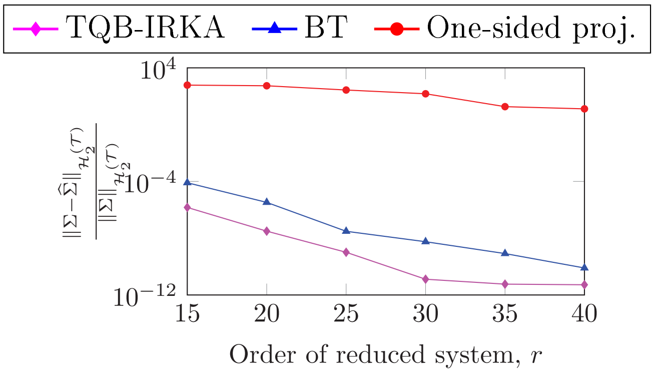


Figure 5.11.: The FitzHugh-Nagumo system: comparison of the truncated $\mathcal{H}_2$-norm of the error system, having obtained reduced-order systems of different orders using various methods.

of the error systems, we could not achieve this in the case of two-sided projection. For POD, we preserve the cubic nonlinearity in the reduced-order system; hence, the truncated $\mathcal{H}_2$-norm definition in Lemma 5.4 does not apply. Thus, we cannot compute the truncated $\mathcal{H}_2$-norm of the error system in the cases of the two-sided projection and POD, thereby these methods are not included in Figure 5.11.

5.4. Conclusions and Outlook

In this paper, we have investigated the optimal model reduction problem for quadratic-bilinear control systems. We have first defined the $\mathcal{H}_2$-norm for quadratic-bilinear systems based on the kernels of the underlying Volterra series and introduced a truncated $\mathcal{H}_2$-norm. We have then derived the first-order necessary conditions to be satisfied by a minimizer of the newly defined truncated $\mathcal{H}_2$-norm of the error system. These optimality conditions lead to the proposed model reduction algorithm (TQB-IRKA), which iteratively constructs reduced order models that *approximately* satisfy the optimality conditions. We have also discussed the efficient computation of the reduced Hessian, utilizing the Kronecker product structure of the Hessian of the QB system. Via several numerical examples, we have shown that TQB-IRKA outperforms the one-sided interpolation method, performs better than the two-sided projection in the majority of the cases, and is comparable to balanced truncation. Furthermore, unlike POD, since TQB-IRKA only depends on the state space quantities and not a specific choice of input, it outperforms POD for input functions that were not in the training set. Even for inputs which are used to train POD, TQB-IRKA still yields satisfactory performance, but is not better than POD as expected. Especially for MIMO QB systems, TQB-IRKA and BT are the preferred methods of choice to construct reduced-orders since the current framework of two-sided subspace interpolatory projection method is only applicable to SISO systems and the extension of the one-sided interpolatory projection method to MIMO QB systems yields reduced models whose dimension increases quadratically with the number of inputs. Moreover, our numerical experiments reveal that in terms of stability, the reduced-order systems via TQB-IRKA and BT are more robust as compared to the one-sided and two-sided interpolatory projection methods although we do not have any theoretical justification of this observation yet.

So far, it is not clear how to choose an appropriate order of reduced-order system in TQB-IRKA unlike in balanced truncation. Therefore, it would be a promising contribution if a priory error estimation can be derived, allowing us to determine a suitable order of a reduced-order system. Additionally, even though a stable random initialization of TQB-IRKA has performed well in all of our numerical examples, a more educated but cheaper initial guess, for example via the two-sided interpolatory method [25], can further improve the convergence of TQB-IRKA and the quality of the obtained reduced-order systems. Even though we have investigated the efficient computation of the reduced Hessian by utilizing the Kronecker product structure of the Hessian of the QB system, further research in this direction using even more sophisticated tools from tensor theory would prove significant in accelerating the iteration steps in TQB-IRKA. Furthermore, it is worthwhile to further investigate the convergence of $\mathcal{H}_2$ iterative schemes such as TQB-IRKA, and the asymptotic stability of the reduced-order systems upon convergence. A natural extension of TQB-IRKA to quadratic-bilinear descriptor systems still remains an open problem.

CHAPTER 6

INTERPOLATION-BASED MODEL ORDER REDUCTION FOR BILINEAR DESCRIPTOR SYSTEMS

Contents

6.1. Introduction

In this chapter, we turn our attention to model order reduction for descriptor systems. Precisely, we study interpolation-based model order reduction for bilinear systems subject to algebraic constraints, which are referred to as bilinear differential algebraic equations (DAEs), or bilinear descriptor systems. In general, a bilinear descriptor system is of the form

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{6.1}$$

where $x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the state, input, and output vectors at time t, respectively. The matrices E, A, N_k, $k \in \{1,\ldots,m\}$, B, C and D are all real with dimensions determined by those of $x(t)$, $u(t)$ and $y(t)$. The matrix E is considered to be singular, but it is assumed that the matrix pencil $\alpha E - \beta A$ is regular, that is,

$$\det(\alpha E - \beta A) \neq 0, \quad \text{for some} \quad (\alpha, \beta) \in \mathbb{C}^2.$$

The generalized eigenvalues of the matrix pencil $\lambda E - A$ are defined by pairs $(\alpha_i, \beta_i) \in \mathbb{C}^2 \backslash \{0, 0\}$ such that $\det(\alpha_i E - \beta_i A) = 0$. The pairs corresponding to $\beta_i \neq 0$ are the finite eigenvalues of the matrix pencil, given as $\lambda_i = \alpha_i / \beta_i$, and on the other hand, the pairs corresponding to $\beta_i = 0$, are called infinite eigenvalues of the matrix pencil. Additionally, we assume that the matrix pencil $\lambda E - A$ is *c-stable*, that is, all the finite eigenvalues of the matrix pencil lie in the open left half plane. These assumptions are made in order to ensure the existence and uniqueness of smooth solutions to the dynamical system for sufficiently smooth inputs. For more details, we refer to [97].

Moreover, if the matrix pencil $\lambda E - A$ is regular, then there exist nonsingular matrices X and Y, transforming the pencil into the Weierstrass canonical form [19, 124]:

$$E = X \begin{bmatrix} I_{n_f} & 0 \\ 0 & \bar{N} \end{bmatrix} Y, \qquad A = X \begin{bmatrix} J & 0 \\ 0 & I_{n_\infty} \end{bmatrix} Y,$$

where the Jordan matrix J is such that its eigenvalues coincide with the finite eigenvalues of the matrix pencil, and $\bar{N}$ is a nilpotent matrix corresponding to the infinite eigenvalues. If the index of nilpotent matrix $\bar{N}$ is $\nu > 0$, then $\bar{N}^{\nu} = 0$ and $\bar{N}^{\nu-1} \neq 0$. This nilpotency index is often called the (Kronecker) *index* of the matrix pencil $\lambda E - A$. Moreover, n_f and n_∞ denote the dimensions of deflating subspaces of $\lambda E - A$ corresponding to the finite and infinite eigenvalues. For more details, see, e.g., [97].

When it comes to defining an index of a general nonlinear system, there are many notions of indices for descriptor systems such as the *differentiation index*, the *tractability index*, the *Kronecker index* of nonlinear DAEs, see, e.g., [9, 97, 100]. Determining these indices of a system especially for nonlinear systems might be very complicated. Moreover, it is still not clear how these index concepts might help in the model reduction framework for nonlinear settings. We, however, know from linear cases that the block structure of matrices E and A can be utilized in the model reduction process. Here, we also aim at extending these ideas for linear DAEs to bilinear DAEs, where we take an advantage of the block structure of matrices E and A, or in other words, the weierstrass canonical form of pencil $\lambda E - A$, see, e.g., [80, 125]. This way, one can arguably define the simplest notion of index in the case of bilinear systems, which is based on the index of the matrix pencil $\lambda E - A$ or the Kronecker index of the matrix pencil. More precisely, in this thesis, our focus lies on bilinear DAEs that have the matrix pencil $\lambda E - A$ of index-1 and index-2 and to show how the existing interpolatory techniques for bilinear ODEs can be extended to such bilinear DAEs.

Coming back to MOR problem for bilinear systems, many model reduction techniques for linear systems have been extended to bilinear systems with $E{=}I$ or E being invertible. Gramians-based approaches have been discussed in a great detail in Chapter 3, and interpolation-based model reduction techniques have also been successfully extended from the linear case to the bilinear case, see, e.g., [12, 40, 106], where interpolation of the leading k subsystems is considered. In [133], the Gramian-based Wilson conditions for $\mathcal{H}_2$-optimality were extended from linear systems [132] to bilinear systems.

Later, the analog problem of determining an $\mathcal{H}_2$-optimal reduced-order system for bilinear systems was considered in [21], where first-order necessary conditions for $\mathcal{H}_2$-optimality are derived by taking derivatives of the $\mathcal{H}_2$-norm of the error system with respect to the entries of the realization of the reduced-order system. Based on these conditions, the bilinear iterative rational Krylov algorithm (B-IRKA) was proposed which upon convergence leads to a locally $\mathcal{H}_2$-optimal reduced-order system. Moreover, recently, a new framework of interpolation for bilinear systems, the so-called multipoint interpolation, was considered, which interpolates the whole underlying Volterra series at predefined frequency points [59, 60], and therein also, first-order necessary conditions for $\mathcal{H}_2$-optimality were also proposed but in terms of the pole-residue formulation. It is also shown that a reduced-order system, satisfying the $\mathcal{H}_2$ optimality conditions in the pole-residues form, satisfies also the optimality conditions derived in [21].

However, there are ample challenges when it comes to model reduction of bilinear descriptor systems with singular E, and it is necessary to study this case due to its omnipresence in applications [97]. In this chapter, we focus on interpolatory model reduction techniques for bilinear descriptor systems with singular matrix E. The interpolation conditions for bilinear systems with $E = I$ can be readily extended to singular E by just replacing I by E. However, it is shown in [80] that directly extending the interpolation conditions for linear ODEs to linear DAEs may lead to an unbounded error in the $\mathcal{H}_2$-norm due to the mismatch of the polynomial part of the system. This observation immediately holds for bilinear descriptor systems as well. As a consequence, we need to pay a special attention to the polynomial part of the bilinear system along with interpolation.

Our primary focus lies in extending the existing interpolation methods for bilinear ODEs such as subsystem interpolation method, see, e.g., [12, 40], the Volterra series interpolation [61] to bilinear DAEs, having special structures, while paying a special attention to polynomial part of the bilinear system along with interpolation. Furthermore, we aim to study how to construct $\mathcal{H}_2$-optimal reduced-order systems for the specially structured bilinear DAEs, extending the work done in [21, 61] for bilinear ODE systems.

The structure of the chapter is as follows. In the subsequent section, we provide a detailed overview of interpolation-based model reduction techniques for bilinear ODEs. This includes subsystem interpolation [40], Volterra interpolation [61] and interpolation-based $\mathcal{H}_2$-optimal approximation [21, 61]. We then discuss challenges in extending these techniques to bilinear DAEs. In Section 6.3, we extend the subsystem interpolation method to a special class of bilinear DAEs, having index-1 matrix pencil $\lambda E - A$, which, along with interpolation, also focuses on retaining the polynomial part of the bilinear system. Furthermore, we discuss the Volterra series interpolation of the same class of bilinear DAEs and investigate their $\mathcal{H}_2$-optimal model reduction problem in the subsequent section. In Section 6.5, we discuss how $\mathcal{H}_2$-optimal model reduction for bilinear ODEs can be employed to another specially structured bilinear DAEs, having index-2 matrix pencil $\lambda E - A$. Finally, we conclude the chapter with our contributions and future research topics.

6.2. Interpolation-Based MOR for Bilinear ODE Systems

In this section, we briefly provide an overview of the subsystem interpolation and Volterra series interpolation for bilinear ODEs and later review first-order necessary conditions for $\mathcal{H}_2$-optimality. We consider a bilinear ODE system of the form

$$\Sigma_B : \begin{cases} \dot{x}(t) = Ax(t) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \\ y(t) = Cx(t), \quad x(0) = 0, \end{cases} \tag{6.2}$$

where the dimensions of $A, N_k \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{p\times n}$. Our goal is to construct a reduced-order system

$$\widehat{\Sigma}_B : \begin{cases} \dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t) + \sum_{k=1}^{m} \widehat{N}_k \widehat{x}(t) u_k(t) + \widehat{B}u(t), \\ \widehat{y}(t) = \widehat{C}\widehat{x}(t), \quad \widehat{x}(0) = 0, \end{cases} \tag{6.3}$$

where the dimensions of $\widehat{A}, \widehat{N}_k \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$, $\widehat{B} \in \mathbb{R}^{n\times\widehat{n}}$ and $C \in \mathbb{R}^{\widehat{n}\times n}$ with $\widehat{n} \ll n$, ensuring $y \approx \widehat{y}$ for all admission input functions $u \in L_2^m[0, \infty)$.

For simplicity, we denote the system (6.2) by Σ_B. Moreover, for ease of notation, we stick to single-input single-output (SISO) bilinear systems while discussing subsystem interpolation and Volterra series interpolation; however, they are also applicable to bilinear multi-input multi-output (MIMO) systems. In the case of SISO bilinear systems, we denote $N_1 =: N$. Recall from Section 3.2, the output $y(t)$ of a SISO bilinear system Σ_B can be described by a nonlinear mapping of the input $u(t)$:

$$y(t) = \sum_{k=1}^{\infty} \int_0^t \int_0^{t_1} \cdots \int_0^{t_{k-1}} g_k(t_1, t_2, \dots, t_k) u(t - t_1 - t_2 \cdots - t_k) \cdots u(t - t_k) dt_1 \cdots dt_k,$$

where g_k is the regular Volterra *kernel*, whose corresponding multi-variate transfer functions can be given by

$$G_k(s_1, s_2, \dots, s_k) = C(s_k I - A)^{-1} N \cdots (s_2 I - A)^{-1} N (s_1 I - A)^{-1} B.$$

For more details, we refer to Section 3.2, where we have collected the relevant control theoretic concepts for bilinear systems. Interpolatory-based model reduction techniques for bilinear systems have been widely studied in the literature; see, e.g., [12, 40, 61, 106]. In this following, we first show how to construct a reduced-order system, whose the first k subsystems interpolate the corresponding original subsystem.

6.2.1. Subsystem interpolation for bilinear ODEs

The idea of interpolation of subsystems of bilinear systems mainly relies on the moments of the multi-variate transfer functions. Thus, we first define the multi-moments for bilinear systems.

Definition 6.1:
Let $G_k(s_1,\dots,s_k)$ be kth-order multi-variate transfer function of the SISO bilinear system (6.2). Performing the Neumann series expansion of $G_k(s_1,\dots,s_k)$ for $(s_1,\dots,s_k)$ around the expansion points $(\sigma_1,\dots,\sigma_k)$ leads to

$$G_k(s_1,\dots,s_k) = \sum_{l_k=1}^{\infty}\cdots\sum_{l_1=1}^{\infty} m(l_1,\dots,l_k)(s_1-\sigma_1)^{l_1-1}\cdots(s_k-\sigma_k)^{l_k-1},$$

where

$$m(l_1,\dots,l_k) = (-1)^k C^T(A-\sigma_k I_n)^{-l_k}N\cdots(A-\sigma_2 I_n)^{-l_2}N(A-\sigma_1 I_n)^{-l_1}B.$$

The $m(l_1,\dots,l_k)$ are called multi-moments of the bilinear system, corresponding to the kth-order multi-variate transfer function at $(\sigma_1,\dots,\sigma_k)$. ♢

The idea of subsystem interpolation is to construct a reduced-order system such that the multi-moments at a specific set of expansion points match with those of the original system. In this direction, the subsystem interpolation was first investigated in [106] for SISO bilinear systems, and a methodology to construct reduced-order systems was proposed, matching the multi-moments at infinity, i.e., $\sigma_1 = \sigma_2 = \cdots = \sigma_k = \infty$. Then, the interpolating the multi-moments at zero, i.e., $\sigma_1 = \sigma_2 = \sigma_k = 0$, was considered in [12]. Finally, the problem of determining a reduced-order system, matching multi-moments at any given interpolation points $\sigma_i \in \mathbb{C}$ was explored in [40], which is outlined in the following theorem, and we aim at extending this method to bilinear DAEs later in Section 6.3.

Theorem 6.2 ([40]):
Consider arbitrary interpolation points $\sigma_j, \mu_j \in \mathbb{C}$ such that $sI_n - A$ and $sI_{\widehat{n}} - \widehat{A}$ are invertible for $s = \sigma_j, \mu_j,\ j \in \{1,\dots,k\}$. Define the projection matrices V and W as follows:

$$\begin{aligned}
\operatorname{range}\left(V^{(1)}\right) &= \mathcal{K}_q\left((\sigma_1 I - A)^{-1}, (\sigma_1 I - A)^{-1}B\right),\\
\operatorname{range}\left(V^{(i)}\right) &= \mathcal{K}_q\left((\sigma_i I - A)^{-1}, (\sigma_i I - A)^{-1}NV^{(i-1)}\right), \quad i \in \{2,\dots,k\},\\
\operatorname{range}\left(W^{(1)}\right) &= \mathcal{K}_q\left((\mu_1 I - A)^{-T}, (\mu_1 I - A)^{-T}C^T\right),\\
\operatorname{range}\left(W^{(i)}\right) &= \mathcal{K}_q\left((\mu_i I - A)^{-T}, (\mu_i I - A)^{-T}N^T W^{(i-1)}\right), \quad i \in \{2,\dots,k\},
\end{aligned}$$

$$\operatorname{range}(V) = \bigcup_{i=1}^{k}\left\{\operatorname{range}\left(V^{(i)}\right)\right\}, \quad \operatorname{range}(W) = \bigcup_{i=1}^{k}\left\{\operatorname{range}\left(W^{(i)}\right)\right\},$$

where $\mathcal{K}_q(\mathcal{A},\mathcal{B}) = \operatorname{span}(\mathcal{B}, \mathcal{A}\mathcal{B},\dots,\mathcal{A}^{q-1}\mathcal{B})$ denotes the Krylov subspace. Assuming V and W are of full column rank and reduced matrices are construed as:

$$\begin{aligned}
\widehat{E} &= W^T E V, & \widehat{A} &= W^T A V, & \widehat{N} &= W^T N V,\\
\widehat{B} &= W^T B, & \widehat{C} &= CV,
\end{aligned}$$

then the multi-moments of multi-variate transfer functions of the reduced-order system $\widehat{m}(l_1,\ldots,l_i)$ match those of the corresponding multi-variate transfer functions of the original system as follows:

$$m(l_1,\ldots l_i) = \widehat{m}(l_1,\ldots,l_i), \quad \text{for } i \in \{1,\ldots,k\} \text{ and } l_{\{1,\ldots,k\}} \in \{1,\ldots,q\}. \qquad \diamondsuit$$

Although the interpolation of subsystems have been successfully applied to various applications to construct reduced-order systems, the main setback of this method is that the dimension of the reduced-order system can increase rather rapidly. For instance, if the subspaces V and W are determined by using the first k subsystems as shown in Theorem 6.2, then an interpolating reduced-order system will have the dimension of order $r = q + q^2 + \cdots + q^k$. This grows even more rapidly in the case of MIMO systems. However, in practice, we observe that interpolations of the first two subsystems result in reduced-order systems, which can replicate the important dynamics of the original bilinear system.

Recently, a novel interpolation problem of the bilinear system was proposed in [59]. This is motivated by the fact that the response of a bilinear system is given by an infinite series, the so-called Volterra series of a bilinear system, and it is linked with the inverse Laplace transform of the kth-order multi-variate transfer function. Therefore, it would be also interesting to enforce the interpolation of the whole Volterra series, rather than interpolating each subsystem, separately.

6.2.2. Multi-point Volterra series interpolation for bilinear ODEs

Next, we outline the multi-point interpolation of the Volterra series problem statement for the bilinear system (6.2). For this, we consider two sets of interpolation points $\sigma_j, \mu_j \in \mathbb{C}$, for $j \in \{1,\ldots,\widehat{n}\}$, along with matrices $U, S \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$ and define the weighted Volterra series as

$$\zeta_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\sum_{l_2=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\ldots,l_{k-1},j} G_k(\sigma_{l_1},\sigma_{l_2},\ldots,\sigma_j) \tag{6.4}$$

and

$$\varphi_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\sum_{l_2=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,\ldots,l_{k-1},j} G_k(\mu_j,\mu_{l_1},\ldots,\mu_{l_{k-1}}), \tag{6.5}$$

where $\eta_{l_1,\ldots,l_{k-1},j}$ and $\vartheta_{l_1,\ldots,l_{k-1},j}$ are the weights associated to each subsystem in the Volterra series and are defined in terms of the elements of the matrices U and S as follows:

$$\begin{aligned} \eta_{l_1,\ldots,l_{k-1},j} &= u_{j,l_{k-1}} u_{l_{k-1},l_{k-2}} \cdots u_{l_2,l_1} \quad \text{for} \quad k \geq 2 \quad \text{and} \quad \eta_{l_1} = 1, \\ \vartheta_{l_1,\ldots,l_{k-1},j} &= s_{j,l_{k-1}} s_{l_{k-1},l_{k-2}} \cdots s_{l_2,l_1} \quad \text{for} \quad k \geq 2 \quad \text{and} \quad \vartheta_{l_1} = 1. \end{aligned} \tag{6.7}$$

The goal of the new interpolation framework is to construct a reduced-order system $\widehat{\Sigma}_B$ of dimension $\widehat{n}$ such that the following interpolation conditions are satisfied for each $j \in \{1, \ldots, \widehat{n}\}$:

$$\zeta_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \sum_{l_2=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\ldots,l_{k-1},j} \widehat{G}_k(\sigma_{l_1}, \sigma_{l_2}, \ldots, \sigma_j) \tag{6.8}$$

and

$$\varphi_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \sum_{l_2=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,\ldots,l_{k-1},j} \widehat{G}_k(\mu_j, \mu_{l_1}, \ldots, \mu_{l_{k-1}}), \tag{6.9}$$

where $\widehat{G}_k(\mu_{l_1}, \ldots, \mu_k)$ is the kth-order multi-variate transfer function associated with the reduced-order system. Similar to the linear case, the reduced-order system matrices are constructed via projection matrices V and W, assuming W^TV being invertible, as follows:

$$\begin{aligned} \widehat{A} &= (W^TV)^{-1}W^TAV, & \widehat{N} &= (W^TV)^{-1}W^TNV, \\ \widehat{B} &= (W^TV)^{-1}W^TB, & \widehat{C} &= CV. \end{aligned} \tag{6.10}$$

Then, the problem of identifying these projection matrices is considered in [59], which provides us a reduced-order system such that the interpolation conditions (6.8) and (6.9) are satisfied. The following theorem suggests the choice of such projection matrices.

Theorem 6.3 ([59]):
Consider a SISO bilinear system Σ_B of dimension n and the interpolation points σ_j, $\mu_j \in \mathbb{C}$, $j \in \{1, \ldots, \widehat{n}\}$, along with matrices $U, S \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$. Let the projection matrices V and W be the solutions of the following Sylvester equations

$$V\Omega - AV - NVU^T = B\mathbb{1}_{\widehat{n}}^T \tag{6.11}$$

and

$$W\Xi - A^TW - N^TWS^T = C^T\mathbb{1}_{\widehat{n}}^T, \tag{6.12}$$

where $\Omega = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\widehat{n}})$, $\Xi = \operatorname{diag}(\mu_1, \ldots, \mu_{\widehat{n}})$, and $\mathbb{1}_{\widehat{n}}^T$ is the vector of ones in $\mathbb{R}^{\widehat{n}}$. Furthermore, assume that $W^TV \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ is invertible and a reduced-order system $\widehat{\Sigma}_B$ of order $\widehat{n}$ is computed using the projection matrices V and W as shown in (6.10), then the interpolation conditions (6.8) and (6.9) are fulfilled. ♢

6.2.3. $\mathcal{H}_2$-optimal model reduction for bilinear ODEs

Here, we present a construction of an $\mathcal{H}_2$-optimal reduced-order system for bilinear ODEs. There exist mainly two approaches in the literature for it. One approach is to extend the structured-orthogonality conditions for linear systems [79] to bilinear systems. In [79], it is shown for linear systems that a particular type of conditions for Hilbert-space orthogonality is equivalent to all other derived first-order necessary conditions for linear systems in the literature. These structure-orthogonality conditions for bilinear systems are studied in [60]. However, therein, it is concluded that in general, it is not possible to construct a reduced-order system, satisfying such orthogonality conditions. This is because we require an infinite dimensional bilinear realization, which can satisfy such orthogonality conditions. This is undesirable, especially when we are aiming at constructing low-order models.

Another approach is a conventional one, which first involves deriving the error expression, that is, the $\mathcal{H}_2$-norm of the error system and then derive the optimality conditions which minimize the error expression. Initially, this problem was considered in [133], where first-order necessary conditions for optimality were derived which minimize the desired error expression. However, it was not an easy task to construct a reduced-order system from these derived optimality conditions. Later on, the analog problem was considered in [21], where it is shown how to write the $\mathcal{H}_2$-norm of the error using Kronecker product properties, leading to computable necessary conditions for optimality. In the following theorem, we first note down the $\mathcal{H}_2$-norm of the error system.

Theorem 6.4 ([21]):
Let Σ_B and $\widehat{\Sigma}_B$ be the original system (6.2) and a reduced-order system (6.3), respectively. Then, the $\mathcal{H}_2$-norm of the error system can be given by

$$\begin{aligned}
&\|\Sigma_B - \widehat{\Sigma}_B\|_{\mathcal{H}_2}\\
&= \mathcal{J}_p^T \left(\begin{bmatrix} C & -\widetilde{C} \end{bmatrix} \otimes \begin{bmatrix} C & -\widehat{C} \end{bmatrix} \right) \times\\
&\left(- \begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\widehat{n}} \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_{\widehat{n}} \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \widehat{A} \end{bmatrix} - \sum_{k=1}^{m} \begin{bmatrix} N_k & 0 \\ 0 & \widetilde{N}_k \end{bmatrix} \otimes \begin{bmatrix} N_k & 0 \\ 0 & \widehat{N}_k \end{bmatrix} \right)^{-1} \times\\
&\left(\begin{bmatrix} B \\ \widetilde{B} \end{bmatrix} \otimes \begin{bmatrix} B \\ \widehat{B} \end{bmatrix} \right) \mathcal{J}_m
\end{aligned} \tag{6.13}$$

where $R\Lambda R^{-1}$ is the spectral decomposition of $\widehat{A}$, and $\widetilde{B}$, $\widetilde{C}$ and $\widetilde{N}_k$ are defined as $R^{-1}\widehat{B}$, CR and $R^{-1}\widehat{N}_k R$, respectively. ◊

Having had the error expression (6.13), the next goal is to determine a reduced-order system of order $\widehat{n}$ that solves

$$\min_{\widehat{\Sigma}_B \text{ of order } \widehat{n}} \|\Sigma_B - \widehat{\Sigma}_B\|_{\mathcal{H}_2}. \tag{6.14}$$

To solve the above minimization problem, one can consider Λ, $\widetilde{B}$, $\widetilde{C}$ and $\widetilde{N}_k$ as optimization parameters. Then, first-order necessary conditions for an $\mathcal{H}_2$-optimal approximation can be derived by differentiating the error expressions with respect to the optimization parameters, which we summarize in the following theorem.

Theorem 6.5 ([21]):
Let Σ_B and $\widehat{\Sigma}_B$ be the original and reduced-order systems. Furthermore, let $R\Lambda R^{-1}$ is the spectral decomposition of $\widehat{A}$, and $\widetilde{B}$, $\widetilde{C}$ and $\widetilde{N}_k$ are defined as $R^{-1}\widehat{B}$, CR and $R^{-1}\widehat{N}_k R$, respectively. Then, if $\widehat{\Sigma}_B$ is a locally $\mathcal{H}_2$-optimal approximation of Σ_B, then the following conditions need to be satisfied:

$$
\begin{aligned}
\mathcal{J}_p^T \left(e_i e_j^T \otimes C\right) \mathcal{X} \left(\widetilde{B} \otimes B\right) \mathcal{J}_m &= \mathcal{J}_p^T \left(e_i e_j^T \otimes \widehat{C}\right) \widehat{\mathcal{X}} \left(\widetilde{B} \otimes \widehat{B}\right) \mathcal{J}_m, \\
\mathcal{J}_p^T \left(\widetilde{C} \otimes C\right) \mathcal{X} \left(e_j e_i^T \otimes B\right) \mathcal{J}_m &= \mathcal{J}_p^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{\mathcal{X}} \left(e_j e_i^T \otimes \widehat{B}\right) \mathcal{J}_m, \\
\mathcal{J}_p^T \left(\widetilde{C} \otimes C\right) \mathcal{X} \left(e_i e_i^T \otimes I_n\right) \mathcal{X} \left(\widetilde{B} \otimes B\right) \mathcal{J}_m &= \\
&\mathcal{J}_p^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{\mathcal{X}} \left(e_i e_i^T \otimes I_{\hat{n}}\right) \widehat{\mathcal{X}} \left(\widetilde{B} \otimes \widehat{B}\right) \mathcal{J}_m, \\
\mathcal{J}_p^T \left(\widetilde{C} \otimes C\right) \mathcal{X} \left(e_i e_i^T \otimes N_k\right) \mathcal{X} \left(\widetilde{B} \otimes B\right) \mathcal{J}_m &= \\
&\mathcal{J}_p^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{\mathcal{X}} \left(e_i e_i^T \otimes \widehat{N}_k\right) \widehat{\mathcal{X}} \left(\widetilde{B} \otimes \widehat{B}\right) \mathcal{J}_m
\end{aligned}
$$

in which

$$
\begin{aligned}
\mathcal{X} &:= \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \widetilde{N}_k \otimes N_k\right)^{-1}, \\
\widehat{\mathcal{X}} &:= \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \widehat{A} - \sum_{k=1}^{m} \widetilde{N}_k \otimes \widehat{N}_k\right)^{-1}.
\end{aligned}
\qquad \diamond
$$

Like the $\mathcal{H}_2$-optimality conditions for linear systems [79], in the bilinear setting as well, the optimality conditions involve the parameter of the reduced-order systems which are not available beforehand. However, Benner and Breiten in [21] have proposed an iterative scheme for bilinear systems which upon convergence leads to reduced-order systems, satisfying the optimality conditions in Theorem 6.5. We outline steps in Algorithm 6.1 to construct reduced-order systems using fix-point iterations, which extends the Iterative Rational Krylov Algorithm (IRKA) for linear systems [79] to bilinear systems.

