Stability Analysis of Inexact Linear Solves in Model Order Reduction

Ph.D. Thesis

By

Rajendra Choudhary

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

August 2019

Stability Analysis of Inexact Linear Solves in Model Order Reduction

A THESIS

Submitted in partial fulfillment of the requirements for the award of the degree
of
DOCTOR OF PHILOSOPHY

by
Rajendra Choudhary

DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY INDORE
August 2019

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Stability Analysis of Inexact Linear Solves in Model Order Reduction** in the partial fulfillment of the requirements for the award of the degree of **Doctor Of Philosophy** and submitted in the **Discipline of Computer Science & Engineering, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from January 2014 to August 2019 under the supervision of Dr. Kapil Ahuja, Associate Professor, Indian Institute of technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

06/08/19
Signature of the student with date
(RAJENDRA CHOUDHARY)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

6/8/19
Signature of Thesis Supervisor with date
(DR. KAPIL AHUJA)

RAJENDRA CHOUDHARY has successfully given his Ph.D. Oral Examination held on **25/10/2019**.

Signature of Chairperson (OEB)	Signature of External Examiner	Signature(s) of Thesis Supervisor(s)
Date: 25/10/2019	Date: 25.10.19	Date: 25/10/19
Signature of PSPC Member #1	Signature of PSPC Member #2	Signature of Convener, DPGC
Date: 25.10.2019	Date: 25-10-2019	Date: 25/10/2019

Signature of Head of Discipline
Date: 25/10/2019

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Kapil Ahuja for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. With his enthusiasm, inspiration, and great efforts to explain things clearly and simply, he helped make mathematics fun for me. He gently corrected me at every stage: coming up with the relevant ideas, implementing them efficiently, presenting them with less ambiguity, and writing concisely.

Besides my advisor, I would like to thank my PSPC committee members: Dr. Sk. Safique Ahmad, and Dr. Aruna Tiwari, not only for their insightful comments and encouragement, but also for the hard questions that motivated me to widen my research from various perspectives.

My sincere thanks also goes to Prof. Dr. Peter Benner, who provided me an opportunity to join their team at Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany, as an intern.

I would like to thank my fellow senior researchers Mr. Pramod Mane, Mr. Rajat Saxena, Dr. Vipul Mishra and Dr. Robin Singh Bhadoria for the stimulating discussions that we had during my Ph.D. I would also like to thank my fellow lab mates Navneet, Aditya and Rohit for the sleepless nights that we spent working together before deadlines. Thanks to my friends and juniors Piyush, Sadaf, Saumya, Omprakash, Rudresh, Nikhil, Mayank, Chandan, Iyappan, Gyan, Ram, Aaditya, Animesh, Akhileshji and Pratibha, for all the fun we have had together. I also thank Mr. Shailendra Verma for his cooperation and support in miscellaneous matters.

Last, but not the least, I would like to express my deepest gratitude to my family and my friends. Especially, my maternal uncles Mr. Prakash Rathore, Mr. Dinesh Rathore and Mr. Mukesh Rathore, my brother Mr. Hemant Patidar, and my friends Subhadeep Karan, Ankit Parsai, Akshat Sarmandal, Ambar Dixit, Amritesh Arya for supporting me throughout my life and specially during my time as a Ph.D. scholar.

Rajendra Choudhary

Dedicated to

my wonderful parents Yashwant Choudhary and Sadhana Choudhary, my sister Jyoti Parmar and brother-in-law Vivek Parmar.

ABSTRACT

A dynamical system describes a relation between two or more measurable quantities by a set of differential equations. We focus on first-order non-parametric as well as parametric dynamical systems with varying linearity (linear and bilinear). In general, dynamical systems corresponding to real-world applications are extremely large in size. Simulation and computation with such systems require a large amount of space and time. By using Model Order Reduction (MOR) techniques, these large dynamical systems are reduced into a smaller size, which makes the simulation and computation easier. MOR can be done in many ways, i.e., by using balanced truncation, Hankel approximations or Krylov projection. Projection methods obtain the reduced model by projecting the original full model on a lower dimensional subspace and are quite popular. Interpolation is usually used to obtain the subspaces involved in the projection. Thus, these methods are referred to as interpolatory projection based MOR algorithms, which we specifically focus on.

In most of these MOR algorithms, people often use direct methods like LU-factorization, etc., to solve the arising linear systems, which have a high time complexity (cubic in terms of the system size). A common solution to this scaling problem is to use iterative methods like Krylov subspace methods, etc., which have a reduced time complexity (between linear and quadratic in terms of the system size), where nnz is the number of nonzeros in the system matrix). Although iterative methods are cheap, they are inexact too. Hence, studying the stability of the underlying MOR algorithms with respect to such approximate (inexact) linear solves becomes important.

One of the first works that performed such a stability analysis focused on popular MOR algorithms for first-order *non-parametric linear* dynamical systems. Here, the authors briefly mention that their analysis would be easily carried from the first-order to the second-order case. Some researchers expanded this stability analysis to reducing second-order *non-parametric linear* dynamical systems. Apart from this, a different kind of stability analysis for MOR of second-order *non-parametric linear* dynamical systems has also been done in literature. In this, the authors first show that the SOAR

algorithm (second order Arnoldi) is unstable with respect to the machine precision errors (and not inexact linear solves). Then, they propose a Two-level orthogonal Arnoldi (TOAR) algorithm that cures this instability of SOAR.

Since our focus is on first-order systems, we extend the stability analysis done for the reduction of *non-parametric linear* dynamical systems to the reduction of the following classes of dynamical systems: *non-parametric bilinear* and *parametric linear*. Our analyses can be easily extended to MOR of *parametric bilinear* dynamical systems, leading to coverage of most of the existing MOR algorithms.

The innovative aspects of this work are as follows: capturing the behavior of bilinear terms in the stability conditions, providing two different sets of constraints for achieving backward stable algorithms, and easily satisfying the extra-orthogonality constraints imposed while achieving stability.

Publications and Presentations from Thesis

Journal Papers

1. **Rajendra Choudhary** and Kapil Ahuja, "Stability Analysis of Bilinear Iterative Rational Krylov Algorithm", *Linear Algebra and its Applications*, Elsevier, Vol. 538, pp. 56-88, 2018.

2. **Rajendra Choudhary** and Kapil Ahuja, "Inexact Linear Solves in Model Reduction of Bilinear Dynamical Systems", *IEEE ACCESS*, Vol. 7, pp. 72297-72307, 2019.

Conference Presentations

3. **Rajendra Choudhary** and Kapil Ahuja, "Stability of Linear Solves in Multipoint Volterra Series Interpolatory Bilinear Model Reduction", *Ramanujan Conclave-2015*, IIT Indore, Dec 22-23, 2015 (Poster),

   and

   Industry Academia Conclave (IAC), IIT Indore, Feb 18-20, 2016 (Poster).

4. **Rajendra Choudhary** and Kapil Ahuja, "Stability Analysis of Bilinear Iterative Rational Krylov Algorithm", *Reduced Basis Summer School*, Germany, Oct 4-7, 2016,

   and

   International Conference on Mathematical Modelling, Differential Equations, Scientific Computing & Applications (ICMMDESCA), IIT Kanpur, Mar 27-29, 2016.

5. **Rajendra Choudhary** and Kapil Ahuja, "Stability Analysis of Iterative Linear Solves in Truncated Bilinear Iterative Rational Krylov Algorithm", 2^{nd} *Cyber-Physical Systems Symposium (CyPhySS)*, Indian Institute of Science Bangalore, July 11-12, 2018 (Poster).

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

SISO	Single Input Single Output
MIMO	Multiple Input Multiple Output
IRKA	Iterative Rational Krylov Algorithm
BIRKA	Bilinear Iterative Rational Krylov Algorithm
TBIRKA	Truncated Bilinear Iterative Rational Krylov Algorithm
PMOR	Parametric Model Order Reduction
IPMOR	Interpolatory Parametric Model Order Reduction
SOAR	Second Order Arnoldi
TOAR	Two-level Orthogonal Arnoldi
BiCG	BiConjugate Gradient
RBiCG	Recycling BiConjugate Gradient
AIRGA	Adaptive Rational Global Arnoldi

Notations

$\mathbb{R},\ \mathbb{C}$	fields of real and complex numbers
$\mathbb{R}^n,\ \mathbb{C}^n$	vector space of real/complex n-tuples
$\mathbb{R}^{m\times n},\ \mathbb{C}^{m\times n}$	matrix space of real/complex $m \times n$ tuples
$\mathcal{A}$	matrix $\in \mathbb{R}^{m\times n}$
$\mathcal{A}^{-1}$	inverse of matrix $\mathcal{A}$
$\mathcal{A}^T$	transpose of matrix $\mathcal{A}$
$\|\mathcal{A}\|$ or $\|\mathcal{A}\|_2$	2-norm of matrix $\mathcal{A}$
$\|\mathcal{A}\|_F$	Frobenius norm of matrix $\mathcal{A}$
I_n	identity matrix of size $n \times n$
$Range\,(\mathcal{A})$	subspace spanned by the columns of matrix $\mathcal{A}$
$orth\,(\mathcal{A})$	orthonormal subspace spanned by the columns of a matrix $\mathcal{A}$
$vec\,(\mathcal{A})$	vectorization of matrix $\mathcal{A}$
$\mathcal{K}^q\,(\mathcal{A},\ b)$	Krylov subspace spanned by $\{b,\ \mathcal{A}b,\ \ldots,\ \mathcal{A}^{q-1}b\}$
$\mathcal{O}\,(x)$	order of x, where x is a real number
$P \otimes Q$	Kronecker product between two matrices P and Q
$\zeta\,(p)$	a parametric bilinear dynamical system
$\zeta_r\,(p)$	a reduced parametric bilinear dynamical system
ζ	a non-parametric bilinear dynamical system
ζ_r	a reduced non-parametric bilinear dynamical system
$\widetilde{\zeta}$	a perturbed non-parametric bilinear dynamical system
ζ^M	a truncated non-parametric bilinear dynamical system, where M is the truncation index
$H_k\,(s_1,\ s_2,\ \ldots,\ s_k;\ p)$	k^{th} order transfer function of the parametric bilinear dynamical system, where $s_1,\ s_2,\ \ldots,\ s_k$ are the frequencies and p is the set of parameters
$\|\zeta\|_{H_2}$	H_2-norm of non-parametric bilinear dynamical system
$k\,(\zeta)$	condition number of ζ

CHAPTER 1

INTRODUCTION

A dynamical system describes a relation between two or more measurable quantities by a set of differential equations of many orders, however, we focus only on the first-order. The system may be non-parametric/ parametric or linear/ nonlinear, and can be described both in the time domain and in the frequency domain. In the time domain, a v parameters Multiple Input Multiple Output (MIMO) bilinear dynamical system with m inputs and q outputs is represented as follows [12, 20]:

$$\zeta(p): \begin{cases} E(p)\dot{x}(t) &= A\,(p)\,x(t) + \sum_{j=1}^{m} N_j\,(p)\,x(t)u_j(t) + B\,(p)\,u(t), \\ y(t) &= C\,(p)\,x(t), \end{cases} \tag{1.1}$$

where $p = [p_1 \ \ldots \ p_v]^T \in \mathbb{R}^v$, $E(p),\ A(p) \in \mathbb{R}^{n\times n},\ N_j(p) \in \mathbb{R}^{n\times n}$ for $j = 1,\ \ldots, m$, $B(p) \in \mathbb{R}^{n\times m}$, and $C(p) \in \mathbb{R}^{q\times n}$. Also, $u(t) = [u_1(t) \ \ldots \ u_m(t)]^T$: $\mathbb{R} \to \mathbb{R}^m$, $x(t)$: $\mathbb{R} \to \mathbb{R}^n$ and $y(t)$: $\mathbb{R} \to \mathbb{R}^q$ represent the input, the state and the output of the dynamical system, respectively. We make no assumption on the structure of the system matrices. It is not possible to write the transfer function of a complete bilinear dynamical system, therefore, in [41, 24, 26] the authors represent the bilinear dynamical system in the frequency domain by a series of subsystem transfer functions, i.e.,

$$\zeta(p) = \lim_{k\to\infty} \zeta^k(p), \tag{1.2}$$

where $\zeta^k(p) = \{H_1(s_1;\ p),\ H_2(s_1,\ s_2;\ p),\ \ldots,\ H_k\,(s_1,\ s_2,\ \ldots,\ s_k;\ p)\}$; $s_1,\ s_2,\ \ldots,\ s_k$ are the frequencies and p is the set of parameters. Here, $H_k\,(s_1,\ s_2,\ \ldots,\ s_k;\ p)$ is the k^{th} order transfer function of the parametric bilinear dynamical system and is defined as [24, 26]:

$$\begin{aligned} H_k(s_1,\ s_2,\ \ldots,\ s_k;\ p) &= C(p)\,(s_kE(p) - A(p))^{-1}\,\bar{N}(p) \\ &\cdot \left[I_m \otimes (s_{k-1}E(p) - A(p))^{-1}\right]\left(I_m \otimes \bar{N}(p)\right) \\ &\vdots \\ &\cdot \left[\underbrace{I_m \otimes \ldots \otimes I_m}_{k-2 \text{ times}} \otimes (s_2E(p) - A(p))^{-1}\right]\left(\underbrace{I_m \otimes \ldots \otimes I_m}_{k-2 \text{ times}} \otimes \bar{N}(p)\right) \\ &\cdot \left[\underbrace{I_m \otimes \ldots \otimes I_m}_{k-1 \text{ times}} \otimes (s_1E(p) - A(p))^{-1}\right]\left(\underbrace{I_m \otimes \ldots \otimes I_m}_{k-1 \text{ times}} \otimes B(p)\right), \end{aligned} \tag{1.3}$$

where $\bar{N}(p) = [N_1(p)\ \ldots\ N_m(p)]$, I_m is the identity matrix of size m, and $\otimes$ denotes Kronecker product (defined later).

If in (1.1), the matrix $N(p)$ is a zero matrix, then the system is a parametric linear dynamical system. That is, a v parameter MIMO linear dynamical system is represented as

$$\begin{aligned} E(p)\dot{x}(t) &= A(p)x(t) + B(p)u(t), \\ y(t) &= C(p)x(t). \end{aligned} \tag{1.4}$$

The transfer function of the linear dynamical system in the frequency domain is defined as follows:

$$H(s;\ p) = C(p)(sE(p) - A(p))^{-1}B(p). \tag{1.5}$$

Also, if the system matrices above are independent of the parameter (p), then this refers to a non-parametric dynamical system (bilinear or linear as the case may be).

In general, dynamical systems corresponding to the real world applications are extremely large in size. Simulation and computation with such systems requires large amount of space and time. By using model order reduction (MOR) techniques [28, 47, 51, 5, 16, 45, 29, 17], these large dynamical systems are reduced into a smaller size, which makes the simulation and computation easier. MOR can be done in many ways,

i.e., by using balanced truncation [47], Hankel approximations or Krylov projection [28, 29, 17]. Projection methods obtain the reduced model by projecting the original full model on a lower dimensional subspace, and are quite popular. In literature, there are several techniques of projecting a dynamical system [28, 29, 17, 5, 8, 30, 12, 24, 26]. The Petrov-Galerkin projection is one such projection technique that gives nice properties in the reduced model. Interpolation is usually used to obtain the subspaces involved in the Petrov-Galerkin projection.

Based upon the theory of Petrov-Galerkin based interpolatory model reduction, authors in [11, 30, 19] have proposed Iterative Rational Krylov Algorithm (IRKA) for model reduction of *non-parametric linear* dynamical systems. IRKA provides the reduced model that is optimal (the kind of optimality is discussed in the next section). Similar to IRKA, authors in [12, 17] have proposed Bilinear Iterative Rational Krylov Algorithm (BIRKA) for model reduction of *non-parametric bilinear* dynamical systems.

BIRKA's biggest drawback is that it does not scale well in time (with respect to increase in the size of the input dynamical system). To overcome this drawback, researchers have proposed other efficient algorithms for MOR of *non-parametric bilinear* dynamical systems. This includes TBIRKA (Truncated Bilinear Iterative Rational Krylov Algorithm) [24, 26], balanced truncation based [13], Gramian based [50], moment-matching based [8], and implicit Volterra series based [1]. TBIRKA forms the base of all these efficient algorithms, which is a cheaper variant of BIRKA.

Analogous to the non-parametric case of [30] (IRKA as above), MOR algorithms for reducing *parametric linear* dynamical systems have also been proposed (also generically termed as Parametric Model Order Reduction algorithms or PMOR algorithms). For e.g., interpolatory PMOR algorithm (IPMOR) [9], piecewise $\mathcal{H}_2-$optimal interpolatory PMOR [9], multi-parameter and multi-frequency moment-matching based PMOR algorithm [36], PMOR using extended balanced truncation [44], PMOR with $\mathcal{H}_2-$error using radial basis functions [14], etc. IPMOR's theory feeds into the other algorithms listed above.

Recently, *parametric bilinear* dynamical systems are also being studied extensively

[15]. For reducing such systems, in [20], authors have proposed an interpolatory parametric bilinear MOR method.

The main computational bottleneck in reducing larger models (or dynamical systems) is solving large sparse linear systems of equations. The reason for this is that typically, model reducers use direct solvers like LU-factorization, Gaussian elimination, etc., to solve such linear systems of equations, which have a high time complexity ($\mathcal{O}(n^3)$, where n is the original system size) [43, 23]. A common solution to this scaling problem is to use iterative methods like the Krylov subspace methods, etc., which have a reduced time complexity (i.e., $\mathcal{O}(n \times nnz)$, where nnz is the number of nonzeros in the system matrix) [43, 23]. Although iterative methods are cheap, they are inexact too, i.e., they solve linear systems of equations up to a certain stopping tolerance. Hence, studying stability of the underlying MOR algorithms with respect to such approximate (inexact) linear solves becomes important [23, 48]. In other words, we need to check that small errors in linear solves does not substantially deteriorate the quality of the reduced model.

One of the first works that performed such a stability analysis focused on popular MOR algorithms for first-order *non-parametric linear* dynamical systems [10]. Here, the authors briefly mention that their analysis would be easily carried from the first-order to the second-order case. Some researchers expanded this stability analysis to reducing second-order *non-parametric linear* dynamical systems [46]. Apart from this, a different kind of stability analysis for MOR of second-order *non-parametric linear* dynamical systems has also been done in literature [37]. In this, the authors first show that the SOAR algorithm (second order Arnoldi) is unstable with respect to the machine precision errors (and not inexact linear solves). Then, they propose a Two-level orthogonal Arnoldi (TOAR) algorithm that cures this instability of SOAR.