Another way to formulate the $\mathcal{H}_2$-optimal problem is by using the pole-residue formulation. The $\mathcal{H}_2$-norm of a SISO bilinear system in terms of the pole-residue form is given by a weighted sum over all possible combinations of point evaluations at the mirror image of eigenvalues of the matrix A (the poles of the system) and then is summed over all multi-variate transfer functions, for details, see Proposition 3.9. Utilizing this, one can obtain the $\mathcal{H}_2$-norm of the error system, which has two parts: one part contains

Algorithm 6.1: Bilinear iterative rational Krylov algorithm (B-IRKA) [21].

```
Input: The system matrices: $A$, $N_k$, $B$, $C$.
Make an initial guess of $\Lambda$, $\widetilde{B}$, $\widetilde{N}_k$ and $\widetilde{C}$.
while relative change in $\{\lambda_i\} >$ tol convergence do
    Solve for $V$ and $W$:
        $V(-\Lambda) + AV + \sum_{k=1}^{m} N_k V \widetilde{N}_k^T + B\widetilde{B}^T = 0,$
        $W(-\Lambda) + A^T W + \sum_{k=1}^{m} N_k^T W \widetilde{N}_k + C^T\widetilde{C} = 0.$
    Perform:
        $V = \operatorname{orth}(V)$ and $W = \operatorname{orth}(V)$.
    Compute the reduced matrices:
        $\widehat{A} = (W^TV)^{-1}W^TAV,$    $\widehat{N}_k = (W^TV)^{-1}W^TN_kV,$
        $\widehat{B} = (W^TV)^{-1}W^TB,$    $\widehat{C} = CV.$
    Determine the spectral decomposition of $\widehat{A} =: R\Lambda R^{-1}$.
    Define $\widetilde{B}$, $\widetilde{C}$ and $\widetilde{N}_k$ are defined as $R^{-1}\widehat{B}$, $CR$ and $R^{-1}\widehat{N}_kR$, respectively.
Output: $\widehat{A}$, $\widehat{N}_k$, $\widehat{B}$, $\widehat{C}$.
```

a weighted sum of the difference of the multi-variate transfer functions of the original and reduced-order systems, computed at the mirror image of the pole of the original system across the imaginary axis, and the second part also has the similar structure but computed at the mirror image of the poles of the reduced-order system across the imaginary axis. Analogous to the linear case [7], the aim is to eliminate the error in the $\mathcal{H}_2$-norm of the error system, due to the mismatch at the reduced-order system singularities. This leads to the following first-order necessary conditions for optimality in the pole-residue formulation for a SISO bilinear system:

$$\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}}\widehat{\phi}_{l_1,\dots,l_k}\left(G_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_{l_k}) - \widehat{G}_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_{l_k})\right) = 0 \tag{6.15}$$

and

$$\begin{aligned}\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}}\widehat{\phi}_{l_1,\dots,l_k}&\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}G_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_{l_k})\right)\\ &= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}}\widehat{\phi}_{l_1,\dots,l_{l_k}}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}\widehat{G}_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_{l_k})\right),\end{aligned} \tag{6.16}$$

where the $\widehat{\lambda}_i$'s are the zeros of $\det(sI_{\widehat{n}} - \widehat{A})$, and $\widehat{\phi}_{l_1,\dots,l_k}$ are the residues of the kth-order multi-variate transfer functions $\widehat{G}_k(s_1, s_2, \dots, s_k)$ computed at $(s_1, \dots, s_k) =$

$(-\widehat{\lambda}_{l_1}, \ldots, -\widehat{\lambda}_{l_k})$, as defined in (3.16); the operator $\frac{\partial}{\partial s_j} G_k(-\widehat{\lambda}_{l_1}, \ldots, -\widehat{\lambda}_{l_k})$ denotes the partial derivative of $G_k(s_1, \ldots, s_k)$ with respect to s_j, evaluated at $(s_1, \ldots, s_k) = (-\widehat{\lambda}_{l_1}, \ldots, -\widehat{\lambda}_{l_k})$.

The connection between the $\mathcal{H}_2$-optimality conditions in the pole-residue formulations and the multi-point Volterra series interpolation was established in [59]. That is, the $\mathcal{H}_2$-optimality conditions are equivalent to the multi-point Volterra interpolation if the interpolation points Ω and Ξ in (6.11)–(6.12) are chosen to be the mirror images of the poles of the reduced-order system across the imaginary axis, i.e., $\Omega = \Xi = -\Lambda$, respectively, where $\Lambda = R^{-1}\widehat{A}R$; the matrices U and S are given by the bilinear term $\widehat{N}$ as $U = R^{-1}\widehat{N}R$ and $S = R^T\widehat{N}^TR^{-T}$, and the vector $\mathbb{1}_{\widehat{n}}$ in (6.11) and (6.12) is replaced with $R^{-1}\widehat{B}$ and $\widehat{C}R$, respectively. For details, we refer to [59, 60].

These optimality conditions in the pole-residue formulation play an important role while studying an $\mathcal{H}_2$-optimal model reduction problem for bilinear DAEs. This is because we do not have a nice expression for the error system, e.g., in a Kronecker product form, for bilinear descriptor systems, in contrast to the case of bilinear ODEs. Therefore, it is not so easy to derive the optimality conditions for bilinear DAEs with respect to the realization of the reduced-order system as done in [21]. However, since the pole-residue formulation requires information of the transfer function of a bilinear system which can be easily determined in the case of descriptor systems as well, it will be relatively easier to proceed further in the direction of the pole-residue formulation for bilinear DAEs.

Remark 6.6:
The subsystem interpolation and the multi-point interpolation of the underlying Volterra series as discussed in Subsystems 6.2.1–6.2.2, respectively can be extended to bilinear DAEs straightforwardly by replacing I_n with E while computing the projection matrices V and W. This yields a reduced-order system which satisfies the interpolation conditions. However, directly extending the interpolation conditions to descriptor systems without any modifications may lead to poor reduced-order systems with the $\mathcal{H}_2$-norm error blowing up, occurring due to the unmatched polynomial part of the system. This statement is based on the analysis in [80] for linear DAEs. $\diamondsuit$

Motivated by the work done in [80] for linear DAEs, we pay a special attention to the polynomial part of the bilinear descriptor system in the upcoming sections of this chapter along with interpolation.

6.3. Subsystem Interpolation for Index-1 Bilinear Descriptor Systems

In this section, we extend the subsystem based interpolatory model order reduction technique of SISO bilinear DAEs, having a special structure in the semi-explicit form as follows:

$$E_{11}\dot{x}_1(t) + E_{12}\dot{x}_2(t) = A_{11}x_1(t) + A_{12}x_2(t) + N_{11}x_1(t)u(t) + N_{12}x_2(t)u(t) + B_1u(t), \tag{6.17a}$$

$$0 = A_{21}x_1(t) + A_{22}x_2(t) + N_{21}x_1(t)u(t) + N_{22}x_2(t)u(t) + B_2u(t), \tag{6.17b}$$

$$y(t) = C_1x_1(t) + C_2x_2(t), \tag{6.17c}$$

where $x_1(t) \in \mathbb{R}^{n_1}$ and $x_2(t) \in \mathbb{R}^{n_2}$, and all other matrices are of appropriate sizes. Furthermore, it is assumed that A_{22} is invertible as is $E_{11} - E_{12}A_{22}^{-1}A_{21}$. Thus, the system (6.17) has an index-1 structure in case $N_{ij} = 0$, $\{i, j\} \in \{1, 2\}$. Moreover, we assume that the initial condition of the system (6.17) is consistent. For simplicity of notion, we consider SISO bilinear DAEs (6.17), but all results can be extended to MIMO bilinear DAEs.

The multi-variate transfer functions for bilinear DAEs can be determined by the exponential growth approach analogous to bilinear ODEs [111]. The structure of the multi-variate transfer function corresponding to the kth subsystem of (6.17) in the regular form is given by

$$H_k(s_1, \dots, s_k) = C(s_kE - A)^{-1}N(s_{k-1}E - A)^{-1}N \cdots N(s_1E - A)^{-1}B, \tag{6.18}$$

where

$$E = \begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix},\ A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},\ N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix},\ B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},\ C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}.$$

As mentioned at the end of the previous section, the subsystem of the leading kth subsystems of the bilinear DAEs can be achieved by using the subspaces V and W as shown in Theorem 6.2 by just replacing I_n with E. However, the direct extension to bilinear DAEs may lead to unbounded error in $\mathcal{H}_2$-norm, which is due to the unmatched polynomial part of the system. Hence, we need to take care of the polynomial part of the bilinear system as well, together with interpolation, which is our main focus.

6.3.1. Polynomial part of bilinear DAEs, having index-1 matrix pencil

As a first step, we aim at determining the polynomial of the kth-order transfer function (6.18) explicitly. The next lemma shows that each subsystem of the system (6.17) has a constant polynomial part.

Lemma 6.7:
Let $H_k(s_1, \dots, s_k) =: H_k(\mathcal{S}_k)$ be the multi-variate transfer functions of the bilinear DAEs, which are defined as in (6.18). Assume that the matrices A_{22} and

$E_{11}-E_{12}A_{22}^{-1}A_{21}$ in (6.17) are both nonsingular. Then, the polynomial part of $H_k(\mathcal{S}_k)$ is constant and is given by

$$D_k = C(MN)^{k-1}MB, \tag{6.19}$$

where M is as defined as

$$M = \lim_{s\to\infty}(sE-A)^{-1} = \begin{bmatrix} 0 & E_A^{-1}E_{12}A_{22}^{-1} \\ 0 & -A_{22}^{-1}\left(I + A_{21}E_A^{-1}E_{12}A_{22}^{-1}\right) \end{bmatrix} \tag{6.20}$$

with $E_A = E_{11} - E_{12}A_{22}^{-1}A_{21}$ and $s := 2\pi\imath f$ is the Laplace variable in which f is the frequency and $\imath$ is the imaginary unit. $\lozenge$

Proof. Let $F(\mathcal{S}_k) := F(s_1,\dots,s_k)$ be the multi-variable function

$$F(\mathcal{S}_k) = (s_kE-A)^{-1}N(s_{k-1}E-A)^{-1}N\cdots N(s_1E-A)^{-1}B, \tag{6.21}$$

then the polynomial part of $H_k(\mathcal{S}_k)$ is given by

$$D_k = C\lim_{\mathcal{S}_k\to\infty}F(\mathcal{S}_k). \tag{6.22}$$

Note that for $k=1$, Eq. (6.21) yields

$$\lim_{\mathcal{S}_1\to\infty}F(s_1) = \lim_{s_1\to\infty}(s_1E-A)^{-1}B.$$

Then, using (6.20), we obtain

$$\lim_{\mathcal{S}_1\to\infty}F(s_1) = \lim_{s_1\to\infty}(s_1E-A)^{-1}B = MB. \tag{6.23}$$

It is easy to see from (6.22) that (6.19) holds for $k=1$ (analog to the linear case [80]). Now, for $k=j\geq 1$, assume that

$$\lim_{\mathcal{S}_j\to\infty}F(\mathcal{S}_j) = (MN)^{j-1}MB. \tag{6.24}$$

Then, we need to show that the above equation holds for $k=j+1$ as well. First, note that

$$F(\mathcal{S}_{j+1}) = (s_{j+1}E-A)^{-1}NF(\mathcal{S}_j).$$

Taking the limit $\mathcal{S}_{j+1}\to\infty$, we have

$$\begin{aligned}\lim_{\mathcal{S}_{j+1}\to\infty}F(\mathcal{S}_{j+1}) &= \lim_{s_{j+1}\to\infty}(s_{j+1}E-A)^{-1}N\lim_{\mathcal{S}_j\to\infty}F(\mathcal{S}_j)\\ &= \lim_{s_{j+1}\to\infty}(s_{j+1}E-A)^{-1}N(MN)^{j-1}MB,\end{aligned}$$

where the last equation follows from (6.24). Now, we define $\mathcal{B}_{MN} := N(MN)^{j-1}MB$ and use (6.20) to obtain

$$\lim_{\mathcal{S}_{j+1}\to\infty} F(\mathcal{S}_{j+1}) = \lim_{s_{j+1}\to\infty}(s_{j+1}E - A)^{-1}\mathcal{B}_{MN} = M\mathcal{B}_{MN} = (MN)^j MB.$$

Hence, we obtain

$$D_k = C \lim_{\mathcal{S}_{j+1}\to\infty} F(\mathcal{S}_{j+1}) = C(MN)^k MB,$$

thus, concluding the proof. □

6.3.2. Subsystem interpolation while retaining the polynomial part

Since we now have the polynomial part of each subsystem, the next goal is to construct a reduced-order system that retains the polynomial part of each subsystem associated with the original bilinear system, in addition to the interpolation of subsystems. As a first step, we assume the structure of the kth-order multi-variate transfer function of a reduced bilinear system as follows:

$$\widehat{H}(s_1,\ldots,s_k) = \widehat{C}(s_k\widehat{E} - \widehat{A})^{-1}\widehat{N}(s_{k-1}\widehat{E} - \widehat{A})^{-1}\widehat{N}\cdots\widehat{N}(s_1\widehat{E} - \widehat{A})^{-1}\widehat{B} + D_k, \qquad (6.25)$$

where $\widehat{E}$, $\widehat{A}$, $\widehat{N} \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $\widehat{B}$, $\widehat{C}^T \in \mathbb{R}^{\hat{n}}$ with $\hat{n} \ll n$. Moreover, the matrix $\widehat{E}$ is assumed to be nonsingular, and D_k is the polynomial part of the kth subsystem of the original bilinear system. A major advantage of considering the structure of the kth-order transfer function of the reduced-order system as in (6.25) is that the reduced bilinear system ensures the matching of polynomial parts of the subsystems corresponding to those of the original bilinear system. Now, we need to develop conditions which guarantee the interpolation between kth-order multi-variate transfer functions of the original and reduced-order systems as well. For this, in the next theorem, we show how to construct reduced-order matrices $\widehat{E}$, $\widehat{A}$, etc., ensuring the desired goal.

Theorem 6.8:
Consider arbitrary interpolation points $\sigma_j, \mu_j \in \mathbb{C}$ such that $sE - A$ and $s\widehat{E} - \widehat{A}$ are invertible for $s = \sigma_j, \mu_j,\ j \in \{1,\ldots,k\}$. Define the projection matrices V and W as follows:

$$\begin{aligned}
\operatorname{range}\left(V^{(1)}\right) &= \mathcal{K}_q\left((\sigma_1 E - A)^{-1}, (\sigma_1 E - A)^{-1}B\right),\\
\operatorname{range}\left(V^{(i)}\right) &= \mathcal{K}_q\left((\sigma_i E - A)^{-1}, (\sigma_i E - A)^{-1}NV^{(i-1)}\right), \quad i \in \{2,\ldots,k\},\\
\operatorname{range}\left(W^{(1)}\right) &= \mathcal{K}_q\left((\mu_1 E - A)^{-T}, (\mu_1 E - A)^{-T}C^T\right),\\
\operatorname{range}\left(W^{(i)}\right) &= \mathcal{K}_q\left((\mu_i E - A)^{-T}, (\mu_i E - A)^{-T}N^T W^{(i-1)}\right), \quad i \in \{2,\ldots,k\},
\end{aligned}$$

$$\operatorname{range}(V) = \bigcup_{i=1}^{k}\operatorname{range}\left(V^{(i)}\right), \quad \operatorname{range}(W) = \bigcup_{i=1}^{k}\operatorname{range}\left(W^{(i)}\right).$$

Moreover, define the intermediate matrices as follows:

$$\begin{aligned} \widetilde{E} &= E, & \widetilde{A} &= A + L_A, & \widetilde{N} &= N - L_N, \\ \widetilde{B} &= B - L_B, & \widetilde{C} &= C - L_C, \end{aligned}$$

where L_A, L_N, L_B, L_C are solutions to the following equations:

$$W^T L_B = \left[D_1(e_1^q)^T, D_2(e_1^{q^2})^T, \ldots, D_k(e_1^{q^k})^T \right]^T, \tag{6.26a}$$

$$L_C V = \left[D_1(e_1^q)^T, D_2(e_1^{q^2})^T, \ldots, D_k(e_1^{q^k})^T \right], \tag{6.26b}$$

$$L_A V = \left[L_B(e_1^q)^T, L_N \left[V^{(1)}(I_q \otimes (e_1^q)^T), \ldots, V^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T)) \right] \right], \tag{6.26c}$$

$$W^T L_A = \left[L_C^T(e_1^q)^T, L_N^T [W^{(1)}(I_q \otimes (e_1^q)^T), \ldots, W^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T)] \right]^T, \tag{6.26d}$$

in which D_k is the polynomial part of the kth-order multi-variate transfer function of the original system, and e_1^l is the first column of the identity matrix of size $l \times l$. Then, the projection of the intermediate system results in a reduced-order system:

$$\begin{aligned} \widehat{E} &= W^T \widetilde{E} V, & \widehat{A} &= W^T \widetilde{A} V, & \widehat{N} &= W^T \widetilde{N} V, \\ \widehat{B} &= W^T \widetilde{B}, & \widehat{C} &= \widetilde{C} V \end{aligned}$$

that satisfies

$$H_k(\mathcal{S}_k) = \widehat{H}_k(\mathcal{S}_k) + \mathcal{O}\left((s_1 - \mu_1)^q \cdots (s_k - \mu_k)^q (s_1 - \sigma_1)^q \cdots (s_k - \sigma_k)^q \right). \qquad \diamondsuit$$

Proof. Consider the first subsystem at $s_1 = \sigma_1$:

$$H_1(\sigma_1) - \widehat{H}_1(\sigma_1) = C \left((\sigma_1 E - A)^{-1} B - V(\sigma_1 \widehat{E} - \widehat{A})^{-1} \widehat{B} \right) \\ + L_C V(\sigma_1 \widehat{E} - \widehat{A})^{-1} \widehat{B} - D_1. \tag{6.27}$$

Since $V(\sigma_1 \widehat{E} - \widehat{A})^{-1} \widehat{B} = V(\sigma_1 \widehat{E} - \widehat{A})^{-1} W^T (B - L_B)$ and from (6.26c),

$$L_A (\sigma_1 E - A)^{-1} B = L_B,$$

we obtain

$$V(\sigma_1 \widehat{E} - \widehat{A})^{-1} \widehat{B} = V(\sigma_1 \widehat{E} - \widehat{A})^{-1} W^T \left((\sigma_1 E - A) - L_A \right) (\sigma_1 E - A)^{-1} B.$$

Now, introducing an oblique projector $P_\sigma = V(\sigma \widehat{E} - \widehat{A})^{-1} W^T \left((\sigma E - A) - L_A \right)$ and utilizing $P_\sigma z = z$ for $z \in \text{range}(V)$, we thus get

$$V(\sigma_1 \widehat{E} - \widehat{A})^{-1} \widehat{B} = P_{\sigma_1} (\sigma_1 E - A)^{-1} B = (\sigma_1 E - A)^{-1} B. \tag{6.28}$$

Using the above relation in (6.27), we obtain

$$H_1(\sigma_1) - \widehat{H}_1(\sigma_1) = L_C(\sigma_1 E - A)^{-1}B - D_1.$$

From (6.26b), $L_C(\sigma_1 E - A)^{-1}B = D_1$; thus, $H_1(\sigma_1) = \widehat{H}_1(\sigma_1)$. Similarly, one can show that $H_1(\mu_1) = \widehat{H}_1(\mu_1)$. Now, we consider the second subsystem, which is:

$$\begin{aligned} H_2(\sigma_1,\sigma_2) - \widehat{H}_2(\sigma_1,\sigma_2) &= C(\sigma_2 E - A)^{-1}N(\sigma_1 E - A)^{-1}B \\ &\quad - \widehat{C}(\sigma_2\widehat{E} - \widehat{A})^{-1}\widehat{N}(\sigma_1\widehat{E} - \widehat{A})^{-1}\widehat{B} - D_2 \\ &= C\Big((\sigma_2 E - A)^{-1}N(\sigma_1 E - A)^{-1}B \\ &\quad - V(\sigma_2\widehat{E} - \widehat{A})^{-1}\widehat{N}(\sigma_1\widehat{E} - \widehat{A})^{-1}\widehat{B}\Big) \\ &\quad + L_C V(\sigma_2\widehat{E} - \widehat{A})^{-1}\widehat{N}(\sigma_1\widehat{E} - \widehat{A})^{-1}\widehat{B} - D_2. \end{aligned} \tag{6.29}$$

Hence,

$$\begin{aligned} V\underbrace{(\sigma_2\widehat{E}-\widehat{A})^{-1}\widehat{N}(\sigma_1\widehat{E}-\widehat{A})^{-1}\widehat{B}}_{\widehat{z}} &= V(\sigma_2\widehat{E}-\widehat{A})^{-1}W^T(N-L_N)V(\sigma_1\widehat{E}-\widehat{A})^{-1}\widehat{B} \\ &= V(\sigma_2\widehat{E} - \widehat{A})^{-1}W^T(N - L_N)(\sigma_1 E - A)^{-1}B, \end{aligned}$$

where the last equation follows from (6.28). Moreover, from (6.26c), it implies that

$$L_A(\sigma_2 E - A)^{-1}N(\sigma_1 E - A)^{-1}B = L_N(\sigma_1 E - A)^{-1}B.$$

Thus,

$$\begin{aligned} V\widehat{z} &= V(\sigma_2\widehat{E} - \widehat{A})^{-1}W^T((\sigma_2 E - A) - L_A)(\sigma_2 E - A)^{-1}N(\sigma_1 E - A)^{-1}B \\ &= P_{\sigma_2}(\sigma_2 E - A)N(\sigma_2 E - A)^{-1}B = (\sigma_2 E - A)N(\sigma_1 E - A)^{-1}B. \end{aligned}$$

Using the above relation in (6.29), we obtain

$$H_2(\sigma_1,\sigma_2) - \widehat{H}_2(\sigma_1,\sigma_2) = L_C(\sigma_2 E - A)^{-1}N(\sigma_1 E - A)^{-1}B - D_2.$$

Utilizing (6.26b), we have $L_C(\sigma_2 E - A)^{-1}N(\sigma_1 E - A)^{-1}B = D_2$. Thus, $H_2(\sigma_1,\sigma_2) = \widehat{H}_2(\sigma_1,\sigma_2)$. Similarly, we can prove that $H(\mu_1,\mu_2) = \widehat{H}(\mu_1,\mu_2)$. Analogously, we can also deal with higher subsystems and higher derivatives. □

6.3.3. Computational issues

In the previous section, we have shown how to construct a reduced-order system whose subsystems do not only interpolate those of the original systems but also retains their polynomial parts. However, the main bottleneck in computing reduced-order systems is that we require the computation of L_A, L_N, L_B and L_C to determine the intermediate matrices, e.g., $\widetilde{E}$, $\widetilde{A}$, which might be computationally expensive. Therefore, we next show how to determine a reduced-order system without computing these intermediate matrices.

Before we proceed further, we discuss the condition for the existence of a simultaneous solution of two given linear systems in the following lemma, which helps us in determining the condition for the existence of the solutions of (6.26a)–(6.26d).

Lemma 6.9:
Consider the matrices $\mathcal{A}_i, \mathcal{B}_i \in \mathbb{R}^{n\times m}$, $i \in \{1,2\}$ and $X \in \mathbb{R}^{n\times n}$, where $n \geq m$, satisfying the following two linear equations:

$$\mathcal{A}_1^T X = \mathcal{B}_1^T, \tag{6.30a}$$

$$X\mathcal{A}_2 = \mathcal{B}_2. \tag{6.30b}$$

If $\mathcal{A}_1^T\mathcal{B}_2 = \mathcal{B}_1^T\mathcal{A}_2$, then there exists an X that satisfies both (6.30a) and (6.30b), else it is not possible to determine an X, satisfying both (6.30a) and (6.30b) simultaneously. ♢

Proof. We first recall an important property of the Kronecker product and vectorization

$$\operatorname{vec}\left(\widetilde{X}\widetilde{Y}\widetilde{Z}\right) = (\widetilde{Z}^T \otimes \widetilde{X})\operatorname{vec}\left(\widetilde{Y}\right).$$

Using the vec $(\cdot)$ operation on both sides of (6.30a) and (6.30b) leads to

$$\begin{bmatrix} I_n \otimes \mathcal{A}_1^T \\ \mathcal{A}_2^T \otimes I_n \end{bmatrix} X_v = \begin{bmatrix} \operatorname{vec}\left(\mathcal{B}_1^T\right) \\ \operatorname{vec}\left(\mathcal{B}_2\right) \end{bmatrix}, \tag{6.31}$$

where $X_v := \operatorname{vec}(X)$. Next, we define a matrix

$$M = \begin{bmatrix} \mathcal{A}_2^T \otimes I_m & 0 \\ \mathcal{P} \otimes I_m & 0 \\ 0 & I_m \otimes \mathcal{A}_1^T \\ 0 & I_m \otimes \mathcal{Q} \end{bmatrix},$$

where $\mathcal{P}, \mathcal{Q} \in \mathbb{R}^{(n-m)\times n}$ such that the matrix M is invertible. Multiplying M on both sides of (6.31) yields

$$\underbrace{\begin{bmatrix} \mathcal{A}_2^T \otimes \mathcal{A}_1^T \\ \mathcal{P} \otimes \mathcal{A}_1^T \\ \mathcal{A}_2^T \otimes \mathcal{A}_1^T \\ \mathcal{A}_2^T \otimes \mathcal{Q} \end{bmatrix}}_{\mathcal{A}} X_v = \begin{bmatrix} (\mathcal{A}_2^T \otimes I_m)\operatorname{vec}\left(\mathcal{B}_1^T\right) \\ (\mathcal{P} \otimes I_m)\operatorname{vec}\left(\mathcal{B}_1^T\right) \\ (I_m \otimes \mathcal{A}_1^T)\operatorname{vec}\left(\mathcal{B}_2\right) \\ (I_m \otimes \mathcal{Q})\operatorname{vec}\left(\mathcal{B}_2\right) \end{bmatrix} = \underbrace{\begin{bmatrix} \operatorname{vec}\left(\mathcal{B}_1^T\mathcal{A}_2\right) \\ (\mathcal{P} \otimes I_m)\operatorname{vec}\left(\mathcal{B}_1^T\right) \\ \operatorname{vec}\left(\mathcal{A}_1^T\mathcal{B}_2\right) \\ (I_m \otimes \mathcal{Q})\operatorname{vec}\left(\mathcal{B}_2\right) \end{bmatrix}}_{\mathcal{B}} \tag{6.32}$$

Now, by using the Kronecker-Capelli theorem [119], the linear system (6.32) has a solution if and only if:

$$\mathsf{rank}\mathcal{A} = \mathsf{rank}[\mathcal{A}, \mathcal{B}]. \tag{6.33}$$

Clearly, the first and third row blocks of the matrix $\mathcal{A}$ are the same. Therefore, $\mathcal{P}$ and $\mathcal{Q}$ can be chosen such that the rank of the matrix $\mathcal{A}$ is equal to $m(2n - m)$ which is equal to the number of rows, having removed the third row block of the matrix $\mathcal{A}$. In order to have the same rank for the matrix $[\mathcal{A}, \mathcal{B}]$, the first and the third row blocks of the matrix $\mathcal{B}$ should also be the same. This leads to the following condition:

$$\mathcal{B}_1^T \mathcal{A}_2 = \mathcal{A}_1^T \mathcal{B}_2. \tag{6.34}$$

One can verify that if the above condition (6.34) is fulfilled, then $\mathsf{rank}[\mathcal{A}, \mathcal{B}]$ is also equal to $m(2n - m)$. This means that the system (6.32) has a solution if $m \leq n$. On the other hand, if the condition (6.34) is not satisfied, then $\mathsf{rank}[\mathcal{A}, \mathcal{B}]$ is also equal to $m(2n - m) + 1$, implying that the system (6.32) does not have any solution. Hence, it is not possible to determine an X that satisfies both (6.30a) and (6.30b). □

Coming back to the computational issues related to the solutions of (6.26a)–(6.26d), first note that L_B and L_C are independent of other unknowns, e.g., L_A and L_N; therefore, they can be easily computed. However, the main issue lies in the computation of L_A and L_N. These matrices require the simultaneous solution of (6.26c) and (6.26d) for given L_B and L_C. Next, using Lemma 6.9, we derive a necessary and sufficient condition, called the *compatibility condition* to ensure the existence of the simultaneous solution. This follows by equating the right-hand sides of (6.26c) and (6.26d) after pre-multiplying by W^T and post-multiplying by V, respectively:

$$\begin{aligned} &W^T \left[L_B(e_1^q)^T, L_N \left[V^{(1)}(I_q \otimes (e_1^q)^T), \ldots, V^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T))\right]\right] \\ &\qquad = \left[L_C^T(e_1^q)^T, L_N^T \left[W^{(1)}(I_q \otimes (e_1^q)^T), \ldots, W^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T)\right]\right]^T V. \end{aligned} \tag{6.35}$$

The following theorem suggests a choice of L_N, guaranteeing that the above compatibility condition is satisfied.

Theorem 6.10:
Let the projection matrices V and W be defined as in Theorem 6.8 and assume L_B and L_C fulfill the conditions (6.26a) and (6.26b), respectively. Moreover, let $L_N \in \mathbb{R}^{n \times n}$ satisfy

$$W^T L_N V = \mathcal{T} \begin{bmatrix} D_2 & \cdots & D_{k+1} \\ \vdots & \ddots & \vdots \\ D_{k+1} & \cdots & D_{2k} \end{bmatrix} \mathcal{T}^T, \tag{6.36}$$

where $\mathcal{T} = \sum\limits_{i=0}^{k-1} (e_{1+q^i}^{\widehat{n}}) \otimes \left(e_{i+1}^k\right)^T$, and $\widehat{n} = q + \cdots + q^k$ is the order of the reduced-order system, and D_j is the polynomial part of the jth subsystem. In other words, L_N

satisfies

$$\left(W^{(l)}\right)^T L_N \left(V^{(m)}\right)^T = e_1^{q^l} D_{m+l+1} \left(e_1^{q^m}\right)^T, \quad \text{for } \{l, m\} \in \{1, \dots, k\}. \tag{6.37}$$

Then, the compatibility condition (6.35) is satisfied. ◊

Proof. Consider the first row of the block matrix given in (6.35):

$$\left(W^{(1)}\right)^T \left[L_B (e_1^q)^T, L_N \left[V^{(1)}(I_q \otimes (e_1^q)^T), \dots, V^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T)\right]\right] = e_1^q L_C V.$$

To show that the above equation holds, we use (6.26a) and (6.36):

$$\begin{aligned}
&\left(W^{(1)}\right)^T \left[L_B (e_1^q)^T, L_N \left[V^{(1)}(I_q \otimes (e_1^q)^T), \dots, V^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T)\right]\right] \\
&\quad = [e_1^q D_1 (e_1^q)^T, e_1^q D_2 (e_1^q)^T (I_q \otimes (e_1^q)^T), \dots, e_1^q D_k (e_1^{q^{k-1}})^T (I_{q^{k-1}} \otimes (e_1^q)^T)] \\
&\quad = e_1^q [D_1 (e_1^q)^T, D_2 (e_1^q)^T \otimes (e_1^q)^T, \dots, D_k (e_1^{q^{k-1}})^T \otimes (e_1^q)^T] \\
&\quad = e_1^q [D_1 (e_1^q)^T, D_2 (e_1^{q^2})^T, \dots, D_k (e_1^{q^k})^T] \\
&\quad = e_1^q L_C V,
\end{aligned}$$

where the last equality follows from (6.26b). Now, we consider the ith row of the block matrix in (6.35). That is,

$$\begin{aligned}
&\left(W^{(i)}\right)^T \left[L_B (e_1^q)^T, L_N \left[V^{(1)}(I_q \otimes (e_1^q)^T), \dots, V^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T))\right]\right] \\
&\qquad = \left[L_N^T W^{(i-1)} (I_{q^{i-1}} \otimes (e_1^q)^T)\right]^T V.
\end{aligned}$$

To prove the above relation, we again make use of (6.26a) and (6.36). Thus, we obtain

$$\begin{aligned}
&\left(W^{(i)}\right)^T \left[L_B (e_1^q)^T, L_N \left[V^{(1)}(I_q \otimes (e_1^q)^T), \dots, V^{(k-1)}(I_{q^{k-1}} \otimes (e_1^q)^T)\right]\right] \\
&\qquad = \Big[e_1^{q^i} D_i (e_1^q)^T, e_1^{q^i} D_{i+1} (e_1^q)^T (I_q \otimes (e_1^q)^T), \dots, \\
&\qquad\qquad\qquad e_1^{q^i} D_{i+k-1} (e_1^{q^{k-1}})^T (I_{q^{k-1}} \otimes (e_1^q)^T)\Big] \\
&\qquad = (I_{q^{i-1}} \otimes (e_1^q)^T)^T e_1^{q^{i-1}} \left[D_i (e_1^q)^T, D_{i+1} (e_1^{q^2})^T, \dots, D_{i+k-1} (e_1^{q^k})^T\right] =: \mathcal{R}.
\end{aligned}$$

Using the condition on L_N given in (6.36) or in (6.37), we obtain

$$\begin{aligned}
\mathcal{R} &= (I_{q^{i-1}} \otimes (e_1^q)^T)^T \left[(W^{(i-1)})^T L_N V^{(1)}, (W^{(i-1)})^T L_N V^{(2)}, \dots, (W^{(i-1)})^T L_N V^k\right] \\
&= (I_{q^{i-1}} \otimes (e_1^q)^T)^T \left[(W^{(i-1)})^T L_N V^{(1)}, (W^{i-1})^T L_N V^{(2)}, \dots, (W^{(i-1)})^T L_N V^{(k)}\right] \\
&= \left[L_N^T W^{(i-1)} (I_{q^{i-1}} \otimes (e_1^q)^T)^T\right]^T V.
\end{aligned}$$

This means that each row of the block matrix corresponding to the left and right-hand sides of the compatibility condition given in (6.35) are equal. Therefore, if L_N is chosen to satisfy the assumption (6.36), then it is ensured that (6.35) holds. □

Algorithm 6.2: Subsystem interpolation MOR for bilinear DAEs, having index-1 matrix pencil $\lambda E - A$.

```
**Input:** $E$, $A$, $N$, $B$, $C$, $[\sigma_1, \cdots, \sigma_k]$, $[\mu_1, \cdots, \mu_k]$, $q$.
**Output:** $\widehat{E}$, $\widehat{A}$, $\widehat{N}$, $\widehat{B}$, $\widehat{C}$.
Construct $V$ and $W$ according to Theorem 6.8.
Compute the polynomial part of the $k$th-order subsystem:
    $D_k = C(MN)^{k-1}MB$.
Identify the expression of $W^T L_B$, $L_C V$ , $W^T L_A V$ and $W^T L_N V$ as:
    $W^T L_B = \left[D_1(e_1^q)^T, D_2(e_1^{q^2})^T, \dots, D_k(e_1^{q^k})^T\right]^T =: R_B,$
    $W^T L_B = \left[D_1(e_1^q)^T, D_2(e_1^{q^2})^T, \dots, D_k(e_1^{q^k})^T\right] =: R_C,$
    $W^T L_N V = \mathfrak{T} \begin{bmatrix} D_2 & \cdots & D_{k+1} \\ \vdots & \ddots & \vdots \\ D_{k+1} & \cdots & D_{2k} \end{bmatrix} \mathfrak{T}^T =: R_N,$
    where $\mathfrak{T} = \sum_{i=0}^{k-1} e_{1+q^i}^{\widehat{n}} \otimes \left(e_{i+1}^k\right)^T$, $\widehat{n} = \sum_{i=1}^k q^i$, and
    $W^T L_A V = \big[R_B(e_1^q)^T, R_N(:, 1:q)(I_q \otimes (e_1^q)^T), \dots,$
        $R_N(:, q + \cdots + q^{k-1} + (1:q^k))(I_{q^{k-1}} \otimes (e_1^q)^T)\big] =: R_A.$
Compute the reduced model as:
    $\widehat{E} = W^T E V$,    $\widehat{A} = W^T A V + R_A$,    $\widehat{N} = W^T N V - R_N$,
    $\widehat{B} = W^T B - R_B$,    $\widehat{C} = CV - R_C$.
```

Remark 6.11:
It is interesting to see that to compute a reduced-order system, we do not need to compute explicitly the matrices L_A, L_N, L_B and L_C. We rather require the expressions for $W^T L_B$, $L_C V$, $W^T L_A V$ and $W^T L_N V$. One can substitute $W^T L_B$ and $L_C V$ directly from (6.26a) and (6.26b). The expression of $W^T L_N V$ can be easily identified by using (6.36). Similarly, one can obtain the expression of $W^T L_A V$ without explicitly computing L_A by pre-multiplying (6.26c) by W^T and using (6.36) and (6.26a). ◊

Now, we summarize the complete methodology of computing a reduced-order system for the system (6.17) in Algorithm 6.2.

Remark 6.12:
As shown in [40], a two-sided projection method might lead to much better approximation, since more multi-moments are matched for higher order subsystems. The same holds for the proposed modified Krylov subspace technique for the structured bilinear DAEs as well. To see this, we consider an example similar to the one used

in [40]. Let us assume the projection subspaces V and W, depending on the first two subsystems are as follows:

$$\begin{aligned}\operatorname{span}(V) &= \operatorname{span}\big\{A^{-1}B, \dots, (A^{-1}E)^5A^{-1}B, A^{-1}NA^{-1}B, A^{-1}N(A^{-1}E)A^{-1}B\big\}\,,\\ \operatorname{span}(W^T) &= \operatorname{span}\big\{CA^{-1}, \dots, C(A^{-1}E)^5A^{-1}, CA^{-1}NA^{-1}, C(A^{-1}E)A^{-1}NA^{-1}\big\}\,.\end{aligned}$$

According to Theorem 6.8, the reduced-order system preserves 12 multi-moments of the first subsystem

$$C(A^{-1}E)^{l_1}A^{-1}B + D\delta(l_1) = \widehat{C}^T(\widehat{A}^{-1}\widehat{E})^{l_1}\widehat{A}^{-1}\widehat{B} + D_1\delta(l_1),$$

where $l_1 = 0, \dots, 11$. For the second subsystem, 29 multi-moments are matched

$$\begin{aligned}C(A^{-1}E)^{l_2}A^{-1}N(A^{-1}E)^{l_1}A^{-1}B = \widehat{C}^T&(\widehat{A}^{-1}\widehat{E})^{l_2}\widehat{A}^{-1}\widehat{N}(\widehat{A}^{-1}\widehat{E})^{l_1}\widehat{A}^{-1}\widehat{B}\\ &+ D_2\delta(l_1)\delta(l_2),\end{aligned}$$

where $l_1, l_2 = 0, 1, \dots, 5$ or $l_1 = 6,\ l_2 = 0, 1$ and $l_1 = 0, 1,\ l_2 = 6$. For the third subsystem, 37 multi-moments are matched

$$\begin{aligned}&C(A^{-1}E)^{l_3}A^{-1}N\cdots N(A^{-1}E)^{l_1}A^{-1}B\\ &\qquad= \widehat{C}^T(\widehat{A}^{-1}\widehat{E})^{l_3}\widehat{A}^{-1}\widehat{N}\cdots\widehat{N}(\widehat{A}^{-1}\widehat{E})^{l_1}\widehat{A}^{-1}\widehat{B} + D_3\delta(l_1)\delta(l_2)\delta(l_3),\end{aligned}$$

where $l_1 = 0, 1, \dots, 5, l_2 = 0, l_3 = 0, 1$ or $l_1 = 0, 1, l_2 = 0, l_3 = 2, 3, 4, 5$ or $l_1 = 0, 1, l_2 = 1, l_3 = 0, 1$. For the fourth subsystem, 4 multi-moments are matched

$$\begin{aligned}&C(A^{-1}E)^{l_4}A^{-1}N\cdots N(A^{-1}E)^{l_1}A^{-1}B\\ &\qquad= \widehat{C}(\widehat{A}^{-1}\widehat{E})^{l_4}\widehat{A}^{-1}\widehat{N}\cdots\widehat{N}(\widehat{A}^{-1}\widehat{E})^{l_1}\widehat{A}^{-1}\widehat{B} + D_4\delta(l_1)\delta(l_2)\delta(l_3)\delta(l_4),\end{aligned}$$

where $l_1 = 0, 1, l_2 = 0, l_3 = 0, l_4 = 0, 1$. ◊

Remark 6.13:
There are some scenarios, where the polynomial part of higher subsystems of bilinear DAEs are zero. For instance, if the structures of E and N in (6.17) are either as:

$$E = \begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad N = \begin{bmatrix} N_{11} & N_{12} \\ 0 & 0 \end{bmatrix}, \tag{6.38}$$

or as:

$$E = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & 0 \end{bmatrix}, \tag{6.39}$$

where $E_{11}, N_{11} \in \mathbb{R}^{n_1\times n_1}$ and $E_{12}, N_{12}, N_{21}^T \in \mathbb{R}^{n_1\times n_2}$, then $D_k = 0$ for $k > 1$ and $\mathcal{P} = D_1 = CMB$, where M is defined as in (6.20). This can simplify Algorithm 6.2. For example, in these cases, L_N would be zero; thus $R_N = 0$ in Algorithm 6.2, and $R_B = e_1^{\widehat{n}}D_1^T$, $R_C = D_1(e_1^{\widehat{n}})^T$ and $R_A = e_1^{\widehat{n}}D_1(e_1^{\widehat{n}})^T$, where $\widehat{n} = q + \cdots + q^k$. ◊

6.3.4. Time-domain representation of the reduced-order system

Till now, we have shown how to achieve interpolation for the leading k multi-variate transfer functions of the original and reduced-order systems along with matching their polynomial parts. However, our interest lies in determining a time-domain bilinear system, whose kth-order multi-variate transfer function is given by (6.25). Therefore, in this subsection, we derive the time-domain representation of a reduced bilinear system whose kth-order multi-variate transfer function is of the form given in (6.25). The following theorem summarizes our results.

Theorem 6.14:
Given a bilinear system, whose kth order multi-variate transfer function has the form given in (6.25). Then, the time-domain representation of this bilinear system can be written as

$$\begin{aligned} \widehat{E}\dot{\widehat{x}}(t) &= \widehat{A}\widehat{x}(t) + \widehat{N}\widehat{x}(t)u(t) + \widehat{B}u(t), \\ \widehat{y}(t) &= \widehat{C}\widehat{x}(t) + \sum_{k=1}^{\infty} D_k u^k(t). \end{aligned} \tag{6.40}$$

◊

Proof. We begin with the kth-order multi-variate transfer function

$$\begin{aligned} \widehat{H}_k(s_1,\dots,s_k) = &\quad \widehat{C}(s_k\widehat{E}-\widehat{A})^{-1}\widehat{N}(s_{k-1}\widehat{E}-\widehat{A})^{-1}\widehat{N}\cdots\widehat{N}(s_1\widehat{E}-\widehat{A})^{-1}\widehat{B} + D_k \\ = &\quad \widetilde{C}(s_kI_{\widehat{n}}-\widetilde{A})^{-1}\widetilde{N}(s_{k-1}I_{\widehat{n}}-\widetilde{A})^{-1}\widetilde{N}\cdots\widetilde{N}(s_1I_{\widehat{n}}-\widetilde{A})^{-1}\widetilde{B} + D_k, \end{aligned} \tag{6.41}$$

where

$$\widetilde{A} = \widehat{E}^{-1}\widehat{A}, \quad \widetilde{N} = \widehat{E}^{-1}\widehat{N}, \quad \widetilde{B} = \widehat{E}^{-1}\widehat{B} \quad \text{and} \quad \widetilde{C} = \widehat{C}. \tag{6.42}$$

By utilizing the multi-variate inverse Laplace transform on (6.41), we obtain the regular Volterra kernel as:

$$h_k(t_1,t_2,\dots,t_k) = \widetilde{C}e^{\widetilde{A}t_k}\widetilde{N}e^{\widetilde{A}t_{k-1}}\widetilde{N}\cdots\widetilde{N}e^{\widetilde{A}t_1}\widetilde{B} + D_k\delta(t_k)\delta(t_{k-1})\cdots\delta(t_1). \tag{6.43}$$

As discussed in [111], the output $\widehat{y}(t)$ of a nonlinear system can be described in terms of the Volterra kernel $h_k(t_1,t_2,\dots,t_k)$ and input $u(t)$ as follows:

$$\widehat{y}(t) = \sum_{k=1}^{\infty}\int_0^{t_1}\int_0^{t_2}\cdots\int_0^{t_k} h_k(t_1,t_2,\dots,t_k)u(t-\sum_{i=1}^{k}t_i)\cdots u(t-t_k)dt_k\cdots dt_1.$$

Substituting (6.43) in the above equation, we can write

$$\widehat{y}(t) = \widehat{y}^{(1)}(t) + \widehat{y}^{(2)}(t),$$

where

$$\widehat{y}^{(1)}(t)=\sum_{k=1}^{\infty}\int_0^{t_1}\int_0^{t_2}\cdots\int_0^{t_k}\widetilde{C}e^{\widetilde{A}t_k}\widetilde{N}\cdots\widetilde{N}e^{\widetilde{A}t_2}\widetilde{N}e^{\widetilde{A}t_1}\widetilde{B}u(t-\sum_{i=1}^{k}t_i)\cdots u(t-t_k)dt_k\cdots dt_1,$$
$$\widehat{y}^{(2)}(t)=\sum_{k=1}^{\infty}\int_0^{t_1}\int_0^{t_2}\cdots\int_0^{t_k}D_k\delta(t_k)\delta(t_{k-1})\cdots\delta(t_1)u(t-\sum_{i=1}^{k}t_i)\cdots u(t-t_k)dt_k\cdots dt_1.$$

The response $\widehat{y}^{(1)}(t)$ is simply the Volterra series representation of a bilinear ODE system with zero initial condition [111]. This means that corresponding to $\widehat{y}^{(1)}(t)$, we have

$$\begin{aligned}\dot{\widehat{x}}(t) &= \widetilde{A}\widehat{x}(t)+\widetilde{N}\widehat{x}(t)u(t)+\widetilde{B}u(t),\\ \widehat{y}^{(1)}(t) &= \widetilde{C}\widehat{x}(t), \qquad \widehat{x}(0)=0.\end{aligned} \tag{6.44}$$

For $\widehat{y}^{(2)}(t)$, we use the properties of the Dirac delta function [46] which leads to

$$\widehat{y}^{(2)}(t)=\sum_{k=1}^{\infty}D_k u(t)\cdots u(t)=\sum_{k=1}^{\infty}D_k\left(u(t)\right)^k.$$

By combining the responses $\widehat{y}^{(1)}(t)$ and $\widehat{y}^{(2)}(t)$ and substituting the expression for $\widetilde{A},\widetilde{N},\widetilde{B}$ and $\widetilde{C}$ from (6.42), we obtain a bilinear system as in (6.40), and this proves the theorem. □

Since the output equation in (6.40) contains the sum of an input dependent infinite series, we need to compute the summation at each time step. This increases the computational cost, which may destroy the effect of the model reduction procedure. In the following, we discuss some cases, where this infinite summation can be computed cheaply.

Case 1: As noted in Remark 6.13, if the matrices E and N in (6.17) have special structures, then the polynomial part of the first subsystem is non-zero and all others have zero polynomial parts, i.e., $D_k = 0$ for $k \geq 2$ and $D_1 \neq 0$. Thus, $\sum\limits_{k=1}^{\infty} D_k u^k(t)$ reduces to $D_1u(t)$, which is computationally cheap.

Case 2: There are some applications where the input $u(t)$ can be considered constant or unity ($u(t) = \alpha$ or $u(t) = 1$). These scenarios may appear, for example in the parameter varying systems [20]. In such a case

$$\sum_{k=1}^{\infty} D_k u^k(t) = D_1\alpha + D_2\alpha^2 + D_3\alpha^3\cdots.$$

Substituting the expression of D_k from Lemma 6.7 in the above equation, we get

$$\begin{aligned}\sum_{k=1}^{\infty} D_k u^k(t) &= (CMB+D)\alpha + C(MN)MB\alpha^2 + C(MN)^2MB\alpha^3 + \cdots \\ &= \alpha C(I + \alpha MN + \alpha^2(MN)^2 + \cdots)MB + \alpha D.\end{aligned}$$

Now, if we assume $\|\alpha MN\|_2 < 1$, then we have

$$\sum_{k=1}^{\infty} D_k u^k(t) = (C(I-\alpha MN)^{-1}MB + D)\alpha.$$

Thus, we can identify an expression of the convergent series for constant inputs.

Case 3: In this case, we assume convergence for $\|D_k\|$, i.e., $\sum\limits_{k=j+1}^{\infty} \|D_k\| < \tau \ll 1$. Then, for bounded inputs, we can truncate the infinite summation after the jth term. That is

$$\sum_{k=1}^{\infty} D_k u^k(t) \approx \sum_{k=1}^{j} D_k u^k(t).$$

Thus, we can save the computations associated with $\sum\limits_{k=j+1}^{\infty} D_k u^k(t)$.

6.3.5. Interpolation of multi-input multi-output bilinear DAEs

Thus far, we have concentrated on an interpolation-based model reduction problem for SISO bilinear DAEs for simplicity of notation. However, it can be extended to MIMO bilinear systems, but the notation becomes much more difficult to handle. Therefore, we consider interpolation of the first 2 subsystems only in order to give a glimpse of how the proposed methodology can be applied to MIMO bilinear DAEs, but nonetheless one can consider interpolation of the leading first k subsystem as well. We consider a MIMO bilinear system as follows:

$$\begin{aligned}E_{11}\dot{x}_1(t) + E_{12}\dot{x}_2(t) &= A_{11}x_1(t) + A_{12}x_2(t) + \sum_{k=1}^{m}\left(N_{11}^{(k)}x_1(t) + N_{12}^{(k)}x_2(t)\right)u_k(t) + B_1u(t), \\ 0 &= A_{21}x_1(t) + A_{22}x_2(t) + \sum_{k=1}^{m}\left(N_{21}^{(k)}x_1(t) + N_{22}^{(k)}x_2(t)\right)u_k(t) + B_2u(t), \\ y(t) &= C_1x_1(t) + C_2x_2(t) + Du(t),\end{aligned} \tag{6.45}$$

where $x_1(t) \in \mathbb{R}^{n_1}$, $x_2(t) \in \mathbb{R}^{n_2}$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are respectively the state, input, and output vectors. The number p and m denotes the number of outputs and inputs, respectively. Moreover, the matrix pencil $\lambda E - A$ is of index-1, and $u_k(t)$ denotes the kth component of $u(t)$. Thus, the leading four subsystems of (6.45) can be given as follows:

$$\begin{aligned}
H_1(\mathcal{S}_1) &= C(s_1E - A)^1 B,\\
H_2(\mathcal{S}_2) &= \left[H_2^{(1)}(\mathcal{S}_2),\ \ldots,\ H_2^{(m)}(\mathcal{S}_2)\right],\\
H_3(\mathcal{S}_3) &= \left[H_3^{(1,1)}(\mathcal{S}_3),\ \ldots,\ H_3^{(1,m)}(\mathcal{S}_3),\ H_3^{(2,1)}(\mathcal{S}_3),\ \ldots,\ H_3^{(m,m)}(\mathcal{S}_3)\right],\\
H_4(\mathcal{S}_4) &= \left[H_4^{(1,1,1)}(\mathcal{S}_4),\ \ldots,\ H_4^{(1,1,m)}(\mathcal{S}_4),\ H_4^{(2,1,1)}(\mathcal{S}_4),\ \ldots,\ H_4^{(m,m,m)}(\mathcal{S}_4)\right],
\end{aligned}$$

where

$$E = \begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix},\ A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},\ N_k = \begin{bmatrix} N_{11}^{(k)} & N_{12}^{(k)} \\ N_{21}^{(k)} & N_{22}^{(k)} \end{bmatrix},\ B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}.$$

and

$$H_i^{(l_1,\ldots,l_{i-1})}(\mathcal{S}_i) = C\phi(s_i)N_{l_{i-1}}\phi(s_{i-1})\cdots N_{l_1}\phi(s_1)B \tag{6.46}$$

with $\phi(s_i) := (s_iE - A)^{-1}$. Moreover, we denote the polynomial parts of $H_i^{(l_1,\ldots,l_{i-1})}(\mathcal{S}_i)$ by $D_i^{(l_1,\ldots,l_{i-1})}$, which can be given as follows by using Lemma 6.7:

$$\begin{aligned}
D_1 &= CMB + D,\\
D_i^{(l_1,\ldots,l_{i-1})} &= CMN_{l_{i-1}}M\cdots N_{l_1}MB,
\end{aligned} \tag{6.47}$$

where M is the same as defined in (6.20). Furthermore, we need to use a more general nested structure to determine the projection matrices. For this, we assume arbitrary interpolation points $\sigma_i, \mu_i \in \mathbb{C}$ such that $sE - A$ and $s\widehat{E} - \widehat{A}$ are invertible for $s = \sigma_i, \mu_i$, and define the projection matrices V and W as follows:

$$\begin{aligned}
\text{range}\left(V^{(1)}\right) &= \mathcal{K}_\alpha\left((\sigma_1E - A)^{-1}E, (\sigma_1E - A)^{-1}B\right),\\
\text{range}\left(V_k^{(2)}\right) &= \mathcal{K}_\alpha\left((\sigma_iE - A)^{-1}E, (\sigma_iE - A)^{-1}N_kV^{(1)}\right),\quad k \in \{1,\ldots,m\},\\
\text{range}\left(W^{(1)}\right) &= \mathcal{K}_\beta\left((\mu_1E - A)^{-T}E^T, (\mu_1E - A)^{-T}C^T\right),\\
\text{range}\left(W_k^{(2)}\right) &= \mathcal{K}_\alpha\left((\mu_iE - A)^{-T}E^T, (\mu_iE - A)^{-T}N_k^TW^{(1)}\right),\quad k \in \{1,\ldots,m\},
\end{aligned}$$

$$\begin{aligned}
\text{range}\,(V) &= \text{range}\left(V^{(1)}\right) + \bigcup_{k=1}^{m}\left\{\text{range}\left(V_k^{(2)}\right)\right\},\\
\text{range}\,(W) &= \text{range}\left(W^{(1)}\right) + \bigcup_{k=1}^{m}\left\{\text{range}\left(W_k^{(2)}\right)\right\}.
\end{aligned}$$

In order to ensure the same number of columns in V and W, we choose α and β such that $m\alpha = p\beta$, where p and m are the numbers of outputs and inputs, respectively. Next, we consider $L_A, L_{N^{(k)}}, L_B$ and L_C which are solutions to the following set of equations:

$$\begin{aligned}
W^T L_B &= \left[(e_1^{\alpha})^T \otimes D_1, (e_1^{\alpha^2})^T \otimes D_2^{(1)}, \ldots, (e_1^{\alpha^2})^T \otimes D_2^{(m)}\right]^T, \\
L_C V \ &= \left[D_1 \otimes (e_1^{\beta})^T, D_2^{(1)} \otimes (e_1^{\beta^2})^T, \ldots, D_2^{(m)} \otimes (e_1^{\beta^2})^T\right], \\
L_A V \ &= \left[L_B \otimes (e_1^{\alpha})^T, L_{N^{(1)}} V^{(1)}(I_{m\alpha} \otimes (e_1^{\alpha})^T), \ldots, L_{N^{(m)}} V^{(1)}(I_{m\alpha} \otimes (e_1^{\alpha})^T)\right], \\
W^T L_A &= \left[L_C^T \otimes (e_1^{\beta})^T, L_{N^{(1)}}^T W^{(1)}(I_{p\beta} \otimes (e_1^{\alpha})^T), \ldots, L_{N^{(m)}}^T W^{(1)}(I_{p\beta} \otimes (e^{\alpha})^T)\right]^T.
\end{aligned}$$

Then, the reduced-order system can be determined as follows:

$$\begin{aligned}
&\widehat{E} = W^T E V, && \widehat{A} = W^T(A + L_A)V, && \widehat{N}_k = W^T(N_k - L_{N^{(k)}})V, \\
&\widehat{B} = W^T(B - L_B), && \widehat{C} = (C - L_C)V.
\end{aligned}$$

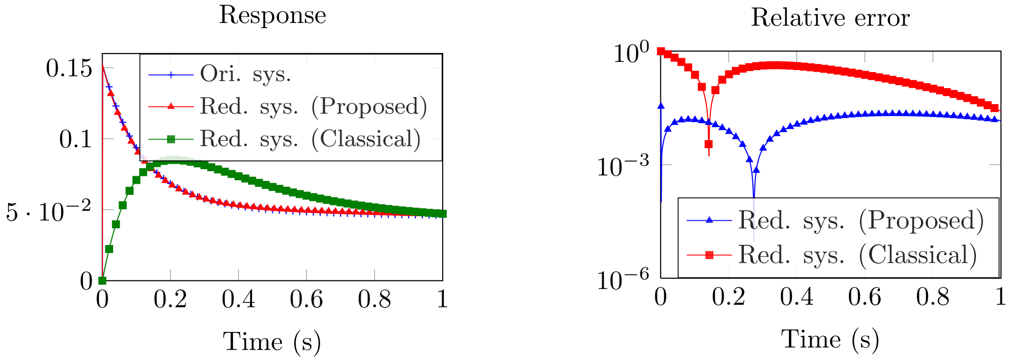
Similar to SISO bilinear systems, the explicit computation of the matrices L_A, $L_{N^{(i)}}$, L_B and L_C can also be avoided in order to determine reduced-order systems, which can be done analogously as shown for the SISO case.

6.3.6. Numerical experiments

We present numerical results for model reduction of the structured bilinear DAE systems using different approaches. The reduced-order system can be computed either by direct implementation of Theorem 6.2, without matching the polynomial part in the reduced-order system (classical interpolatory technique) or by our proposed methodology which achieves the matching of the polynomial part in addition to interpolation. All the numerical results were simulated in MATLAB Version 8.0.0.783(R2012b) 64-bit (glnza64) on Intel® Core™2 Quad CPU Q9550 @ 2.83GHz 6 MB cache, 4GB RAM, openSUSE Linux 12.04.

An artificial example

The bilinear DAE system, that is to be reduced, is generated randomly of order $n = 100$ and with partitioning $n_1 = 90,\ n_2 = 10$. It is ensured that the matrix pencil $\lambda E - A$ is of index-1. The polynomial parts of the first 4 subsystems of the bilinear system are $D_1 = 0.1472, D_2 = 5{\cdot}10^{-3}, D_3 = 1.92{\cdot}10^{-4}, D_4 = 7.35{\cdot}10^{-6}$, where D_k is the polynomial part of the kth subsystem. The interpolation points are selected as $\sigma = \mu = [0, 0.5]$ with multiplicity $q = 1$ resulting in a reduced-order system of order $\widehat{n} = 4$. We truncate the infinite summation in Theorem 6.14 after 4 terms since $\|D_i\|$ decreases exponentially.



(a) Transient response.

(b) Relative error.

Figure 6.1.: An artificial example of index-1 bilinear system: comparison of the original and reduced-order systems for an input $u(t) = e^{-10t}$.

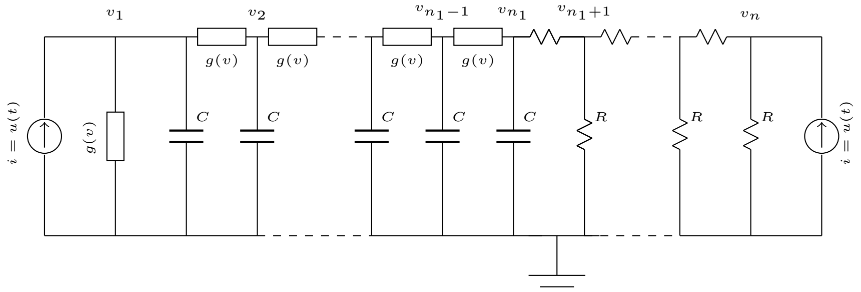


Figure 6.2.: Nonlinear transmission line circuit, having index-1 matrix pencil $\lambda E - A$.

We compute the reduced-order systems by using the classical interpolation technique and the proposed methodology, having the same interpolation points and multiplicities. The time-domain responses of the actual and the reduced bilinear systems, obtained by using the implicit Euler method, are shown in Figure 6.1a for an exponential input. The relative errors associated with the two approaches are shown in Figure 6.1b.

Certainly, the reduced-order system obtained from the direct implementation shows completely different dynamics whereas the proposed methodology captures the dynamics of the original system well.

Nonlinear RC circuit

As a second example, we consider a nonlinear RC circuit that represents a modified form of the transmission line circuit proposed in [78]. The circuit includes resistors, capacitors and diodes as shown in Figure 6.2.