Before performing the stability analysis of the above discussed algorithms, we revisit the theory of the different model reduction algorithms in the next chapter (Chapter 2). Recall, our focus is only on first-order systems. With focus on *non-parametric bilinear* MOR, we first perform the stability analysis of BIRKA (in Chapter 3). We prove that under mild assumptions, BIRKA is backward stable. The most

novel contributions here are capturing the behavior of the bilinear terms ($N_j(p)$) for $j = 1, \ldots, m$ from (1.1) in the conditions for stability as well as analyzing the invertibility of all involved matrices.

We also compute the expression for conditioning of the problem and perturbation (introduced as part of stability analysis) to get the accuracy of the reduced system. Finally, we support all our results by numerical experiments.

Next, we extend the earlier stability analysis of BIRKA to more efficient algorithms for MOR of *non-parametric bilinear* dynamical system, specifically TBIRKA (in Chapter 4). The approach here is slightly different, which forms our most novel contribution.

In BIRKA stability analysis, a single expression for bilinear dynamical system norm is used (involving a Volterra series). In TBIRKA stability analysis, a similar single expression (involving truncated Volterra series) leads to one set of stability conditions. Alternatively, because of truncation, the bilinear dynamical system here can be represented by a finite set of functions (which was not possible in-case of BIRKA because of the need of infinite such function there) leading to another set of stability conditions. Depending upon the input dynamical system, one set of conditions may be more easy to satisfy than the other.

We also compute the expression for conditioning of the problem as well as perturbation, both of which are different than their respective expressions in BIRKA, leading to computation of accuracy of the reduced system. We support all our conjectures (including two ways of achieving a backward stable TBIRKA) by numerical experiments.

Finally, with focus on MOR of *parametric linear* dynamical systems, we perform stability analysis of the IPMOR algorithm (in Chapter 5). Besides deriving the conditions for stability, expressions for accuracy of the reduced system, and numerical experiments, our most novel contribution here is achieving a backward stable IPMOR.

To achieve this, we first categorize the involved orthogonality conditions into different classes. Second, we adapt the underlying iterative solves (here BiConjugate Gradient or BiCG [43, 49]) to satisfy these orthogonalities. Finally, and third, we

derive a new variant of the Recycling BiCG [3, 4] so that these orthogonalities can be achieved with no code changes to the iterative solver (for a end user or a model reducer here) as well as cheaply (extra orthogonality cost offset by savings because of recycling). We give our conclusions and discuss future work in Chapter 6. For the rest of this dissertation we use the terms and notations as listed below.

a. The H_2-norm is a functional norm defined as [5, 10, 24]

$$\|H_k\|_{H_2}^2 = \left(\frac{1}{2\pi}\right)^k \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \|H_k\left(i\omega_1,\ \dots,\ i\omega_k\right)\|_F^2 \, d\omega_1 \dots d\omega_k, \tag{1.6}$$

where i denotes $\sqrt{-1}$. Here, we assume that all H_2-norms computed further exist. In other words, the improper integrals defined by the H_2-norm give finite value. This is a reasonable assumption because this happens often in practice (see [10], where stability analysis of IRKA is done).

b. The $H_\infty-$norm is also a functional norm, defined as [5, 10, 24]

$$\|H_k\|_{H_\infty} = \max_{\omega_1,\ \dots,\ \omega_k \in \mathbb{R}} \|H_k\left(i\omega_1,\ \dots,\ i\omega_k\right)\|_2 . \tag{1.7}$$

c. The Kronecker product between two matrices P (of size $m \times n$), and Q (of size $s \times t$) is defined as

$$P \otimes Q = \begin{bmatrix} p_{11}Q & \cdots & p_{1n}Q \\ \vdots & \ddots & \vdots \\ p_{m1}Q & \cdots & p_{mn}Q \end{bmatrix},$$

where p_{ij} is an element of matrix P and order of $P \otimes Q$ is $ms \times nt$.

d. In literature [51, 12], the H_2-norm of a bilinear dynamical system is defined as

$$\|\zeta\|_{H_2}^2 = vec(I_p)^T \left(C \otimes C\right) \left(-A \otimes I_n - I_n \otimes A - \sum_{j=1}^{m} N_j \otimes N_j\right)^{-1} \left(B \otimes B\right) vec(I_m), \tag{1.8}$$

where I_p and I_m are identity matrices of size p and m, respectively. Also, in

[24, 26], the H_2-norm of a truncated bilinear dynamical system is defined as

$$\left\|\zeta^M\right\|_{H_2}^2 = vec(I_p)^T (C \otimes C) \sum_{k=0}^{M} \left((-A \otimes I_n - I_n \otimes A)^{-1} \sum_{j=1}^{m} N_j \otimes N_j \right)^k \\ (-A \otimes I_n - I_n \otimes A)^{-1} (B \otimes B)\, vec(I_m), \tag{1.9}$$

where M is the truncation index. If the type of norm is not written, then in the case of functional norm it is a H_2-norm. In the case of matrices it is a 2-norm.

e. *vec* operator on a matrix P is defined as

$$vec(P) = (p_{11},\ \ldots,\ p_{m1},\ p_{12},\ \ldots,\ p_{m2},\ \ldots\ \ldots,\ p_{1n},\ \ldots,\ p_{mn})^T.$$

f. Also, $\mathbb{R}$ denotes the set of real numbers and $\mathbb{F}$ denotes the discrete subset of real numbers.

CHAPTER 2

BACKGROUND

A reduced dynamical system can be obtained by a projection-type framework. A matrix $P \in \mathbb{R}^{n \times n}$ is a projector (onto a subspace $\mathcal{V} \subset \mathbb{R}^n$) if $Range(P) = \mathcal{V}$ and $P^2 = P$. If $P = P^T$, then P is an orthogonal projector (i.e., Galerkin projection), otherwise an oblique projector (i.e., Petrov-Galerkin projection) [17].

According to the Petrov-Galerkin projection, the residual of a dynamical system obtained after projecting on a lower dimensional subspace, is made orthogonal to some other subspace defined by a test basis. Let η_i denote the residual of this dynamical system, then according to the Petrov-Galerkin condition, $\eta_i \perp L$, where L denotes any test subspace.

The subspace on which we project, and the orthogonal subspace are not known to us. We can arbitrarily pick these subspaces, but then we cannot guarantee a good input-output behavior from the reduced model. For the reduced model to provide a high fidelity approximation to the input-output behavior of the original full model, we use interpolation to obtain these subspaces. In [30], authors give an algorithm for model reduction of non-parametric linear dynamical systems called IRKA (Iterative Rational Krylov Algorithm). IRKA is a Petrov-Galerkin based interpolatory model reduction algorithm. Here authors have focused on Hermite interpolation of the transfer function to obtain these subspaces. Hermite interpolation is a popular method from

interpolatory theory, where a function and its derivative are interpolated. Here, the transfer function of the original full model $H(s)$ (and its derivative) and reduced model (and its derivative) are interpolated at a set of interpolation points. For a certain type of linear dynamical systems, IRKA locally converges to a local minimum of the underlying H_2-optimization problem [25]. For H_2-optimality discussion in this case we refer the reader to [30] and [25].

Now, we discuss H_2-optimality in the non-parametric bilinear case. Here also, we apply Petrov-Galerkin based interpolatory MOR to a non-parametric bilinear dynamical system. This is a short summary of the original work in [12] and [24]. After reduction, the non-parametric bilinear dynamical system (1.1)[1] can be represented as [12, 24]

$$\zeta_r : \begin{cases} \dot{x}_r(t) &= A_r x_r(t) + \sum_{j=1}^{m} N_{j_r} x_r(t) u_j(t) + B_r u(t), \\ y_r(t) &= C_r x_r(t), \end{cases} \tag{2.1}$$

where $A_r,\ N_{j_r} \in \mathbb{R}^{r\times r},\ B_r \in \mathbb{R}^{r\times m}$ and $C_r \in \mathbb{R}^{p\times r}$ for $j = 1,\ \ldots,\ m$ with $r \ll n$. Here, the input $u\,(t)$ is the same (maps from $\mathbb{R}$ to $\mathbb{R}^m$) while state $x_r\,(t)$ maps from $\mathbb{R}$ to $\mathbb{R}^r$ (instead of $\mathbb{R}^n$ earlier). We want ζ_r to approximate ζ in an appropriate norm, and hence, $y_r(t)$ should be nearly equal to $y(t)$ for all admissible inputs. Let the two r-dimensional subspaces, $\mathcal{V}_r$ and $\mathcal{W}_r$ be chosen in such a way that $\mathcal{V}_r = Range(V_r)$ and $\mathcal{W}_r = Range(W_r)$, where $V_r \in \mathbb{R}^{n\times r}$ and $W_r \in \mathbb{R}^{n\times r}$ are matrices. We project the original full model (1.1)[1] to a lower dimensional subspace, i.e., $x\,(t) \approx V_r x_r(t)$, and enforce the Petrov-Galerkin condition [12, 24]

$$W_r^T \left(V_r \dot{x}_r(t) - A V_r x_r(t) - \sum_{j=1}^{m} N_j V_r x_r(t) u_j(t) - B u(t) \right) = 0,$$
$$y(t) = C V_r x_r(t).$$

[1]Here, we have taken a non-parametric system, and hence, the system matrices are independent of the parameters (p). As in the original non-parametric bilinear MOR papers [12] and [24], we take $E = I_n$.

Comparing the above equations with (2.1), we get

$$A_r = \left(W_r^T V_r\right)^{-1} W_r^T A V_r, \quad N_{j_r} = \left(W_r^T V_r\right)^{-1} W_r^T N_j V_r,$$
$$B_r = \left(W_r^T V_r\right)^{-1} W_r^T B, \text{ and } C_r = C V_r,$$

where $\left(W_r^T V_r\right)$ is assumed to be invertible. Obtaining such an invertible matrix is not hard [12]. Different selection of the subspaces $\mathcal{V}_r$ and $\mathcal{W}_r$ give different reduced models, but we choose the subspaces $\mathcal{V}_r$ and $\mathcal{W}_r$ by enforcing interpolation. There are two different ways of doing interpolation, i.e., subsystem interpolation and Volterra series interpolation [24, 26]. These are explained below.

A bilinear system can be represented by a series of subsystem transfer functions. If we apply certain interpolation conditions on a finite number of subsystems, then it is called *subsystem interpolation* [24]. In this approach we interpolate the each of the subsystem transfer function expression (1.3), without the parameter p.

Another way is *Volterra series interpolation*. A non-parametric bilinear dynamical system ζ can also be represented by following Volterra series, which non-linearly relates all admissible inputs $u(t)$ to outputs $y(t)$:

$$y(t) = \sum_{k=1}^{\infty} \int_0^{t_1} \int_0^{t_2} \dots \int_0^{t_k} h_k\left(t_1,\ t_2,\ \dots,\ t_k\right) \left(u\left(t - \sum_{i=1}^{k} t_i\right) \otimes \dots \otimes u\left(t - t_k\right)\right) dt_k \cdots dt_1.$$

In this Volterra series representation, the term

$$h_k(t_1,\ t_2,\ \dots,\ t_k) = C e^{A t_k} \bar{N} \left(I_m \otimes e^{A t_{k-1}}\right) \left(I_m \otimes \bar{N}\right) \cdots \left(\underbrace{I_m \otimes \dots \otimes I_m}_{k-2 \text{ times}} \otimes e^{A t_2}\right) \left(\underbrace{I_m \otimes \dots \otimes I_m}_{k-2 \text{ times}} \otimes \bar{N}\right) \cdot \left(\underbrace{I_m \otimes \dots \otimes I_m}_{k-1 \text{ times}} \otimes e^{A t_1}\right) \left(\underbrace{I_m \otimes \dots \otimes I_m}_{k-1 \text{ times}} \otimes B\right),$$

is called the degree k Volterra kernel, where $\bar{N} = [N_1,\ \dots,\ N_m]$. A degree k Volterra kernel in frequency domain is equivalent to the k^{th} order transfer function of the bilinear dynamical system (see (1.3) without parameter p). Here, interpolation is

done on a weighted sum of all Volterra kernel transfer functions given by (1.3). We refer the reader to [24, 42] for a detailed discussion on the definition of the Volterra series, the Volterra kernels, and the subsequent derivations.

As the subsystem interpolation approach is unable to satisfy any optimality conditions [24] (error between the original full model and the reduced model is minimum in some norm), we focus on the Volterra series interpolation. We need to know how to build V_r and W_r such that the conditions of the Volterra series interpolation are satisfied. We also need to decide where to interpolate so that we get an optimal reduced model. Here, we focus on H_2-optimality.

The following error system expression is differentiated for getting the H_2-optimality conditions [12, 24]:

$$\begin{aligned}
&\|\zeta-\zeta_r\|_{H_2} = vec(I_{2p})^T\left(\begin{bmatrix} C & -\breve{\breve{C}}\end{bmatrix}\otimes\begin{bmatrix} C & -\breve{C}\end{bmatrix}\right)\\
&\left(-\begin{bmatrix} A & 0\\ 0 & \Lambda\end{bmatrix}\otimes\begin{bmatrix} I_n & 0\\ 0 & I_r\end{bmatrix}-\begin{bmatrix} I_n & 0\\ 0 & I_r\end{bmatrix}\otimes\begin{bmatrix} A & 0\\ 0 & \breve{A}\end{bmatrix}-\sum_{j=1}^{m}\begin{bmatrix} N_j & 0\\ 0 & \breve{\breve{N}}_j^T\end{bmatrix}\otimes\begin{bmatrix} N_j & 0\\ 0 & \breve{N}_j\end{bmatrix}\right)^{-1}\\
&\left(\begin{bmatrix} B\\ \breve{\breve{B}}^T\end{bmatrix}\otimes\begin{bmatrix} B\\ \breve{B}\end{bmatrix}\right)vec(I_{2m}),
\end{aligned}\tag{2.2}$$

where $\breve{A}$, $\breve{B}$, $\breve{C}$ and $\breve{N}_j$ are the initial guesses for the reduced system. Also, $\breve{A} = R\Lambda R^{-1}$, $\breve{\breve{B}} = \breve{B}^T R^{-T}, \breve{\breve{C}} = \breve{C}R$ and $\breve{\breve{N}}_j = R^T\breve{N}_j^T R^{-T}$. Performing interpolation on the inverse images of the reduced system poles helps achieve H_2-optimality. Theorem 1 below summarizes this, where the poles of the transfer function of every reduced subsystem (say H_{r_k}) are computed (say represented by $\lambda_{l_1},\ \lambda_{l_2},\ \ldots,\ \lambda_{l_k}$), inverted (leading to $-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}$), and finally, interpolation is performed at these points.

Theorem 1. *[24, 26] Let $\zeta = (A,\ N_j,\ B,\ C)$ be a non-parametric bilinear system of order n, where $j = 1,\ \ldots,\ m$. Let $\zeta_r = (A_r,\ N_{r_j},\ B_r,\ C_r)$ be an H_2-optimal approximation of order r. Then, ζ_r satisfies the following multi-point Volterra series*

interpolation conditions:

$$\sum_{k=1}^{\infty}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}H_k\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right)\ =$$
$$\sum_{k=1}^{\infty}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}H_{r_k}\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right),\quad \text{and}$$

$$\sum_{k=1}^{\infty}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}H_k\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right)\right)=$$
$$\sum_{k=1}^{\infty}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}H_{r_k}\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right)\right),$$

where $\phi_{l_1,\ l_2,\ \ldots,\ l_k}$ *and* $\lambda_{l_1},\ \lambda_{l_2},\ \ldots,\ \lambda_{l_k}$ *are residues and poles of the transfer function* H_{r_k} *associated with* ζ_r*, respectively.*

Algorithm 2.1 BIRKA [12]

1: Given an input bilinear dynamical system $A,\ N_1,\ \ldots,\ N_m,\ B,\ C$.

2: Select an initial guess for the reduced system as $\check{A},\ \check{N}_1,\ \ldots,\ \check{N}_m,\ \check{B},\ \check{C}$. Also select stopping tolerance *btol*.

3: while $\big($relative change in eigenvalues of $\check{A} \geqslant btol\big)$

a. $R\Lambda R^{-1} = \check{A},\ \check{\check{B}} = \check{B}^T R^{-T},\ \check{\check{C}} = \check{C}R,\ \check{\check{N}}_j = R^T\check{N}_j R^{-T}$ for $j = 1,\ \ldots,\ m$.

b. $vec\,(V) = \left(-\Lambda\otimes I_n - I_r\otimes A - \sum_{j=1}^{m}\check{\check{N}}_j^T\otimes N_j\right)^{-1}\left(\check{\check{B}}^T\otimes B\right)\ vec(I_m)$.

c. $vec\,(W) = \left(-\Lambda\otimes I_n - I_r\otimes A^T - \sum_{j=1}^{m}\check{\check{N}}_j\otimes N_j^T\right)^{-1}\left(\check{\check{C}}^T\otimes C^T\right)\ vec(I_q)$.

d. $V_r = orth\,(V)\,,\ W_r = orth\,(W)$.

e. $\check{A} = (W_r^TV_r)^{-1}W_r^TAV_r,\quad \check{N}_j = \left(W_r^TV_r\right)^{-1}W_r^TN_jV_r,$

$\check{B} = \left(W_r^TV_r\right)^{-1}W_r^TB,\quad \check{C} = CV_r.$

4: $A_r = \check{A},\quad N_{j_r} = \check{N}_k,\quad B_r = \check{B},\quad C_r = \check{C}$.

Obtaining the residues and the poles of the H_2-optimal reduced model is not possible since we do not have such a system. In [12] the authors propose Bilinear Iterative Rational Krylov Algorithm (BIRKA), which at convergence, ensures that the conditions of Theorem 1 are satisfied. BIRKA gives a locally H_2-optimal reduced model. Algorithm 2.1 lists BIRKA. Next, we study other efficient algorithms for non-parametric bilinear MOR.

2.1 Other Efficient Bilinear MOR Algorithms (Non-parametric)

As mentioned earlier, BIRKA is a computationally expensive algorithm. Hence, next, we first look at its cheaper variant called TBIRKA (Truncated Bilinear Iterative Rational Krylov Algorithm) [24, 26]. TBIRKA is similar to BIRKA in most of the aspects, except that it performs a truncated Volterra series interpolation. Here, instead of ζ in (1.1)-(1.2)[1], authors work with ζ^M, where M is the truncation index (i.e., $k = M$ in (1.2)[1]). Thus, a truncated non-parametric bilinear dynamical system ζ^M is represented as

$$\zeta^M = \left\{H_1\left(s_1\right),\ H_2\left(s_1,\ s_2\right),\ H_3\left(s_1,\ s_2,\ s_3\right),\ \ldots H_M\left(s_1,\ \ldots,\ s_M\right)\right\}, \tag{2.3}$$

with $H_k\left(s_1,\ \ldots,\ s_k\right)$ for $k \in \{1,\ \ldots,\ M\}$ is given by (1.3)[1].