All the resistances and capacities are set to 1, and all the diodes ensure $i_D = e^{40v_D} + v_D - 1$, where i_D represents the current and v_D is the voltage across the diodes. The input $u(t)$ is the current source i and the output $y(t)$ represents the average voltage over all nodes ranging from 1 to n. Using Kirchhoff's current law at each node, we have

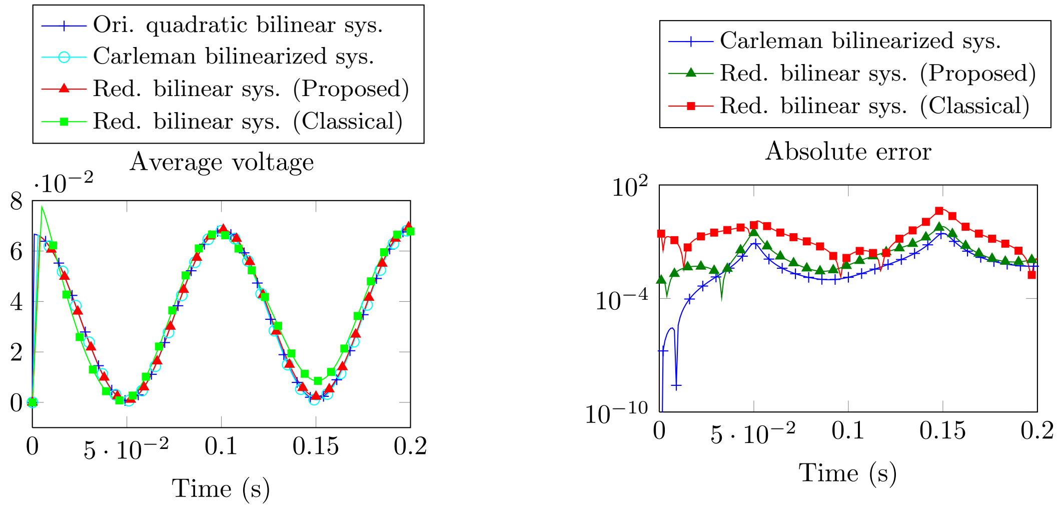
$$\begin{aligned}
\dot{v}_1 &= -2v_1 + v_2 + 2 - e^{40v_1} - e^{40(v_1 - v_2)} + u(t), \\
\dot{v}_k &= -2v_k + v_{k-1} + v_{k+1} + e^{40(v_{k-1} - v_k)} - e^{40(v_k - v_{k+1})}, \quad (2 \leq k \leq n_1 - 1) \\
\dot{v}_{n_1} &= -2v_{n_1} + v_{n_1-1} + v_{n_1+1} - 1 + e^{40(v_{n_1-1} - v_{n_1})}, \\
0 &= 3v_k - v_{k-1} - v_{k+1}, \qquad\qquad\qquad\qquad\qquad (n_1 + 1 \leq k \leq n - 1) \\
0 &= -2v_n + v_{n-1} + u(t).
\end{aligned}$$

In order to represent the above nonlinear system as a quadratic-bilinear system, we set v_1 to $v_{k,k+1}$ $(v_{k,k+1} = v_k - v_{k+1})$, $k\ in\{1, \ldots, n_1 - 1\}$, and v_{n_1+1} to v_n as the state variables, and perform some changes of variables by defining $y_1 = e^{40v_1} - 1$ and $y_k = e^{40(v_{k-1,k})} - 1,\ 2 \leq k \leq n_1$. Together with the differential equations of all y_k, one gets the following set of equations:

$$\begin{aligned}
\dot{v}_1 &= -v_1 - v_{1,2} - y_1 - y_2 + u(t), \\
\dot{v}_{1,2} &= -v_1 - 2v_{1,2} + v_{2,3} - y_1 - 2y_2 + y_3 + u(t), \\
\dot{v}_{k,k+1} &= -2v_{k,k+1} + v_{k-1,k} + v_{k+1,k+2} + y_k - 2y_{k+1} + y_{k+2}, \quad (2 \leq k \leq n_1 - 2) \\
\dot{v}_{n_1-1,n_1} &= -2v_{n_1-1,n_1} + v_{n_1-2,n_1-1} + v_{n_1} - v_{n_1+1} + y_{n_1-1} - 2y_{n_1}, \\
0 &= 3v_k - v_{k-1} - v_{k+1}, \qquad\qquad\qquad\qquad (n_1 + 1 \leq k \leq n - 1) \\
0 &= -2v_n + v_{n-1} + u(t), \\
\dot{y}_1 &= 40(y_1 + 1)(-v_1 - v_{1,2} - y_1 - y_2 + u(t)), \\
\dot{y}_2 &= 40(y_2 + 1)(-v_1 - 2v_{1,2} + v_{2,3} - y_1 - 2y_2 + y_3 + u(t)), \\
\dot{y}_k &= 40(y_k + 1)(-2v_{k-1,k} + v_{k-2,k-1} + v_{k,k+1} + y_{k-1} - 2y_k + y_{k+1}), \\
\dot{y_{n_1}} &= 40(y_{n_1} + 1)(-2v_{n_1-1,l_1} + v_{n_1-2,n_1-1} + v_{n_1} - v_{n_1-1} + y_{n_1-1} - 2y_{n_1}).
\end{aligned}$$

In the above set of equations, we fixed v_{n_1} to $v_1 - \sum_{k=2}^{n_1} v_{k-1,k}$. This means that the circuit can be modelled by a quadratic-bilinear descriptor system of order $\widetilde{n} = n_1 + n$, having an index-1 matrix pencil associated with the quadratic system of index-1. Next, we utilize the Carleman bilinearization, ensuring that the resulting bilinearized system also has an index-1 matrix pencil [70]. The order of the bilinearized DAE system is $\mathcal{N} = (n_1 + n)(2n_1 + 1)$.

For our experiment, we choose $n_1 = 10$ and $n = 30$. The bilinearized system is, therefore, of order $\mathcal{N} = 840$. The polynomial part of the first subsystem of the bilinearized system is $D_1 = 0.0333$ and higher order subsystems have zero polynomial parts. Using Theorem 6.2, we compute the projection matrices such that the reduced-order system guarantees interpolation of the first two subsystems at $\sigma = \mu = [10, 50, 300]$. The multiplicities of all the interpolation points are set to 1. The reduced-order systems of the bilinearized system are computed using the classical and the proposed methodology



(a) Transient response.

(b) Absolute error.

Figure 6.3.: A nonlinear RC circuit: comparison of the original, Carleman bilinearized and the reduced-order systems for an input $u(t) = \cos(20\pi t) + 1$.

using the same interpolation points and multiplicities, since we do not have specific criteria yet to choose these interpolation points and their multiplicities which can ensure a stable reduced-order system for both the modified and the classical method. For our result, it is possible to get stable reduced-order systems using this methodology for the same interpolation points and same multiplicities in the case of one-sided projection, i.e., $W = V$.

The time responses of the resulting reduced-order bilinear systems are shown in Figure 6.3a by utilizing the implicit Euler method, and also the absolute errors ($|y - \widehat{y}|$) are shown in Figure 6.3b. Clearly, the proposed interpolatory technique shows a substantial improvement in the transient response of the system.

6.3.7. Conclusions

In this section, we have studied subsystem interpolation method for bilinear descriptor systems, having the matrix pencil $\lambda E - A$ of index-1, with a particular attention to their polynomial parts. An expression that explicitly identifies the polynomial part of each subsystem associated with the bilinear system has been derived. This extends the expression for the polynomial part of linear index-1 DAE systems discussed in [80] to bilinear systems. Also, we have derived conditions on interpolatory subspaces that not only guarantee interpolation of the first k subsystems but also retain the polynomial part of the bilinear system. We have also discussed the related computational issues. By means of a couple of numerical examples, we have shown the efficiency of the proposed model reduction technique.

However, we have observed in the RC circuit example that two-sided interpolation does not preserve the stability of the reduced-order systems, and it depends on a choice of interpolation points. Therefore, it would be interesting to study the stability problem in the future. Furthermore, the quality of the reduced-order system highly depends on the choice of interpolation points. Thus, the next question arising from here is how to choose these interpolation points which can yield reduced-order systems that are optimal in some measure, e.g., $\mathcal{H}_2$-measure. We aim at answering some of these questions in the subsequent section.

6.4. Multipoint Volterra Series Interpolation and $\mathcal{H}_2$-Optimal Model Reduction for Index-1 Bilinear Descriptor Systems

In the previous section, we have shown how to construct reduced-order systems, interpolating the first k multi-variate transfer functions and retaining their polynomial parts. In this section, we aim at extending the multi-point Volterra series interpolation (see Subsection 6.2.2 for bilinear ODEs) to SISO bilinear DAEs, having index-1 matrix pencil $\lambda E - A$ (6.17). Then, based on it, we futher study $\mathcal{H}_2$-optimal approximations of such bilinear DAEs.

6.4.1. Multipoint Volterra series interpolation for bilinear DAEs

We first define a multi-point Volterra interpolation problem for bilinear systems. For this, we consider two sets of interpolation points $\sigma_j, \mu_j \in \mathbb{C}$, $j \in \{1, 2, \ldots, \widehat{n}\}$, along with matrices $U, S \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$ and define the weighted Volterra series as follows:

$$\nu_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j} H_k(\sigma_{l_1}, \sigma_{l_2}, \ldots, \sigma_j) \tag{6.48}$$

and

$$\gamma_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,l_2,\ldots,l_{k-1},j} H_k(\mu_j, \mu_{l_1}, \ldots, \mu_{l_{k-1}}), \tag{6.49}$$

where $H_k(s_1, \ldots, s_k)$ are the multi-variate transfer function of the original system, and $\eta_{l_1,l_2,\ldots,l_{k-1},j}$ and $\vartheta_{l_1,l_2,\ldots,l_{k-1},j}$ are the weights, defined in terms of the elements of the matrix U and S, respectively, as follows:

$$\begin{aligned} \eta_{l_1,\ldots,l_{k-1},j} &= u_{j,l_{k-1}} u_{l_{k-1},l_{k-2}} \cdots u_{l_2,l_1} \quad \text{for} \quad k \geq 2 \quad \text{and} \quad \eta_{l_1} = 1, \\ \vartheta_{l_1,\ldots,l_{k-1},j} &= s_{j,l_{k-1}} s_{l_{k-1},l_{k-2}} \cdots s_{l_2,l_1} \quad \text{for} \quad k \geq 2 \quad \text{and} \quad \vartheta_{l_1} = 1. \end{aligned} \tag{6.51}$$

It is assumed that ν_j and γ_j converge for each $j \in \{1, 2, \dots, \widehat{n}\}$. The goal of the multi-point Volterra series interpolation is to determine a reduced-order system, with its kth-order multi-variate transfer function being of the form (6.25), so that the following are satisfied for each $j \in \{1, 2, \dots, \widehat{n}\}$:

$$\nu_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_{k-1},j} \widehat{H}_k(\sigma_{l_1}, \sigma_{l_2}, \dots, \sigma_j) \tag{6.53}$$

and

$$\gamma_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,l_2,\dots,l_{k-1},j} \widehat{H}_k(\mu_j, \mu_{l_1}, \dots, \mu_{l_{k-1}}). \tag{6.55}$$

As a first step in this direction, we establish the relation between the weighted Volterra series and the generalized Sylvester equation for the bilinear DAEs in the following lemma, similar to the case of bilinear ODEs in [59, Lemma 3.1].

Lemma 6.15:
Consider a SISO bilinear DAE (6.17) and let $\sigma_j, \mu_j \in \mathbb{C}$, $j \in \{1, 2, \dots, \widehat{n}\}$, be two sets of interpolation points. Given matrices $U, S \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$, and assume the following series:

$$v_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_{k-1},j} (\sigma_j E - A)^{-1} N \cdots (\sigma_{l_2} E - A)^{-1} N (\sigma_{l_1} E - A)^{-1} B \tag{6.56}$$

and

$$w_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,l_2,\dots,l_{k-1},j} (\sigma_j E - A)^{-T} N^T \cdots (\sigma_{l_2} E - A)^{-T} N^T (\sigma_{l_1} E - A)^{-T} C^T$$

converge for each $j \in \{1, 2, \dots, \widehat{n}\}$. Then, the matrices V and W, whose jth columns are v_j and w_j, respectively, solve the following generalized Sylvester equations:

$$EV\Omega - AV - NVU^T = B\mathbb{1}_{\widehat{n}}^T \tag{6.57}$$

and

$$E^T W \Xi - A^T W - N^T W S^T = C^T \mathbb{1}_{\widehat{n}}^T, \tag{6.58}$$

respectively, where $\Omega = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_{\widehat{n}})$ and $\Xi = \operatorname{diag}(\mu_1, \mu_2, \dots, \mu_{\widehat{n}})$. ◊

Proof. The lemma can be proven by extending the proof [59, Lemma 3.1] for $E = I$ to $E \neq I$. For completeness, we provide a complete proof here. Consider $V^{(1)} \in \mathbb{R}^{n\times\widehat{n}}$, solving

$$EV^{(1)}\Lambda - AV^{(1)} = B\mathbb{1}_{\widehat{n}}^T,$$

and let $V^{(k)}$ solve

$$EV^{(k)}\Lambda - AV^{(k)} = NV^{(k-1)}U^T.$$

Then, $v_{1,j} = (\sigma_j E - A)^{-1} B$, and $v_{k,j} = (\sigma_j E - A)^{-1} f_{k-1,j}$, where $v_{k,j}$ is the jth column of the matrix $V^{(k)}$ and $f_{k-1,j}$ is the jth column of $NV^{(k-1)}U^T$. Next assume that

$$f_{k-1,j} = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_{k-1},j} N(\sigma_{k-1}E - A)^{-1}N\cdots(\sigma_{l_2}E - A)^{-1}N(\sigma_{l_1}E - A)^{-1}B, \tag{6.59}$$

then we need to show that the above equation also holds for k as well. Assuming the expression for $f_{k-1,j}$ holds as in (6.59), we obtain

$$\begin{aligned} v_{k,j} &= (\sigma_j E - A)^{-1} f_{k-1,j} \\ &= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_{k-1},j} N(\sigma_{k-1}E - A)^{-1}N\cdots(\sigma_{l_2}E - A)^{-1}N(\sigma_{l_1}E - A)^{-1}B. \end{aligned}$$

Thus, it yields

$$\begin{aligned} f_{k,j} &= \sum_{l_k=1}^{\widehat{n}} u_{j,l_k} N v_{k,l_k} \\ &= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}} u_{j,l_k}\eta_{l_1,l_2,\dots,l_{k-1},j} N(\sigma_k E - A)^{-1}N\cdots(\sigma_{l_2}E - A)^{-1}N(\sigma_{l_1}E - A)^{-1}B \\ &= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_k,j} N(\sigma_k E - A)^{-1}N\cdots(\sigma_{l_2}E - A)^{-1}N(\sigma_{l_1}E - A)^{-1}B. \end{aligned}$$

As we know, $v_j = \sum_{k=1}^{\infty} v_{k,j}$ and it converges by the assumption; hence, $V = \sum_{k=1}^{\infty} V^{(k)}$ which can be now verified that it is the solution of (6.57). Analogously, we can show that W solves (6.58). □

Next, in the following theorem, we discuss the construction of a reduced-order system with required modifications so that (6.53) and (6.55) are satisfied.

Theorem 6.16:
Consider the SISO bilinear DAE (6.17) of order n. Assume for some $\widehat{n} < n$ that two sets of interpolation points $\sigma_j \in \mathbb{C}$ and $\mu_j \in \mathbb{C}$, $j \in \{1, 2, \dots, \widehat{n}\}$ and matrices $U, S \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$ such that $\sigma(U)\cap\sigma(S) = \emptyset$, where $\sigma(\cdot)$ denotes the spectrum of a matrix. Let the matrices V and W be the solutions of (6.57) and (6.58), respectively, and

L_A, L_N, L_B and L_C be the solutions to

$$L_A V + L_N V U^T + L_B \mathbb{1}_{\widehat{n}}^T = 0, \tag{6.60a}$$
$$L_A^T W + L_N^T W S^T + L_c^T \mathbb{1}_{\widehat{n}}^T = 0, \tag{6.60b}$$
$$W^T L_B + [\alpha_1, \alpha_2, \ldots, \alpha_{\widehat{n}}]^T = 0, \tag{6.60c}$$
$$L_C V + [\beta_1, \beta_2, \ldots, \beta_{\widehat{n}}] = 0, \tag{6.60d}$$

where

$$\alpha_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,l_2,\ldots,l_{k-1},j} D_k$$

and

$$\beta_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j} D_k$$

with D_k being the polynomial part of the kth-order multi-variate transfer function, see Lemma 6.7. If matrices of a reduced-order system are computed as

$$\begin{aligned} &\widehat{E} = W^T E V, && \widehat{A} = W^T(A + L_A)V, && \widehat{N} = W^T(N + L_N)V, \\ &\widehat{B} = W^T(B + L_B), && \widehat{C} = (C + L_C)V, && \end{aligned} \tag{6.61}$$

then the interpolation conditions (6.53) and (6.55) are satisfied for each $j \in \{1, \ldots, \widehat{n}\}$. Furthermore, if $\widehat{E}$ is invertible, then the polynomial part of each subsystem is also matched. ♢

Proof. We begin with the Sylvester equation, determining the projection matrix V

$$EV\Omega - AV - NVU^T - B\mathbb{1}_{\widehat{n}}^T = 0. \tag{6.62}$$

Subtracting (6.60a) from (6.62) yields

$$EV\Omega - (A + L_A)V - (N + L_N)VU^T - (B + L_B)\mathbb{1}_{\widehat{n}}^T = 0.$$

Premultiplying the above equation by W^T, we obtain

$$W^T \left(EV\Omega - (A + L_A)V - (N + L_N)VU^T - (B + L_B)\mathbb{1}_{\widehat{n}}^T\right) = 0.$$

This implies

$$\widehat{E}\Omega - \widehat{A} - \widehat{N}U^T - \widehat{B}\mathbb{1}_{\widehat{n}}^T = 0.$$

From the above equation, it follows that $\Psi = I_{\widehat{n}}$ solves the following projected Sylvester equation:

$$\widehat{E}\Psi\Omega - \widehat{A}\Psi - \widehat{N}\Psi U^T - \widehat{B}\mathbb{1}_{\widehat{n}}^T = 0.$$

The above projected Sylvester equation has a structure similar to the one in Lemma 6.15. So, using Lemma 6.15, the jth column of Ψ, denoted by ψ_j, can be given as

$$\psi_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j}(\sigma_j\widehat{E}-\widehat{A})^{-1}\widehat{N}\cdots(\sigma_{l_2}\widehat{E}-\widehat{A})^{-1}\widehat{N}(\sigma_{l_1}\widehat{E}-\widehat{A})^{-1}\widehat{B}. \tag{6.63}$$

Now, we multiply ψ_j by $\widehat{C}$ to obtain

$$\widehat{C}\psi_j = (C+L_C)V\psi_j = CV\psi_j + L_CV\psi_j. \tag{6.64}$$

Since the vector ψ_j is the jth column of the identity matrix, $V\psi_j$ gives the jth column of the matrix V, given in (6.15) and multiplication with C gives

$$CV\psi_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j}H_k(\sigma_{l_1},\sigma_{l_2},\ldots,\sigma_j) = \nu_j. \tag{6.65}$$

By (6.60d), we get

$$L_cV\psi_j = -\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j}D_k. \tag{6.66}$$

Finally, we substitute (6.65), (6.66) and the expression for ψ_j from (6.63) in (6.64) to have

$$\begin{aligned}\nu_j &= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j}\widehat{C}(\sigma_j\widehat{E}-\widehat{A})^{-1}\widehat{N}\cdots(\sigma_{l_2}\widehat{E}-\widehat{A})^{-1}\\ &\qquad\times\widehat{N}(\sigma_{l_1}\widehat{E}-\widehat{A})^{-1}\widehat{B} + \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j}D_k\\ &= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\ldots,l_{k-1},j}\widehat{H}_k(\sigma_{l_1},\sigma_{l_2},\ldots,\sigma_j).\end{aligned}$$

Using a similar argument, we can prove

$$\gamma_j = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,l_2,\ldots,l_{k-1},j}\widehat{H}_k(\mu_j,\mu_{l_1},\ldots,\mu_{l_{k-1}}).$$

Since we have assumed the form of the kth-order multi-variate transfer function of the reduced-order system as shown in (6.25) and $\widehat{E}$ being invertible, this means that the polynomial parts of each subsystem of the original and reduced-order systems are equal to D_k. This concludes the proof. □

Remark 6.17:
In Theorem 6.16, it is assumed that the matrices U and S do not have any common eigenvalue in order to have simultaneous solutions of the set of equations (6.60a)–(6.60d) for the matrices L_A, L_N, L_B and L_C. If the matrices U and S have common eigenvalues, then this leads to numerical issues which we discuss later in the section.$\Diamond$

Theorem 6.16 shows how to choose the projection matrices and to obtain a reduced-order system with the required modifications which not only interpolate the underlying Volterra series but also retain the polynomial part of each subsystem. Meanwhile, we also like to highlight an important aspect that the reduced-order system matrices obtained from Theorem 6.16 are not obtained via projection of the original system matrices (6.17). They are rather obtained via projection of another bilinear system (intermediate bilinear system) of order n whose kth-order multi-variate transfer function is given by

$$\widetilde{H}(s_1, s_2, \ldots, s_k) = \widetilde{C}(s_k\widetilde{E} - \widetilde{A})^{-1}\widetilde{N} \cdots (s_2\widetilde{E} - \widetilde{A})^{-1}\widetilde{N}(s_1\widetilde{E} - \widetilde{A})^{-1}\widetilde{B} + D_k, \tag{6.67}$$

where

$$\begin{aligned} \widetilde{E} &= E, & \widetilde{A} &= A + L_A, & \widetilde{N} &= N + L_N, \\ \widetilde{B} &= B + L_B, & \widetilde{C} &= C + L_C. \end{aligned} \tag{6.68}$$

Interestingly, we project the intermediate bilinear system using the projection matrices V and W, depending on the original bilinear system matrices, as opposed to the intermediate bilinear system matrices. So next, to resolve this discrepancy, we show the formulation of the reduced-order system, obtained in Theorem 6.16, in a standard projection framework using the intermediate bilinear system. We reveal that the projection matrices obtained using the original and intermediate bilinear system matrices are exactly the same.

Proposition 6.18:
For some $\widehat{n} < n$, we consider two sets of interpolation points $\sigma_j, \mu_j \in \mathbb{C}$, $j \in \{1, \ldots, \widehat{n}\}$, and matrices $U, S \in \mathbb{C}^{\widehat{n}\times\widehat{n}}$ such that $\sigma(U) \cap \sigma(S) = \emptyset$. Let the matrices V and W be the solutions of (6.57) and (6.58), respectively, and let the projection matrices $\widetilde{V}$ and $\widetilde{W}$ be the solutions to

$$\widetilde{E}\widetilde{V}\Omega - \widetilde{A}\widetilde{V} - \widetilde{N}\widetilde{V}U^T = \widetilde{B}\mathbb{1}_{\widehat{n}}^T \tag{6.69}$$

and

$$\widetilde{E}^T\widetilde{W}\Omega - \widetilde{A}\widetilde{W} - \widetilde{N}^T\widetilde{W}S^T = \widetilde{C}^T\mathbb{1}_{\widehat{n}}^T, \tag{6.70}$$

respectively. Then, $\widetilde{V} = V$ and $\widetilde{W} = W$ also solve (6.69) and (6.70), respectively. $\Diamond$

Proof. We begin by proving that the matrix V also satisfies (6.69). Consider

$$\begin{aligned}
&\widetilde{E}V\Omega - \widetilde{A}V - \widetilde{N}VU^T \\
&\quad = EV\Omega - AV - L_AV - NVU^T - L_NVU^T \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(substituting for } \widetilde{A} \text{ and } \widetilde{N} \text{ from (6.68))} \\
&\quad = (EV\Omega - AV - NVU^T) - (L_AV + L_NVU^T)
\end{aligned}$$

From (6.57), $EV\Omega - AV - NVU^T = B\mathbb{1}_{\widehat{n}}^T$ and using the relation between L_A, L_N and L_B from (6.60a), we get

$$\widetilde{E}V\Omega - \widetilde{A}V - \widetilde{N}VU^T = B\mathbb{1}_{\widehat{n}}^T + L_B\mathbb{1}_{\widehat{n}}^T = \widetilde{B}\mathbb{1}_{\widehat{n}}^T.$$

An analogous argument can be given for (6.70) as well. This proves the assertion. □

Based on this investigation, we propose the following corollary.

Corollary 6.19:
The reduced-order system, determined in Theorem 6.16, coincides with the reduced-order system obtained from the intermediate bilinear system, whose kth-order multivariate transfer function is given in (6.67), via the projection subspaces $\widetilde{V}$ and $\widetilde{W}$ in a standard projection framework. ◊

6.4.2. $\mathcal{H}_2$-optimal model reduction for bilinear DAEs, having index-1 matrix pencil

So far, we have shown how to determine a reduced-order system with appropriate modifications so that the multi-point interpolation of the underlying Volterra series can be achieved together with retaining the polynomial part of each subsystem. As the subsystem interpolation method, the quality of the reduced-order system, obtained via Volterra interpolation is highly dependent on the choice of interpolation points as well as the matrices U and S. Next, we discuss first-order necessary conditions for $\mathcal{H}_2$-optimality of bilinear DAEs (6.17), having index-1 matrix pencil $\lambda E - A$. First-order necessary conditions, in terms of the pole-residues of the multi-variate transfer functions, for bilinear ODEs were derived in [59] by minimizing the error in the $\mathcal{H}_2$-norm of the error system. Here, we also consider the analog first-order necessary conditions for optimality for bilinear DAEs which are as follows:

$$\begin{aligned}
&\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}} \widehat{\phi}_{l_1,l_2,\ldots,l_k} H_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k}) \\
&\quad = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}} \widehat{\phi}_{l_1,l_2,\ldots,l_k} \widehat{H}_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k})
\end{aligned} \tag{6.71}$$

and

$$\begin{aligned}&\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}}\widehat{\phi}_{l_1,\ldots,l_k}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}H_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k})\right)\\&=\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}}\widehat{\phi}_{l_1,\ldots,l_k}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}\widehat{H}_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k})\right),\end{aligned}\tag{6.72}$$

where $\widehat{\phi}_{l_1,\ldots,l_k}$ and $\widehat{\lambda}_{l_i}$ are the residues and poles, respectively, of the transfer functions $\widehat{H}_k(s_1, s_2, \ldots, s_k)$. In this regard, we first establish the connection between the multi-point interpolation of the Volterra series interpolation conditions and the pole-residues of the kth-order multi-variate transfer function of the reduced-order system.

Lemma 6.20:
Let $H_k(s_1, s_2, \ldots, s_k)$ and $\widehat{H}_k(s_1, s_2, \ldots, s_k)$ be the kth-order multi-variate transfer functions of the original and reduced-order systems as shown in (6.18) and in (6.25), respectively. Decompose $Y\widehat{A}Z = \Omega = \operatorname{diag}(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_{\widehat{n}})$ and $Y\widehat{E}Z = I_{\widehat{n}}$, where $\{\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_{\widehat{n}}\}$ are the eigenvalues of the matrix pencil $\lambda\widehat{E} - \widehat{A}$ and the columns of $Z = [z_1, z_2, \ldots, z_{\widehat{n}}]$ and $Y = [y_1, y_2, \ldots, y_{\widehat{n}}]$ are the right and left eigenvectors, respectively.

Moreover, define $\mathcal{B} = Y\widehat{B}$, $\mathcal{N} = Y\widehat{N}Z$ and $\mathcal{C} = \widehat{C}Z$, and let $\widehat{\phi}_{l_1,\ldots,l_k}$ be the residues corresponding to the kth-order multi-variate transfer function $\widehat{H}_k(s_1, \ldots, s_k)$. Assume that the projection matrices V and W solve

$$EV(-\Omega) - AV - NV\mathcal{N}^T = B\mathcal{B}^T, \tag{6.73}$$

$$E^TW(-\Omega) - A^TW - N^TW\mathcal{N} = C^T\mathcal{C}, \tag{6.74}$$

respectively. Then,

$$\mathcal{C}\,(CV)^T = \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}}\widehat{\phi}_{l_1,\ldots,l_k}H_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k}). \qquad \Diamond$$

Proof. We begin by comparing (6.73) and (6.57) which readily shows that these two equations are equivalent after setting

$$U = \mathcal{N}, \quad \mathbb{1}_{\widehat{n}} = \mathcal{B} \quad \text{and} \quad \sigma_j = -\widehat{\lambda}_j, \quad j \in \{1, 2, \ldots, \widehat{n}\}.$$

By applying Lemma 6.15, we can write the jth column of V, v_j, as

$$\begin{aligned}v_j = &\sum_{k=2}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}}\eta_{l_1,\ldots,l_{k-1},j}\mathcal{B}_{l_1}(\widehat{\lambda}_jE - A)^{-1}N\cdots\\&\times(\widehat{\lambda}_{l_2}E - A)^{-1}N(\widehat{\lambda}_{l_1}E - A)^{-1}B + \mathcal{B}_j(\widehat{\lambda}_jE - A)^{-1}B,\end{aligned}\tag{6.75}$$

where $\eta_{l_1,\dots,l_{k-1},j} = \mathcal{N}(j, l_{k-1})\mathcal{N}(l_{k-1}, l_{k-2})\cdots\mathcal{N}(l_2, l_1)$ for $k \geq 2$ by the definition of $\eta_{l_1,\dots,l_{k-1}}$ in (6.7), and $\mathcal{B}_i$ is the ith element of $\mathcal{B}$. Multiplying (6.75) by C yields

$$Cv_j = \sum_{k=2}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},j}\mathcal{B}_{l_1}H_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_j) + \mathcal{B}_jH_1(-\widehat{\lambda}_j).$$

Hence,

$$(CV)^T = \begin{bmatrix} \sum\limits_{k=2}^{\infty}\sum\limits_{l_1=1}^{\widehat{n}}\cdots\sum\limits_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},1}\mathcal{B}_{l_1}H_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_1) + \mathcal{B}_1H_1(-\lambda_1) \\ \sum\limits_{k=2}^{\infty}\sum\limits_{l_1=1}^{\widehat{n}}\cdots\sum\limits_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},2}\mathcal{B}_{l_1}H_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_2) + \mathcal{B}_2H_1(-\lambda_2) \\ \vdots \\ \sum\limits_{k=2}^{\infty}\sum\limits_{l_1=1}^{\widehat{n}}\cdots\sum\limits_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},\widehat{n}}\mathcal{B}_{l_1}H_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_{\widehat{n}}) + \mathcal{B}_{\widehat{n}}H_1(-\lambda_r) \end{bmatrix}.$$

Next, we premultiply the above equation by $\mathcal{C} = [\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{\widehat{n}}]$, where $\mathcal{C}_i$ is the ith element of $\mathcal{C}$. This yields

$$\mathcal{C}(CV)^T = \sum_{k=2}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_k=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},l_k}\mathcal{C}_{l_k}\mathcal{B}_{l_1}H_k(-\widehat{\lambda}_{l_1},\dots,-\widehat{\lambda}_{l_k}) + \sum_{l_k=1}^{\widehat{n}} \mathcal{C}_{l_k}\mathcal{B}_{l_k}H_1(-\lambda_{l_k}). \tag{6.76}$$

Now, we recall the expression for the residues $\widehat{\phi}_{l_1,\dots,l_k}$ of the kth-order multi-variate transfer function of the reduced-order system which are given as:

$$\begin{aligned} \widehat{\phi}_{l_k} &= \mathcal{C}_{l_k}\mathcal{B}_{l_k}, \\ \widehat{\phi}_{l_1,\dots,l_k} &= \mathcal{C}_{l_k}\eta_{l_1,\dots,l_{k-1},l_k}\mathcal{B}_{l_1}, \quad \text{for} \quad k \geq 2. \end{aligned}$$

Lastly, we substitute the above relation in (6.76), leading to the desired result. □

Our next task is to obtain a reduced-order system that satisfies the necessary conditions for optimality (6.71) and (6.72). The following theorem reveals the choice of a reduced-order system, ensuring first-order necessary conditions for $\mathcal{H}_2$-optimality.