Similar to BIRKA, in TBIRKA also, we have to differentiate an error system expression for getting the H_2-optimality conditions [24, 26]:

$$\begin{aligned}\left\|\zeta^M-\zeta_r^M\right\|_{H_2} = vec(I_{2p})^T\left(\begin{bmatrix}C & -\check{\check{C}}\end{bmatrix}\otimes\begin{bmatrix}C & -\check{C}\end{bmatrix}\right)\\ \sum_{k=0}^{M}\left[\left(-\begin{bmatrix}A & 0\\ 0 & \Lambda\end{bmatrix}\otimes\begin{bmatrix}I_n & 0\\ 0 & I_r\end{bmatrix}-\begin{bmatrix}I_n & 0\\ 0 & I_r\end{bmatrix}\otimes\begin{bmatrix}A & 0\\ 0 & \check{A}\end{bmatrix}\right)^{-1}\right.\\ \left.\sum_{j=1}^{m}\begin{bmatrix}N_j & 0\\ 0 & \check{\check{N}}_j^T\end{bmatrix}\otimes\begin{bmatrix}N_j & 0\\ 0 & \check{N}_j\end{bmatrix}\right]^k\\ \left(-\begin{bmatrix}A & 0\\ 0 & \Lambda\end{bmatrix}\otimes\begin{bmatrix}I_n & 0\\ 0 & I_r\end{bmatrix}-\begin{bmatrix}I_n & 0\\ 0 & I_r\end{bmatrix}\otimes\begin{bmatrix}A & 0\\ 0 & \check{A}\end{bmatrix}\right)^{-1}\\ \left(\begin{bmatrix}B\\ \check{\check{B}}^T\end{bmatrix}\otimes\begin{bmatrix}B\\ \check{B}\end{bmatrix}\right)vec(I_{2m}),\end{aligned}\tag{2.4}$$

where as earlier $\check{A},\ \check{B},\ \check{C}$ and $\check{N}_j$ are the initial guesses for the reduced system. Also, $\check{A} = R\Lambda R^{-1},\ \check{\check{B}} = \check{B}^T R^{-T}, \check{\check{C}} = \check{C}R$ and $\check{\check{N}}_j = R^T\check{N}_j^T R^{-T}$. Here, also, interpolation is performed on the inverse images of the reduced system poles to achieve the H_2-optimality. The following Theorem 2 summarizes this, which is similar to Theorem 1 of BIRKA case, except that the interpolation is performed on a truncated Volterra series.

Theorem 2. *[24, 26] Let $\zeta = (A, N_j, B, C)$ be an order n bilinear system and ζ^M be the polynomial system determined by ζ. Let $\zeta_r = (A_r, N_{j_r}, B_r, C_r)$ be a bilinear system of order r, and define ζ_r^M as the polynomial system determined by ζ_r. Suppose that ζ_r^M is an H_2-optimal approximation to ζ^M. Then ζ_r^M satisfies*

$$\begin{aligned}&\sum_{k=1}^{M}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}H_k\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right) =\\ &\sum_{k=1}^{M}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}H_{r_k}\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right), \quad \text{and}\end{aligned}$$

$$\sum_{k=1}^{M}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}H_k\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right)\right) =$$
$$\sum_{k=1}^{M}\sum_{l_1=1}^{r}\cdots\sum_{l_k=1}^{r}\phi_{l_1,\ l_2,\ \ldots,\ l_k}\left(\sum_{j=1}^{k}\frac{\partial}{\partial s_j}H_{r_k}\left(-\lambda_{l_1},\ -\lambda_{l_2},\ \ldots,\ -\lambda_{l_k}\right)\right),$$

where $\phi_{l_1,\ l_2,\ \ldots,\ l_k}$ *and* $\lambda_{l_1},\ \lambda_{l_2},\ \ldots,\ \lambda_{l_k}$ *are residues and poles of the transfer function* H_{r_k} *associated with* ζ_r^M*, respectively.*

Algorithm 2.2 lists TBIRKA.

Both BIRKA and TBIRKA in turn require solving large sparse linear systems of equations. If we compare Algorithm 2.1 and 2.2, we realize that the number of linear solves at each step of the `While` loop in the former is 2 systems of size $nr \times nr$ and in the latter is $2M$ systems of size $nr \times nr$. This makes it seem that TBIRKA is more expensive than BIRKA. However, TBIRKA is implemented in such a way that the Kronecker products are avoided making it more efficient than BIRKA. For further details on this see Chapter 4 in [24] and Section 5.3 in [26]. These implementation details do not affect our analysis, and hence, we use Algorithm 2.2 in the current form as our base.

Apart from TBIRKA, this class of efficient MOR algorithms also includes balanced truncation based [13], Gramian based [50], moment-matching based [8], and implicit Volterra series based [1] MOR algorithms. For generality, we explore the last two further, i.e., moment-matching based and implicit Volterra series based. Both of these algorithms are proposed for SISO systems[2].

The moment-matching based projection method [8] is a single sided projection method, i.e., $V = W$[3], with

$$span\{V\} = span\{\bigcup_{k=1}^{r} span\{V_k\}\},$$

[2] A SISO non-parametric bilinear dynamical system is represented by (1.1), where system matrices are free from parameters and $B = b \in \mathbb{R}^{n\times 1}$, $C = c \in \mathbb{R}^{1\times n}$, and $j = 1$ (i.e., $N_j = N$). As earlier, we have $E = I_n$.

[3] Here, V and W actually mean V_r and W_r as discussed for BIRKA and TBIRKA, respectively. This is because V and W with subscript r here signifies another set of intermeditory/ sub matrices.

Algorithm 2.2 TBIRKA [24, 26]

1: Given an input bilinear dynamical system $A,\ N_1,\ \ldots,\ N_m,\ B,\ C$.

2: Select an initial guess for the reduced system as $\check{A},\ \check{N}_j,\ \ldots,\ \check{N}_j,\ \check{B},\ \check{C}$. Also select the truncation index M and stopping tolerance *tbtol*.

3: While $\big($relative change in eigenvalues of $\check{A} \geqslant tbtol\big)$

a. $R\Lambda R^{-1} = \check{A},\ \check{\check{B}} = \check{B}^T R^{-T},\ \check{\check{C}} = \check{C}R,\ \check{\check{N}}_j = R^T \check{N}_j R^{-T}$ for $j = 1,\ \ldots,\ m$.

b. Compute

$$vec\,(\mathbf{V}_1) = (-\Lambda \otimes I_n - I_r \otimes A)^{-1} \left(\check{\check{B}}^T \otimes B\right) vec(I_m),$$
$$vec\,(\mathbf{W}_1) = \left(-\Lambda \otimes I_n - I_r \otimes A^T\right)^{-1} \left(\check{\check{C}}^T \otimes C^T\right) vec(I_q).$$

c. For $k = 2, \ldots, M$, solve

$$vec\,(\mathbf{V}_k) = (-\Lambda \otimes I_n - I_r \otimes A)^{-1} \sum_{j=1}^{m} \left(\check{\check{N}}_j^T \otimes N_j\right) vec\,(\mathbf{V}_{k-1}),$$
$$vec\,(\mathbf{W}_k) = \left(-\Lambda \otimes I_n - I_r \otimes A^T\right)^{-1} \sum_{j=1}^{m} \left(\check{\check{N}}_j \otimes N_j^T\right) vec\,(\mathbf{W}_{k-1}).$$

d. $\mathbf{V} = \sum_{k=1}^{M} \mathbf{V}_k,\ \mathbf{W} = \sum_{k=1}^{M} \mathbf{W}_k$.

e. $\mathbf{V}_r = orth\,(\mathbf{V}),\ \mathbf{W}_r = orth\,(\mathbf{W})$.

f. $\check{A} = (\mathbf{W}_r^T \mathbf{V}_r)^{-1} \mathbf{W}_r^T A \mathbf{V}_r,\ \check{N}_j = \left(\mathbf{W}_r^T \mathbf{V}_r\right)^{-1} \mathbf{W}_r^T N_j \mathbf{V}_r,$

$\check{B} = \left(\mathbf{W}_r^T \mathbf{V}_r\right)^{-1} \mathbf{W}_r^T B, \quad \check{C} = C\mathbf{V}_r.$

4: $A_r = \check{A}, \quad N_{j_r} = \check{N}_j, \quad B_r = \check{B}, \quad C_r = \check{C}$.

$$span\{V_1\} = \mathcal{K}^q\left(A^{-1},\ A^{-1}b\right), \quad \text{and}$$
$$span\{V_k\} = \mathcal{K}^q\left(A^{-1},\ A^{-1}NV_{k-1}\right),$$

for $k = 2,\ \ldots,\ M$, where $\mathcal{K}^q$ denotes the standard Krylov subspace of q^{th} order. As evident above, for obtaining the subspace V, we have to solve a sequence of linear systems, whose structure is similar to those arising in TBIRKA algorithm (see Algorithm 2.2).

Implicit Volterra series based MOR algorithm [1] does a weighted interpolation for reduction. Here, the projection matrices are defined as[3]

$$V = [V_1\ \ldots\ V_r] \quad \text{and} \quad W = [W_1\ \ldots\ W_r]$$

where,

$$V_k = \sum_{i=1}^{\infty}\sum_{l_1=1}^{r}\cdots\sum_{l_{i-1}=1}^{r}\eta_{l_1,\ \ldots,\ l_{i-1},k}(\sigma_k E - A)^{-1}N(\sigma_{l_{i-1}}E - A)^{-1}\ldots N(\sigma_{l_1}E - A)^{-1}b,$$
$$W_k = \sum_{i=1}^{\infty}\sum_{l_1=1}^{r}\cdots\sum_{l_{i-1}=1}^{r}\eta_{l_1,\ \ldots,\ l_{i-1},k}(\sigma_k E - A)^{-T}N^T(\sigma_{l_{i-1}}E - A)^{-T}\ldots N^T(\sigma_{l_1}E - A)^{-T}c^T,$$

for $k = 1,\ \ldots, r$. Also, $\sigma_{l_1},\ \ldots,\ \sigma_{l_i}$ are the set of interpolation points and $\eta_{l_1,\ \ldots,\ l_{i-1},k}$ are weights defined in terms of the elements of a $r \times r$ matrix. These weights can be calculated from Lemma 3.1 in [1]. Again, here also, linear systems of equations similar to those arising in TBIRKA (Algorithm 2.2) are to be solved. Since TBIRKA forms the basis of all algorithms in this class, we specifically focus on it for our stability analysis. Next, we look at MOR of parametric linear dynamical systems.

2.2 Parametric Model Order Reduction

Until now, we have focused only on non-parametric dynamical systems (linear or bilinear as the case). Parametric dynamical systems are more challenging and vibrant area of study. As parametric dynamical systems are recent, their MOR algorithms are also contemporary. In general, they are referred to as Parametric Model Order Reduction (PMOR) algorithms.

Here, we first look at parametric linear dynamical system as given by (1.4)-(1.5). Often, the system matrices have parametric dependence as follows [9, 15]:

$$\begin{aligned} E(p) &= E_0 + e_1(p) E_1 + \cdots + e_M(p) E_M, \\ A(p) &= A_0 + f_1(p) A_1 + \cdots + f_M(p) A_M, \\ B(p) &= B_0 + g_1(p) B_1 + \cdots + g_M(p) B_M, \\ C(p) &= C_0 + h_1(p) C_1 + \cdots + h_M(p) C_M, \end{aligned}$$

where $e_i,\ f_i,\ g_i$ and h_i (for $i = 1,\ \ldots,\ M$) are parameter dependent functions and $M \in \mathbb{R}$. However, for rest of this dissertation, we do not assume that our system matrices have such a structure. Let the reduced parametric linear dynamical system be represented as [9, 15]

$$\begin{aligned} E_r(p)\,\dot{x}_r(t) &= A_r(p)\,x_r(t) + B_r(p)\,u(t), \\ y_r(t) &= C_r(p)\,x_r(t), \end{aligned} \tag{2.5}$$

where $E_r(p),\ A_r(p) \in \mathbb{R}^{r\times r}$, $B_r(p) \in \mathbb{R}^{r\times m}$ and $C_r(p) \in \mathbb{R}^{q\times r}$ for $r \ll n$. Here, the input $u(t)$ is the same (maps from $\mathbb{R}$ to $\mathbb{R}^m$) while state $x_r(t)$ maps from $\mathbb{R}$ to $\mathbb{R}^r$ (instead of $\mathbb{R}^n$ earlier). We want $H_r(s;p)$ to approximate $H(s;p)$ in an appropriate norm, and hence, $y_r(t)$ should be nearly equal to $y(t)$ for all admissible inputs, where

$$H_r(s;\ p) = C_r(p)(sE_r(p) - A_r(p))^{-1}B_r(p). \tag{2.6}$$

Analogous to the non-parametric case of [30, 6], in [9], authors propose a set of PMOR algorithms for reducing parametric linear dynamical systems. We focus on the interpolatory projection based PMOR algorithm (Algorithm 4.1 in [9]), called Interpolatory PMOR (IPMOR) because it forms the foundation of all the other algorithms of [9].

Let the two r-dimensional subspaces, $\mathcal{V}_r$ and $\mathcal{W}_r$, be chosen in such a way that $\mathcal{V}_r = Range(V)$ and $\mathcal{W}_r = Range(W)$[3], where $V \in \mathbb{R}^{n\times r}$ and $W \in \mathbb{R}^{n\times r}$ are matrices. Again, as earlier, we project the original full model (1.4) to a lower dimensional subspace, i.e.,

$$\begin{aligned} W^T\left(E(p)\,V\dot{x}_r(t) - A(p)\,V x_r(t) - B(p)\,u(t)\right) &= 0, \\ y_r(t) &= C(p)\,V x_r(t). \end{aligned}$$

Comparing the above equations with (2.5), we get

$$A_r(p) = \left(W^T V\right)^{-1} W^T A(p) V, \quad E_r(p) = \left(W^T V\right)^{-1} W^T E(p) V,$$
$$B_r(p) = \left(W^T V\right)^{-1} W^T B(p), \quad \text{and} \quad C_r(p) = C(p) V,$$

where $p \in \{p^1, \ldots, p^L\}$.

Similar to IRKA in IPMOR also, the subspaces $\mathcal{V}_r$ and $\mathcal{W}_r$ are computed by performing interpolation. That is,

$$\begin{aligned} \mathcal{V}_r &= \underset{\substack{i=1,\ ...,\ K \\ j=1,\ ...,\ L}}{span} \left\{ \left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-1} B\left(p^j\right) \mathbb{r}_{ij} \right\}, \quad \text{and} \\ \mathcal{W}_r &= \underset{\substack{i=1,\ ...,\ K \\ j=1,\ ...,\ L}}{span} \left\{ \left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-T} C\left(p^j\right)^T \mathbb{l}_{ij} \right\}, \end{aligned} \tag{2.7}$$

where $\sigma_1, \ldots, \sigma_K \in \mathbb{C}$ are the points where interpolation is performed (also referred as shifts or frequencies); $p^1, \ldots, p^L \in \mathbb{R}^v$ are parameter vectors; $\mathbb{r}_{11}, \ldots, \mathbb{r}_{1L}, \ldots\ldots, \mathbb{r}_{K1}, \ldots, \mathbb{r}_{KL}$ are right tangential direction vectors with $\mathbb{r}_{ij} \in \mathbb{R}^{m\times 1}$; and $\mathbb{l}_{11}, \ldots, \mathbb{l}_{1L}, \ldots\ldots, \mathbb{l}_{K1}, \ldots, \mathbb{l}_{KL}$ are left tangential direction vectors with $\mathbb{l}_{ij} \in \mathbb{R}^{q\times 1}$. Here, the reduced system size is r, which is equals to $K \times L$. Thus, the projection matrices are built as follows:

$$\begin{aligned} V &= \begin{bmatrix} V_1(p^1) & \cdots & V_K(p^1) & \cdots\cdots & V_1\left(p^L\right) & \cdots & V_K\left(p^L\right) \end{bmatrix}, \quad \text{and} \\ W &= \begin{bmatrix} W_1(p^1) & \cdots & W_K(p^1) & \cdots\cdots & W_1\left(p^L\right) & \cdots & W_K\left(p^L\right) \end{bmatrix}, \end{aligned} \tag{2.8}$$

where from (2.7),

$$\begin{aligned} V_i\left(p^j\right) &= \left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-1} B\left(p^j\right) \mathbb{r}_{ij} \quad \text{and} \\ W_i\left(p^j\right) &= \left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-T} C\left(p^j\right)^T \mathbb{l}_{ij}. \end{aligned} \tag{2.9}$$

Algorithm 2.3 lists IPMOR algorithm. A total of $2KL$ linear systems have to be solved in the IPMOR algorithm.

Apart from the IPMOR algorithm, this class of PMOR algorithms also includes piecewise $\mathcal{H}_2-$optimal interpolatory PMOR [9], multi-parameter and multi-frequency moment-matching based PMOR algorithm [36], PMOR using extended balanced truncation [44], PMOR with $\mathcal{H}_2-$error using radial basis functions [14]. For generality,

Algorithm 2.3 IPMOR Algorithm [9]

1: Given an input parametric linear dynamical system $A(p),\ E(p),\ B(p), C(p)$.

2: Select an initial guess for interpolation points $\sigma_1,\ \ldots,\ \sigma_K \in \mathbb{C}$, parameter vectors $p^1,\ \ldots,\ p^L \in \mathbb{R}^v$, right tangent directions $\{\mathbb{r}_{11},\ \ldots,\ \mathbb{r}_{1L},\ \mathbb{r}_{21},\ \ldots, \mathbb{r}_{KL}\} \subset \mathbb{C}^m$, and left tangent directions $\{\mathbb{l}_{11},\ \ldots,\ \mathbb{l}_{1L},\ \mathbb{l}_{21},\ \ldots, \mathbb{l}_{KL}\} \subset \mathbb{C}^q$. The order of the reduced model will be $r = K \times L$.

3: For $j = 1,\ \ldots,\ L$,

for $i = 1,\ \ldots,\ K$, compute

$$V_i\left(p^j\right) = \left(\sigma_i E(p^j) - A(p^j)\right)^{-1} B\left(p^j\right) \mathbb{r}_{ij}$$
$$W_i\left(p^J\right) = \left(\sigma_i E(p^j) - A(p^j)\right)^{-T} C^T\left(p^j\right) \mathbb{l}_{ij}.$$

4: Set

$$V = \left[V_1\left(p^1\right) \ \cdots \ V_K\left(p^1\right) \ \cdots\cdots \ V_1\left(p^L\right) \ \cdots \ V_K\left(p^L\right)\right] \quad \text{and}$$
$$W = \left[W_1\left(p^1\right) \ \cdots \ W_K\left(p^1\right) \ \cdots\cdots \ W_1\left(p^L\right) \ \cdots \ W_K\left(p^L\right)\right].$$

5: $A_r(p) = (W^T V)^{-1} W^T A(p) V$, $E_r(p) = \left(W^T V\right)^{-1} W^T E(p) V$,
$B_r(p) = \left(W^T V\right)^{-1} W^T B(p)$, $C_r(p) = C(p) V$.

we explore the first two algorithms of the previous list in more detail, i.e., piecewise $\mathcal{H}_2-$optimal interpolatory PMOR and multi-parameter and multi-frequency moment-matching based PMOR algorithm.

In [9], authors extend the IPMOR algorithm to a piecewise H_2-optimal interpolatory PMOR algorithm. Here, for each parameter vector, IRKA (Algorithm 4.1 in [30]) is executed to get the subspaces, i.e., for parameter vector p^j we obtain V_j and W_j, where $j = 1,\ \ldots,\ L$. Finally, we concatenate all the piecewise subspaces to get the final subspaces V and W[3], i.e.,

$$V = [V_1 \ \ldots \ V_L] \quad \text{and} \quad W = [W_1 \ \ldots \ W_L]$$

These subspaces V and W give the piecewise H_2-optimal reduced system.

Similar to IPMOR algorithm, in [36], multi-parameter and multi-frequency moment-matching based PMOR algorithm is derived. Here, the projection matrices V and W^3 are defined as

$$span\{V\} = \mathcal{K}^r\left(\mathbb{M}_1, \mathbb{F}\right) \quad \text{and} \quad span\{W\} = \mathcal{K}^r\left(\mathbb{M}_2, \mathbb{L}\right),$$

where

$$\begin{aligned}
\mathbb{F} &= rowspan\left\{\left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-1} B\left(p^j\right)\right\}_{i=1,\ j=1}^{K,L}, \\
\mathbb{L} &= rowspan\left\{\left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-T} C^T\left(p^j\right)\right\}_{i=1,\ j=1}^{K,L}, \\
\mathbb{M}_1 &= rowspan\left\{\left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-1} E\left(p^j\right)\right\}_{i=1,\ j=1}^{K,L}, \\
\mathbb{M}_2 &= rowspan\left\{\left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^{-T} E^T\left(p^j\right)\right\}_{i=1,\ j=1}^{K,L},
\end{aligned}$$

and as earlier, $\mathcal{K}^r$ denotes the standard Krylov subspace of r^{th} order. Both the above algorithms require solving sequences of linear system of equations as those arising in IPMOR.

Since IPMOR forms the basis of algorithms in this class, we specifically focus on it for our stability analysis. Next, we revisit the standard backward stability definitions and also describe its meaning in our context.

2.3 Backward Stability Analysis

In general, numerical algorithms for a problem are continuous in nature but, a digital computer solves them in a discrete manner. The reason is limitation on the representation of real / complex numbers. Since complex numbers can be represented by real numbers, we focus on latter only. Let $fd : \mathbb{R} \rightarrow \mathbb{F}$ be a function giving a finite approximation to a real number. It provides rounded equivalent as [48]

$$fd(x) = x(1 + \epsilon_{machine}) \text{ for all } x \in \mathbb{R},$$

where $\epsilon_{machine}$ is the machine precision. Also, for every operation between any two finite numbers, the result is exact up to a relative error, i.e., for all x, y $\in \mathbb{F}$

$$fd(x \oplus y) = (x \oplus y)(1 + \epsilon_{machine}),$$

where $\oplus$ can be any of the following operation: $+$, $-$, $*$, and $/$.

Consider a continuous mathematics algorithm $f : X \rightarrow Y$. Say executing this algorithm on a digital computer (that uses finite precision arithmetic) is represented as $\widetilde{f} : X \rightarrow Y$. To check how good the approximated algorithm $\widetilde{f}$ is, one usually computes the accuracy of $\widetilde{f}$. We say an algorithm $\widetilde{f}$ is accurate if [48]

$$\frac{\left\| f(x) - \widetilde{f}(x) \right\|}{\|f(x)\|} = \mathcal{O}(\epsilon_{machine}),$$

where $x \in X$. From the above equation, we find that computing accuracy is not possible since we do not know $f(x)$. A more easier parameter to check the goodness of $\widetilde{f}$ is stability. According to [48], "A stable algorithm gives nearly the right answer to nearly the right question", which although is useful but does not provide a handle on the accuracy. Backward stability is more useful notion in this context. To quote [48], "A backward stable algorithm gives exactly the right answer to nearly the right question". Mathematically, an algorithm f is backward stable if [48]

$$\begin{aligned} \widetilde{f}(x) = f(\widetilde{x}) \quad & \text{for some } \widetilde{x} \text{ with} \\ \frac{\|x - \widetilde{x}\|}{\|x\|} &= \mathcal{O}(\epsilon_{machine}). \end{aligned}$$

This notion of backward stability is useful since one can easily compute accuracy of the result/ output for a backward stable algorithm. The theorem below summarizes this result.

Theorem 3. *[48] If $f : X \rightarrow Y$ is a backward stable algorithm, and $k(x)$ is the condition number of the problem, then the relative error*

$$\frac{\left\| f(x) - \widetilde{f}(x) \right\|}{\|f(x)\|} = \mathcal{O}\left(k(x)\ \epsilon_{machine}\right),$$

where $\epsilon_{machine}$ is the machine precision (or perturbation in x).

All MOR algorithms discussed earlier, require solving sequences of linear system of equations. For such systems, as mentioned earlier, iterative methods are preferred

because of the reduced complexity. Iterative methods are inexact in nature, which means they do not solve linear systems, say `Ax = b`, exactly. Instead `Ax = b` + δ is solved, where δ is the final residual related to the stopping tolerance. Our aim is to find that if one uses an iterative solver (also called inexact solver from now on) in MOR algorithms, then are these algorithms stable with respect to the error introduced by the inexact solves? As earlier, we check for backward stability. For IRKA, the backward stability analysis has been done in [30].

Let in a MOR algorithm, V_r and W_r be calculated exactly, and g be the functional representation of the interpolation process that uses V_r and W_r in this MOR algorithm (i.e., the exact MOR algorithm). Similarly, let $\widetilde{V}_r$ and $\widetilde{W}_r$ be calculated inexactly (i.e., by an iterative solver), and $\widetilde{g}$ be the functional representation of the interpolation process that uses $\widetilde{V}_r$ and $\widetilde{W}_r$ in this MOR algorithm (i.e., the inexact MOR algorithm). Then, from the backward stability definition, this MOR algorithm is backward stable if

$$\widetilde{g}(\Pi) = g(\widetilde{\Pi}) \quad \text{for some } \widetilde{\Pi} \text{ with} \tag{2.10}$$

$$\frac{\|\Pi - \widetilde{\Pi}\|_{H_2 \text{ or } H_\infty}}{\|\Pi\|_{H_2 \text{ or } H_\infty}} = \mathcal{O}(\|F\|), \tag{2.11}$$

where Π and $\widetilde{\Pi}$ denote the original full model and the perturbed full model, respectively, corresponding to the error in the linear solves for $\widetilde{V}_r$ and $\widetilde{W}_r$ in the inexact MOR algorithm. This perturbation is denoted by F.

In the subsequent chapters, we look at the above two conditions for stability in the earlier discussed MOR algorithms for specific types of dynamical systems. As earlier, in the non-parametric bilinear case original full model is represented by ζ (i.e., $\Pi \equiv \zeta$) and after Volterra series truncation, we represent the same original full model by ζ_M (i.e., $\Pi \equiv \zeta_M$), where M is the truncation index. Similarly, in the parametric linear case, the original full model is represented by its transfer function (i.e., $\Pi \equiv H(s;\ p)$).

CHAPTER 3

STABILITY ANALYSIS OF BIRKA

Let the original full order model be represented as $\zeta : A,\ N_1,\ \ldots,\ N_m,\ B,\ C$. Recall from Algorithm 2.1, the following:

$$\begin{aligned} vec\,(V) &= \left(-\Lambda \otimes I_n - I_r \otimes A - \sum_{j=1}^{m} \check{\check{N}}_j^T \otimes N_j\right)^{-1} \left(\check{\check{B}}^T \otimes B\right)\ vec(I_m) \quad \text{and} \\ vec\,(W) &= \left(-\Lambda \otimes I_n - I_r \otimes A^T - \sum_{j=1}^{m} \check{\check{N}}_j \otimes N_j^T\right)^{-1} \left(\check{\check{C}}^T \otimes C^T\right)\ vec(I_p). \end{aligned} \tag{3.1}$$

Also, let the residuals associated with iterative solves for computing $vec(\widetilde{V})$ and $vec(\widetilde{W})$ be $vec(R_B)$ and $vec(R_C)$, respectively. Then, the above equations lead to

$$\left(-\Lambda \otimes I_n - I_r \otimes A - \sum_{j=1}^{m} \check{\check{N}}_j^T \otimes N_j\right) vec\left(\widetilde{V}\right) = \left(\check{\check{B}}^T \otimes B\right)\ vec(I_m) + vec\,(R_B) \ \text{ and} \tag{3.2}$$

$$\left(-\Lambda \otimes I_n - I_r \otimes A^T - \sum_{j=1}^{m} \check{\check{N}}_j \otimes N_j^T\right) vec\left(\widetilde{W}\right) = \left(\check{\check{C}}^T \otimes C^T\right)\ vec(I_p) + vec\,(R_C)\,. \tag{3.3}$$

Let $\widetilde{V}_r = orth(\widetilde{V})$ and $\widetilde{W}_r = orth(\widetilde{W})$. The Petrov-Galerkin projection connects the reduced model matrices (obtained by inexact BIRKA) to the original full model

matrices as

$$\begin{aligned}\widetilde{A}_r &= \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T A\widetilde{V}_r,\quad \widetilde{N}_{j_r} = \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T N_j\widetilde{V}_r,\\ \widetilde{B}_r &= \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T B,\ \text{and}\ \ \widetilde{C}_r = C\widetilde{V}_r,\end{aligned} \tag{3.4}$$

where this reduced model is represented as $\widetilde{\zeta}_r : \widetilde{A}_r,\ \widetilde{N}_{1_r},\ \ldots,\ \widetilde{N}_{m_r},\ \widetilde{B}_r, \widetilde{C}_r$.

By the backward stability definition, next we find a perturbed full model whose exact interpolation will give the reduced model as obtained by inexact interpolation of the original full model. Let the perturbed full model be represented as $\widetilde{\zeta} : \widetilde{A},\ \widetilde{N}_1,\ \ldots,\ \widetilde{N}_m,\ \widetilde{B},\ \widetilde{C}$ or $\widetilde{\zeta} : A+F,\ N_1+E_1,\ \ldots,\ N_m+E_m,\ B+G,\ C+H$, where $F,\ E_1,\ \ldots,\ E_m,\ G,\ H$ are the constant perturbation matrices. Then, we have

$$\begin{aligned}&\left(-\Lambda\otimes I_n - I_r\otimes(A+F) - \sum_{j=1}^{m}\check{\check{N}}_j^T\otimes(N_j+E_j)\right)vec\left(\widetilde{V}\right)\\ &\quad= \left(\check{\check{B}}^T\otimes(B+G)\right)\ vec(I_m)\quad\text{and}\\ &\left(-\Lambda\otimes I_n - I_r\otimes(A+F)^T - \sum_{j=1}^{m}\check{\check{N}}_j\otimes(N_j+E_j)^T\right)vec\left(\widetilde{W}\right)\\ &\quad= \left(\check{\check{C}}^T\otimes(C+H)^T\right)\ vec(I_p),\end{aligned} \tag{3.5}$$

or

$$\begin{aligned}\left(-\Lambda\otimes I_n - I_r\otimes A - \sum_{j=1}^{m}\check{\check{N}}_j^T\otimes N_j\right)vec\left(\widetilde{V}\right) &= \left(\check{\check{B}}^T\otimes B\right)\ vec(I_m)\\ &+ \left(\check{\check{B}}^T\otimes G\right)\ vec(I_m)\\ &+ \left(I_r\otimes F + \sum_{j=1}^{m}\check{\check{N}}_j^T\otimes E_j\right)vec\left(\widetilde{V}\right)\quad\text{and}\end{aligned} \tag{3.6}$$

$$\begin{aligned}\left(-\Lambda\otimes I_n - I_r\otimes A^T - \sum_{j=1}^{m}\check{\check{N}}_j\otimes N_j^T\right)vec\left(\widetilde{W}\right) &= \left(\check{\check{C}}^T\otimes C^T\right)\ vec(I_p)\\ &+ \left(\check{\check{C}}^T\otimes H^T\right)\ vec(I_p)\\ &+ \left(I_r\otimes F^T + \sum_{j=1}^{m}\check{\check{N}}_j\otimes E_j^T\right)vec\left(\widetilde{W}\right).\end{aligned} \tag{3.7}$$

As earlier, $\widetilde{V}_r = orth(\widetilde{V})$ and $\widetilde{W}_r = orth(\widetilde{W})$. Using the Petrov-Galerkin projection to connect the reduced model matrices (obtained by exact BIRKA) with the perturbed

full model matrices we get

$$\begin{aligned}\widehat{A}_r &= \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T\left(A+F\right)\widetilde{V}_r, \quad \widehat{N}_{j_r} = \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T\left(N_j+E_j\right)\widetilde{V}_r,\\ \widehat{B}_r &= \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T\left(B+G\right), \quad \text{and} \quad \widehat{C}_r = \left(C+H\right)\widetilde{V}_r,\end{aligned} \tag{3.8}$$

where this reduced model is represented as $\widehat{\zeta}_r : \widehat{A}_r,\ \widehat{N}_{1_r},\ \ldots,\ \widehat{N}_{m_r},\ \widehat{B}_r, \widehat{C}_r$. To satisfy the backward stability's first condition (2.10), we equate the reduced models in (3.4) and (3.8). That is,

$$\begin{aligned}\widehat{A}_r &= \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^T\left(A+F\right)\widetilde{V}_r = \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^TA\widetilde{V}_r + \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^TF\widetilde{V}_r\\ &= \widetilde{A}_r + \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^TF\widetilde{V}_r.\end{aligned}$$

Similarly, $\widehat{N}_{j_r} = \widetilde{N}_{j_r} + \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^TE_j\widetilde{V}_r$, $\widehat{B}_r = \widetilde{B}_r + \left(\widetilde{W}_r^T\widetilde{V}_r\right)^{-1}\widetilde{W}_r^TG$ and $\widehat{C}_r = \widetilde{C}_r + H\widetilde{V}_r$.

From the above, we note that if $\widetilde{W}_r^TF\widetilde{V}_r = 0$, then $\widehat{A}_r = \widetilde{A}_r$. Similarly, if $\widetilde{W}_r^TE_j\widetilde{V}_r = 0$, then $\widehat{N}_{j_r} = \widetilde{N}_{j_r}$; if $\widetilde{W}_r^TG = 0$, then $\widehat{B}_r = \widetilde{B}_r$; and if $H\widetilde{V}_r = 0$, then $\widehat{C}_r = \widetilde{C}_r$. Using the Petrov-Galerkin framework for the inexact solves in (3.2) and (3.3), we can easily achieve some of the above relations. We discuss this next.

The Petrov-Galerkin Framework for Inexact Solves

The Petrov-Galerkin framework by definition implies finding the solution of a linear system of equation, such that its residual at every point is orthogonal to some other suitable subspace [49]. In our context, we define the Petrov-Galerkin framework as below.

$$\begin{aligned}&\text{Find} \quad \widetilde{V} \in \mathcal{P}_r \quad \text{such that} \quad R_B \perp \mathcal{Q}_r \quad \text{and}\\ &\text{find} \quad \widetilde{W} \in \mathcal{Q}_r \quad \text{such that} \quad R_C \perp \mathcal{P}_r,\end{aligned} \tag{3.9}$$

where $\mathcal{P}_r$ and $\mathcal{Q}_r$ are any two r-dimensional subspaces of $\mathbb{C}^n$; $\widetilde{V}$ and R_B satisfy (3.2); and $\widetilde{W}$ and R_C satisfy (3.3).

Comparing (3.2) with (3.6) and (3.3) with (3.7), we get the following equations:

$$vec\,(R_B) = \left(\breve{\tilde{B}}^T \otimes G\right)\ vec(I_m) + \left(I_r \otimes F + \sum_{j=1}^{m} \breve{\tilde{N}}_j^T \otimes E_j\right) vec\left(\tilde{V}\right) \qquad \text{and}$$

$$vec\,(R_C) = \left(\breve{\tilde{C}}^T \otimes H^T\right)\ vec(I_p) + \left(I_r \otimes F^T + \sum_{j=1}^{m} \breve{\tilde{N}}_j \otimes E_j^T\right) vec\left(\widetilde{W}\right) \quad \text{or}$$

$$R_B = G\breve{\tilde{B}} + F\tilde{V} + \sum_{j=1}^{m} E_j \tilde{V} \breve{\tilde{N}}_j \quad \text{and} \quad R_C = H^T\breve{\tilde{C}} + F^T\widetilde{W} + \sum_{j=1}^{m} E_j^T \widetilde{W} \breve{\tilde{N}}_j^T. \tag{3.10}$$

Next, we consider perturbations in $A,\ N_j,\ B$ and C individually, and use the Petrov-Galerkin framework discussed above. First, if we take the perturbation F in A only, then (3.10) is equivalent to

$$R_B = F\tilde{V} \qquad \text{and} \qquad R_C^T = \widetilde{W}^T F. \tag{3.11}$$

In the above, if we multiply $\widetilde{W}^T$ from left in the first equation and $\tilde{V}$ from right in the second equation, then we get

$$\widetilde{W}^T R_B = \widetilde{W}^T F\tilde{V} \qquad \text{and} \qquad R_C^T\tilde{V} = \widetilde{W}^T F\tilde{V}.$$

From the Petrov-Galerkin framework (3.9), $\widetilde{W} \perp R_B$ and $\tilde{V} \perp R_C$, and hence,

$$\widetilde{W}^T F\tilde{V} = 0.$$

We also have[1]

$$\widetilde{W}_r^T F\tilde{V}_r = 0. \tag{3.12}$$

Similarly, if we take the perturbation E_j in *any one* N_j matrix, then (3.10) is equivalent to

$$R_B = E_j\tilde{V}\breve{\tilde{N}}_j \qquad \text{and} \qquad R_C^T = \breve{\tilde{N}}_j\widetilde{W}^T E_j.$$

[1]Since $\tilde{V}_r = orth(\tilde{V})$ and $\widetilde{W}_r = orth(\widetilde{W})$, we have $\tilde{V} = \tilde{V}_r\ Z_1$ and $\widetilde{W} = \widetilde{W}_r\ Z_2$, where Z_1 and Z_2 are lower triangular matrices. Here $\widetilde{W}^T F\tilde{V} = 0$ implies $Z_2^T\left(\widetilde{W}_r^T F\tilde{V}_r\right) Z_1 = 0$. If $\tilde{V}$ and $\widetilde{W}$ are full ranked then, Z_1 and Z_2 are invertible and we have $\widetilde{W}_r^T F\tilde{V}_r = 0$. This full rank assumption exists in original BIRKA as well (see Lemma 5.2 in [12]).