Theorem 6.21:
Let $H_k(s_1, s_2, \dots, s_k)$ and $\widehat{H}_k(s_1, s_2, \dots, s_k)$ be the kth-order multi-variate transfer functions of the original and reduced-order bilinear systems, respectively, and assume the projection matrices V and W are given by (6.73) and (6.74), respectively. Also, assume that L_A, L_N, L_B and L_C satisfy the following set of equations:

$$L_AV + L_NV\mathcal{N}^T + L_B\mathcal{B}^T = 0, \tag{6.77a}$$

$$L_A^TW + L_N^TW\mathcal{N} + L_c^T\mathcal{C} = 0, \tag{6.77b}$$

$$W^TL_B + [\alpha_1, \alpha_2, \dots, \alpha_{\widehat{n}}]^T = 0, \tag{6.77c}$$

$$L_CV + [\beta_1, \beta_2, \dots, \beta_{\widehat{n}}] = 0, \tag{6.77d}$$

where

$$\alpha_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \vartheta_{l_1,l_2,\dots,l_{k-1},j} \mathcal{C}_{l_1} D_k \tag{6.78}$$

and

$$\beta_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_{k-1},j} \mathcal{B}_{l_1} D_k \tag{6.79}$$

with

$$\begin{aligned} \eta_{l_1,\dots,l_{k-1},j} &= \mathcal{N}(j, l_{k-1})\mathcal{N}(l_{k-1}, l_{k-2}) \cdots \mathcal{N}(l_2, l_1) \quad \text{for} \quad k \geq 2, \\ \vartheta_{l_1,\dots,l_{k-1},j} &= \mathcal{N}(l_{k-1}, j)\mathcal{N}(l_{k-2}, l_{k-1}) \cdots \mathcal{N}(l_1, l_2) \quad \text{for} \quad k \geq 2. \end{aligned}$$

If the reduced-order system matrices are computed as shown in (6.61), then first-order necessary conditions for $\mathcal{H}_2$-optimality (6.71) and (6.72) are satisfied along with retaining the polynomial part of each subsystem. ♢

Proof. We begin by recalling Lemma 6.15 that provides us the formulation of the jth column of the identity matrix, ψ_j, see (6.63),

$$\psi_j = \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,l_2,\dots,l_{k-1},j} \mathcal{B}_{l_1} (\sigma_j \widehat{E} - \widehat{A})^{-1} \widehat{N} \cdots (\sigma_{l_2} \widehat{E} - \widehat{A})^{-1} \widehat{N} (\sigma_{l_1} \widehat{E} - \widehat{A})^{-1} \widehat{B}.$$

Now, we multiply the above equation by $\widehat{C}$ to get

$$\widehat{C}\Psi = (C + L_C)V = CV + L_C V. \tag{6.80}$$

Transposing (6.80) and premultiplying by $\mathcal{C}$ lead to

$$\mathcal{C}(\widehat{C}\Psi)^T = \mathcal{C}(CV)^T + \mathcal{C}(L_C V)^T.$$

Next, we substitute $L_C V$ given in (6.77d) and employ (6.79) which on simplification yields

$$\mathcal{C}(CV)^T = \mathcal{C} \begin{bmatrix} \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},j} \mathcal{B}_{l_1} \widehat{H}_k(-\widehat{\lambda}_{l_1}, \dots, -\widehat{\lambda}_1)) \\ \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},j} \mathcal{B}_{l_1} \widehat{H}_k(-\widehat{\lambda}_{l_1}, \dots, -\widehat{\lambda}_2)) \\ \vdots \\ \sum_{k=1}^{\infty} \sum_{l_1=1}^{\widehat{n}} \cdots \sum_{l_{k-1}=1}^{\widehat{n}} \eta_{l_1,\dots,l_{k-1},j} \mathcal{B}_{l_1} \widehat{H}_k(-\widehat{\lambda}_{l_1}, \dots, -\widehat{\lambda}_{\widehat{n}})) \end{bmatrix}.$$

Using Lemma 6.20 and simple algebra gives us

$$\sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \widehat{\phi}_{l_1,\ldots,l_{k-1},j} H_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k})$$
$$= \sum_{k=1}^{\infty}\sum_{l_1=1}^{\widehat{n}}\cdots\sum_{l_{k-1}=1}^{\widehat{n}} \widehat{\phi}_{l_1,\ldots,l_{k-1},j} \widehat{H}_k(-\widehat{\lambda}_{l_1},\ldots,-\widehat{\lambda}_{l_k}).$$

The second necessary condition (6.72) can be easily obtained in a similar fashion as shown in [59, Thm. 4.2] by tracing the terms corresponding to $W(:,j)^TV(:,j)$, for $j = 1, 2, \ldots, \widehat{n}$. □

Clearly, still, the computation of the reduced-order system realization involves the matrices L_A, L_N, L_B and L_C which are not readily available. In what follows, we show how to compute the reduced-order system without explicitly computing these matrices and their related computational issues.

Computational issues

Now, we discuss the computational issues related to determining the realization of the reduced-order system. It is interesting to note that we do not need the matrices L_A, L_N, L_B and L_C explicitly, but we rather require expressions for W^TL_AV, W^TL_NV, W^TL_B and L_CV to determine the reduced-order system. The expressions for W^TL_B and L_CV are given in (6.77c) and (6.77d), respectively, which are

$$\begin{aligned} W^TL_B &= -[\mathcal{C}^TD_1 + \mathcal{N}^T\mathcal{C}^TD_2 + (\mathcal{N}^T)^2\mathcal{C}^TD_3 + \cdots],\\ L_CV &= -[D_1\mathcal{B}^T + D_2\mathcal{B}^T\mathcal{N}^T + D_3\mathcal{B}^T(\mathcal{N}^T)^2 + \cdots]. \end{aligned}$$

In order to determine the expressions for W^TL_AV and W^TL_NV, we premultiply (6.77a) and (6.77b) by W^T and V^T, respectively, and obtain

$$W^TL_AV + W^TL_NV\mathcal{N}^T + W^TL_B\mathcal{B}^T = 0, \tag{6.81}$$
$$V^TL_A^TW + V^TL_N^TW\mathcal{N} + V^TL_C^T\mathcal{C} = 0. \tag{6.82}$$

Now, we subtract (6.81) from the transpose of (6.82), leading to the following Sylvester equation in W^TL_NV:

$$\mathcal{N}^T(W^TL_NV) - (W^TL_NV)\mathcal{N}^T + \mathcal{C}^TL_CV - W^TL_B\mathcal{B}^T = 0. \tag{6.83}$$

In order to have a unique solution of the above Sylvester equation, the matrix $\mathcal{X} := I_{\widehat{n}} \otimes \mathcal{N}^T - \mathcal{N} \otimes I_{\widehat{n}}$ should be invertible. But, it is easy to see that the matrix contains zero eigenvalues. It implies that if $\operatorname{vec}(\mathcal{F}) \in \operatorname{range}(\mathcal{X})$, where $\mathcal{F} := \mathcal{C}^TL_CV - W^TL_B\mathcal{B}^T$, where $\operatorname{vec}(\cdot)$ denotes the vectorization of a matrix by stacking the columns of the matrix on top of each other, then (6.83) has infinitely many solutions, otherwise it has no solution. Practically, it is difficult to ensure in each iteration of an iterative scheme such as B-IRKA for all possible bilinear systems that $\operatorname{vec}(\mathcal{F}) \in \operatorname{range}(\mathcal{X})$. However, if one assumes that $D_k = 0$ for $k \geq 3$, then the equation (6.83) boils down to

$$\mathcal{N}^T(W^TL_NV + \mathcal{C}^TD_2\mathcal{B}^T) - (W^TL_NV + \mathcal{C}^TD_2\mathcal{B}^T)\mathcal{N}^T = 0.$$

This implies W^TL_NV has infinite solutions which are as follows:

$$W^TL_NV = -\mathcal{C}^TD_2\mathcal{B}^T + \mathcal{Y},$$

where $\operatorname{vec}(\mathcal{Y}) \in \operatorname{null}(\mathcal{X})$. For simplicity, we take $\mathcal{Y} = 0$ to avoid some additional computations. Moreover, if we compute W^TL_NV, having $\mathcal{Y} \neq 0$, then we seldom observe the convergence of a fixed point iterative scheme. This probably happens due to the fact that the computation of $\mathcal{Y}$ does not take into account the realization of the reduced-order system anymore. It rather depends only on the null space of the matrix $\mathcal{X}$, which might be creating some numerical instability in the iteration process of B-IRKA. Therefore, we recommend to set $\mathcal{Y} = 0$; thus, W^TL_NV can be computed easily. The expression for W^TL_AV can be simply computed by inserting the expressions for W^TL_B and W^TL_NV in (6.81).

Remark 6.22:
As we have noted above, the Sylvester equation (6.83) either does not have a unique solution or even has no solution. However, it is possible to determine a solution if $D_k = 0 \ \forall\ k \geq 3$.

In case of $D_k \neq 0$ for some $k \geq 3$, Eq. (6.83), in general, does not have any solution. This implies that it is not possible to obtain a reduced-order system, satisfying the necessary conditions for optimality. Nevertheless, here we set W^TL_NV equal to $-\mathcal{C}^TD_2\mathcal{B}^T$ which often may be a good choice as D_k generally decreases fast. ◊

Now, we sketch the iterative scheme in Algorithm 6.3 based on our theoretical discussions for the class of bilinear DAEs (6.17).

Remark 6.23:
As noted in Remark 6.13, there exist particular structures of E and N, when higher order systems with $k \geq 2$, all have zero polynomial parts, i.e., $D_k = 0 \ \forall\ k \geq 2$. ◊

Remark 6.24:
The expressions for $\mathcal{R}_B$ and $\mathcal{R}_C$ require the summation of the infinite series. However, D_i generally decreases fast; therefore, one can consider only the leading terms which may approximate the infinite summation very well. In case D_k does not decay, we can always choose a factor $0 < \gamma < 1$ that scales N and B when multiplying the input with $(\frac{1}{\gamma})$. The dynamics of the system do not change by doing so. This way, one can ensure the decay of the D_k's. However, in all applications we consider in the next section, $D_k = 0 \ \forall\ k \geq 3$. ◊

Algorithm 6.3: B-IRKA for bilinear DAEs, having an index-1 matrix pencil.

```
Input: $E, A, N, B, C$.
Output: $\widehat{E}, \widehat{A}, \widehat{N}, \widehat{B}, \widehat{C}$.
Make an initial guess of $\Omega, \mathcal{B}, \mathcal{N}$ and $\mathcal{C}$.
while no convergence do
    Solve for $V$ and $W$:
        $EV(-\Omega) + AV + NV\mathcal{N}^T + B\mathcal{B}^T = 0,$
        $E^TW(-\Omega) + A^TW + N^TW\mathcal{N} + C^T\mathcal{C} = 0.$
    Compute the expressions for
        $W^TL_B = -\sum_{k=1}^{\infty} (\mathcal{N}^T)^{k-1}\mathcal{C}^TD_k =: \mathcal{R}_B,$
        $L_CV \;\; = -\sum_{k=1}^{\infty} D_k\mathcal{B}^T(\mathcal{N}^T)^{k-1} =: \mathcal{R}_C.$
    Compute the expression for $W^TL_NV =: \mathcal{R}_N$,
        $\mathcal{R}_N = -\mathcal{C}^TD_2\mathcal{B}^T.$
    Determine the expression for $W^TL_AV =: \mathcal{R}_A$,
        $\mathcal{R}_A = -\mathcal{R}_N\mathcal{N}^T - \mathcal{R}_B\mathcal{B}^T.$
    Compute the reduced-order system matrices:
        $\widehat{E} = W^TEV, \;\; \widehat{A} = W^TAV + \mathcal{R}_A, \;\; \widehat{N} = W^TNV + \mathcal{R}_N,$
        $\widehat{B} = W^TB + \mathcal{R}_B, \;\; \widehat{C} = CV + \mathcal{R}_C.$
    Determine $Y$ and $Z$ such that $Y\widehat{A}Z = \Omega$, $Y\widehat{E}Z = I_{\widehat{n}}$.
    Compute $\mathcal{N} = Y\widehat{N}Z$, $\mathcal{B} = Y\widehat{B}$ and $\mathcal{C} = \widehat{C}Z$.
```

Remark 6.25:
For simplicity of notation, we have shown B-IRKA (Algorithm 6.3) for SISO bilinear DAEs. Nevertheless, it can be applied to MIMO bilinear systems as well. In the MIMO case, the polynomial part of the kth subsystem, D_k is a matrix of size $D_k \in \mathbb{R}^{p\times m^k}$, where p and m are the numbers of outputs and inputs, respectively. Let us consider D_k consisting of m^{k-1} column blocks of size $p \times m$, and we denote the $\left(p_1 + \sum_{i=2}^{k-1} m^{i-1}(p_i - 1)\right)$th column block of D_k as $D_k^{p_1,\dots,p_{k-1}} \in \mathbb{R}^{p\times m}$, $p_i \in \{1, \dots, m\}$, which can be written as

$$D_k^{p_1,\dots,p_{k-1}} = C(MN_{p_{k-1}})\cdots(MN_{p_1})MB.$$

Then, the expressions for $\mathcal{R}_B, \mathcal{R}_C, \mathcal{R}_{N^i}$ and $\mathcal{R}_A$ in Algorithm 6.3 can be determined as follows:

$$
\begin{aligned}
\mathcal{R}_B &= -\sum_{k=1}^{\infty}\sum_{p_1=1}^{m}\cdots\sum_{p_{k-1}=1}^{m}(\mathcal{N}_{p_{k-1}}\cdots\mathcal{N}_{p_1})^T\mathcal{C}^T D_k^{p_1,\dots,p_{k-1}},\\
\mathcal{R}_C &= -\sum_{k=1}^{\infty}\sum_{p_1=1}^{m}\cdots\sum_{p_{k-1}=1}^{m} D_k^{p_1,\dots,p_{k-1}}\mathcal{B}^T(\mathcal{N}_{p_{k-1}}\cdots\mathcal{N}_{p_1})^T,\\
\mathcal{R}_{N^i} &= -\mathcal{C}^T D_2^i\mathcal{B}^T,\\
\mathcal{R}_A &= -\sum_{i=1}^{m}\mathcal{R}_{N^i}\mathcal{N}_i^T - \mathcal{R}_B\mathcal{B}^T.
\end{aligned}
$$

Furthermore, to solve for the projection matrices V and W in the case of MIMO, we need to replace the $NV\mathcal{N}^T$ and $N^TW\mathcal{N}$ terms at step 4 in Algorithm 6.3 with $\sum_{i=1}^m N_iV\mathcal{N}_i^T$ and $\sum_{i=1}^m N_i^TW\mathcal{N}_i$, respectively. ◊

Thus far, we have presented how to obtain the realization of the reduced-order system that aim at satisfying first-order necessary conditions for $\mathcal{H}_2$-optimality together with retaining the polynomial part of each subsystem, by assuming the structure of the kth-order transfer function of the reduced-order system as in (6.25). So, the time-domain bilinear systems, whose kth-order transfer function is given by (6.25), has been discussed in Subsection 6.3.4.

6.4.3. Numerical experiments

In this section, we illustrate the performance of the proposed B-IRKA (Algorithm 6.3) for bilinear DAEs using various numerical examples. We also compare it with the reduced bilinear systems, obtained by using POD-based approximation, the Loewner method for bilinear systems [69, 92], and by applying IRKA to the corresponding linear part [80, Algo. 5.2] and then project bilinear terms. The stopping criterion for Algorithm 6.3 is chosen based on the relative change of the norm of the poles of the reduced-order system. If the relative change becomes smaller than *tol*, then we stop the iteration, where *tol* is chosen as the square-root of the machine precision. Moreover, the initialization of the algorithm is done by choosing arbitrary interpolation points and tangential directions. We also consider a scaling factor for smooth convergence of B-IRKA as discussed in [21, 59]. In order to employ the Loewner method for bilinear systems, we take the samples of the transfer functions $H_k(s_1,\dots,s_k)$ of the bilinear systems at l chosen logarithmically spaced frequencies $\omega_i \in [\omega_a,\omega_b]$:

$$
[\jmath\omega_l,\dots,\jmath\omega_l], \qquad [(\jmath\omega_1,\jmath\omega_1),\dots,(\jmath\omega_l,\jmath\omega_l)],
$$

where $\jmath = \sqrt{-1}$. We obtain a set of left nodes $[\mu_i,(\mu_i,\mu_i)]$ and right nodes $[\lambda_i,(\lambda_i,\lambda_i)]$ by using an alternative partition for ω_i. This leads to Loewner and shifted Loewner

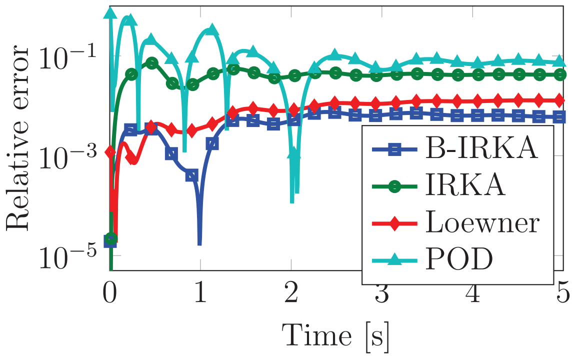


Figure 6.4.: A nonlinear RC circuit: a comparison of relative errors between the original and reduced-order systems obtained by using various methods for an input $u(t) = \cos(2\pi t)e^{-t} + 1$.

matrices of dimension $2l$. See [69, 92] for more detailed insights in the Loewner method for bilinear systems. All the simulations are carried out in MATLAB version 8.0.0.783(R2012b)64-bit(glnza64) on an Intel® Core™2 Quad CPU Q9550 @2.83GHz 6MB cache, 4GB RAM, openSUSE Linux 12.04.

A nonlinear RC circuit

As a first example, we consider the same nonlinear RC circuit example which we have considered in the previous section. Next, we determine the reduced-order systems by employing Algorithm 6.3 and linear IRKA by choosing the scaling factor $\gamma = 0.5$. We take $l = 50$ samples logarithmically between frequencies $[1, 2000]$ (rad/sec) in order to determine Loewner and shifted Loewner matrices. Furthermore, for POD-based approximation, we determine 1000 snapshots of the original solution for the input excitation $u(t) = \cos(2\pi t)$. All the reduced-order systems are of the order $r = 5$. To illustrate the accuracy of the reduced-order systems, we determine the time-domain response for the input $u(t) = \cos(2\pi t)e^{-t} + 1$ and show the relative errors in Figure 6.4.

Evidently, the reduced-order system obtained by using B-IRKA replicates the input-output behavior of the original system better as compared to the reduced-order system obtained by using IRKA, Loewner for bilinear systems. Since the projection subspace of POD corresponds to the training input $u(t) = \cos(2\pi t)$, the POD-based approximation does not approximate the transient response very well even for the sightly different input $u(t) = \cos(2\pi t)e^{-t} + 1$.

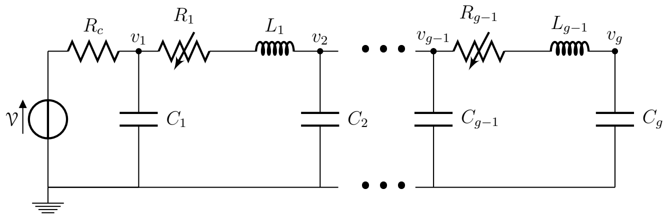


Figure 6.5.: An RLC circuit diagram with variable resistors.

A parametric RLC circuit

Next, we consider an RLC circuit as shown in Figure 6.5 whose first node has three branches, connected to the voltage source $\mathcal{V}$ via a constant resistance R_c, to a variable resistance, and to ground via a capacitor. The last, nth, node of the circuit is grounded via a capacitor. All other nodes also have three branches; the first one is grounded via a capacitor; the second one is connected to an inductor, and the third one is connected to a variable resistor as shown in Figure 6.5.

Using the Kirchhoff's voltage law at each node, we obtain the following system of equations:

$$\begin{aligned} C_j\frac{d}{dt}v_j(t) &= i_j(t) - i_{j+1}(t), && j \in \{1,2,\dots,g-1\},\\ L_j\frac{d}{dt}i_{j+1}(t) &= -R_j i_{j+1}(t) + v_{j+1}(t) - v_j(t), && j \in \{1,2,\dots,g-1\},\\ C_g\frac{d}{dt}v_g(t) &= i_g(t),\\ 0 &= v_1(t) + i_1(t)R_c - \mathcal{V}(t). \end{aligned}$$

Here, we set all the capacitors C, inductors L, and the resistance R_C equal to 1. We also assume that the variable resistances vary linearly with the parameter p as follows:

$$R_j = \mathcal{R}_j(1+p).$$

Also, we consider $\mathcal{R}_j = 1$. Combining all these equations and utilizing the parametric relation of the variable resistance, we obtain the following parametric linear system:

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + pA_1x(t) + Bu(t),\\ y(t) &= Cx(t), \end{aligned} \tag{6.84}$$

where $x(t)$ is the state vector containing the voltage at each node and current through resistances. The input $u(t)$ is the voltage source, and the quantity of interest $y(t)$ is the current through the voltage source. We set $g = 250$, leading to a linear parametric descriptor system of order $n = 500$ which has the structure of matrices E and A as

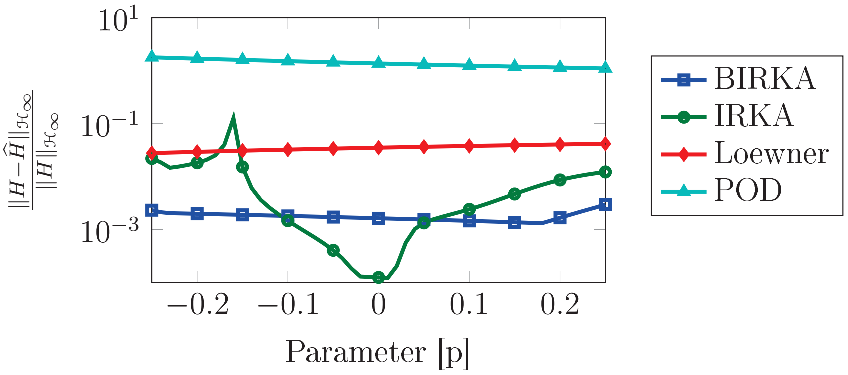


Figure 6.6.: A parametric RLC circuit: comparison of the relative $\mathcal{H}_\infty$-norm.

mentioned in (6.17). It has been shown in [20] that a special class of linear parametric systems can be treated as bilinear systems, by rewriting the parameter p as an input to the system. Therefore, we can write the system (6.84) as a bilinear system with two inputs $\widetilde{u}(t) = [u(t), p]^T$ as follows:

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + \sum_{i=1}^{2} N_i x(t)\widetilde{u}(t) + \widetilde{B}\widetilde{u}(t), \\ y(t) &= Cx(t), \end{aligned} \tag{6.85}$$

where $N_1 = 0, N_2 = A_1$, and $\widetilde{B} = \begin{bmatrix} B, & \mathbf{0} \end{bmatrix}$. The polynomial part of the first subsystem of the transformed bilinear system (6.85) is equal to 1 and all other subsystems have zero polynomial parts, i.e., $D_1 = 0$ and $D_k = 0, \forall\, k \geq 2$. We determine reduced bilinear systems by using B-IRKA and IRKA. We choose the scaling factor $\gamma = 0.1$ for a smooth convergence of B-IRKA. We take $l = 200$ samples logarithmically between frequencies $[10^{-6}, 10^{4}]$ (rad/sec) to compute Loewner and shifted Loewner matrices. Furthermore, for POD-based approximation, we determine 1000 snapshots of the actual solution for the input excitation as used in the first example. We set the order of all reduced bilinear systems to $r = 15$.

Next, these computed reduced bilinear systems can be again rewritten as reduced parametric linear systems. To determine the accuracy of the reduced-order systems, we compare the $\mathcal{H}_\infty$-norm of transfer functions of the original and reduced-order systems by varying the parameter p which is shown in Figure 6.6.

Figure 6.6 clearly shows that the reduced-order system obtained by using B-IRKA outperforms the ones obtained by using IRKA for a wide range of the parameter. On the other hand, reduced-order systems obtained by using the Loewner method and POD fail to capture the dynamics. This may be because of not treating the polynomial part of the system properly. We also like to mention that the projection matrices computed by using IRKA capture the dynamics of the original system very well in the

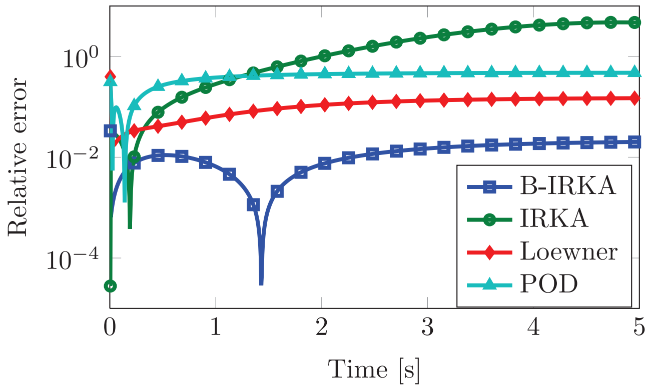


Figure 6.7.: An artificial example: comparison of the relative errors between the original and reduced-order systems for an input $u(t) = e^{0.1t}/10$.

vicinity of the parameter $p = 0$. This is why one can see a drop in the relative error in Figure 6.6 around the parameter $p = 0$ for the reduced-order system obtained by IRKA.

An artificial example

Lastly, we consider a simple artificial example of order $n = 6$, having matrices E, A, B, C as follows:

$$E = \operatorname{diag}\left([1, 1, 1, 1, 0, 0]\right), \quad A = \operatorname{diag}\left([-1, -2, -3, -4, 1, 1]\right),$$
$$C = B^T = [1, \ldots, 1]$$

and a bilinear N is such that its $(i + 1, i)$ entries are 1, $i \in \{1, \ldots, 5\}$ and all other entries are zeros. The polynomial parts of the first two subsystems are $D_1 = -2$, $D_2 = 1$ and all other subsystems have zero polynomial parts, unlike the previous two examples where D_2 is zero as well. This still fulfills the requirement to obtain a reduced-order system, satisfying $\mathcal{H}_2$-optimality conditions as stated in Remark 6.22. Next, we determine reduced-order systems via B-IRKA and IRKA. Here, we choose the scaling factor to be $\gamma = 0.1$. We take the same frequency samples as taken in the first example to compute the reduced-order system via the Loewner method. Also, for a POD-based approximation, 1000 samples of the true solutions are taken for the actuation input $u(t) = \cos(2\pi t)$. We set the order of reduced-order systems to $r = 2$. In order to observe the accuracy of the reduced-order systems, we perform time-domain simulations for a new control input $u(t) = 0.1e^{0.1t}$ and plot the relative errors between the original and reduced-order systems in Figure 6.7.

Figure 6.7 indicates that the polynomial part of system plays a significant role in the dynamics of the system which is preserved by B-IRKA along with interpolation, unlike for the other methods. We also observe that as the input is changed to a different input than the training one, the POD-based approximation fails to replicate the dynamics of the system. The figure indicates that the reduced-order system obtained via the proposed B-IRKA performs better when compared to the other methods.

6.4.4. Conclusions

In this section, we have extended the multi-point Volterra series interpolation to a family of bilinear DAEs (6.17) with the polynomial part of its kth order multi-variate transfer function being constant. We have presented the modified interpolation conditions which not only achieve multi-point interpolation of the underlying Volterra series but also retain the polynomial part of each subsystem. Based on first-order necessary conditions for $\mathcal{H}_2$-optimality, we have proposed an iterative rational Krylov algorithm, the so-called B-IRKA for such bilinear DAEs, which converges to a locally $\mathcal{H}_2$-optimal reduced-order system if it converges. Using various numerical examples, we have demonstrated the efficiency of the proposed methodology and compared it with reduced-order systems obtained by using IRKA, the Loewner method for bilinear systems and POD-based approximation.

6.5. $\mathcal{H}_2$-Model Reduction for Index-2 Bilinear Descriptor Systems

In this section, we discuss an interpolatory-based model order reduction technique for another important structured bilinear DAEs, which are of the form

$$\begin{aligned} E_{11}\dot{x}_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + \sum_{k=1}^{m} N_k x_1(t)u_k(t) + B_1u(t),\\ 0 &= A_{21}x_1(t) + B_2u(t),\\ y(t) &= C_1x_1(t) + C_2x_2(t) + Du(t), \end{aligned} \tag{6.86}$$

where $x_1(t) \in \mathbb{R}^{n_1}$, $x_2(t) \in \mathbb{R}^{n_2}$ are the generalized states; $y(t) \in \mathbb{R}^p$ and $u(t) \in \mathbb{R}^m$ are the output and input vectors of the system, respectively, and all the matrices are of appropriate dimensions. It is assumed that E_{11} and $A_{21}E_{11}^{-1}A_{12}$ are invertible. This implies that the dynamical system (6.86) is a Hessenberg index-2 differential algebraic system [81] in the case of $N_k = 0$. Generally, these special bilinear systems (6.86) arise from linearized boundary control Navier-Stokes equations or constraint RLC circuits. As a motivating example, we consider a constraint transmission circuit as shown in Figure 6.8.

The above transmission circuit contains nonlinear diodes, $g(v) = e^{40v_D} + v_D - 1$, where v_D is the voltage difference across the nodes. Using Kirchhoff's current law, we can

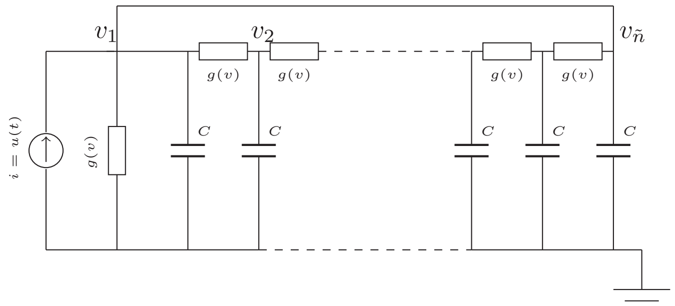


Figure 6.8.: A constraint nonlinear transmission line (index-2).

model the dynamics of the circuit as a quadratic-bilinear DAE, having an index-2 matrix pencil $\lambda E - A$ (a detailed modeling is presented later in the numerical subsection). Nonetheless, such quadratic systems, having an index-2 matrix pencil, can be approximated as bilinear systems via Carleman bilinearization [71]. The approximated bilinear systems have a similar structure as (6.86). Later in this section, we present some more constraint circuit examples which have the same structure as in (6.86). Therefore, there is a need to develop efficient model reduction techniques for such bilinear structured systems.

In Section 6.2, we have collected interpolation-based model reduction techniques of bilinear ODEs, which can be extended to bilinear DAEs having index-2 matrix pencil $\lambda E - A$ (6.86), including interpolation based $\mathcal{H}_2$-optimal model reduction. Our main goal here is to employ B-IRKA (Algorithm 6.1) for bilinear ODEs to bilinear DAEs (6.86), resulting in reduced-order systems that are locally $\mathcal{H}_2$-optimal. In Subsection 6.5.1, we present the transformation of the bilinear DAEs (6.86) into equivalent ODE systems using projectors. This allows us to employ the version of B-IRKA which is extended in [42] from the $E = I$ case to the $E \neq I$ case. However, the direct implementation of B-IRKA requires the explicit computation of the projectors which is highly undesirable. Therefore, we show how to apply B-IRKA without explicit computation of the projectors. In the end, we illustrate the efficiency of the proposed methodology by means of numerical examples.