Again in the above, if we multiply $\widetilde{W}^T$ from left in the first equation and $\widetilde{V}$ from right in the second equation, then we get

$$\widetilde{W}^T R_B = \widetilde{W}^T E_j \widetilde{V} \breve{N}_j \qquad \text{and} \qquad R_C^T \widetilde{V} = \breve{N}_j \widetilde{W}^T E_j \widetilde{V}.$$

Using the Petrov-Galerkin framework (3.9) in above we get

$$\widetilde{W}^T E_j \widetilde{V} \breve{N}_j = 0 \qquad \text{and} \qquad \breve{N}_j \widetilde{W}^T E_j \widetilde{V} = 0.$$

To achieve the desired result, i.e., $\widetilde{W}_r^T E_j \widetilde{V}_r = 0$, we need $\breve{N}_j$ to be invertible. This cannot always be guaranteed. Thus, we drop the perturbation analysis with N_j matrices.

Finally, if we only take the perturbations G and H, in the matrices B and C, respectively, then (3.10) is equivalent to

$$R_B = G\breve{B} \qquad \text{and} \qquad R_C^T = \breve{C}^T H.$$

As in the last two paragraphs, multiplying by $\widetilde{W}^T$ from left in the first equation above, multiplying by $\widetilde{V}$ from right in the second equation above, and using the Petrov-Galerkin framework (3.9) we get

$$\widetilde{W}^T G \breve{B} = 0 \qquad \text{and} \qquad \breve{C}^T H \widetilde{V} = 0.$$

As above, to achieve the desired result, i.e., $\widetilde{W}_r^T G = 0$ and $H\widetilde{V}_r = 0$, we need $\breve{B}$ and $\breve{C}$ to be invertible. This cannot always be guaranteed because these are non-square matrices. Thus, we drop the perturbation analysis with B and C matrices both.

Hence, (3.12) implies that if we consider the perturbation in A matrix only and use a Petrov-Galerkin framework for the inexact linear solves, then

$$\widehat{A}_r = \widetilde{A}_r, \quad \widehat{N}_{j_r} = \widetilde{N_{j_r}}, \quad \widehat{B}_r = \widetilde{B}_r, \quad \text{and} \quad \widehat{C}_r = \widetilde{C}_r \quad \text{or}$$
$$\widetilde{g}(\zeta) = g(\widetilde{\zeta}).$$

The theorem below summarizes this.

Theorem 4. *If the inexact linear solves in BIRKA (line 3b. and 3c. of Algorithm 2.1) are solved using the Petrov-Galerkin framework* (3.9)*, then BIRKA satisfies the first condition of backward stability with respect to these solves, i.e.,* (2.10)*.*

Next, we look at the second condition of stability in BIRKA.

3.1 Second Condition of Backward Stability

Next, we show that the *second condition* of backward stability, given in (2.11), is also satisfied. According to (2.11), the difference between the original full model and the perturbed full model should be order of the perturbation, i.e.,

$$\frac{\|\zeta - \widetilde{\zeta}\|_{H_2 \text{ or } H_\infty}}{\|\zeta\|_{H_2 \text{ or } H_\infty}} = \mathcal{O}(\|F\|),$$

where H_2-norm is defined in (1.8). We satisfy the above condition in the absolute sense, since ζ is independent of F. That is,

$$\left\|\zeta - \widetilde{\zeta}\right\|_{H_2}^2 = \mathcal{O}\Big(\|F\|_2\Big).$$

Consider the error system $\zeta^{err} = \zeta - \widetilde{\zeta}$ whose matrices are defined as follows [12, 24]:

$$A^{err} = \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix},\ N_j^{err} = \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix},\ B^{err} = \begin{bmatrix} B \\ B \end{bmatrix},\ \text{and} \quad C^{err} = \begin{bmatrix} C & -C \end{bmatrix}.$$

The H_2-norm of this error system is

$$\begin{aligned} \|\zeta^{err}\|_{H_2}^2 =& vec(I_{2p})^T \left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix}\right) \\ & \times \left(-\begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \right. \\ & \left. - \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix} \right) vec(I_{2m}), \end{aligned} \tag{3.13}$$

$$
\begin{aligned}
=&vec(I_{2p})^T \left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix} \right) \\
&\times \left(- \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \right. \\
&\left. - \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \right)^{-1} \\
&\times \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix} \right) vec(I_{2m}).
\end{aligned}
$$

Let

$$
\begin{aligned}
\widehat{C} &= \left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix} \right), \\
\widehat{Q} &= \left(- \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \right. \\
&\qquad \left. - \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \right), \qquad (3.14) \\
\widehat{F} &= \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix}, \qquad (3.15) \\
\widehat{\widehat{F}} &= \left(I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n} \right), \text{ and} \qquad (3.16) \\
\widehat{B} &= \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix} \right).
\end{aligned}
$$

Then, the norm of this error system is

$$
\begin{aligned}
\|\zeta^{err}\|_{H_2}^2 &= vec(I_{2p})^T \widehat{C} \left(\widehat{Q} - \widehat{\widehat{F}} \right)^{-1} \widehat{B} \; vec(I_{2m}), \qquad (3.17) \\
&= vec(I_{2p})^T \widehat{C} \widehat{Q}^{-1} \left(I_{4n^2} - \widehat{\widehat{F}} \widehat{Q}^{-1} \right)^{-1} \widehat{B} \; vec(I_{2m}).
\end{aligned}
$$

If $\left\|\widehat{\widehat{F}}\widehat{Q}^{-1}\right\|_2 < 1$, then by the Neumann series we get that

$$\begin{aligned}
\|\zeta^{err}\|_{H_2}^2 &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{4n^2} - \widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}),\\
&= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{4n^2} + \widehat{\widehat{F}}\widehat{Q}^{-1} + \left(\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^2 + \dots\right)\widehat{B}\ vec(I_{2m}),\\
&= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{B}\ vec(I_{2m})\\
&\quad + vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(I_{4n^2} + \widehat{\widehat{F}}\widehat{Q}^{-1} + \left(\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^2 + \dots\right)\widehat{B}\ vec(I_{2m}).
\end{aligned}$$

Since $\|\zeta - \zeta\|_{H_2}^2 = vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{B}\ vec(I_{2m}) = 0$, the above equation simplifies to

$$\|\zeta^{err}\|_{H_2}^2 = vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(I_{4n^2} + \widehat{\widehat{F}}\widehat{Q}^{-1} + \left(\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^2 + \dots\right)\widehat{B}\ vec(I_{2m}). \tag{3.18}$$

Bounding the right hand side of the above equation we get the following:

$$\begin{aligned}
&\left|vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(I_{4n^2} - \widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m})\right|\\
&\qquad \leqslant \left\|vec(I_{2p})^T\right\|\left\|\widehat{C}\widehat{Q}^{-1}\right\|\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|\left\|\left(I_{4n^2} - \widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\right\|\left\|\widehat{B}\right\|\left\|vec(I_{2m})\right\|.
\end{aligned}$$

Using Lemma 2.3.3 from [27], we get that the right hand side above is bounded by

$$\left\|vec(I_{2p})^T\right\|\left\|\widehat{C}\widehat{Q}^{-1}\right\|\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|\left(\frac{1}{1 - \left\|\widehat{\widehat{F}}\widehat{Q}^{-1}\right\|}\right)\left\|\widehat{B}\right\|\left\|vec(I_{2m})\right\|.$$

Substituting the above two results in (3.18) we get

$$\|\zeta^{err}\|_{H_2}^2 \leqslant \left\|vec(I_{2p})^T\right\|\left\|\widehat{C}\widehat{Q}^{-1}\right\|\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|\left(\frac{1}{1 - \left\|\widehat{\widehat{F}}\widehat{Q}^{-1}\right\|}\right)\left\|\widehat{B}\right\|\left\|vec(I_{2m})\right\|. \tag{3.19}$$

Let $\left\|\widehat{Q}^{-1}\right\| < 1$, which is defined by the original system (further analyzed in Section 3.2.1) and $\left\|\widehat{\widehat{F}}\right\| < 1$, which is related to the residuals of linear solves (further analyzed

in Section 3.2.2). Then, using the matrix norm property we have the following:

$$\left\|\widehat{\widehat{F}}\widehat{Q}^{-1}\right\| \leqslant \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \text{ or}$$
$$\frac{1}{1-\left\|\widehat{\widehat{F}}\widehat{Q}^{-1}\right\|} \leqslant \frac{1}{1-\left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}.$$

Substituting the above in (3.19) we get

$$\|\zeta^{err}\|_{H_2}^2 \leqslant \left\|vec(I_{2p})^T\right\| \left\|\widehat{C}\widehat{Q}^{-1}\right\| \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}\right) \left\|\widehat{B}\right\| \|vec(I_{2m})\| \quad (3.20)$$

or

$$\|\zeta^{err}\|_{H_2}^2 \leqslant \mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right). \quad (3.21)$$

Next, we relate $\left\|\widehat{\widehat{F}}\right\|$ and $\|F\|$. From (3.16) we know

$$\widehat{\widehat{F}} = \left(I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n}\right).$$

Taking norms on both the sides of the above equation, and applying the triangle inequality property ($\|X+Y\| \leqslant \|X\| + \|Y\|$) we get

$$\left\|\widehat{\widehat{F}}\right\| = \left\|I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n}\right\| \leqslant \left\|I_{2n} \otimes \widehat{F}\right\| + \left\|\widehat{F} \otimes I_{2n}\right\|.$$

Further, using the norm distribution property of Kronecker product ($\|X \otimes Y\| = \|X\| \|Y\|$) [35, 34], we have the following:

$$\left\|\widehat{\widehat{F}}\right\| \leqslant \|I_{2n}\| \left\|\widehat{F}\right\| + \left\|\widehat{F}\right\| \|I_{2n}\|,$$
$$\leqslant \mathcal{O}\left(\left\|\widehat{F}\right\|\right).$$

From (3.15) we know $\widehat{F} = \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix}$. Using the definitions of all the commonly used matrix norms (Frobenius, 2, 1 and ∞) [38] we get

$$\mathcal{O}\left(\|\widehat{\widehat{F}}\|\right) \leqslant \mathcal{O}\left(\|F\|\right). \quad (3.22)$$

Substituting the above in (3.21) we get

$$\|\zeta^{err}\|_{H_2}^2 = \left\|\zeta - \widetilde{\zeta}\right\|_{H_2}^2 \leqslant \mathcal{O}\Big(\|F\|\Big).$$

Thus, we have satisfied the second condition of backward stability. The theorem below summarizes this.

Theorem 5. *If $\widehat{Q}$ defined in* (3.14) *is invertible, $\left\|\widehat{Q}^{-1}\right\| < 1$, and $\left\|\widehat{\widehat{F}}\right\| < 1$, where $\widehat{\widehat{F}}$ is defined in* (3.16)*, then BIRKA satisfies the second condition of backward stability with respect to the inexact linear solves, i.e.,* (2.11)*.*

The hypotheses of this theorem are usually easy to satisfy, and are discussed in the next section. The corollary below summarizes our stability result.

Corollary 1. *Assuming the hypotheses of Theorem 4 and Theorem 5 are satisfied, then BIRKA is backward stable with respect to the inexact linear solves.*

In the next section, we analyze all the involved matrices and accuracy of the reduced system.

3.2 Analysis

Next, we analyze our assumptions and results from the previous sections. First, we revisit the assumed invertibility of all relevant matrices (in Section 3.2.1). Second, we derive the expression for accuracy of the reduced system, in-terms of the residuals of the linear solves as well as the conditioning of the bilinear system (in Section 3.2.2).

3.2.1 Invertibility of Involved Matrices

Until now, we have assumed invertibility of eight matrices. Most of these invertibility assumptions directly come from the control system theory as well as the model reduction theory of bilinear systems. We have also assumed invertibility of few newly proposed matrices. In this subsection, we summarize/ analyze all these assumptions in the order of appearance of the corresponding matrix in this chapter. We first summarize the invertibility assumptions from literature.

(a) We assume invertibility of $(s_k I_n - A)$ and $(sI_n - A)$ in (1.3) and (1.5)[1], respectively. These come from the transfer function definitions. Please see Section 2 of [26] and Section 1 of [30], respectively.

(b) In the H_2-norm definition of a bilinear dynamical system (1.8), we assume that $\big(- A \otimes I_n - I_n \otimes A - \sum_{j=1}^{m} N_j \otimes N_j \big)$ is invertible. This is a standard definition. Please see Theorem 3.4 of [12].

(c) We assume invertibility of $\left(\widetilde{W}_r^T \widetilde{V}_r\right)$. As mentioned earlier, this is easy to enforce and come from BIRKA. Please see Algorithm 2 of [12] or Algorithm 1 of [26].

(d) In (2.2), we assume the middle term, i.e.,

$$\left(- \begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_r \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_r \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \check{A} \end{bmatrix} - \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & \check{\check{N}}_j^T \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & \check{N}_j \end{bmatrix} \right)$$

is invertible. This comes from the H_2-norm of the error system $(\zeta - \zeta_r)$. Please see Corollary 4.1 of [12] or Theorem 4.5 of [24].

(e) We assume invertibility of $\big(- \Lambda \otimes I_n - I_r \otimes A - \sum_{j=1}^{m} \check{\check{N}}_j^T \otimes N_j \big)$ in Algorithm 2.1. This again comes from BIRKA. Please see Algorithm 2 of [12] or Algorithm 1 of [26].

During the backward stability analysis of BIRKA, we assume invertibility of some newly proposed matrices. Next, we analyze these matrices. Note that below, we discuss the matrix in (b) before the matrix in (c) although the latter appears first in this chapter. This is done for ease of exposition.

(a) In IRKA [30], $(sI - A)$ is inverted to form the projection subspace. Hence, in the backward stability analysis of IRKA, invertibility of the corresponding perturbed matrix $(sI - (A + F))$ is assumed (see Theorem 4.1 of [10]). As discussed in (e) above, in BIRKA, $\big(- \Lambda \otimes I_n - I_r \otimes A - \sum_{j=1}^{m} \check{\check{N}}_j^T \otimes N_j \big)$ is inverted to form

the projection subspace. Hence, we assume invertibility of the corresponding perturbed matrix $\big(-\Lambda\otimes I_n - I_r\otimes(A+F) - \sum_{j=1}^{m}\breve{N}_j^T\otimes(N_j+E_j)\big)$ in (3.5).

(b) We assume invertibility of $\widehat{Q}$ given in (3.14). Also listed below for easy access.

$$\widehat{Q} = -\begin{bmatrix} A & 0\\ 0 & A\end{bmatrix}\otimes\begin{bmatrix} I_n & 0\\ 0 & I_n\end{bmatrix} - \begin{bmatrix} I_n & 0\\ 0 & I_n\end{bmatrix}\otimes\begin{bmatrix} A & 0\\ 0 & A\end{bmatrix} - \sum_{j=1}^{m}\begin{bmatrix} N_j & 0\\ 0 & N_j\end{bmatrix}\otimes\begin{bmatrix} N_j & 0\\ 0 & N_j\end{bmatrix}.$$

This is one of the most important assumption in obtaining a backward stable BIRKA (see Corollary 1). Hence, here we relate this invertibility assumption with the underlying bilinear dynamical system. If we define $A_2 = \begin{bmatrix} A & 0\\ 0 & A\end{bmatrix}$, $I_{2n} = \begin{bmatrix} I_n & 0\\ 0 & I_n\end{bmatrix}$, $N_{2j} = \begin{bmatrix} N_j & 0\\ 0 & N_j\end{bmatrix}$ and $\widehat{Q} = Q_1\otimes Q_2$, where $Q_1, Q_2 \in \mathbb{R}^{2n\times 2n}$ are any two matrices, then $\widehat{Q}$ can be rewritten as

$$-A_2\otimes I_{2n} - I_{2n}\otimes A_2 - \sum_{j=1}^{m} N_{2j}\otimes N_{2j} = Q_1\otimes Q_2 \qquad \text{or}$$

$$-(A_2\otimes I_{2n})\,vec(I_{2n}) - (I_{2n}\otimes A_2)\,vec(I_{2n}) - \sum_{j=1}^{m}(N_{2j}\otimes N_{2j})\,vec(I_{2n}) = (Q_1\otimes Q_2)\,vec(I_{2n}) \quad \text{or}$$

$$-A_2^T - A_2 - \sum_{j=1}^{m} N_{2j}N_{2j}^T = Q_2Q_1^T \qquad \text{or}$$

$$-\begin{bmatrix} A & 0\\ 0 & A\end{bmatrix}^T - \begin{bmatrix} A & 0\\ 0 & A\end{bmatrix} - \sum_{j=1}^{m}\begin{bmatrix} N_j & 0\\ 0 & N_j\end{bmatrix}\begin{bmatrix} N_j & 0\\ 0 & N_j\end{bmatrix}^T = Q_2Q_1^T \qquad \text{or}$$

$$\begin{bmatrix} -A^T - A - \sum_{j=1}^{m} N_jN_j^T & 0\\ 0 & -A^T - A - \sum_{j=1}^{m} N_jN_j^T\end{bmatrix} = Q_2Q_1^T.$$

If $\left(-A^T - A - \sum_{j=1}^{m} N_jN_j^T\right)$ is invertible, then Q_1 and Q_2 are invertible. This implies that $\widehat{Q} = (Q_1\otimes Q_2)$ is invertible. Consider the following generalized

Lyapunov equation used in the derivation of BIRKA [12, 13]:

$$AP + PA^T + \sum_{j=1}^{m} N_j P N_j^T = -BB^T.$$

If the solution of this equation is the identity matrix (i.e., $P = I_n$), then the left hand side matrix in this Lyapunov equation is $\left(A^T + A + \sum_{j=1}^{m} N_j N_j^T \right)$, which needs to be invertible for invertibility of $\widehat{Q}$.

(c) In (3.13) and (3.17), we assume invertibility of

$$\left(- \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} - \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \right)$$

and $\left(\widehat{Q} - \widehat{\widehat{F}} \right)$, respectively, both of which represent the same matrix (i.e., $\widehat{Q}$ with perturbation). This matrix is invertible if $\left(- (A+F)^T - (A+F) - \sum_{j=1}^{m} N_j N_j^T \right)$ is invertible.

3.2.2 Accuracy of the Reduced System

Assume that BIRKA satisfies the hypotheses of Corollary 1, i.e., it is backward stable with respect to the inexact linear solves. Then, from Theorem 3 we get that

$$\frac{\|g(\zeta) - \widetilde{g}(\zeta)\|_{H_2}}{\|g(\zeta)\|_{H_2}} = \mathcal{O}\left(k(\zeta)\, \|F\|\right),$$

where, as earlier (recall (2.10)-(2.11)), g denotes exact BIRKA, $\widetilde{g}$ denotes inexact BIRKA, ζ is the original full model, $k(\zeta)$ is the condition number of ζ (discussed below), and F is the perturbation in ζ.

If we define, $g(\zeta) = \zeta_r$, and $\widetilde{g}(\zeta) = \widetilde{\zeta}_r$, then the above equation can be rewritten as

$$\frac{\left\|\zeta_r - \widetilde{\zeta}_r\right\|_{H_2}}{\|\zeta_r\|_{H_2}} = \mathcal{O}\left(k(\zeta)\, \|F\|\right).$$

Here, we are looking at the reduced systems obtained at line 3e. of Algorithm 2.1, i.e., at the end of every iterative step of BIRKA. Thus, accuracy of the reduced system is dependent on the conditioning of the problem as well as the perturbation. Next, we look at both these quantities separately.

First, we want to compute conditioning of our bilinear system with respect to performing the inexact linear solves on lines 3b. and 3c. of Algorithm 2.1. Since for backward stability we equate the reduced model obtained by performing inexact BIRKA on the original full model (ζ) and performing exact BIRKA on the perturbed full model ($\widetilde{\zeta}$), these inexact linear solves are captured by $\widetilde{\zeta}$. Thus, the condition number of our bilinear system with respect to computing the H_2-norm of the error system $\zeta_{err} = \zeta - \widetilde{\zeta}$ will give us a *good approximation* to the condition number that we want to compute (with respect to computing the H_2-norm of $\widetilde{\zeta}_r - \zeta$ or $\widetilde{\zeta}_r - \zeta_r$). Similar behavior has been observed for linear dynamical systems (see Theorem 3.1 and 3.3 in [10]).

Recall, the condition number by definition means relative change in the output (for us this is $\frac{\left\|\zeta - \widetilde{\zeta}\right\|_{H_2}}{\|\zeta\|_{H_2}}$) with respect to the relative change in the input (for us this is $\frac{\|F\|}{\|A\|}$ since we are perturbing the A matrix). Hence, from (3.20) we have

$$\left\|\zeta - \widetilde{\zeta}\right\|_{H_2} \leqslant \left\|vec\,(I_{2p})^T\right\| \left\|\widehat{C}\widehat{Q}^{-1}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{B}\right\| \|vec(I_{2m})\| \frac{\left\|\widehat{\widehat{F}}\right\|}{1 - \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}, \tag{3.23}$$

where $\left\|\widehat{Q}^{-1}\right\| < 1$ and $\left\|\widehat{\widehat{F}}\right\| < 1$. Since $\left\|\widehat{\widehat{F}}\right\| < 1$, then we also have

$$\frac{1}{1 - \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|} \leqslant \frac{1}{1 - \left\|\widehat{Q}^{-1}\right\|}.$$

Using above, (3.23) can be rewritten as

$$\left\|\zeta-\widetilde{\zeta}\right\|_{H_2} \leqslant \left\|vec\left(I_{2p}\right)^T\right\| \left\|\widehat{C}\widehat{Q}^{-1}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{B}\right\| \|vec(I_{2m})\| \frac{\left\|\widehat{\widehat{F}}\right\|}{1-\left\|\widehat{Q}^{-1}\right\|} \quad or$$

$$\frac{\left\|\zeta-\widetilde{\zeta}\right\|_{H_2}}{\|\zeta\|_{H_2}} \leqslant \frac{\left\|vec\left(I_{2p}\right)^T\right\| \left\|\widehat{C}\widehat{Q}^{-1}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{B}\right\| \|vec(I_{2m})\| \|A\|}{\|\zeta\|_{H_2}} \frac{1}{1-\left\|\widehat{Q}^{-1}\right\|} \frac{\left\|\widehat{\widehat{F}}\right\|}{\|A\|}.$$

From (3.22), we know $\|\widehat{\widehat{F}}\| \leqslant \|F\|$. Hence, the above inequality is equivalent to

$$\frac{\left\|\zeta-\widetilde{\zeta}\right\|_{H_2}}{\|\zeta\|_{H_2}} \leqslant k\left(\zeta\right) \frac{\|F\|}{\|A\|},$$

where

$$k\left(\zeta\right) = \frac{\left\|vec\left(I_{2p}\right)^T\right\| \left\|\widehat{C}\widehat{Q}^{-1}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{B}\right\| \|vec(I_{2m})\| \|A\|}{\|\zeta\|_{H_2}} \frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}. \tag{3.24}$$

In the numerical experiments section, for both our problems, we show that this condition number is fairly small[2]. In other words, both our problems are well-conditioned with respect to computing the H_2-norm of the error system ζ_{err}. Note that $\left\|\widehat{Q}^{-1}\right\| < 1$ and $\left\|\widehat{\widehat{F}}\right\| < 1$ as assumed here come from the assumptions for backward stability of BIRKA (see Corollary 1), and hence, we do not need any extra assumptions.

Second, we relate the perturbation F with the residuals R_B and R_C given in (3.2) and (3.3), respectively. Recall that we are considering the perturbation F in A matrix, and hence, this F should satisfy both the equations in (3.11). That is,

$$R_B = F\widetilde{V} \qquad \text{and} \qquad R_C^T = \widetilde{W}^T F. \tag{3.25}$$

From the assumptions for backward stability of BIRKA (Corollary 1), we know that we need to use a Petrov-Galerkin framework, i.e.,

$$\widetilde{W} \perp R_B \qquad \text{and} \qquad \widetilde{V} \perp R_C, \tag{3.26}$$

[2]If the problem is ill-conditioned (i.e., the condition number is large), then we cannot get a good handle on the accuracy of the reduced system.

where $\widetilde{V}$ and $\widetilde{W}$ are again given in (3.2) and (3.3), respectively. Using (3.26), we get that

$$F = R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T + \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T, \tag{3.27}$$

satisfies (3.25). This is assuming $\left(\widetilde{W}^T\widetilde{V}\right)$ is nonsingular, which has already been discussed in the previous subsection. The theorem below gives a bound on this F. This theorem is similar to Theorem 4.2 from [10] in the linear case.

Theorem 6. *Let R_B and $\widetilde{V}$ be defined as in* (3.2)*, R_C and $\widetilde{W}$ be defined as in* (3.3)*, and F be defined as in* (3.27)*. Define $R_B = [R_{B_1},\ R_{B_2},\ \ldots,\ R_{B_r}]$ and $R_C = [R_{C_1},\ R_{C_2},\ \ldots,\ R_{C_r}]$ and assume $\widetilde{W}^T\widetilde{V}$ is nonsingular. Then, the perturbation F satisfies*

$$\|F\|_2 \leqslant \|F\|_F \leqslant \sqrt{r}\left\{\max_i \|R_{B_i}\| \left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\| + \max_i \|R_{C_i}\| \left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|\right\}.$$

Proof. Note that

$$\begin{aligned} F &= R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T + \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T. \\ \|F\|_F &= \left\|R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T + \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T\right\|_F \\ \|F\|_F &\leqslant \left\|R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|_F + \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T\right\|_F. \end{aligned}$$

Consider the first term from the above expression as

$$\begin{aligned} \left\|R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|_F &\leqslant \|R_B\|_F \left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\| \\ &\leqslant \sqrt{r}\ \max_i \|R_{B_i}\| \left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|. \end{aligned}$$

Similarly, taking the second term as

$$\begin{aligned} \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T\right\|_F &\leqslant \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\| \|R_C\|_F \\ &\leqslant \sqrt{r}\ \max_i \|R_{C_i}\| \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\|. \end{aligned}$$

Finally, we get

$$\|F\|_2 \leqslant \|F\|_F \leqslant \sqrt{r}\left\{\max_i \|R_{B_i}\| \left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\| + \max_i \|R_{C_i}\| \left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|\right\}.$$

□

In the expression of $\|F\|$ above, we see that the norm of the perturbation is proportional to the norm of the two residuals obtained while solving the two set of linear systems ($\|R_B\|$ and $\|R_C\|$) as well as the norm of two other quantities $\left(\left\| (\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T \right\| \right.$ and $\left. \left\| \widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1} \right\| \right)$. These two quantities are very less dependent on accuracy of the linear systems we solve. They are also not sensitive to different initializations of BIRKA as well as different reduced system sizes. This behavior is similar to the related quantities obtained in the stability analysis of IRKA [10]. We support this argument with numerical experiments in Section 3.3.2.

To summarize, $\left\| \zeta_r - \widetilde{\zeta}_r \right\|_{H_2}$ is proportional to $k(\zeta)$ and $\|F\|$. The problem is usually well conditioned, and $\|F\|$ is directly proportional to $\|R_B\|$ and $\|R_C\|$. Thus, as we iteratively solve the linear systems arising in BIRKA more accurately (i.e., reduce the stopping tolerance of the linear solver), we get a more accurate reduced system. This is very useful in deciding on when to stop the linear solver. If we need a very accurate reduced system, then we need to iterate more in the linear solver, else we can stop earlier. We support this with numerical experiments in the next section.

3.3 Numerical Experiments

We perform experiments to support the conjecture, as discussed above, on two models. First, we use a flow model [18] in Section 3.3.1, and then we use a heat transfer model [12, 13] in Section 3.3.2. These models give us both SISO as well as MIMO bilinear dynamical systems of sizes varying from 100 to $40,000$.

The resulting linear systems to be solved vary from 600×600 to $2,00,000 \times 2,00,000$. For solving the linear systems while computing V and W by a direct method (exact BIRKA), we use a backslash in Matlab. This uses Gaussian elimination as the underlying algorithm. The most popular iterative methods for solving the sparse linear systems of equations are the Krylov subspace methods [43]. As discussed in the start of this chapter, for a backward stable BIRKA with respect to the inexact linear solves, we need to use a linear solver based upon the Petrov-Galerkin framework (Theorem 4 and Corollary 1). Since the Biconjugate Gradient (BiCG) algorithm [4]

is an iterative linear solver based upon this framework, we use it for solving the linear systems while computing V and W by an iterative method (inexact BIRKA), i.e., $\widetilde{V}$ and $\widetilde{W}$.

We implement our codes in MATLAB (2015a), and test on a machine with the following configuration: Intel Xeon(R) CPU E5-1620 V3 @ 3.50 GHz., frequency 1200 MHz., 8 CPU, 64 GB RAM.

3.3.1 A Flow Model

We first do experiments on a "flow model" [18], which consists of a one dimensional viscid Burgers equation. That is,

$$\frac{\partial w}{\partial t} + w\frac{\partial w}{\partial x} = \frac{\partial}{\partial x}\left(v\frac{\partial w}{\partial x}\right), \qquad \text{for } (x,\ t) \in (0,\ L) \times (0,\ T),$$
$$w\,(0,\ t) = u\,(t), \qquad \text{for } t \in (0,\ T),$$

where $w(x,\ t)$ is the velocity at a particular point x and a time t; and $v(x,\ t)$ is the viscosity coefficient that we take as a constant (v). We perform spatial semi-discretization of the above equation with equidistant step size $h = \dfrac{L}{N+1}$, where N is the number of interior points in the interval (0, L). Further, using Carleman bilinearization [12, 18], we obtain a bilinear dynamical system of order $N \times N^2$. We briefly show these steps below.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_i \\ \cdot \\ \cdot \\ w_N \end{bmatrix} = \begin{bmatrix} \dfrac{-w_1 w_2}{2h} + \dfrac{v}{h^2}(w_2 - 2w_1) \\ \dfrac{-w_2}{2h}(w_3 - w_1) + \dfrac{v}{h^2}(w_3 - 2w_2 + w_1) \\ \cdot \\ \cdot \\ \dfrac{-w_i}{2h}(w_{i+1} - w_{i-1}) + \dfrac{v}{h^2}(w_{i+1} - 2w_i + w_{i-1}) \\ \cdot \\ \cdot \\ \dfrac{-w_N w_{N-1}}{2h} + \dfrac{v}{h^2}(-2w_N + w_{N-1}) \end{bmatrix} + \begin{bmatrix} \dfrac{w_1}{2h} + \dfrac{v}{h^2} \\ 0 \\ \cdot \\ \cdot \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix} u$$

or

$$\frac{\mathrm{d}w}{\mathrm{d}t} = f(w) + g(w)u,$$

where $\omega = \begin{bmatrix} \omega_1, & \omega_2, \ldots, & \omega_N \end{bmatrix}^T$; and $f(w)$ and $g(w)$ can be written in Kronecker product form as below.

$$f(w) = A_1 w + \frac{1}{2} A_2 (w \otimes w),$$
$$g(w) = B_0 + B_1 w,$$

where $B_0 \in \mathbb{R}^{N \times 1}$; $A_1, B_1 \in \mathbb{R}^{N \times N}$ are the Jacobians of $f(w)$ and $g(w)$, respectively; and $A_2 \in \mathbb{R}^{N \times N^2}$ is the second derivative of $f(w)$. Let

$$\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t} \qquad \text{and} \qquad \dot{\omega} = \frac{\mathrm{d}\omega}{\mathrm{d}t}.$$

Finally, we get the bilinear system of order $N + N^2$ as

$$\dot{x} = \begin{bmatrix} A_1 & \frac{1}{2} A_2 \\ 0 & A_1 \otimes I + I \otimes A_1 \end{bmatrix} x + \begin{bmatrix} B_1 & 0 \\ B_0 \otimes I + I \otimes B_0 & 0 \end{bmatrix} xu + \begin{bmatrix} B_0 \\ 0 \end{bmatrix} u,$$
$$y = \frac{1}{N} \begin{bmatrix} \underbrace{1 \cdots 1}_{\text{N times}} & \underbrace{0 \cdots \cdots 0}_{N^2 \text{ times}} \end{bmatrix} x,$$

where

$$x = \begin{bmatrix} w \\ w \otimes w \end{bmatrix} \qquad \text{and} \qquad \dot{x} = \begin{bmatrix} \dot{w} \\ \dot{w} \otimes w + w \otimes \dot{w} \end{bmatrix}.$$

We refer the reader to [18] for exact structure of $A_1,\ A_2,\ B_0$ and B_1.

For our experiments, we take N = 10, L = 1 and $v = 0.1$ that gives us a SISO bilinear dynamical system of size 110. We initialize the input system in BIRKA by random matrices based upon similar setup in [12] and [24]. The stopping tolerance for BIRKA is taken as 10^{-6}, and we reduce this model to size 6. Both of these are again chosen based upon similar values in [12] and [24]. This leads to solving the linear systems of size 660×660. While using BiCG we use two different stopping tolerances (10^{-2} and 10^{-8}). Ideally, we should obtain a more accurate reduced model when using the smaller BiCG tolerance.

First, let us look at the remaining assumptions for backward stability of BIRKA (see Theorem 5 and Corollary 1). $\widehat{Q}$ is invertible here. We also have $\left\| \widehat{Q}^{-1} \right\|$ less than one (i.e., 1.6051×10^{-3}). Finally, $\left\| \widehat{\widehat{F}} \right\|$, at the end of the first BIRKA step, for the BiCG

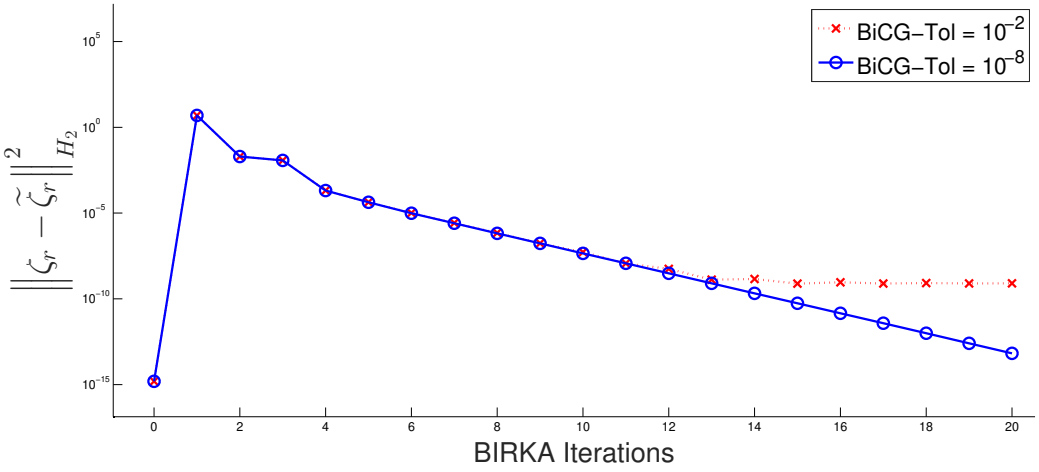

Figure 3.1: Accuracy of the reduced system plotted at each BIRKA iteration for the two different stopping tolerances in BiCG; flow model of size 110. Here, the x-axis is in the linear scale and the y-axis is in the log scale.

stopping tolerance of 10^{-2} and 10^{-8} is 3.0675×10^{-1} and 2.4596×10^{-4}, respectively, both of which are also less than one. These values are less than one at the end of all the other BIRKA steps as well. The condition number for our problem, as defined in (3.24), is 1.2125×10^{-2}. This shows that the flow model is well-conditioned.

The accuracy results are given in Figure 3.1 and Table 3.1. In Figure 3.1, we have accuracy of the reduced system $\left(\left\|\zeta_r - \widetilde{\zeta}_r\right\|_{H_2}\right)$ on the y-axis in the log scale and the BIRKA iterations on the x-axis in the linear scale. Table 3.1 gives the corresponding data. From Figure 3.1, we do not observe any difference in the values of $\left(\left\|\zeta_r - \widetilde{\zeta}_r\right\|_{H_2}\right)$ for the two BiCG tolerances at the starting BIRKA iterations. The dotted line, which corresponds to the BiCG stopping tolerance 10^{-2} and the solid line, which corresponds to the BiCG stopping tolerance 10^{-8} coincide.