6.5.1. Transformation into bilinear ODEs and model reduction

We begin with the case $B_2 = 0$ in (6.86), i.e.,

$$E_{11}\dot{x}_1(t) = A_{11}x_1(t) + A_{12}x_2(t) + \sum_{k=1}^{m} N_k x_1(t)u_k(t) + B_1u(t), \tag{6.87a}$$

$$0 = A_{21}x_1(t), \quad x_1(0) = 0, \tag{6.87b}$$

$$y(t) = C_1x_1(t) + C_2x_2(t) + Du(t), \tag{6.87c}$$

where the dimensions of the matrices are the same as in (6.86). As a first step, we transform the bilinear DAE (6.87) into an equivalent bilinear ODE system. The special structure of bilinear DAE (6.87) allows us to decouple the system into an algebraic and a differential part. In other words, we can obtain an ODE system for $x_1(t)$ which does not involve the variable $x_2(t)$ and an algebraic equation which evaluates $x_2(t)$ as a function of $x_1(t)$, e.g., see [3, 85, 84]. Using (6.87b), we know that $A_{21}\frac{d}{dt}x_1(t) = 0$. If (6.87a) is multiplied by $A_{21}E_{11}^{-1}$ from the left-side, it yields

$$0 = A_{21}E_{11}^{-1}\big(A_{11}x_1(t) + A_{12}x_2(t) + \sum_{k=1}^{m} N_k x_1(t)u_k(t) + B_1u(t)\big),$$

thus implying

$$x_2(t) = -(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}\big(A_{11}x_1(t) + \sum_{k=1}^{m} N_k x_1(t)u_k(t) + B_1u(t)\big). \tag{6.88}$$

Substituting the expression for $x_2(t)$ from (6.88) in (6.87a) results in

$$E_{11}\dot{x}_1(t) = \Pi A_{11}x_1(t) + \sum_{k=1}^{m} \Pi N_k x_1(t)u_k(t) + \Pi B_1u(t), \quad x_1(0) = 0, \tag{6.89a}$$

$$y(t) = \mathcal{C}x_1(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)} x_1(t)u_k(t) + \mathcal{D}u(t), \tag{6.89b}$$

where

$$\mathcal{C} = C_1 - C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}A_{11}, \quad \mathcal{C}_N^{(k)} = -C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}N_k,$$
$$\mathcal{D} = D - C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}B_1$$

and

$$\Pi = I - A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}. \tag{6.90}$$

In what follows, for simplicity, we further assume that $A_{21} = A_{12}^T$ and E_{11} is symmetric. However, $A_{21} \neq A_{12}^T$ and $E_{11} \neq E_{11}^T$ can be treated in the current bilinear framework as well by easily extending the arguments used in [80].

Note that Π is the discrete *Helmholtz* projector that is commonly used to transform Stokes type DAEs into ODEs [80, 84, 85] and that has the following properties:

$$\Pi^2 = \Pi,\ E_{11}\Pi = \Pi^T E_{11},\ \ker(\Pi) = \operatorname{range}(A_{12}),\ \text{and}\ \operatorname{range}(\Pi) = \ker\left(A_{12}^T E_{11}^{-1}\right).$$

Using these properties of Π, one can derive that

$$A_{12}^T z = 0 \quad \text{if and only if} \quad \Pi^T z = z. \tag{6.91}$$

By construction, a solution $x_1(t)$ of (6.87) fulfills $A_{12}^T x_1(t) = 0$; therefore, in (6.89), we can replace $x_1(t)$ with $\Pi^T x_1(t)$ and, using $\Pi = \Pi^2$ and $E_{11}\Pi = \Pi^T E_{11}$, we obtain the following equivalent system

$$\Pi E_{11}\Pi^T \dot{x}_1(t) = \Pi A_{11}\Pi^T x_1(t) + \sum_{k=1}^{m} \Pi N_k \Pi^T x_1(t) u_k(t) + \Pi B_1 u(t), \tag{6.92a}$$

$$y(t) = \mathcal{C}\Pi^T x_1(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)} \Pi^T x_1(t) u_k(t) + \mathcal{D}u(t), \qquad x_1(0) = 0. \tag{6.92b}$$

The above dynamical system (6.92) lies in the $n_1 - n_2$ dimensional null space of Π. Therefore, as in [85], we can decompose the projector Π as

$$\Pi = \phi_1 \phi_2^T, \tag{6.93}$$

with $\phi_1, \phi_2 \in \mathbb{R}^{n_1 \times n_1 - n_2}$ satisfying

$$\phi_1^T \phi_2 = I.$$

This decomposition allows us to write (6.92) in the following form

$$\phi_2^T E_{11}\phi_2 \dot{\widetilde{x}}_1(t) = \phi_2^T A_{11}\phi_2 \widetilde{x}_1(t) + \sum_{k=1}^{m} \phi_2^T N_k \phi_2 \widetilde{x}_1(t) u_k(t) + \phi_2^T B_1 u(t), \tag{6.94a}$$

$$y(t) = \mathcal{C}\phi_2 \widetilde{x}_1(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)} \phi_2 \widetilde{v}u(t) + \mathcal{D}u(t), \tag{6.94b}$$

where $\widetilde{x}_1(t) = \phi_1^T x_1(t)$ and $\widetilde{x}_1(0) = 0$. Thus, a model reduction problem of the system (6.94) is equivalent to a model reduction problem of the system (6.87). However, the advantage of the system (6.94) is that $\phi_2^T E_{11}\phi_2$ is nonsingular, allowing us to employ Algorithm 6.1 if bilinear terms are neglected in the output equation. This leads a locally $\mathcal{H}_2$-optimal reduced-order system upon convergence. Unfortunately, to determine the system matrices of (6.94), we require the explicit computation of the basis matrix ϕ_2, which is not readily available. Moreover, it might also appear that the realization of the system (6.94) becomes dense after multiplication with ϕ_2, making the computation of the reduced-order systems expensive. To overcome this, in what follows, we show how to avoid the explicit computation of ϕ_2 in the application of B-IRKA.

Remark 6.26:
In this work, we neglect the nonlinear terms and the control part in the output equation in (6.94) as far as the computation of the projection matrices is concerned. We focus on the linear relation between the state vector and the output. Nonetheless, the bilinear terms in the output equation are projected afterward. ♢

Algorithm 6.4: B-IRKA for bilinear DAEs, having index-2 matrix pencil (involving projector).

```
Input: $E_{11}$, $A_{11}$, $N_k$, $B_1$, $\mathcal{C}$, $\mathcal{C}_N^{(k)}$.
Make an initial guess of $\widehat{E}$, $\widehat{A}$, $\widehat{N}_k$, $\widehat{B}$, $\widehat{\mathcal{C}}$.
while no convergence do
    Compute nonsingular matrices $Y$ and $Z$ such that $Y\widehat{A}Z = \Lambda$ and $Y\widehat{E}Z = I_{\hat{n}}$.
    Define $\widetilde{B} = \widehat{B}^T Y^T$,  $\widetilde{C} = \widehat{C}Z$  and  $\widetilde{N}_k = Z^T \widehat{N}_k^T Y^T$.
    Determine
        $L = -(I_{\hat{n}} \otimes \phi_2)\left(\Lambda \otimes (\phi_2^T E_{11}\phi_2) + I_{\hat{n}} \otimes (\phi_2^T A_{11}\phi_2) + \sum_{k=1}^{m} \widetilde{N}_k^T \otimes (\phi_2^T N_k \phi_2)\right)^{-1} (I_{\hat{n}} \otimes \phi_2^T)$.
    Determine the projection matrices $\mathcal{V}$ and $\mathcal{W}$:
        $\operatorname{vec}(\mathcal{V}) = L(\widetilde{B}^T \otimes B)\mathcal{I}_m$,
        $\operatorname{vec}(\mathcal{W}) = L^T(\widetilde{C}^T \otimes \mathcal{C}^T)\mathcal{I}_p$.
    Compute the reduced-order system matrices:
        $\widehat{E} = \mathcal{W}^T E_{11}\mathcal{V}$,   $\widehat{A} = \mathcal{W}^T A_{11}\mathcal{V}$,   $\widehat{N}_k = \mathcal{W}^T N_k \mathcal{V}$,
        $\widehat{B} = \mathcal{W}^T B_1$,   $\widehat{C} = \mathcal{C}\mathcal{V}$,   $\mathcal{C}_N^{(k)} = \widehat{\mathcal{C}}_N^{(k)}\mathcal{V}$.
Output: $\widehat{E}^{opt} = \widehat{E}$, $\widehat{A}^{opt} = \widehat{A}$, $\widehat{N}_k^{opt} = \widehat{N}_k^{opt}$, $\widehat{B}^{opt} = \widehat{B}$, $\widehat{\mathcal{C}}^{opt} = \widehat{\mathcal{C}}$, $\widehat{\mathcal{C}}_N^{(k)opt} = \widehat{\mathcal{C}}_N^{(k)}$.
```

6.5.2. Computational issues

We consider the following associated bilinear ODE system to compute the projection matrices $\mathcal{V}$ and $\mathcal{W}$:

$$\phi_2^T E_{11}\phi_2\dot{\widetilde{x}}_1(t) = \phi_2^T A_{11}\phi_2\widetilde{x}_1(t) + \sum_{k=1}^{m} \phi_2^T N_k\phi_2\widetilde{x}_1(t)u_k(t) + \phi_2^T B_1 u(t), \tag{6.95a}$$

$$\widetilde{y}(t) = \mathcal{C}\phi_2\widetilde{x}_1(t), \quad \widetilde{x}_1(0) = 0. \tag{6.95b}$$

In the view of resolving the computational issues, we first aim at determining $\mathcal{V}$ and $\mathcal{W}$ such that the system matrices $E_{11}, A_{11}, N_k, B_1, \mathcal{C}$ and $\mathcal{C}_N$ can be directly reduced using the projection matrices as shown in Algorithm 6.4 which is a straightforward application of B-IRKA (Algorithm 6.1) to the system (6.95). However, note that the Sylvester equations to compute V and W are now written as linear systems using Kroneker products.

We notice that the projection matrices can be directly applied to the original system matrices, but in order to compute the projection matrices $\mathcal{V}$ and $\mathcal{W}$, we still require the matrix ϕ_2 explicitly. Therefore, our next goal is to construct the matrices $\mathcal{V}$ and $\mathcal{W}$ without resorting to ϕ_2.

Lemma 6.27:
Let ϕ_2 be the matrix as defined in (6.93) and $\mathcal{F}$ be a matrix such that $(I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)$ is invertible. Define $\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{F}}$ as follows:

$$\begin{aligned} \mathcal{X}_{\mathcal{F}}^{\mathcal{I}} &:= (I_{\widehat{n}} \otimes \phi_2)\Big((I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)\Big)^{-1}(I_{\widehat{n}} \otimes \phi_2^T), \\ \mathcal{X}_{\mathcal{F}} &:= (I_{\widehat{n}} \otimes \Pi)\mathcal{F}(I_{\widehat{n}} \otimes \Pi^T), \end{aligned} \tag{6.97}$$

where Π is defined in (6.90). Then, the matrices $\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}$ and $\mathcal{X}_{\mathcal{F}}$ satisfy the following relation:

$$\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}\mathcal{X}_{\mathcal{F}} = (\mathcal{X}_{\mathcal{F}}\mathcal{X}_{\mathcal{F}}^{\mathcal{I}})^T = I_{\widehat{n}} \otimes \Pi^T. \qquad \Diamond$$

Proof. We begin with

$$\mathcal{X}_{\mathcal{F}}^{\mathcal{I}}\mathcal{X}_{\mathcal{F}} = (I_{\widehat{n}} \otimes \phi_2)\Big((I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)\Big)^{-1}(I_{\widehat{n}} \otimes \phi_2^T)(I_{\widehat{n}} \otimes \Pi)\mathcal{F}(I_{\widehat{n}} \otimes \Pi^T).$$

We decompose $\Pi = \phi_1\phi_2^T$ and use properties of the Kronecker product to get

$$\begin{aligned} \mathcal{X}_{\mathcal{F}}^{\mathcal{I}}\mathcal{X}_{\mathcal{F}} &= (I_{\widehat{n}} \otimes \phi_2)\Big((I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)\Big)^{-1}(I_{\widehat{n}} \otimes \phi_2^T)(I_{\widehat{n}} \otimes \phi_1)(I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \Pi^T) \\ &= (I_{\widehat{n}} \otimes \phi_2)\Big((I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)\Big)^{-1}(I_{\widehat{n}} \otimes \phi_2^T\phi_1)(I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \Pi^T). \end{aligned}$$

Since $\phi_2^T\phi_1 = I$ from (6.93), we obtain

$$\begin{aligned} \mathcal{X}_{\mathcal{F}}^{I}\mathcal{X}_{\mathcal{F}} &= (I_{\widehat{n}} \otimes \phi_2)\Big((I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)\Big)^{-1}(I_{\widehat{n}} \otimes \phi_2^T)\mathcal{F}(I_{\widehat{n}} \otimes \phi_2)(I_{\widehat{n}} \otimes \phi_1^T) \\ &= (I_{\widehat{n}} \otimes \phi_2)(I_{\widehat{n}} \otimes \phi_1^T) = I_{\widehat{n}} \otimes \Pi^T. \end{aligned}$$

A similar argument can be given for the other equality. □

Using Lemma 6.27 and properties of the Kronecker product, we observe that the projection matrices $\mathcal{V}$ and $\mathcal{W}$, computed in step 7 of Algorithm 6.4, satisfy:

$$(I_{\widehat{n}} \otimes \Pi)\mathcal{F}(I_{\widehat{n}} \otimes \Pi^T)\operatorname{vec}(\mathcal{V}) = (I_{\widehat{n}} \otimes \Pi)(\widetilde{B}^T \otimes B), \tag{6.98a}$$

$$(I_{\widehat{n}} \otimes \Pi)\mathcal{F}^T(I_{\widehat{n}} \otimes \Pi^T)\operatorname{vec}(\mathcal{W}) = (I_{\widehat{n}} \otimes \Pi)(\widetilde{C}^T \otimes \mathcal{C}), \tag{6.98b}$$

where $\mathcal{F} = -\Big(\Lambda \otimes E_{11} + I_{\widehat{n}} \otimes A_{11} + \sum_{k=1}^{m} \widetilde{N}_k^T \otimes N_k\Big)$. Note that $\Pi_\otimes = I_{\widehat{n}} \otimes \Pi$ is also an oblique projector. It can be verified that $(\Pi_\otimes)^2 = \Pi_\otimes$, $\ker(\Pi_\otimes) = \operatorname{range}(I_{\widehat{n}} \otimes A_{12})$, and $\operatorname{range}(\Pi_\otimes) = \ker\big(I_{\widehat{n}} \otimes A_{12}^T E_{11}^{-1}\big)$. Using these properties, it can be shown that

$$(I_{\widehat{n}} \otimes A_{12}^T)Z = 0 \qquad \text{if and only if} \qquad (I_{\widehat{n}} \otimes \Pi^T)Z = Z. \tag{6.99}$$

In the following lemma, we show a way of circumventing the explicit computation of Π to solve (6.98) for $\operatorname{vec}(\mathcal{V})$ or $\operatorname{vec}(\mathcal{W})$ and reveal the connection between the solutions of (6.98) and saddle point problems.

Algorithm 6.5: B-IRKA for bilinear DAEs, having an index-2 matrix pencil.

```
Input: $E_{11}, A_{11}, N_k, B_1, \mathcal{C}, \mathcal{C}_N^{(k)}$.
Make an initial choice of $\widehat{E}, \widehat{A}, \widehat{N}_k, \widehat{B}, \widehat{C}$.
while no convergence do
    Compute nonsingular matrices $Y$ and $Z$ such that $Y\widehat{A}Z = \Lambda$ and $Y\widehat{E}Z = I_{\hat{n}}$.
    Define $\widetilde{B} = \widehat{B}^T Y^T, \quad \widetilde{C} = \widehat{C}Z \quad \text{and} \quad \widetilde{N}_k = Z^T \widehat{N}_k^T Y^T$.
    Determine the projection matrices $\mathcal{V}$ and $\mathcal{W}$:
        $\begin{bmatrix} \mathcal{F} & I_{\hat{n}} \otimes A_{12} \\ I_{\hat{n}} \otimes A_{12}^T & 0 \end{bmatrix} \begin{bmatrix} \operatorname{vec}(\mathcal{V}) \\ \Gamma \end{bmatrix} = \begin{bmatrix} (\widetilde{B}^T \otimes B)\mathcal{I}_m \\ 0 \end{bmatrix},$
        $\begin{bmatrix} \mathcal{F}^T & I_{\hat{n}} \otimes A_{12} \\ I_{\hat{n}} \otimes A_{12}^T & 0 \end{bmatrix} \begin{bmatrix} \operatorname{vec}(\mathcal{W}) \\ \Delta \end{bmatrix} = \begin{bmatrix} (\widetilde{C}^T \otimes \mathcal{C})\mathcal{I}_p \\ 0 \end{bmatrix},$
        where $\mathcal{F} = -(\Lambda \otimes E_{11} + I_{\hat{n}} \otimes A_{11} + \sum_{k=1}^m \widetilde{N}_k^T \otimes N_k)$.
    Perform: $V = \operatorname{orth}(V)$ and $W = \operatorname{orth}(W)$.
    Compute the reduced-order system matrices:
        $\widehat{E} = \mathcal{W}^T E_{11}\mathcal{V}, \quad \widehat{A} = \mathcal{W}^T A_{11}\mathcal{V}, \quad \widehat{N}_k = \mathcal{W}^T N_k \mathcal{V},$
        $\widehat{B} = \mathcal{W}^T B_1, \quad \widehat{C} = \mathcal{C}\mathcal{V}, \quad \mathcal{C}_N^{(k)} = \widehat{\mathcal{C}}_N^{(k)}\mathcal{V}.$
Output:
    $\widehat{E}^{opt} = \widehat{E}, \ \widehat{A}^{opt} = \widehat{A}, \ \widehat{N}_k^{opt} = \widehat{N}_k, \ \widehat{B}^{opt} = \widehat{B}, \ \widehat{\mathcal{C}}^{opt} = \widehat{\mathcal{C}}, \ \widehat{\mathcal{C}}_N^{(k)opt} = \widehat{\mathcal{C}}_N^{(k)}$.
```

Lemma 6.28:
Consider $Z = (I_{\hat{n}} \otimes \Pi^T)Z$ and $(I_{\hat{n}} \otimes \Pi)\mathcal{F}(I_{\hat{n}} \otimes \Pi^T)Z = (I_{\hat{n}} \otimes \Pi)G$. Then, the matrix Z solves

$$\begin{bmatrix} \mathcal{F} & I_{\hat{n}} \otimes A_{12} \\ I_{\hat{n}} \otimes A_{12}^T & 0 \end{bmatrix} \begin{bmatrix} Z \\ \Xi \end{bmatrix} = \begin{bmatrix} G \\ 0 \end{bmatrix}. \tag{6.100}$$

◊

Proof. Since $Z = (I_{\hat{n}} \otimes \Pi^T)Z$, we have $(I_{\hat{n}} \otimes A_{12}^T)Z = 0$ using the properties of $I_{\hat{n}} \otimes \Pi^T$ as stated in (6.99). This implies that the second block of the equation (6.100) is satisfied.

Moreover, $(I_{\hat{n}} \otimes \Pi)\mathcal{F}Z - (I_{\hat{n}} \otimes \Pi)G = 0$ implies that the columns of $\mathcal{F}Z - G$ lie in $\ker(I_{\hat{n}} \otimes \Pi) = \operatorname{range}(I_{\hat{n}} \otimes A_{12})$. Therefore, there exists Ξ, satisfying $\mathcal{F}Z - G = -(I_{\hat{n}} \otimes A_{12})\Xi$, which is nothing but the first block of the equation (6.100). This concludes the proof. □

Using Lemma 6.28, we thus can determine $\operatorname{vec}(\mathcal{V})$ and $\operatorname{vec}(\mathcal{W})$ without explicitly computing Π by solving the corresponding saddle point problems. All these theoretical analyses give rise to Algorithm 6.5 for model reduction of the system (6.94).

Remark 6.29:
As discussed in [85], the general $B_2 \neq 0$ index-2 problems can be brought back to a problem with $B_2 = 0$ type by decomposing $x_1(t)$ as follows:

$$x_1(t) = x_0(t) + x_u(t), \tag{6.101}$$

where $x_u(t) = -\underbrace{E_{11}^{-1}A_{12}(A_{12}^TE_{11}^{-1}A_{12})B_2}_{\Upsilon}u(t)$ and $x_0(t)$ satisfies $A_{12}^Tx_0(t) = 0$. After doing the algebraic calculations as done for the case $B_2 = 0$ case, we get

$$\Pi E_{11}\Pi^T\dot{x}_0(t) = \Pi A_{11}\Pi^T x_0(t) + \sum_{k=1}^{m} \Pi N_k\Pi^T x_0(t)u_k(t) + \Pi\mathcal{B}\widetilde{u}(t), \tag{6.102a}$$

$$\Pi^T x_0(0) = \Pi^T(x_0 - x_u(0)), \tag{6.102b}$$

$$y(t) = \mathcal{C}\Pi^T x_0(t) + \sum_{k=1}^{m} \mathcal{C}_N^{(k)}\Pi^T x_0(t)u_k(t) + \mathcal{D}\widetilde{u}(t) - C_2(A_{12}^TE_{11}^{-1}A_{12})^{-1}B_2\dot{u}(t), \tag{6.102c}$$

where

$$\begin{aligned}
&\mathcal{B} = [\mathcal{B}_1, \mathcal{B}_u^{(1)}, \dots, \mathcal{B}_u^{(m)}] \text{ with } \mathcal{B}_u^{(k)} = -N_i\Upsilon, && \widetilde{u}(t) = \Big([1, u(t)^T] \otimes u(t)^T\Big)^T,\\
&\mathcal{C} = C_1 - C_2(A_{12}^TE_{11}^{-1}A_{12})^{-1}A_{11}, && \mathcal{C}_N^{(k)} = -C_2(A_{12}^TE_{11}^{-1}A_{12})^{-1}A_{12}^TE_{11}^{-1}N_k,\\
&\mathcal{D} = \Big[D - C_1\Upsilon - C_2(A_{12}^TE_{11}^{-1}A_{12})^{-1}A_{12}^TE_{11}^{-1}\mathcal{B}_1,\\
&\qquad\qquad C_2(A_{12}^TE_{11}^{-1}A_{12})^{-1}A_{12}^TE_{11}^{-1}[B_u^{(1)}, \dots, B_u^{(m)}]\Big].
\end{aligned}$$

Although the system (6.102) has terms associated with u, $u{\cdot}u_k$ which are functions of eventually u, but we treat them as different inputs of the system as far as a model reduction problem is concerned. Now, it can be easily seen that determining reduced-order systems of the system (6.102) is analogous to the system (6.92). Therefore, Algorithm 6.5 can be readily applied to the system (6.102) to obtain locally $\mathcal{H}_2$-optimal reduced-order systems, having neglected bilinear terms in the output equation. ♢

Remark 6.30:
So far in the analysis, we have assumed that $A_{12} = A_{21}^T$ and E_{11} is symmetric. However, the similar analysis can be carried out if $A_{12} \neq A_{21}^T$ and $E_{11} \neq E_{11}^T$. In such a case, the resulting algorithm would be similar to Algorithm 6.5, and the only differences will occur in steps 6 and 7, determining the projections V and W. These steps modify as follows in case $A_{12} \neq A_{21}^T$ and $E_{11} \neq E_{11}^T$:

$$\begin{bmatrix} \mathcal{F} & I_{\widehat{n}} \otimes A_{12} \\ I_{\widehat{n}} \otimes A_{21} & 0 \end{bmatrix} \begin{bmatrix} \operatorname{vec}(\mathcal{V}) \\ \Gamma \end{bmatrix} = \begin{bmatrix} (\widetilde{B}^T \otimes B)\mathcal{I}_m \\ 0 \end{bmatrix},$$

$$\begin{bmatrix} \mathcal{F}^T & I_{\widehat{n}} \otimes A_{21} \\ I_{\widehat{n}} \otimes A_{12}^T & 0 \end{bmatrix} \begin{bmatrix} \operatorname{vec}(\mathcal{W}) \\ \Delta \end{bmatrix} = \begin{bmatrix} (\widetilde{C}^T \otimes \mathcal{C})\mathcal{I}_p \\ 0 \end{bmatrix},$$

where $\mathcal{F} = -(\Lambda \otimes E_{11} + I_{\widehat{n}} \otimes A_{11} + \sum_{k=1}^{m} \widetilde{N}_k^T \otimes N_k)$. ♢

6.5.3. Numerical experiments

In this subsection, we investigate the efficiency of the proposed iterative algorithm for bilinear DAEs, having an index-2 matrix pencil and compare the quality of the determined reduced-order systems with the ones obtained by using the projection matrices determined by linear IRKA [80, Algo. 6.2]. The stopping criterion for Algorithm 6.5 is based on the relative change in the eigenvalues of the reduced-order system. If this change is below the square root of the machine precision, then the iteration is stopped. We randomly select the initial guess of the reduced-order matrices in Algorithm 6.5 and also choose a scaling factor γ as suggested in [21] for a smooth convergence of B-IRKA. All the simulations are done on a board with 4 Intel® Xeon®E7-8837 CPUs with a 2.67-GHz clock speed using MATLAB 8.0.0.783 (R2012b).

A nonlinear RC circuit

Here, we consider a variant of the constraint transmission line circuit as shown in Figure 6.8 (see the beginning of the section), where it is assumed that the voltages at the first and last nodes are the same. The electrical component, i.e., I-V diode, has nonlinear characteristics $g(v_D) = e^{40v_D} + v_D - 1$, where v_D is the voltage across the node. Using Kirchhoff's current law at each node, we get the following set of equations:

$$\begin{aligned}\dot{v}_1 &= -2v_1 + v_2 + 2 - e^{40v_1} - e^{40(v_1-v_2)} + u(t),\\ \dot{v}_i &= -2v_i + v_{i-1} + v_{i+1} + e^{40(v_{i-1}-v_i)} - e^{40(v_i-v_{i+1})}, \quad 1 < i < \widetilde{n},\\ \dot{v}_{\widetilde{n}} &= -v_{\widetilde{n}} + v_{\widetilde{n}-1} - 1 + e^{40(v_{\widetilde{n}-1}-v_{\widetilde{n}})}\end{aligned} \tag{6.103}$$

with a constraint

$$0 = v_1 - v_{\widetilde{n}}.$$

The system of equations (6.103) can be written as a quadratic-bilinear DAE by appropriately introducing the new state variables, as shown in [78] for a nonlinear RC circuit example and in Subsection 6.3.6. The dynamics of the system, in the state-space representation, is given as follows:

$$\begin{aligned}\dot{x}(t) &= Ax(t) + G\lambda + Hx(t) \otimes x(t) + Nx(t)u(t) + Bu(t),\\ 0 &= G^T x(t),\end{aligned}$$

where $x(t) \in \mathbb{R}^{\widetilde{n}}$ and λ are state vectors that contain voltages at each node and an appropriate Lagrangian multiplier, respectively. We observe the voltage at the first node. Since we have only one constraint in the system dynamics, this allows us to employ Carleman bilinearization to obtain an approximate bilinearized DAE [71]. We set $\widetilde{n} = 15$, leading to a bilinearized system of order $n = 2 \cdot \widetilde{n} + 4 \cdot \widetilde{n}^2 + 1 = 961$. We apply the $\mathcal{H}_2$-optimal model reduction method (Algorithm 6.5) by setting the order of

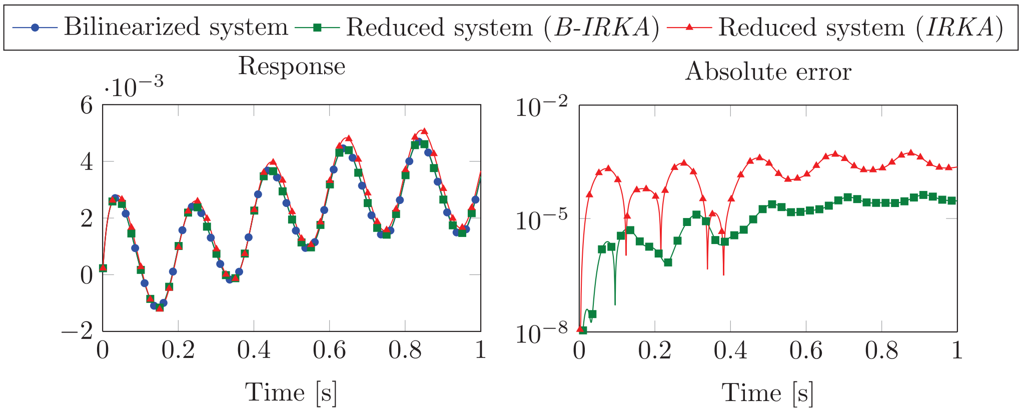


Figure 6.9.: A constraint nonlinear RC ladder: comparison of the transient response of the systems for an input $u(t) = (\sin(10\pi t) + 1)/2$.

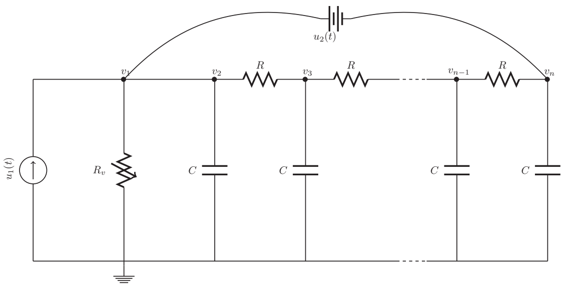


Figure 6.10.: A constraint RC-circuit diagram.

a reduced-order system to $r = 10$. We choose the scaling factor $\gamma = 0.01$ in order to achieve convergence of B-IRKA. We also determine a reduced-order system by using linear IRKA of the same order. In Figure 6.9, we compare the quality of the reduced-order systems with the original system by computing transient responses for an input $u(t) = (\sin(10\pi t) + 1)/2$.

We observe that the reduced-order system obtained by modified B-IRKA captures the dynamics of the system better as compared to the reduced-order system obtained by linear IRKA.

Resistance-varying RC circuit

As our second example, we consider the RC circuit as shown in Figure 6.10 in which the ith node is connected to the $(i-1)$st and the $(i+1)$st nodes via resistances, and connected to the ground via capacitors. Moreover, the first node is connected to the ground via a variable resistance, and the voltage at the first node is influenced by the current (the input u_1). We also add an extra control u_2, controlling the voltage difference between the first and last nodes. Now, we apply Kirchhoff's current law at each node to obtain the following set of ODEs:

$$\begin{aligned} C\dot{v}_1(t) &= \tfrac{1}{R}(-v_1+v_2)+\tfrac{1}{R_v}(0-v_1)+u_1(t),\\ C\dot{v}_i(t) &= \tfrac{1}{R}(-2v_i+v_{i-1}+v_{i+1}), \qquad (2\le i\le n-1),\\ C\dot{v}_n(t) &= \tfrac{1}{R}(-v_n+v_{n-1}) \end{aligned}$$

along with a constraint

$$0 = v_1 - v_n - u_2(t).$$

We set all the capacitors (C) and the constant resistance (R) equal to 1, and consider that the variable resistance R_v varies with respect to the parameter δ as follows:

$$R_v = \frac{R}{1+\delta}.$$

Combining all these equations together, we obtain the dynamics of the RC circuit which are described by the following DAE:

$$\dot{x}(t) = Ax(t) + G^T\lambda(t) + \delta N x(t) + B_1u_1(t), \tag{6.104a}$$
$$0 = Gx(t) + B_2u_2(t), \tag{6.104b}$$
$$y(t) = Cx(t), \tag{6.104c}$$

where $x(t)\in\mathbb{R}^n$ is the state vector containing the voltage at each node, $\lambda\in\mathbb{R}$ is the Lagrange multiplier, $G=[1,0,\dots,0,-1]$ is a constraint matrix, $B_1=[1,0,\dots,0]$ and $B_2=1$. The voltage at the second node is the output of interest, thus yielding $C=[0,1,0,\dots,0]$. For this example, we first transform the system (6.104) into an equivalent system with $\mathcal{B}_2=0$, leading to the following system:

$$\dot{\widetilde{x}}(t) = A\widetilde{x}(t) + \delta N\widetilde{x}(t) + G^T\lambda(t) + \mathcal{B}\widetilde{u}(t), \tag{6.105a}$$
$$0 = G\widetilde{x}(t), \tag{6.105b}$$
$$y(t) = C\widetilde{x}(t) + D\widetilde{u}(t), \tag{6.105c}$$

where $\mathcal{B}=[B_1,\ A\mathcal{G},\ N\mathcal{G}]$ and $D=[0,\ C\mathcal{G},\ 0]$ in which $\mathcal{G}=-G^T(GG^T)^{-1}B_2$, and $\widetilde{u}(t)=[u_1(t),\ u_2(t),\ \delta u_2(t)]$. Now, the system (6.105) can be seen as a linear parameter-varying system in the parameter δ. It is shown in [20] that the special class of parametric systems is closely related to bilinear systems. Therefore, we reformulate the linear system (6.105) appropriately as a bilinear system with four inputs and one output as follows:

$$\dot{\widetilde{x}}(t) = A\widetilde{x}(t) + \sum_{i=1}^{4} N_i\widetilde{x}(t)u_i(t) + G^T\lambda(t) + \mathcal{B}_b\widetilde{u}_b(t), \tag{6.106a}$$
$$0 = G\widetilde{x}(t), \tag{6.106b}$$
$$y(t) = C\widetilde{x}(t) + D_b\widetilde{u}_b(t), \tag{6.106c}$$

where $\left[N_1, N_2, N_3, N_4\right] = \left[0, 0, 0, N\right]$, $\mathcal{B}_b = [\mathcal{B}, 0]$ and $D_b = [D, 0]$ with inputs $\widetilde{u}_b(t) = [\widetilde{u}^T(t), \delta]^T$. We consider $n = 1000$, leading to the order of the system (6.106) $\widetilde{n} = 1001$. Next, we determine reduced bilinear systems of order $r = 15$ by employing Algorithm 6.5 and by using linear IRKA. These reduced bilinear systems again can be rewritten into reduced linear parametric systems. This allows us to determine the quality of the reduced-order systems by comparing the relative H_∞-norm of the error system, i.e., $\frac{\|H-\widehat{H}\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$ by varying parameter values δ as shown in Figure 6.11.

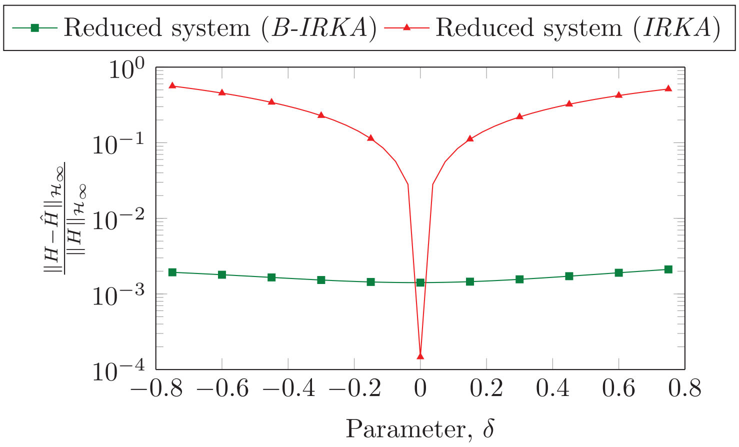


Figure 6.11.: A resistor varying RC circuit: relative H_∞ error versus the parameter δ for the reduced linear parametric systems obtained from B-IRKA and IRKA.

We observe that the reduced parametric system obtained from B-IRKA captures the dynamics of the original system for a wide parameter range much better as compared to the one obtained by using linear IRKA. However, one can see the drop in the relative H_∞ error in Figure 6.11. This is due to an apparent reason that the projection matrices obtained by employing IRKA capture the dynamics of the system quite accurately for $\delta = 0$, but fail to capture the dynamics of the system as the parameter δ moves away from $\delta = 0$. On the contrary, the reduced parametric system obtained from B-IRKA performs quite well over a wide parameter range.

Parameter dependant RLC circuit

Lastly, we consider an RLC circuit as shown in Figure 6.12. The governing equations of the RLC circuit can be written as follows:

$$\begin{aligned} C\tfrac{d}{dt}v_j &= i_j - i_{j-1}, \quad && j \in \{1, \dots, g-1\}, \\ C\tfrac{d}{dt}v_g &= i_g, && \\ L\tfrac{d}{dt}i_j + Ri_j &= v_{j-1} - v_j \quad && j \in \{2, \dots, g\}, \end{aligned}$$

where v_j and i_j are the voltage at the jth node and the current passing through the $(j-1)$st inductor, respectively. Also, $V(t)$ is a control voltage source of the system and i_1 is the current passing through this voltage source. Since the voltage source is connected to the first node via ground, this leads to a constraint $0 = v_1 - V(t)$. We set all capacitors and inductors to 1, and consider variable resistances, depending linearly on the parameter p as follows:

$$R = 1 + p.$$

With these relations, we can write the system in the state-space form as:

$$\begin{aligned} \frac{d}{dt}x_1(t) &= A_{11}x_1(t) + A_{12}x_2(t) + pNx_1(t), \\ 0 &= A_{12}^T x_1(t) + B_2 u(t), \end{aligned} \tag{6.107}$$

where $x_1(t)$ contains the voltages at each node and the currents passing through each inductor, and $x_2(t)$ contains the current through the voltage source. The voltage at the last node is observed. We choose $g = 500$ which results in the order of the system (6.107) $n = 1001$. As a first step, we convert system (6.107) to an equivalent system by using an appropriate change of the state variable so that the constraint equation becomes independent of the input, leading to the following system:

$$\begin{aligned} \frac{d}{dt}\widetilde{x}_1(t) &= A_{11}\widetilde{x}_1(t) + \widetilde{A}_{12}x_2(t) + pN\widetilde{x}_1(t) + \widetilde{B}u(t), \\ 0 &= A_{12}^T\widetilde{x}_1(t), \\ y(t) &= C_1\widetilde{x}_1(t). \end{aligned}$$

Next, we treat the above system as a bilinear system by considering the parameter p as an input to the system. We determine reduced bilinear systems of order $r = 10$, by employing the proposed B-IRKA and IRKA and then convert back to have linear parametric reduced-order systems. In order to compare the quality of the reduced-order systems, we plot the relative H_∞-norm of the error system in Figure 6.13.

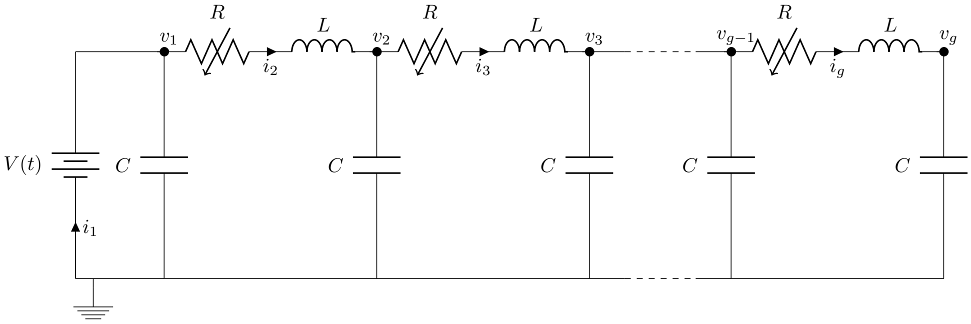


Figure 6.12.: A variable resistance RLC circuit diagram.

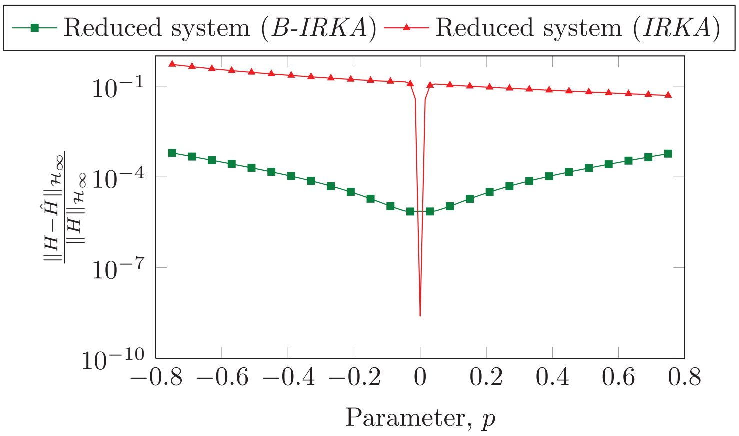


Figure 6.13.: A parametric varying RLC circuit: relative H_∞ error versus the parameter p for the reduced linear parametric systems obtained from B-IRKA and IRKA.

A similar phenomenon, as observed in the previous example can be seen in Figure 6.13, in particular, a drop in the relative H_∞ error for IRKA at $p = 0$. Nevertheless, the reduced-order system, obtained by using B-IRKA, outperforms the one obtained by using IRKA for a wide range of the parameter.

6.5.4. Conclusions

In this section, we have proposed an iterative algorithm for MOR of the structured bilinear DAEs, having an index-2 matrix pencil. This gives rise to locally $\mathcal{H}_2$-optimal reduced-order systems on convergence, if it converges. For this, we have transformed the original bilinear DAE into an equivalent bilinear ODE system by means of projectors. This enabled us to employ bilinear iterative rational Krylov algorithm B-IRKA. Next, in the view of implementation, we have proposed a modified B-IRKA which does not require the undesirable explicit computation of the spectral projector in order to compute reduced-order systems. Finally, we have illustrated the efficiency of the proposed B-IRKA using various constraint electrical circuit examples, showing that reduced-order systems, obtained by using modified B-IRKA, replicate the dynamics of the original system much better as compared to reduced-order systems obtained by using linear IRKA.

6.6. Outlook

Summarizing, in this chapter, we have studied interpolation-based model reduction techniques for special structured bilinear DAEs while paying a particular attention to their polynomial parts. Precisely, we have investigated subsystem interpolation, multipoint Volterra series interpolation and $\mathcal{H}_2$-optimal model order reduction problem for bilinear DAEs, having index-1 matrix pencil $\lambda E - A$. Moreover, we have also studied $\mathcal{H}_2$-optimal model reduction problem for bilinear DAEs, having index-2 matrix pencil.

However, there are still ample open problems to investigate in the future. One follow up question is to extend index-2 bilinear DAEs methodology to the specially structured bilinear DAEs, having index-3 matrix pencil $\lambda E - A$ as done for the linear index-3 DAEs in [1]. These systems certainly appear in modeling of constraint mechanical systems. Moreover, model reduction for bilinear DAEs, having matrix pencil $\lambda E - A$ of a general index ν remains an open problem as a further research topic in this area. In addition to these, an extension of a truncated $\mathcal{H}_2$-optimal model reduction problem [59, 60] for bilinear DAEs would certainly be an attractive problem as well. Balancing-based model reduction for bilinear descriptor systems would also be a useful contribution to model reduction for bilinear DAEs.

CHAPTER 7 CONCLUSIONS AND FUTURE PERSPECTIVES

Contents

7.1. Conclusions

In this thesis, we have investigated model order reduction problems for large-scale nonlinear dynamical control systems. Precisely, we have studied two important classes of the nonlinear systems: the first class of nonlinear systems is bilinear systems which act a bridge between linear systems and fully nonlinear systems, and the second class is quadratic-bilinear (QB) systems, which cover a large class of smooth nonlinear systems; this is due to the fact that a smooth nonlinear affine system, involving combinations of elementary functions like exponential, trigonometric, and polynomial functions, etc., can be rewritten in the QB form. In the past decades, results of balanced truncation and rational interpolation, and their extension were mainly studied for model reduction of linear systems. In recent years, a lot of attention has been paid to model reduction for bilinear systems and to extend the existing ideas for linear systems, and a little attention has been given to QB systems so far. The focus of this thesis thus has been mainly twofold. First was to study balanced truncation model reduction for bilinear control systems [26, 73]. We have also extended rational interpolation-based model reduction for bilinear ODEs, [59, 12, 22, 21, 40] to bilinear systems subject to algebraic constraints. These results have been theoretically explained, and efficiencies of the resulting numerical algorithms have been illustrated by means of several numerical examples. The second main goal was to study balanced truncation and optimal interpolation model reduction techniques to QB system. For this, we have extended the idea of algebraic Gramians for bilinear systems, e.g., [26, 73] to QB systems. We have further investigated $\mathcal{H}_2$-optimal model reduction framework for QB systems by means of tools from tensor theory. We have provided theoretical background of the results and their interpretations, and the results have been verified by means of several numerical simulations.

In Chapter 3, we have investigated widely studied balanced truncation method for bilinear systems, e.g., see [26, 73]. We have provided concrete conditions under which the bounds for energy functionals hold that are given in terms of certain quadratic form of the algebraic Gramians for bilinear systems. Furthermore, we have introduced a notion of *truncated* Gramians for bilinear systems and have given energy functional interpretations of these truncated Gramians. By using a couple of numerical examples, we have also illustrated the benefits of truncated Gramians in the model order reduction framework.

In Chapter 4, we have extended balancing-type model reduction method for more general nonlinear systems, namely, QB control systems. For this, we have first derived the input-output mapping, the so-called Volterra series for QB systems, allowing us to propose algebraic Gramians for the latter systems. We have further provided truncated Gramians concept for QB systems as well. We have provided connections between controllability/observability energy functionals and algebraic Gramians for QB systems. Based on these results, we have proposed a balancing square root algorithm to determine the reduced order systems. Furthermore, we have discussed the computational issues and have studied the Lyapunov stability of the obtained reduced-order systems. Finally, we have illustrated balancing-type model reduction by means of various nonlinear semi-discretized PDEs and have shown the out-performance of the proposed method against the existing interpolation-based techniques [25, 78].

In Chapter 5, we have aimed at extending the ideas of interpolation-based $\mathcal{H}_2$-optimal model reduction for QB control systems. For this purpose, we have first proposed the $\mathcal{H}_2$ measure based on the *kernels* for the underlying Volterra series of QB systems. Furthermore, we have provided a truncated $\mathcal{H}_2$-norm for QB systems which is simpler, and an explicit expression of it can be easily given. We then have shown how to generalize the $\mathcal{H}_2$ optimality interpolation conditions from [21, 59] to QB systems by means of some basic tools from tensor theory. This allowed us to proposed an iterative scheme (TQB-IRKA), upon convergence, constructing reduced-order systems that satisfy the optimality conditions approximately at modest cost. Lastly, we have illustrated the efficiency of the proposed methods by means of several numerical examples and have shown its superiority to the common reduction method for nonlinear systems, proper orthogonal decomposition (POD) and interpolation-based methods and its competitiveness with respect to balanced truncation for QB systems.

We stress that the main advantages of both TQB-IRKA and balanced truncation methods are that these methods are firsly input-independent unlike POD method and reduced-order systems due to these methods are constructed in an automatic fashion as opposed to interpolation-based methods in, e.g., [25, 78].

In all previous chapters, we have focused on ODE nonlinear systems. In Chapter 6, we have investigated interpolation-based MOR for bilinear systems which are subject to algebraic constraints as well. Such systems are referred to as bilinear descriptor systems, or bilinear DAEs. For model reduction of DAEs, the polynomial part also plays a crucial role along with interpolation. Thus, we have aimed at extending subsystem interpolation and multi-point Volterra series interpolation for bilinear DAEs, having index-1 matrix pencil $\lambda E - A$ and proposed modified interpolation conditions, allowing us to retain the polynomial of the systems. Furthermore, we have also investigated the problem of $\mathcal{H}_2$-optimal interpolation for bilinear DAEs, having index-1 and index-2. This allowed us to propose modified versions of the bilinear iterative Krylov algorithms, leading to locally $\mathcal{H}_2$-optimal reduced-order systems *upon convergence*. We have illustrated the efficiency of these proposed methods by means of various numerical examples.

7.2. Future Research Perspectives

Even though we have discussed several aspects of bilinear and QB systems by extending the existing concepts for linear/bilinear systems, there are still many open questions and problems which are worthwhile to investigate in future.

Although we have seen in Chapter 3 that balanced truncation method for bilinear systems produces faithful reduced-order systems, yet an important problem is how to quantify the error between the original and reduced-order systems due to the truncation. Furthermore, an extension of balanced truncation for bilinear systems to descriptor systems still remains an open problem.

In Chapter 4, we have proposed algebraic Gramians for quadratic-bilinear systems and have shown their usage in MOR for QB systems. However, the Gramians solve quadratic Lyapunov equations, which are hard to solve. Therefore, it is very important to develop efficient numerical algorithms to determine low-rank factors of these Lyapunov equations. Moreover, the above open questions for bilinear systems also hold for quadratic-bilinear systems, such as an error bound. Moreover, there are some applications where one might be interested in constructing reduced-order systems which capture the system dynamics between time interval $[0, T]$, where $T < \infty$; therefore, it will be nice to extend time-limited balanced truncation for linear systems, e.g.,[33] and bilinear systems [117] to quadratic-bilinear systems. Furthermore, as we know that bilinear systems have a close relation with a particular class of parametric systems, likewise it would be interesting to study the application of quadratic-bilinear systems in a special class of linear parametric system, for example, when the number of parameters are huge or the system matrices change with respect to the state vectors.

In Chapter 5, we have studied an optimal interpolation-based model reduction technique for QB systems, and as a result, we have proposed an iterative method (TQB-

IRKA). However, the main bottleneck in applying the proposed TQB-IRKA for QB systems is that it requires the computations related to Kronecker products such as $H(V \otimes V)\widetilde{H}^T$, which really slows down the iteration process. Although we have some approaches to compute these terms efficiently, it would be useful to come up with a scheme which allows us to approximate the Hessian as follows:

$$H \approx \sum_{k=1}^{l} \mathcal{A}_k \otimes \mathcal{B}_k;$$

where $\mathcal{A}_k$ and $\mathcal{B}_k$ are matrices of appropriate sizes which would allow us to compute the desired terms very efficiently and fast. Furthermore, for a given order of a reduced-order system, TQB-IRKA upon convergence provides us a reduced-order system, satisfying the optimal conditions approximately; however, it is hard to quantify the quality of the obtained reduced-order system. Therefore, it is important to derive some error estimate, allowing to choose an appropriate order of reduced-order systems. Moreover, a good initial selection of interpolation points and tangential direction can reduce the number of iterations taken for the algorithm to converge. In addition to these, in the $\mathcal{H}_2$-optimal framework for QB systems, we have derived first-order optimal conditions by defining a truncated $\mathcal{H}_2$-norm based on the leading three terms of the Volterra series; however, it would be interesting to consider the higher-order terms to define another truncated $\mathcal{H}_2$-norm and then derive the optimal conditions and compare the quality of the reduced-order systems.

In Chapter 6, we have aimed at extending the interpolation concepts including $\mathcal{H}_2$-optimal interpolation for bilinear ODEs to bilinear DAEs by considering the special structures of the matrix pencils. However, for a given general bilinear DAEs, this is yet an open problem. Furthermore, in order to develop interpolation-based model reduction techniques for bilinear DAEs, we have aimed at utilizing the structure of the pencil matrix $\lambda E - A$, rather than using a proper index of a bilinear system such as differentiation index. Therefore, as a future topic, it would be interesting to study index concepts for bilinear DAEs which can be easily coupled with model reduction problem for bilinear DAEs. What is more, the truncated $\mathcal{H}_2$-optimal interpolation idea would also be an appealing extension to bilinear DAEs.

Beside these, as we have noted, smooth nonlinear systems, containing mono-variate functions, can be transformed into QB systems. Such a transformation is exact but not unique. Therefore, a minimal or optimal transformation of nonlinear systems in a QB form and automatic generation of the transformed systems would be desirable. Furthermore, extensions of balancing-based and interpolation-based optimal model reduction techniques for QB ODEs to descriptor systems would be important as well, due to vast applications, e.g., in flow problems. Last but not least, balancing-type and interpolation-based model reduction for more general nonlinear systems, having, for example, rational terms, higher-order polynomial without rewriting them into a QB form, would also be a significant contribution to nonlinear model order reduction.

Appendices

A. A convergence result

Lemma A.1:
Consider a recurrence formula as follows:

$$x_{k+1} = F(x_k), \quad \forall \quad k \geq 1, \tag{A.1}$$

where $F(x) = ax^2 + bx + c$ and a, b, c are real positive scaler numbers. Moreover, assume that $x_1 = c$. Then, $\lim_{k\to\infty} x_k =: x^*$ is finite if

$$b < 1, \quad \text{and} \tag{A.2a}$$
$$1 > (b-1)^2 - 4ac > 0. \tag{A.2b}$$

Furthermore, x^* is given by the smaller root of the the following quadratic equation:

$$ax^2 + (b-1)x + c = 0, \quad \text{i.e.,}$$

$$x^* = \frac{-(b-1) - \sqrt{(b-1)^2 - 4ac}}{2a}. \tag{A.3}$$

◊

Proof. First, note that the sequence (A.1) contains only real positive numbers. Thus, the equilibrium point must also be a real positive number. Furthermore, the equilibrium points solve the quadratic equation $F(x) - x = 0$, and we denote these equilibrium points by $x^{(1)}$ and $x^{(2)}$ with $x^{(1)} \leq x^{(2)}$. Since a, b and c all are positive, both equilibrium points either can be positive or negative depending on the value of b. To ensure the equilibrium points being positive, the minima of $F(x) - x$ must lie in the right half plane; thus, $b - 1 < 0$, leading to the condition (A.2a).

Furthermore, we consider the derivative of $F(x)$, that is, $F'(x) := 2ax + b$. Since $F'(x)$ is an increasing function and $F'(x) \geq 0 \; \forall x \in [c, x^{(1)}]$, we have for $y \in [c, x^{(1)}]$:

$$\begin{aligned} F'(y) &\leq F'(x^{(1)}) \\ &\leq 2ax^{(1)} + b = 2a\left(\frac{-(b-1) - \sqrt{(b-1)^2 - 4ac}}{2a}\right) + b \leq 1 - \sqrt{(b-1)^2 - 4ac}. \end{aligned}$$

Assuming $1 > (b-1)^2 - 4ac > 0$, we have $F'(y) < 1$, $\forall y \in [c, x^{(1)}]$. Thus, by Banach fix-point theorem, $F(x)$ is a contraction on $[c, x^{(1)}]$, and the fixed point is given by $x^{(1)}$. □

B. Important relations of the Kronecker products

In this section, we provide some relations between Kronecker products, which will simplify necessary conditions for $\mathcal{H}_2$-optimality of QB systems.

Lemma B.1 ([21]):
Consider $f(x) \in \mathbb{R}^{s\times n}$, $A(y) \in \mathbb{R}^{n\times n}$, $G \in \mathbb{R}^{n\times q}$ with $x, y \in \mathbb{R}$ and let $\mathcal{L}(y)$ be defined as

$$\mathcal{L}(y) = -A(y) \otimes I_n - I_n \otimes A(y).$$

If the functions f and A are differentiable with respect to x and y, respectively, then

$$\frac{\partial}{\partial x}\left[(\mathcal{I}_s)^T \left(f(x) \otimes f(x)\right) \mathcal{L}^{-1}(y)(G \otimes G)\mathcal{I}_q\right] \\ = 2(\mathcal{I}_s)^T \left(\left(\frac{\partial}{\partial x} f(x)\right) \otimes f(x)\right) \mathcal{L}^{-1}(y)(G \otimes G)\mathcal{I}_q.$$

Moreover, let $X, Y \in \mathbb{R}^{n\times n}$ be symmetric matrices. Then,

$$\frac{\partial}{\partial y}\left[\operatorname{vec}(X)^T \mathcal{L}^{-1}(y) \operatorname{vec}(Y)\right] = 2 \cdot \operatorname{vec}(X)^T \mathcal{L}^{-1}(y) \left(\frac{\partial}{\partial y} A(y) \otimes I_n\right) \mathcal{L}^{-1}(y) \operatorname{vec}(Y)\Diamond$$

Lemma B.2:
Let $\mathcal{F}$, $\widehat{\mathcal{F}}$ be defined as follows:

$$\mathcal{F} = \begin{bmatrix} I_n & 0 \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \end{bmatrix} \quad \text{and} \quad \widehat{\mathcal{F}} = \begin{bmatrix} 0 & I_r \end{bmatrix} \otimes \begin{bmatrix} 0 & I_r \end{bmatrix},$$

and consider a permutation matrix

$$M = \begin{bmatrix} M_{nnr} & 0 \\ 0 & M_{rnr} \end{bmatrix}, \tag{B.1}$$

where M_{pqr} is defined in (5.28). Moreover, let the two column vectors x and y be partitioned as

$$x = \begin{bmatrix} x_1^T & x_2^T & x_3^T & x_4^T \end{bmatrix}^T \quad \text{and} \quad y = \begin{bmatrix} y_1^T & y_2^T & y_3^T & y_4^T \end{bmatrix}^T,$$

where $x_1, y_1 \in \mathbb{R}^{n^2}$, $x_{\{2,3\}}, y_{\{2,3\}} \in \mathbb{R}^{nr}$, and $x_4, y_4 \in \mathbb{R}^{r^2}$. Then, the following relations hold:

$$(\widehat{\mathcal{F}} \otimes \mathcal{F})T_{(n+r,n+r)}(M \otimes M)(x \otimes y) = T_{(n,r)}(x_3 \otimes y_3), \tag{B.2}$$

$$(\widehat{\mathcal{F}} \otimes \widehat{\mathcal{F}})T_{(n+r,n+r)}(M \otimes M)(x \otimes y) = T_{(r,r)}(x_4 \otimes y_4), \tag{B.3}$$

where $T_{(n,m)}$ is also a permutation matrix given by

$$T_{(n,m)} = I_m \otimes \left[I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n\right] \otimes I_n. \qquad \Diamond$$

Proof. Let us begin by considering the following equation:

$$(\widehat{\mathcal{F}}\otimes\mathcal{F})T_{(n+r,n+r)} = \begin{bmatrix}0 & I_r\otimes\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\end{bmatrix}\left(I_{n+r}\otimes\mathcal{G}\right),$$

where $\mathcal{G} = \begin{bmatrix}I_{n+r}\otimes e_1^{n+r}, \dots, I_{n+r}\otimes e_{n+r}^{n+r}\end{bmatrix}\otimes I_{n+r}$. Next, we split I_{n+r} as $I_{n+r} = \begin{bmatrix}I_n & 0\\ 0 & I_r\end{bmatrix}$, leading to

$$\begin{aligned}(\widehat{\mathcal{F}}\otimes\mathcal{F})T_{(n+r,n+r)} &= \begin{bmatrix}0 & I_r\otimes\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\end{bmatrix}\begin{bmatrix}I_n\otimes\mathcal{G} & 0\\ 0 & I_r\otimes\mathcal{G}\end{bmatrix}\\ &= \begin{bmatrix}0 & \left(I_r\otimes\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\right)\left(I_r\otimes\mathcal{G}\right)\end{bmatrix}\\ &= \begin{bmatrix}0 & I_r\otimes\Big(\left(\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\right)\mathcal{G}\Big)\end{bmatrix}.\end{aligned}\tag{B.4}$$

Now, we investigate the following equation, which is a part of the above equation:

$$\left(\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\right)\mathcal{G}_i =: \mathcal{L}_i,$$

where $\mathcal{G}_i$ is ith block column of the matrix $\mathcal{G}$ given by $\mathcal{G}_i = I_{n+r}\otimes e_i^{n+r}\otimes I_{n+r}$. This yields

$$\begin{aligned}\mathcal{L}_i &= \left(\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\right)\left(I_{n+r}\otimes e_i^{n+r}\otimes I_{n+r}\right)\\ &= \left(\begin{bmatrix}0 & I_r\end{bmatrix}I_{n+r}\right)\otimes\left(\mathcal{F}(e_i^{n+r}\otimes I_{n+r})\right) = \begin{bmatrix}0 & I_r\end{bmatrix}\otimes\left(\mathcal{F}(e_i^{n+r}\otimes I_{n+r})\right).\end{aligned}$$

Assuming that $1\leq i\leq n$, we can write $\mathcal{L}_i$ as

$$\begin{aligned}\mathcal{L}_i &= \begin{bmatrix}0 & I_r\end{bmatrix}\otimes\left(\mathcal{F}\left(\begin{bmatrix}e_i^n\\ 0\end{bmatrix}\otimes I_{n+r}\right)\right) = \begin{bmatrix}0 & I_r\end{bmatrix}\otimes\left(\begin{bmatrix}I_n\otimes\begin{bmatrix}I_n & 0\end{bmatrix} & 0\end{bmatrix}\begin{bmatrix}e_i^n\otimes I_{n+r}\\ 0\end{bmatrix}\right)\\ &= \begin{bmatrix}0 & I_r\end{bmatrix}\otimes\begin{bmatrix}e_i^n\otimes\begin{bmatrix}I_n & 0\end{bmatrix}\end{bmatrix} = \begin{bmatrix}0 & I_r\otimes\left(e_i^n\otimes\begin{bmatrix}I_n & 0\end{bmatrix}\right)\end{bmatrix}.\end{aligned}$$

Subsequently, we assume $n+r\geq i>n$, which leads to

$$\begin{aligned}\mathcal{L}_i &= \begin{bmatrix}0 & I_r\end{bmatrix}\otimes\left(\mathcal{F}\left(\begin{bmatrix}0\\ e_{i-n}^r\end{bmatrix}\otimes I_{n+r}\right)\right)\\ &= \begin{bmatrix}0 & I_r\end{bmatrix}\otimes\left(\begin{bmatrix}I_n\otimes\begin{bmatrix}I_n & 0\end{bmatrix} & 0\end{bmatrix}\begin{bmatrix}0\\ e_{i-n}^r\otimes I_{n+r}\end{bmatrix}\right) = 0.\end{aligned}$$

Thus,

$$\left(\begin{bmatrix}0 & I_r\end{bmatrix}\otimes\mathcal{F}\right)\mathcal{G} = \begin{bmatrix}\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_n, 0\end{bmatrix} =: \mathcal{L}.$$

Inserting the above expression in (B.4) yields

$$(\widehat{\mathcal{F}}\otimes\mathcal{F})T_{(n+r,n+r)} = \begin{bmatrix}0 & I_r\otimes\mathcal{L}\end{bmatrix}.$$

Now, we are ready to investigate the following term:

$$\begin{aligned}(\widehat{\mathscr{F}}\otimes\mathscr{F})T_{(n+r,n+r)}(M\otimes M) &= \begin{bmatrix}0 & I_r\otimes\mathscr{L}\end{bmatrix}\begin{bmatrix}M_{nnr}\otimes M & 0\\ 0 & M_{rnr}\otimes M\end{bmatrix}\\ &= \begin{bmatrix}0 & I_r\otimes\mathscr{L}\end{bmatrix}\begin{bmatrix}M_{nnr}\otimes M & 0\\ 0 & M_{rnr}\otimes M\end{bmatrix}\\ &= \begin{bmatrix}0 & (I_r\otimes\mathscr{L})\left(M_{rnr}\otimes M\right)\end{bmatrix}.\end{aligned}$$