BIRKA gets more consistent as it converges to the ideal interpolation points. Hence, towards the end of the BIRKA iterations (iteration 14 to iteration 20), accuracy of the reduced system for the BiCG stopping tolerance of 10^{-8} is substantially better than accuracy of the reduced system for the BiCG stopping tolerance of 10^{-2}. That is, the solid line should be below the dotted line. This behavior is clearly reflected

in Table 3.1 (see the second and the fourth columns).

BIRKA Iteration	BiCG-Tol of 10^{-2}		BiCG-Tol of 10^{-8}	
	$\left\|\zeta_r - \widetilde{\zeta}_r\right\|_{H_2}^2$	BiCG Iteration Count	$\left\|\zeta_r - \widetilde{\zeta}_r\right\|_{H_2}^2$	BiCG Iteration Count
1	4.9214	91	4.8904	167
2	1.9671×10^{-2}	35	1.9649×10^{-2}	85
3	1.1745×10^{-2}	40	1.1735×10^{-2}	85
4	2.0764×10^{-4}	41	2.0583×10^{-4}	92
5	4.3239×10^{-5}	42	4.2785×10^{-5}	89
6	1.0181×10^{-5}	39	9.8618×10^{-6}	89
7	2.6412×10^{-6}	39	2.5583×10^{-6}	82
8	6.9999×10^{-7}	44	6.5685×10^{-7}	90
9	1.7325×10^{-7}	44	1.7213×10^{-7}	90
10	5.3043×10^{-8}	44	4.4857×10^{-8}	90
11	1.1675×10^{-8}	44	1.1745×10^{-8}	90
12	5.5945×10^{-9}	44	3.0702×10^{-9}	90
13	1.3127×10^{-9}	44	8.0359×10^{-10}	90
14	1.4474×10^{-9}	44	2.1026×10^{-10}	90
15	7.7234×10^{-10}	44	5.5041×10^{-11}	90
16	9.2674×10^{-10}	44	1.4398×10^{-11}	90
17	7.8030×10^{-10}	44	3.7841×10^{-12}	90
18	8.2925×10^{-10}	44	9.8779×10^{-13}	90
19	7.9294×10^{-10}	44	2.5543×10^{-13}	90
20	8.0646×10^{-10}	44	6.6835×10^{-14}	90

Table 3.1: Accuracy of the reduced system and BiCG iterations at each BIRKA step for the two different stopping tolerances in BiCG; flow model of size 110.

In Table 3.1, we observe that BiCG takes exactly same number of iterative steps

from the BIRKA iteration 8 until convergence. That is, for the BiCG stopping tolerance of 10^{-2} it stays at 44, and for the BiCG stopping tolerance of 10^{-8} it stays at 90. The reason for this is that the linear systems change very little from the 8^{th} BIRKA step. This can be inferred by looking at the eigenvalue distribution of the linear system matrices as well as their Frobenius norm.

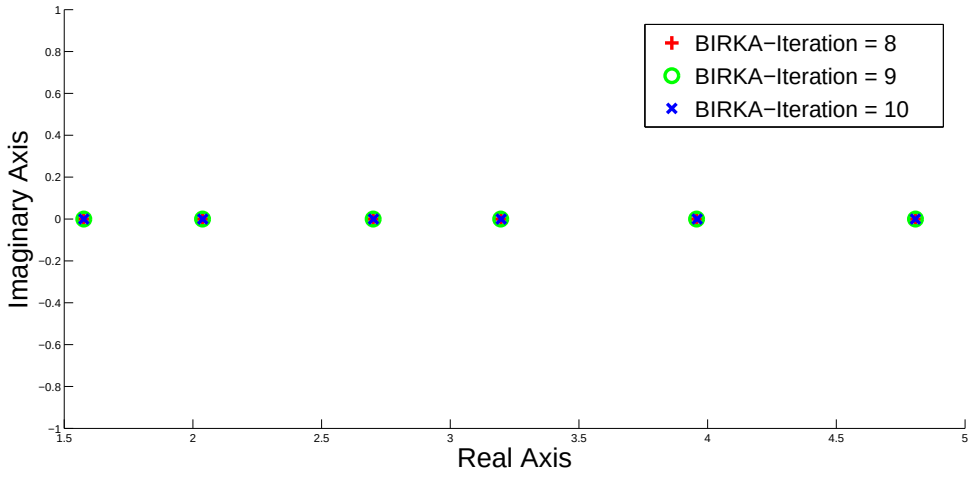

Figure 3.2: The six smallest eigenvalues of the linear systems at the different BIRKA iterations.

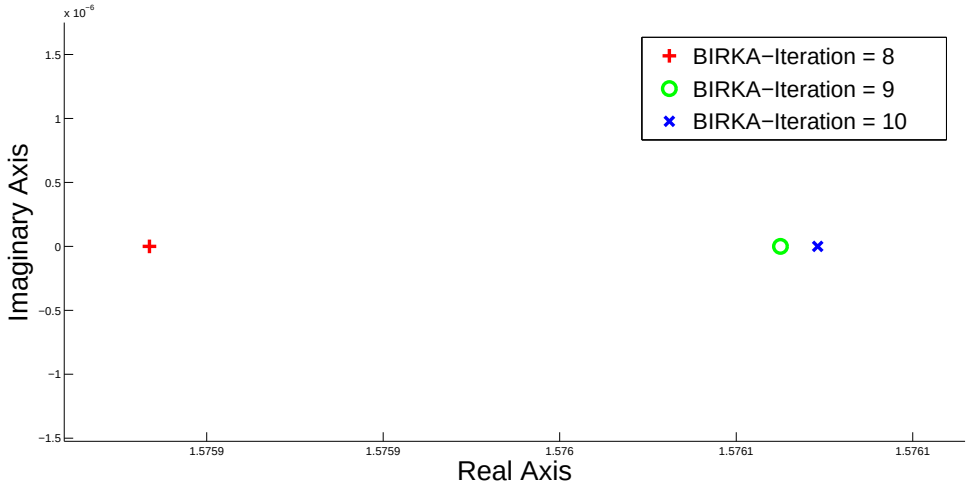

Figure 3.3: Enlarged Figure 3 for the smallest eigenvalue.

Figure 3.2 shows the distribution of the six smallest eigenvalues (in absolute sense) of the linear system matrices corresponding to the BiCG stopping tolerance of 10^{-2}

at the BIRKA steps 8, 9 and 10. Each of these six eigenvalues do not seem to change with respect to the change in the BIRKA steps. However, if we look at any one eigenvalue, specifically, for example the smallest eigenvalue at the three different BIRKA steps, then we observe that it does change, but only slightly (see Figure 3.3). The Frobenius norm of the linear system matrices at the BIRKA steps 8, 9 and 10 are 1.7263×10^3, 1.7264×10^3 and 1.7266×10^3, respectively. Thus, this supports the argument that matrices do not change much.

3.3.2 A Heat Transfer Model

The next set of experiments we do on a heat transfer model as given below [12, 13].

$$\begin{aligned} x_t &= \Delta x \quad \text{in } [0,1] \times [0,1], \\ n.\nabla x &= u_1 (x-1) \quad \text{on } \Gamma_1 := \{0\} \times (0,1), \\ n.\nabla x &= u_2 (x-1) \quad \text{on } \Gamma_2 := (0,1) \times \{0\}, \\ x &= 0 \quad \text{on } \Gamma_3 := \{1\} \times [0,1] \text{ and } \Gamma_4 := [0,1] \times \{1\}, \end{aligned}$$

where $x(l_1,\ l_2,\ t)$ is the temperature at a particular point in the space $(l_1,\ l_2)$ and at a time t; n is the unit outward normal to the domain; u_1 and u_2 are the input variables; and $\Gamma_1,\ \Gamma_2,\ \Gamma_3$, and Γ_4 are the boundaries of the unit square. After spatial discretization of the above equation using K^2 grid points, we obtain a bilinear dynamical system of order $K^2 \times K^2$ with two inputs and one output as shown below.

$$\begin{aligned} \dot{x} &= Ax + u_1 N_1 x + u_2 N_2 x + Bu, \\ y &= Cx, \end{aligned}$$

where, as earlier,

$$\begin{aligned} \dot{x} &= \frac{\mathrm{d}x}{\mathrm{d}t}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \\ A &= \frac{1}{h^2} \left(I_K \otimes T_K + T_K \otimes I_K + E_1 \otimes I_K + I_K \otimes E_K \right), \\ N_1 &= \frac{1}{h} \left(E_1 \otimes I_K \right),\ N_2 = \frac{1}{h} \left(I_K \otimes E_K \right), \\ B &= \begin{bmatrix} \frac{1}{h} \left(e_1 \otimes e \right) & \frac{1}{h} \left(e \otimes e_K \right) \end{bmatrix}, \text{ and } C = \frac{1}{K^2} \left(e \otimes e \right)^T \end{aligned}$$

with I_K being the identity matrix of size K,

$$T_K = \begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & 1 & -2 & 1 \\ & & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{K \times K},$$

$E_j = e_j e_j^T$, the grid size $h = \dfrac{1}{K+1}$, e_j is the j^{th} column of the identity matrix I_K, and $e = [1, \ \ldots, \ 1] \in \mathbb{R}^K$.

We perform experiments on the heat transfer model for three different sizes, i.e., n = 100, 10,000 and 40,000 corresponding to K = 10, 100 and 200, respectively. We initialize the input system in BIRKA by random matrices based upon the similar setup in [12] and [24]. The stopping tolerance for BIRKA is taken as 10^{-3}. The size to which we reduce is different for the different model sizes, and is discussed below. Both these settings (the BIRKA stopping tolerance and the size of reduced system) are chosen based upon similar values in [12, 24]. While using BiCG (unpreconditioned for smaller size and preconditioned for larger sizes), we use two different stopping tolerances (10^{-4} and 10^{-8}). Ideally, as discussed earlier, we should obtain a more accurate reduced model for the smaller stopping tolerance.

We reduce the model of the size 100 to the size 6 as above based upon similar values in [12] and [24]. Hence, the linear systems that are required to be solved are of the size 600×600. As above, we use an unpreconditioned BiCG here. First, let us look at the remaining assumptions for backward stability of BIRKA (see Theorem 5 and Corollary 1). $\widehat{Q}$ is invertible here. We also have $\left\|\widehat{Q}^{-1}\right\|$ less than one (i.e., 5.2893×10^{-4}). Finally, $\left\|\widehat{\widehat{F}}\right\|$, at the end of the first BIRKA step, for the BiCG stopping tolerance of 10^{-4} and 10^{-8} is 1.3370×10^{-1} and 3.4528×10^{-5}, respectively, both of which are also less than one. These values are less than one at the end of all the other BIRKA steps as well. The condition number for our problem, as defined in (3.24), is 2.6653×10^{-2}. This shows that the heat transfer model is well-conditioned.

BIRKA Iteration	$\|R_B\|$	$\|R_C\|$	$\left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|_F$ or $\left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\|_F$	$\|F\|$
1	0.0544	7.7746×10^{-8}	2.4554	0.1337
2	0.0937	1.2331×10^{-7}	2.4526	0.2299
3	0.1223	1.4124×10^{-7}	2.4515	0.2997
4	0.0568	9.8639×10^{-8}	2.4510	0.1392
5	0.0286	4.7669×10^{-8}	2.4508	0.0702
6	0.0319	5.2856×10^{-8}	2.4507	0.0781
7	0.0325	5.7300×10^{-8}	2.4507	0.0797
8	0.0325	6.0807×10^{-8}	2.4507	0.0796
9	0.0325	6.3895×10^{-8}	2.4507	0.0797
10	0.0327	6.6521×10^{-8}	2.4507	0.0801
11	0.0330	6.9071×10^{-8}	2.4507	0.0808

Table 3.2: The perturbation expression quantities (as defined in Theorem 6) for the BiCG stopping tolerance 10^{-4}.

For this model size, we do not give results for supporting the main conjecture (as discussed at the end of Section 3.2; the more accurately we solve the linear systems, the more accurate reduced system we obtain). This is because for a small sized dynamical system we have already reported the data in Section 3.3.1, and we get the similar results here. Here, we do some other analyses corresponding to Theorem 6, i.e., relation between the perturbation and the stopping tolerances.

Table 3.2 lists the values of $\|R_B\|$, $\|R_C\|$, $\left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|$, $\left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\|$ and $\|F\|$ for the BiCG stopping tolerance 10^{-4}, and Table 3.3 gives the same data for the BiCG stopping tolerance 10^{-8}. All these quantities are defined in Theorem 6. It is obvious from these two tables that $\left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|$ and $\left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\|$ are very less sensitive to the BiCG stopping tolerance, while $\|R_B\|$ and $\|R_C\|$ are directly proportional to it. Thus, as conjectured at the end of Section 3.2, the norm of the perturbation $\Big(\|F\|\Big)$

BIRKA Iteration	$\|R_B\|$	$\|R_C\|$	$\left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|_F$ or $\left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\|_F$	$\|F\|$
1	1.4062×10^{-5}	1.3372×10^{-11}	2.4554	3.4528×10^{-5}
2	6.4701×10^{-6}	1.1488×10^{-11}	2.4526	1.5868×10^{-5}
3	7.3663×10^{-6}	9.9444×10^{-12}	2.4515	1.8058×10^{-5}
4	1.1982×10^{-5}	1.6620×10^{-11}	2.4510	2.9369×10^{-5}
5	9.0962×10^{-6}	1.1775×10^{-11}	2.4508	2.2293×10^{-5}
6	4.1159×10^{-6}	6.3212×10^{-12}	2.4507	1.0087×10^{-5}
7	5.2442×10^{-6}	8.2256×10^{-12}	2.4507	1.2852×10^{-5}
8	1.2491×10^{-5}	1.6984×10^{-11}	2.4507	3.0612×10^{-5}
9	1.4070×10^{-5}	3.6218×10^{-11}	2.4507	3.4481×10^{-5}
10	1.1009×10^{-5}	2.7919×10^{-11}	2.4507	2.6981×10^{-5}
11	9.4640×10^{-6}	2.3366×10^{-11}	2.4507	2.3193×10^{-5}

Table 3.3: The perturbation expression quantities (as defined in Theorem 6) for the BiCG stopping tolerance 10^{-8}.

should reduce as we reduce the BiCG stopping tolerance. This is supported by the data in the two tables as well (see columns for $\|F\|$). The values of $\|R_B\|$, which is the residual of the linear systems involving $\widetilde{V}$, for both the BiCG stopping tolerances seem higher that their respective stopping tolerances. The reason for this apparent anomaly is that we are reporting the absolute residuals here. The relative residuals are still less than the respective stopping tolerances.

We also do the sensitivity analysis of $\left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\|$ and $\left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|$ with respect to different random initializations of BIRKA as well as different reduced system sizes. Table 3.4 gives this data at convergence of BIRKA corresponding to the BiCG stopping tolerance of 10^{-4}. As evident from this table, $\left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\|$ and $\left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|$ vary very less.

We reduce the model sizes $10,000$ and $40,000$ to the sizes 6 and 5, respectively, as

Reduced Model Size	$\left\|\left(\widetilde{W}^T\tilde{V}\right)^{-1}\widetilde{W}^T\right\|_F$ or $\left\|\tilde{V}\left(\widetilde{W}^T\tilde{V}\right)^{-1}\right\|_F$				
	Random Initialization 1	Random Initialization 2	Random Initialization 3	Random Initialization 4	Random Initialization 5
4	2.0109	2.0045	2.0048	2.0100	2.0065
5	2.2427	2.2406	2.2413	2.2399	2.2392
6	2.4507	2.4531	2.4511	2.4557	2.4507
7	2.6467	2.6467	2.6468	2.6467	2.6467
8	2.8365	2.8360	2.8366	2.8371	2.8368
9	3.0248	3.0269	3.0193	3.0306	3.0722
10	3.1718	3.1759	3.1768	3.1711	3.2142

Table 3.4: The sensitivity analysis for the heat transfer model of size 100 with respect to random initializations and reduced system sizes.

discussed earlier based upon similar values in [12] and [24]. Hence, the linear systems of size $60,000 \times 60,000$ and $2,00,000 \times 2,00,000$ are required to be solved, respectively. The linear systems arising in the model reduction process of both these size are ill-conditioned. Hence, we use a preconditioned BiCG here. The preconditioner that we use is incomplete LU [22]. The drop tolerance in the preconditioner is taken as 10^{-5} based upon the range given in [22]. The result for the model size $10,000$ is given in Figure 3.4 and the result for the model size $40,000$ is given in Figure 3.5. From both Figure 3.4 and 3.5, it is again evident that we get a more accurate reduced model as we solve the linear systems more accurately (solid line is below the dotted one at all the BIRKA steps).

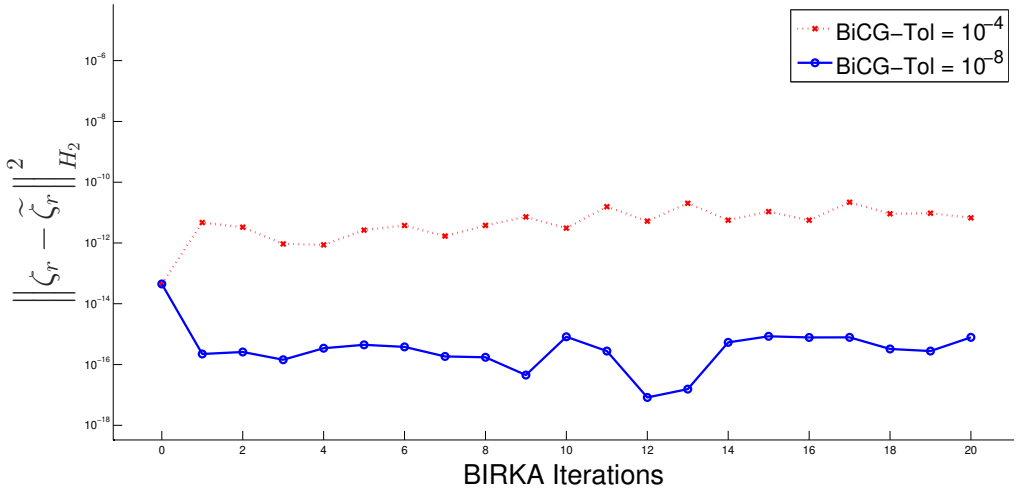

Figure 3.4: Accuracy of the reduced system plotted at each BIRKA iteration for the two different stopping tolerances in BiCG; heat transfer model of size 10,000. Here, the x-axis is in the linear scale and the y-axis is in the log scale.