Further, we consider the second block column of the above relation and substitute for M_{nnr} and M_{rnr} using (5.28) to get

$$\begin{aligned}(I_r\otimes\mathscr{L})\left(M_{rnr}\otimes M\right) &= (I_r\otimes\mathscr{L})\left[I_r\otimes\begin{bmatrix}I_n\\0\end{bmatrix}\otimes M \quad I_r\otimes\begin{bmatrix}0\\I_r\end{bmatrix}\otimes M\right]\\ &= \left[(I_r\otimes\mathscr{L})\left(I_r\otimes\begin{bmatrix}I_n\\0\end{bmatrix}\otimes M\right) \quad (I_r\otimes\mathscr{L})\left(I_r\otimes\begin{bmatrix}0\\I_r\end{bmatrix}\otimes M\right)\right].\end{aligned} \tag{B.5}$$

Our following task is to examine each block column of (B.5). We begin with the first block; this is

$$\begin{aligned}(I_r\otimes\mathscr{L})\left(I_r\otimes\begin{bmatrix}I_n\\0\end{bmatrix}\otimes M\right) = I_r\otimes\left(\mathscr{L}\left(\begin{bmatrix}I_n\\0\end{bmatrix}\otimes M\right)\right) &= I_r\otimes\left(\mathscr{L}\begin{bmatrix}I_n\otimes M\\0\end{bmatrix}\right)\\ &= I_r\otimes\left[\mathscr{L}_1M,\ldots,\mathscr{L}_nM\right].\end{aligned}$$

We next aim to simplify the term $\mathscr{L}_iM$, which appears in the above equation:

$$\begin{aligned}\mathscr{L}_iM &= \begin{bmatrix}0 & I_r\otimes\left[e_j^n\otimes\begin{bmatrix}I_n & 0\end{bmatrix}\right]\end{bmatrix}\begin{bmatrix}M_{nnr} & 0\\ 0 & M_{rnr}\end{bmatrix}\\ &= \begin{bmatrix}0 & \left(I_r\otimes\left[e_j^n\otimes\begin{bmatrix}I_n & 0\end{bmatrix}\right]\right)M_{rnr}\end{bmatrix}\\ &= \left[0 \quad \left(I_r\otimes e_j^n\otimes\begin{bmatrix}I_n & 0\end{bmatrix}\right)\left[I_r\otimes\begin{bmatrix}I_n\\0\end{bmatrix} \quad I_r\otimes\begin{bmatrix}0\\I_r\end{bmatrix}\right]\right]\\ &= \begin{bmatrix}0 & \left(I_r\otimes e_j^n\otimes I_n\right) & 0\end{bmatrix} := \mathscr{X}_i.\end{aligned} \tag{B.6}$$

The second block column of (B.5) can be studied in a similar fashion, and it can be shown that

$$(I_r\otimes\mathscr{L})\left(I_r\otimes\begin{bmatrix}0\\I_r\end{bmatrix}\otimes M\right) = 0.$$

Summing up all these expressions, we obtain

$$(\widehat{\mathscr{F}}\otimes\mathscr{F})T_{(n+r,n+r)}(M\otimes M) = \begin{bmatrix}0 & \left(I_r\otimes\left[\mathscr{X}_1,\ldots,\mathscr{X}_n\right]\right) & 0\end{bmatrix},$$

where $\mathscr{X}_i$ is defined in (B.6). This gives

$$\begin{aligned}(\widehat{\mathscr{F}}\otimes\mathscr{F})T_{(n+r,n+r)}(M\otimes M)(x\otimes y) &= \begin{bmatrix}0 & I_r\otimes\left[\mathscr{X}_1,\ldots,\mathscr{X}_n\right] & 0\end{bmatrix}(x\otimes y)\\ &= \left(I_r\otimes\left[\mathscr{X}_1,\ldots,\mathscr{X}_n\right]\right)(x_3\otimes y).\end{aligned} \tag{B.7}$$

Next, we define another permutation

$$\mathcal{Q} = \left[\underbrace{I_r \otimes I_n \otimes \begin{bmatrix} I_{n^2} \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{\mathcal{Q}_1} \quad \underbrace{I_r \otimes I_n \otimes \begin{bmatrix} 0 \\ I_{nr} \\ 0 \\ 0 \end{bmatrix}}_{\mathcal{Q}_2} \quad \underbrace{I_r \otimes I_n \otimes \begin{bmatrix} 0 \\ 0 \\ I_{nr} \\ 0 \end{bmatrix}}_{\mathcal{Q}_3} \quad \underbrace{I_r \otimes I_n \otimes \begin{bmatrix} 0 \\ 0 \\ 0 \\ I_{r^2} \end{bmatrix}}_{\mathcal{Q}_4}\right],$$

which allows us to write

$$(x_3 \otimes y) = \mathcal{Q} \begin{bmatrix} x_3 \otimes y_1 \\ x_3 \otimes y_2 \\ x_3 \otimes y_3 \\ x_3 \otimes y_4 \end{bmatrix}.$$

Substituting this into (B.7) results in

$$\begin{aligned}&(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)}(M \otimes M)(x \otimes y) \\ &\qquad = \left(I_r \otimes \left[\mathcal{X}_1, \dots, \mathcal{X}_n\right]\right) \left[\mathcal{Q}_1 \quad \mathcal{Q}_2 \quad \mathcal{Q}_3 \quad \mathcal{Q}_4\right] \begin{bmatrix} x_3 \otimes y_1 \\ x_3 \otimes y_2 \\ x_3 \otimes y_3 \\ x_3 \otimes y_4 \end{bmatrix}.\end{aligned}$$

Now, it can be easily verified that $\left(I_r \otimes \left[\mathcal{X}_1, \dots, \mathcal{X}_n\right]\right) \left[\mathcal{Q}_1 \quad \mathcal{Q}_2 \quad \mathcal{Q}_4\right] = 0$. Thus, we obtain

$$\begin{aligned}(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)}(M \otimes M)(x \otimes y) &= \left(I_r \otimes \left[\mathcal{X}_1, \dots, \mathcal{X}_n\right]\right) \mathcal{Q}_3 (x_3 \otimes y_3) \\ &= \left(I_r \otimes \left[\mathcal{X}_1, \dots, \mathcal{X}_n\right]\right) \left(I_r \otimes I_n \otimes \begin{bmatrix} 0 \\ 0 \\ I_{nr} \\ 0 \end{bmatrix} \right) (x_3 \otimes y_3) \\ &= \left(I_r \otimes \left[I_r \otimes e_1^n \otimes I_n, \dots, I_r \otimes e_1^n \otimes I_n\right]\right) (x_3 \otimes y_3) = T_{(n,r)}(x_3 \otimes y_3).\end{aligned}$$

One can prove the relation (B.2) in a similar manner. However, for brevity, we omit it. This concludes the proof. □

BIBLIOGRAPHY

[1] M. I. Ahmad and P. Benner, *Interpolatory model reduction techniques for linear second-order descriptor systems*, in Proc. European Control Conf. ECC 2014, Strasbourg, IEEE, 2014, pp. 1075–1079. 205

[2] M. I. Ahmad, P. Benner, and P. Goyal, *Krylov subspace-based model reduction for a class of bilinear descriptor systems*, J. Comput. Appl. Math., 315 (2017), pp. 303–318. iii

[3] M. I. Ahmad, P. Benner, P. Goyal, and J. Heiland, *Moment-matching based model reduction for Navier–Stokes type quadratic-bilinear descriptor systems*, Z. Angew. Math. Mech., (2017). To appear. 193

[4] M. I. Ahmad, P. Benner, and I. M. Jaimoukha, *Krylov subspace methods for model reduction of quadratic-bilinear systems*, IET Control Theory & Appl., 10 (2016), pp. 2010–2018. 125

[5] S. A. Al-baiyat and M. Bettayeb, *A new model reduction scheme for k-power bilinear systems*, in Proc. 32nd IEEE CDC, IEEE, 1993, pp. 22–27. 32, 35, 36, 45

[6] S. A. Al-Baiyat, M. Bettayeb, and U. M. Al-Saggaf, *New model reduction scheme for bilinear systems*, Int. J. Syst. Sci., 25 (1994), pp. 1631–1642. 32, 79

[7] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*, SIAM Publications, Philadelphia, PA, 2005. 3, 9, 10, 13, 15, 17, 18, 37, 45, 57, 64, 65, 79, 153

[8] A. C. Antoulas, D. C. Sorensen, and S. Gugercin, *A survey of model reduction methods for large-scale systems*, Contemp. Math., 280 (2001), pp. 193–219. 13

[9] U. M. Ascher and L. R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations*, SIAM, Philadelphia, 1998. 145

[10] P. Astrid, S. Weiland, K. Willcox, and T. Backx, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Trans. Autom. Control, 53 (2008), pp. 2237–2251. 56

[11] Z. Bai, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Appl. Numer. Math., 43 (2002), pp. 9–44. 57, 91

[12] Z. Bai and D. Skoogh, *A projection method for model reduction of bilinear dynamical systems*, Linear Algebra Appl., 415 (2006), pp. 406–425. 32, 145, 146, 147, 148, 207

[13] 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*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 667–672. 57, 100

[14] R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826. 107

[15] U. Baur, P. Benner, and L. Feng, *Model order reduction for linear and nonlinear systems: A system-theoretic perspective*, Arch. Comput. Methods Eng., 21 (2014), pp. 331–358. 3, 32, 57

[16] C. A. Beattie and S. Gugercin, *A trust region method for optimal h 2 model reduction*, in Proc. of the Joint 48th IEEE Conference on Decision and Control, and 28th Chinese Control Conference, IEEE, 2009, pp. 5370–5375. 126

[17] C. A. Beattie and S. Gugercin, *Model reduction by rational interpolation*, in Model Reduction and Approximation: Theory and Algorithms, P. Benner, A. Cohen, M. Ohlberger, and K. Willcox, eds., SIAM, Philadelphia, PA, 2017. 3

[18] P. Benner, *Partial stabilization of descriptor systems using spectral projectors*, in Numer. Lin. Alg. in Signals, Sys. and Control, Springer, Netherlands, 2011, pp. 55–76. 14

[19] P. Benner, M. Bollhöfer, D. Kressner, C. Mehl, and T. Stykel, *Numerical Algebra, Matrix Theory, Differential-Algebraic Equations and Control Theory*, Springer, 2015. 144

[20] P. Benner and T. Breiten, *On $\mathcal{H}_2$-model reduction of linear parameter-varying systems*, in Proc. Appl. Math. Mech., vol. 11, 2011, pp. 805–806. 4, 32, 49, 166, 189, 201

[21] ——, *Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 859–885. xxi, 32, 100, 110, 119, 145, 146, 151, 152, 153, 154, 186, 199, 207, 208, 216

[22] ——, *Krylov-subspace based model reduction of nonlinear circuit models using bilinear and quadratic-linear approximations*, 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, Berlin, 2012, Springer-Verlag, pp. 153–159. 38, 207

[23] ——, *Two-sided moment matching methods for nonlinear model reduction*, Preprint MPIMD/12-12, MPI Magdeburg, 2012. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`. 57, 95, 96, 97, 139

[24] ——, *Low rank methods for a class of generalized Lyapunov equations and related issues*, Numer. Math., 124 (2013), pp. 441–470. 33, 44

[25] ——, *Two-sided projection methods for nonlinear model reduction*, SIAM J. Sci. Comput., 37 (2015), pp. B239–B260. xxi, 6, 56, 57, 61, 82, 84, 91, 93, 96, 100, 125, 129, 137, 141, 208

[26] P. Benner and T. Damm, *Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems*, SIAM J. Cont. Optim., 49 (2011), pp. 686–711. 6, 32, 33, 39, 40, 42, 50, 69, 77, 78, 207, 208

[27] P. Benner and P. Goyal, *Multipoint interpolation of Volterra series and $\mathcal{H}_2$-model reduction for a family of bilinear descriptor systems*, Systems Control Lett., 97 (2016), pp. 1–11. iii

[28] ——, *Balanced truncation model order reduction for quadratic-bilinear control systems*, arXiv e-prints 1705.00160, 2017. iii, 103, 125

[29] P. Benner, P. Goyal, and S. Gugercin, *$\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems*, SIAM J. Matrix Anal. Appl., (2018). To appear. iii

[30] P. Benner, P. Goyal, and M. Redmann, *Truncated Gramians for bilinear systems and their advantages in model order reduction*, in P. Benner, M. Ohlberger, T. Patera, G. Rozza, K. Urban (Eds.), Model Reduction of Parametrized Systems, MS&A - Modeling, Simulation and Applications, vol. 17, 2017, pp. 285–300. iii

[31] P. Benner, S. Gugercin, and K. Willcox, *A survey of projection-based model reduction methods for parametric dynamical systems*, SIAM Rev., 57 (2015), pp. 483–531. 3

[32] P. Benner, P. Kürschner, and J. Saak, *Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations*, Electron. Trans. Numer. Anal., 43 (2014), pp. 142–162. 90, 129

[33] ——, *Frequency-limited balanced truncation with low-rank approximations*, SIAM J. Sci. Comput., 38 (2016), pp. A471–A499. 98, 209

[34] P. Benner, V. Mehrmann, and D. C. Sorensen, *Dimension Reduction of Large-Scale Systems*, vol. 45 of LNCSE, Springer-Verlag, Berlin/Heidelberg, Germany, 2005. 3, 13

[35] P. Benner and M. Redmann, *Model reduction for stochastic systems.*, Stoch. PDE: Anal. Comp., 3 (2015), pp. 291–338. 52

[36] P. Benner and J. Saak, *Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey*, GAMM-Mitt., 36 (2013), pp. 32–52. 81, 100

[37] P. Benner, E. Sachs, and S. Volkwein, *Model order reduction for PDE constrained optimization*, in Trends in PDE Constrained Optimization, G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich, eds., vol. 165 of International Series of Numerical Mathematics, Springer International Publishing, 2014, pp. 303–326. 3

[38] B. N. Bond, Z. Mahmood, Y. Li, R. Sredojevic, A. Megretski, V. Stojanovi, Y. Avniel, and L. Daniel, *Compact modeling of nonlinear analog circuits using system identification via semidefinite programming and incremental stability certification*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 29 (2010), pp. 1149–1162. 89

[39] T. Breiten, *Interpolatory Methods for Model Reduction of Large-Scale Dynamical Systems*, PhD thesis, Otto-von-Guericke-Universität, Magdeburg, Germany, 2013. 61

[40] T. Breiten and T. Damm, *Krylov subspace methods for model order reduction of bilinear control systems*, Systems Control Lett., 59 (2010), pp. 443–450. 32, 49, 50, 91, 145, 146, 147, 148, 163, 164, 207

[41] C. Bruni, G. DiPillo, and G. Koch, *On the mathematical models of bilinear systems*, Automatica, 2 (1971), pp. 11–26. 32

[42] A. Bruns and P. Benner, *Parametric model order reduction of thermal models using the bilinear interpolatory rational Krylov algorithm*, Math. Comput. Model. Dyn. Syst., 21 (2015), pp. 103–129. 192

[43] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek, *h_2-norm optimal model reduction for large scale discrete dynamical MIMO systems*, J. Comput. Appl. Math., 233 (2010), pp. 1202–1216. 18

[44] A. Castagnotto, C. Beattie, and S. Gugercin, *Interpolatory methods for $\mathcal{H}_\infty$ model reduction of multi-input/multi-output systems*, in P. Benner, M. Ohlberger, T. Patera, G. Rozza, K. Urban (Eds.), Model Reduction of Parametrized Systems, MS&A - Modeling, Simulation and Applications, Springer International Publishing, Cham., 2016. Accepted. 18

[45] N. Chafee and E. F. Infante, *A bifurcation problem for a nonlinear partial differential equation of parabolic type*, Appl. Anal., 4 (1974), pp. 17–37. 93

[46] S. Chakraborty, *Some applications of Dirac's delta function in statistics for more than one random variable*, Applications Appl. Math., 3 (2008), pp. 42–54. 166

[47] S. Chaturantabut and D. C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput., 32 (2010), pp. 2737–2764. 56, 57, 95

[48] M. Condon and R. Ivanov, *Nonlinear systems-algebraic gramians and model reduction*, COMPEL, 24 (2005), pp. 202–219. 38, 105

[49] ——, *Krylov subspaces from bilinear representations of nonlinear systems*, COMPEL–Int. J. Comp. Math. Electr. Electron. Eng., 26 (2007), pp. 399–406. 32

[50] P. d'Alessandro, A. Isidori, and A. Ruberti, *Realization and structure theory of bilinear dynamical systems*, SIAM J. Cont., 12 (1974), pp. 517–535. 35

[51] T. Damm, *Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations*, Numer. Lin. Alg. Appl., 15 (2008), pp. 853–871. 80

[52] T. Damm and D. Hinrichsen, *Newton's method for a rational matrix equation occurring in stochastic control*, Linear Algebra Appl., 332 (2001), pp. 81–109. 80

[53] Z. Drmač and S. Gugercin, *A new selection operator for the discrete empirical interpolation method–improved a priori error bound and extensions*, SIAM J. Sci. Comput., 38 (2016), pp. A631–A648. 57

[54] V. Druskin and V. Simoncini, *Adaptive rational Krylov subspaces for large-scale dynamical systems*, Systems Control Lett., 60 (2011), pp. 546–560. 18

[55] D. L. Elliott, *Bilinear control systems: matrices in action*, vol. 169, Springer Verlag, 2009. 33

[56] P. Feldmann and R. W. Freund, *Efficient linear circuit analysis by Padé approximation via the Lanczos process*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 14 (1995), pp. 639–649. 18

[57] 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. 32

[58] G. Flagg, C. Beattie, and S. Gugercin, *Convergence of the iterative rational krylov algorithm*, Systems Control Lett., 61 (2012), pp. 688–691. 126

[59] G. Flagg and S. Gugercin, *Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 549–579. 32, 100, 106, 109, 110, 119, 125, 145, 149, 150, 154, 174, 179, 183, 186, 205, 207, 208

[60] G. M. Flagg, *Interpolation methods for the model reduction of bilinear systems*, PhD thesis, Virginia Polytechnic Institute and State University, 2012. 37, 38, 49, 145, 151, 154, 205

[61] G. M. Flagg, C. A. Beattie, and S. Gugercin, *Interpolatory $\mathcal{H}_\infty$ model reduction*, Systems Control Lett., 62 (2013), pp. 567–574. 146, 147

[62] R. W. Freund, *Model reduction methods based on Krylov subspaces*, Acta Numer., 12 (2003), pp. 267–319. 13

[63] K. Fujimoto and J. M. A. Scherpen, *Balanced realization and model order reduction for nonlinear systems based on singular value analysis*, SIAM J. Cont. Optim., 48 (2010), pp. 4591–4623. 32

[64] K. Fujimoto, J. M. A. Scherpen, and W. S. Gray, *Hamiltonian realizations of nonlinear adjoint operators*, Automatica, 38 (2002), pp. 1769–1775. 22, 23

[65] K. Gallivan, A. Vandendorpe, and P. Van Dooren, *Model reduction of MIMO systems via tangential interpolation*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 328–349. 18

[66] ——, *$\mathcal{H}_2$-optimal model reduction of MIMO systems*, Appl. Math. Lett., 21 (2008), pp. 1267–1273. 18

[67] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^{∞}-error norms*, Internat. J. Control, 39 (1984), pp. 1115–1193. 16

[68] G. H. Golub and C. F. Van Loan, *Matrix computations*, vol. 3, Johns Hopkins University Press, 2012. 7, 23

[69] I. V. Gosea and A. C. Antoulas, *Model reduction of linear and nonlinear systems in the Loewner framework: A summary*, in Proc. European Control Conf. 2015, Linz, IEEE, 2015, pp. 345–349. 57, 186, 187

[70] P. Goyal, M. I. Ahmad, and P. Benner, *Model reduction of quadratic-bilinear descriptor systems via Carleman bilinearization*, in Proc. European Control Conf. 2015, Linz, IEEE, 2015, pp. 1177–1182. 32, 171

[71] P. Goyal and P. Benner, *An iterative model order reduction scheme for a special class of bilinear descriptor systems appearing in constraint circuit simulation*, in ECCOMAS Congress 2016, VII European Congress on Computational Methods in Applied Sciences and Engineering, vol. 2, 2016, pp. 4196–4212. iii, 192, 199

[72] L. Grasedyck, *Hierarchical singular value decomposition of tensors*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2029–2054. 7, 23

[73] W. S. Gray and J. Mesko, *Energy functions and algebraic Gramians for bilinear systems*, in Preprints of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, 1998, pp. 103–108. 32, 38, 48, 77, 207, 208

[74] W. S. Gray and J. P. Mesko, *Controllability and observability functions for model reduction of nonlinear systems*, in Proc. Conf. on Information Sci. and Sys., Citeseer, 1996, pp. 1244–1249. 20, 21

[75] W. S. Gray and J. M. A. Scherpen, *On the nonuniqueness of singular value functions and balanced nonlinear realizations*, Systems Control Lett., 44 (2001), pp. 219–232. 32

[76] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera, *Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations*, ESAIM: Math. Model. Numer. Anal., 41 (2007), pp. 575–605. 57, 100

[77] E. J. Grimme, *Krylov projection methods for model reduction*, PhD thesis, Univ. of Illinois at Urbana-Champaign, USA, 1997. 17, 18

[78] C. Gu, *QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems*, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., 30 (2011), pp. 1307–1320. 6, 56, 57, 60, 62, 91, 92, 100, 125, 129, 137, 170, 199, 208

[79] S. Gugercin, A. C. Antoulas, and C. A. Beattie, *$\mathcal{H}_2$ model reduction for large-scale dynamical systems*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 609–638. xxi, 3, 12, 18, 19, 20, 57, 91, 100, 110, 119, 129, 151, 152

[80] S. Gugercin, T. Stykel, and S. Wyatt, *Model reduction of descriptor systems by interpolatory projection methods*, SIAM J. Sci. Comput., 35 (2013), pp. B1010–B1033. 145, 146, 154, 156, 172, 186, 193, 199

[81] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, Heidelberg, New York, 1991. 191

[82] E. Hansen, F. Kramer, and A. Ostermann, *A second-order positivity preserving scheme for semilinear parabolic problems*, Appl. Numer. Math., 62 (2012), pp. 1428–1435. 93

[83] C. Hartmann, B. Schäfer-Bung, and A. Thons-Zueva, *Balanced averaging of bilinear systems with applications to stochastic control*, SIAM J. Cont. Optim., 51 (2013), pp. 2356–2378. 32

[84] J. Heiland, *Decoupling and Optimization of Differential-Algebraic Equations with Application in Flow Control*, PhD thesis, Technische Universität Berlin, Berlin, 2002. 193

[85] M. Heinkenschloss, D. C. Sorensen, and K. Sun, *Balanced truncation model reduction for a class of descriptor systems with applications to the Oseen equations*, SIAM J. Sci. Comput., 30 (2008), pp. 1038–1063. 193, 194, 197

[86] H. V. Henderson and S. R. Searle, *The vec-permutation matrix, the vec operator and Kronecker products: A review*, Linear and multilinear algebra, 9 (1981), pp. 271–288. 27

[87] D. Hinrichsen and A. J. Pritchard, *Mathematical systems theory I: modelling, state space analysis, stability and robustness*, vol. 48, Springer Verlag, 2011. 7, 8

[88] M. Hinze and S. Volkwein, *Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: Error estimates and suboptimal control*, in Dimension Reduction of Large-Scale Systems, P. Benner, V. Mehrmann, and

D. Sorensen, eds., vol. 45 of Lect. Notes Comput. Sci. Eng., Springer-Verlag, Berlin/Heidelberg, Germany, 2005, pp. 261–306. 56, 129

[89] M. Hinze and S. Volkwein, *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition*, Comput. Optim. Appl., 39 (2008), pp. 319–345. 3

[90] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990. 7, 23

[91] C. S. Hsu, U. B. Desai, and C. A. Crawley, *Realization algorithms and approximation methods of bilinear systems*, in 22nd IEEE Conference on Decision and Control, San Antonio, TX, USA, vol. 22, 1983, pp. 783–788. 32

[92] A. C. Ionita, *Lagrange rational interpolation and its application to approximation of large-scale dynamical systems*, PhD thesis, Rice University, 2013. 57, 186, 187

[93] A. Isidori, *Nonlinear Control Systems*, vol. 1, Springer Verlag, 1995. 33

[94] M. Köhler, *On the closest stable descriptor system in the respective spaces RH_2 and RH_∞*, Linear Algebra Appl., 443 (2014), pp. 34–49. 126

[95] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500. xv, 7, 23, 24, 25

[96] K. Kunisch and S. Volkwein, *Proper orthogonal decomposition for optimality systems*, ESAIM: Math. Model. Numer. Anal., 42 (2008), pp. 1–23. 56, 129

[97] P. Kunkel and V. Mehrmann, *Differential-Algebraic Equations: Analysis and Numerical Solution*, European Mathematical Society, 2006. 144, 145, 146

[98] P. Li and L. T. Pileggi, *Compact reduced-order modeling of weakly nonlinear analog and RF circuits*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 24 (2005), pp. 184–203. 91

[99] G. P. McCormick, *Computability of global solutions to factorable nonconvex programs: Part I convex underestimating problems*, Mathematical programming, 10 (1976), pp. 147–175. 60

[100] V. Mehrmann, *Index concepts for differential-algebraic equations*, in Encyclopedia of Appl. Comp. Math., Springer, 2015, pp. 676–681. 145

[101] L. Meier and D. Luenberger, *Approximation of linear constant systems*, IEEE Trans. Autom. Control, 12 (1967), pp. 585–588. 19

[102] R. R. Mohler, *Bilinear Control Processes*, Academic Press, New York, 1973. 32, 33

[103] B. C. Moore, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Trans. Autom. Control, AC-26 (1981), pp. 17–32. 3, 17

[104] N. Nguyen, A. T. Patera, and J. Peraire, *A best points interpolation method for efficient approximation of parametrized functions*, Internat. J. Numer. Methods Engrg., 73 (2008), pp. 521–543. 13

[105] J. R. Phillips, *Projection frameworks for model reduction of weakly nonlinear systems*, in Proc. Design Automation Conf., 2000, pp. 184–189. 57

[106] J. R. Phillips, *Projection-based approaches for model reduction of weakly nonlinear, time-varying systems*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 22 (2003), pp. 171–187. 57, 91, 129, 145, 147, 148

[107] M. Redmann, *Balancing Related Model Order Reduction Applied to Linear Controlled Evolution Equations with Lévy Noise*, PhD thesis, Otto-von-Guericke-Universität, Magdeburg, Germany, 2016. 32, 52

[108] M. Redmann and P. Benner, *Approximation and model order reduction for second order systems with Lévy-noise*, in Dynamical Systems, Differential Equations and Applications, AIMS Proceedings, 2015, pp. 945–953. 50, 51

[109] M. J. Rewieński, *A Trajectory Piecewise-Linear Approach to Model Order Reduction of Nonlinear Dynamical Systems*, PhD thesis, Massachusetts Institute of Technology, 2003. 57

[110] C. W. Rowley, *Model reduction for fluids, using balanced proper orthogonal decomposition*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 997–1013. 57

[111] W. J. Rugh, *Nonlinear System Theory*, The Johns Hopkins University Press, Baltimore, MD, 1981. 32, 33, 34, 50, 101, 155, 165, 166

[112] Y. Saad, *Iterative methods for sparse linear systems*, SIAM, Philadelphia, PA, USA, 2003. 14

[113] J. Saak, M. Köhler, and P. Benner, *M-M.E.S.S.-1.0.1 – The Matrix Equations Sparse Solvers library*. DOI:10.5281/zenodo.50575, 2016. see also: www.mpi-magdeburg.mpg.de/projects/mess. 129

[114] J. M. A. Scherpen, *Balancing for nonlinear systems*, Systems Control Lett., 21 (1993), pp. 143–153. 20, 21, 32, 78

[115] W. H. A. Schilders, H. A. van der Vorst, and J. Rommes, *Model Order Reduction: Theory, Research Aspects and Applications*, Springer-Verlag, Berlin, Heidelberg, 2008. 3, 13

[116] H. Schneider, *Positive operators and an inertia theorem*, Numerische Mathematik, 7 (1965), pp. 11–17. 80

[117] H. R. Shaker and F. Shaker, *Generalized time-limited balanced reduction method*, in American Control Conference (ACC), 2013, IEEE, 2013, pp. 5530–5535. 209

[118] S. D. Shank, V. Simoncini, and D. B. Szyld, *Efficient low-rank solution of generalized Lyapunov equations*, Numer. Math., 134 (2016), pp. 327–342. 33, 44, 49, 80, 82

[119] G. E. Shilov, *Linear Algebra*, Dover Publications, New York, 1977. 161

[120] S. Shokoohi, L. M. Silverman, and P. Van Dooren, *Linear time-variable systems: Balancing and model reduction*, IEEE Trans. Autom. Control, 28 (1983), pp. 810–822. 40, 72

[121] V. Simoncini, *Computational methods for linear matrix equations*, SIAM Rev., 58 (2016), pp. 377–441. 81, 100

[122] T. Siu and M. Schetzen, *Convergence of Volterra series representation and BIBO stability of bilinear systems*, Internat. J. Systems Sci., 22 (1991), pp. 2679–2684. 34, 35

[123] E. D. Sontag, *Mathematical Control Theory*, vol. 6, Springer Verlag, 1998. 8

[124] G. W. Stewart and J. G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990. 144

[125] T. Stykel, *Analysis and Numerical Solution of Generalized Lyapunov Equations*, PhD thesis, Technische Universität Berlin, Berlin, 2002. 13, 145

[126] P. van Dooren, K. A. Gallivan, and P.-A. Absil, $_2$-*optimal model reduction with higher-order poles*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2738–2753. 126

[127] E. Verriest, *Time variant balancing and nonlinear balanced realizations*, in Model Order Reduction, Mathematics in Industry 13, W. H. A. Schilders, H. A. van der Vorst, and J. Rommes, eds., Springer-Verlag, Berlin, 2008, pp. 203–222. 39

[128] E. Verriest and T. Kailath, *On generalized balanced realizations*, IEEE Trans. Autom. Control, 28 (1983), pp. 833–844. 40, 72

[129] D. C. Villemagne and R. E. Skelton, *Model reduction using a projection formulation*, Internat. J. Control, 46 (1987), pp. 2141–2169. 17, 18

[130] K. Willcox and J. Peraire, *Balanced model reduction via the proper orthogonal decomposition*, AIAA J., 40 (2002), pp. 2323–2330. 57

[131] D. Wilson, *Optimum solution of model-reduction problem*, Proc. IEE, 117 (1970), pp. 1161–1165. 126

[132] D. A. Wilson, *Optimum solution of model-reduction problem*, vol. 117, 1970, pp. 1161–1165. 145

[133] L. Zhang and J. Lam, *On H_2 model reduction of bilinear systems*, Automatica J. IFAC, 38 (2002), pp. 205–216. 35, 36, 37, 102, 145, 151

SCHRIFTLICHE EHRENERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; verwendete fremde und eigene Quellen sind als solche kenntlich gemacht. Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert oder verzerrt wiedergegeben.

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadenersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

(Ort, Datum)

Pawan Kumar Goyal