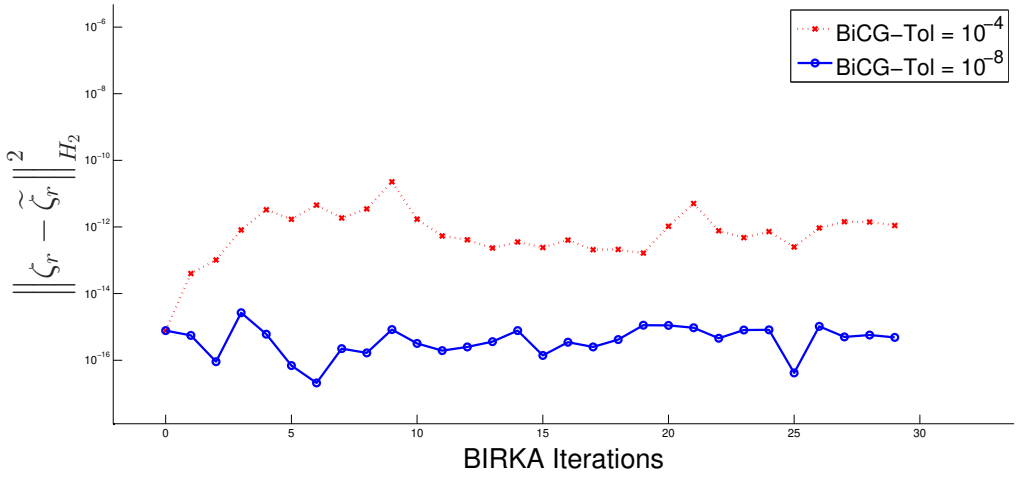

Figure 3.5: Accuracy of the reduced system plotted at each BIRKA iteration for the two different stopping tolerances in BiCG; heat transfer model of size 40,000. Here, the x-axis is in the linear scale and the y-axis is in the log scale.

CHAPTER 4

STABILITY ANALYSIS OF OTHER EFFICIENT ALGORITHMS FOR BILINEAR MOR

As mentioned earlier, we focus on TBIRKA here. The *first condition* is satisfied in a way similar to that of BIRKA except that some extra orthogonality conditions are imposed on the linear solver (discussed below).

Theorem 7. *Let the inexact linear solves in TBIRKA (lines 3b. and 3c. of Algorithm 2.2) are solved satisfying*

$$\begin{aligned} \begin{bmatrix} V_1^T \\ V_2^T \\ \vdots \\ V_M^T \end{bmatrix} \begin{bmatrix} R_{C_1} & R_{C_2} & \cdots & R_{C_M} \end{bmatrix} = 0 \qquad \text{and} \\ \begin{bmatrix} W_1^T \\ W_2^T \\ \vdots \\ W_M^T \end{bmatrix} \begin{bmatrix} R_{B_1} & R_{B_2} & \cdots & R_{B_M} \end{bmatrix} = 0, \end{aligned} \tag{4.1}$$

where V_1 and V_k are given by the first equations of lines 3b. and 3c. of Algorithm 2.2, respectively; R_{C_1} and R_{C_k} are the residuals in the second equations of lines 3b. and 3c. of Algorithm 2.2, respectively; W_1 and W_k are given by the second equations of

lines 3b. and 3c. of Algorithm 2.2, respectively; R_{B_1} *and* R_{B_k} *are the residuals in the first equations of lines 3b. and 3c. of Algorithm 2.2, respectively; and* $k = 2, \ldots, M$*. Then, TBIRKA satisfies the first condition of backward stability with respect to these solves.*

Proof. Follows the same pattern as the proof for Theorem 3 in [21]. □

From the above theorem, we infer that the underlying iterative solver should *firstly* be based upon a Petrov-Galerkin framework to achieve

$$V_k^T R_{C_k} = 0 \quad \text{and} \quad W_k^T R_{B_k} = 0, \tag{4.2}$$

for $k = 1, \ldots, M$. Since BiConjugate Gradient (i.e., BiCG) is one such algorithm [43], we propose its use in TBIRKA. This is exactly same as for BIRKA (Chapter 3 and [21]). *Secondly,* this particular solver should also satisfy the remaining orthogonalities of (4.1).

These orthogonalities have a form similar to the orthogonalities required while reducing second order *linear* dynamical systems ((23) and (24) in [46]; AIRGA algorithm), and can be easily satisfied by using a recycling variant of the underlying iterative solver. In [46], the ideal iterative solver to be used is Conjugate Gradient (i.e., CG) [43] (due to the use of Galerkin projection). Hence, to satisfy the similar orthogonalities there, without any extra cost, the authors use Recycling Conjugate Gradient (i.e., RCG) [39]. Since here BiCG is the ideal iterative solver (as discussed above), we propose the use of Recycling BiConjugate Gradient (i.e., RBiCG) [4, 3], which would ensure that the remaining orthogonalities of (4.1) (besides (4.2)) are satisfied without any extra cost. Similar orthogonalities arise during reduction of parametric dynamical systems (discussed in the next chapter). Hence, we expand upon satisfying such orthogonalities in-detail in the following chapter.

To satisfy the *second condition* of backward stability of TBIRKA, we need to show that

$$\left\| \zeta^M - \widetilde{\zeta}^M \right\|_{H_2} = \mathcal{O}\left(\|F\|_2 \right), \tag{4.3}$$

where ζ^M is the original truncated bilinear dynamical system given by (2.3) or

$$\zeta^M = \big\{H_1\left(s_1\right),\ H_2\left(s_1,\ s_2\right),\ H_3\left(s_1,\ s_2,\ s_3\right),\ \ldots H_M\left(s_1,\ \ldots,\ s_M\right)\big\}, \tag{4.4}$$

with $H_k\left(s_1,\ \ldots,\ s_k\right)$ for $k \in \{1,\ \ldots,\ M\}$ is the k^{th} order transfer function of the corresponding system (defined earlier in (1.3)), and $\widetilde{\zeta}^M$ is the perturbed truncated bilinear dynamical system given by

$$\widetilde{\zeta}^M = \big\{\widetilde{H}_1\left(s_1\right),\ \widetilde{H}_2\left(s_1, s_2\right),\ \widetilde{H}_3\left(s_1,\ s_2,\ s_3\right), \ldots,\ \widetilde{H}_M\left(s_1,\ \ldots,\ s_M\right)\big\}, \tag{4.5}$$

with $\widetilde{H}_k\left(s_1,\ \ldots,\ s_k\right)$ of $k \in \{1,\ \ldots,\ M\}$ is the k^{th} order transfer function of the corresponding system (defined later in Section 4.2), and assuming perturbation F in A matrix of the input dynamical system.

One way to satisfy (4.3) is to use the definition of the H_2-norm of $\zeta^M - \widetilde{\zeta}^M$, i.e., from Lemma 5.1 of [26] (also defined in Chapter 1)

$$\begin{aligned}\left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}^2 = &\left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix}\right) \\ &\sum_{k=0}^{M}\Bigg[\left(-\begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \right. \\ &\left.\begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix}\right)^{-1} \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix}\Bigg]^k \\ &\left(-\begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \right. \\ &\left.\begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix}\right)^{-1} \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix}\right).\end{aligned} \tag{4.6}$$

This approach is followed in satisfying the *second condition* of backward stability for BIRKA, and is one of the ways for satisfying the second condition of stability in TBIRKA as well (Complete system approach; discussed in Section 4.1).

From (4.4) and (4.5), we know that both ζ^M and $\widetilde{\zeta}^M$ are represented by a finite set of transfer functions, respectively. Hence, another way to satisfy (4.3) in-case of TBIRKA, is to show that the norm of the difference between the respective order

transfer functions of (4.4) and (4.5) is equal to the norm of the perturbation. That is, instead of (4.3) we can show that

$$\begin{aligned}
\left\| H_1(s_1) - \widetilde{H}_1(s_1) \right\|_{H_2} &\propto \mathcal{O}\left(\|F\|_2 \right), \\
\left\| H_2(s_1,\ s_2) - \widetilde{H}_2(s_1,\ s_2) \right\|_{H_2} &\propto \mathcal{O}\left(\|F\|_2 \right), \\
&\vdots \\
\left\| H_M(s_1,\ \ldots,\ s_M) - \widetilde{H}_M(s_1,\ \ldots,\ s_M) \right\|_{H_2} &\propto \mathcal{O}\left(\|F\|_2 \right).
\end{aligned} \tag{4.7}$$

This way was not possible in BIRKA because there $M \to \infty$ (see (1.2)-(1.3)[1]). This approach is refereed to as a Subsystem approach, and works only for the SISO systems. We discuss this in Section 4.2.

Note, that in all our subsequent derivations, we assume that all inverses used exist. This is an acceptable assumption because the inverse of matrices arising here are of the form as in [10] and [21] (the papers that discuss stability of IRKA and BIRKA, respectively).

4.1 Complete System Approach

Recall the H_2-norm of the truncated bilinear dynamical system given in (1.9), as

$$\begin{aligned}
\left\| \zeta^M \right\|_{H_2}^2 =& vec(I_p)^T (C \otimes C) \sum_{k=0}^{M} \left((-A \otimes I_n - I_n \otimes A)^{-1} \sum_{j=1}^{m} N_j \otimes N_j \right)^k \\
& (-A \otimes I_n - I_n \otimes A)^{-1} (B \otimes B)\, vec(I_m),
\end{aligned}$$

where M is the truncation index. Since, we are perturbing only A matrix, and hence, we define the error system $\zeta^{M^{err}} = \zeta^M - \widetilde{\zeta}^M$ matrices as follows:
$A^{err} = \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix}$, $N_j^{err} = \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix}$, $B^{err} = \begin{bmatrix} B \\ B \end{bmatrix}$, and $C^{err} = \begin{bmatrix} C & -C \end{bmatrix}$.

Using this notation, the H_2-norm of the error system is given by

$$\begin{aligned}
&\left\|\zeta^{M^{err}}\right\|_{H_2}^2 = \left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}^2 \\
&= vec(I_{2p})^T \left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix}\right) \sum_{k=0}^{M} \Bigg[\Bigg(- \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \\
&\quad \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \Bigg)^{-1} \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \Bigg]^k \\
&\quad \left(- \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \right)^{-1} \\
&\quad \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix} \right) vec(I_{2m}). \\
&= vec(I_{2p})^T \left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix}\right) \sum_{k=0}^{M} \Bigg[\Bigg(- \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \\
&\quad \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \Bigg)^{-1} \\
&\quad \sum_{j=1}^{m} \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \Bigg]^k \Bigg(- \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \\
&\quad - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \Bigg)^{-1} \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix} \right) vec(I_{2m}).
\end{aligned}$$

Let

$$\begin{aligned}
\widehat{C} &= \left(\begin{bmatrix} C & -C \end{bmatrix} \otimes \begin{bmatrix} C & -C \end{bmatrix}\right), \\
\widehat{Q} &= \left(- \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \right),
\end{aligned} \tag{4.8}$$

$$\begin{aligned}
&\widehat{F} = \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix}, \ \widehat{\widehat{F}} = \left(I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n} \right), \\
&\widehat{N} = \sum_{j=1}^{m} \left(\begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \otimes \begin{bmatrix} N_j & 0 \\ 0 & N_j \end{bmatrix} \right), \text{ and } \widehat{B} = \left(\begin{bmatrix} B \\ B \end{bmatrix} \otimes \begin{bmatrix} B \\ B \end{bmatrix} \right).
\end{aligned}$$

Then, the above equation leads to

$$\begin{aligned}\left\|\zeta^{M^{err}}\right\|_{H_2}^2 &= vec(I_{2p})^T\widehat{C}\ \sum_{k=0}^{M}\left[\left(\widehat{Q}-\widehat{\widehat{F}}\right)^{-1}\widehat{N}\right]^k\left(\widehat{Q}-\widehat{\widehat{F}}\right)^{-1}\widehat{B}\ vec(I_{2m}) \quad \text{or}\\ \left\|\zeta^{M^{err}}\right\|_{H_2}^2 &= vec(I_{2p})^T\widehat{C}\ \sum_{k=0}^{M}\left[\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\right]^k\widehat{Q}^{-1}\\ &\qquad \left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}).\end{aligned} \tag{4.9}$$

For different values of k, we get different terms in the above equation (related to the truncation of the Volterra series). In total, we get $M+1$ different terms as we vary the value of k from 0 to M. For $k=0$, we get

$$J_0 = vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}). \tag{4.10}$$

We know that if $\|\mathcal{A}\|_2 < 1$, then the power series expansion of a matrix $\mathcal{A}$ using Neumann series is given as

$$(I-\mathcal{A})^{-1} = I+\mathcal{A}+\mathcal{A}^2+\mathcal{A}^3+\cdots\infty = \sum_{i=0}^{\infty}\mathcal{A}^i.$$

Hence, in (4.10), if $\left\|\widehat{\widehat{F}}\widehat{Q}^{-1}\right\| < 1$, then by Neumann series we get

$$\begin{aligned}J_0 &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m})\\ &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{2n}+\widehat{\widehat{F}}\widehat{Q}^{-1}+\left(\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^2+\cdots\right)\widehat{B}\ vec(I_{2m})\\ &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{B}\ vec(I_{2m})\\ &\quad + vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}).\end{aligned}$$

By the definition of the error system we know $vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{B}\ vec(I_{2m}) = 0$. Hence, the above equation can be re-written as

$$J_0 = vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}).$$

Bounding the right hand side of the above equation we get

$$|J_0| \leqslant \left\|vec(I_{2p})^T\right\|\left\|\widehat{C}\right\|\left\|\widehat{Q}^{-1}\right\|\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|\left\|\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\right\|\left\|\widehat{B}\right\|\left\|vec(I_{2m})\right\|.$$

Using Lemma 2.3.3 from [27] in the above bounding condition we get that since $\left\|\widehat{\hat{F}}\widehat{Q}^{-1}\right\| < 1$ as assumed above, we have

$$|J_0| \leqslant \left\|vec(I_{2p})^T\right\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{\hat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \left(\frac{1}{1-\left\|\widehat{\hat{F}}\widehat{Q}^{-1}\right\|}\right) \left\|\widehat{B}\right\| \|vec(I_{2m})\|. \tag{4.11}$$

Let $\left\|\widehat{Q}^{-1}\right\| < 1$ and $\left\|\widehat{\hat{F}}\right\| < 1$ be defined by the original system and residuals of the linear solves, respectively. Then, by using the matrix norm inequality property we get

$$\left\|\widehat{\hat{F}}\widehat{Q}^{-1}\right\| \leqslant \left\|\widehat{\hat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \text{ or}$$

$$\frac{1}{1-\left\|\widehat{\hat{F}}\widehat{Q}^{-1}\right\|} \leqslant \frac{1}{1-\left\|\widehat{\hat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}.$$

Using this inequality in (4.11) we get

$$|J_0| \leqslant \left\|vec(I_{2p})^T\right\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{\hat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \left(\frac{1}{1-\left\|\widehat{\hat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}\right) \left\|\widehat{B}\right\| \|vec(I_{2m})\|, \tag{4.12}$$

$$\leqslant \mathcal{O}\left(\left\|\widehat{\hat{F}}\right\|\right).$$

Similarly, if we take $k = 1$ in (4.9) we have

$$\begin{aligned}
J_1 &= vec(I_{2p})^T \widehat{C} \left[\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{N}\right] \widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{B}\ vec(I_{2m}) \quad (4.13)\\
&= vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\left(I_{2n} + \widehat{\hat{F}}\widehat{Q}^{-1} + \left(\widehat{\hat{F}}\widehat{Q}^{-1}\right)^2 + \cdots\right) \widehat{N}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1}\\
&\quad \widehat{B}\ vec(I_{2m})\\
&= vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\widehat{N}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{B}\ vec(I_{2m})\\
&\quad + vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\widehat{\hat{F}}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{N}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{B}\ vec(I_{2m})\\
&= vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\widehat{N}\widehat{Q}^{-1}\widehat{\hat{F}}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{B}\ vec(I_{2m})\\
&\quad + vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\widehat{\hat{F}}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{N}\widehat{Q}^{-1}\left(I_{2n} - \widehat{\hat{F}}\widehat{Q}^{-1}\right)^{-1} \widehat{B}\ vec(I_{2m}).
\end{aligned}$$

As earlier, while assigning that $\left\|\widehat{Q}^{-1}\right\| < 1$ and $\left\|\widehat{\widehat{F}}\right\| < 1$, bounding the right hand side of the above equation we get

$$\begin{aligned}\left|J_1\right| &\leqslant \|vec(I_{2p})^T\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \\ &\left(\left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|}\right)+\left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|}\right)^2\right)\left\|\widehat{B}\right\| \|vec(I_{2m})\| \\ &\leqslant \mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right).\end{aligned} \tag{4.14}$$

Similarly, if we take $k = 2$ in (4.9) we get

$$\begin{aligned}J_2 &= vec(I_{2p})^T\widehat{C}\left[\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\right]^2\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}) \\ &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1} \\ &\quad \widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}) \\ &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\left(I_{2n}+\widehat{\widehat{F}}\widehat{Q}^{-1}+\left(\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^2+\cdots\right)\widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1} \\ &\quad \widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}) \\ &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}) \\ &\quad + vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^2\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1} \\ &\quad \widehat{B}\ vec(I_{2m}). \\ &= vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{N}\widehat{Q}^{-1}\left(I_{2n}+\widehat{\widehat{F}}\widehat{Q}^{-1}+\left(\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^2+\cdots\right) \\ &\quad \widehat{N}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}) \\ &\quad + vec(I_{2p})^T\widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^2\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1} \\ &\quad \widehat{B}\ vec(I_{2m}).\end{aligned} \tag{4.15}$$

$$
\begin{aligned}
&= vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\left(\widehat{N}\widehat{Q}^{-1}\right)^2 \widehat{\widehat{F}}\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}) \\
&\quad + vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\left(\widehat{N}\widehat{Q}^{-1}\right)\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right) \\
&\quad \left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}) \\
&\quad + vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^2\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1} \\
&\quad \widehat{B}\ vec(I_{2m}).
\end{aligned}
$$

As earlier, while assigning that $\left\|\widehat{Q}^{-1}\right\| < 1$ and $\left\|\widehat{\widehat{F}}\right\| < 1$, bounding the right hand side of the above equation we get

$$
\begin{aligned}
\left|J_2\right| &\leqslant \left\|vec(I_{2p})^T\right\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{N}\right\|^2 \left\|\widehat{Q}^{-1}\right\|^2 \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \\
&\left(\left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|}\right)+\left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|}\right)^2+\left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|}\right)^3\right) \\
&\left\|\widehat{B}\right\| \left\|vec(I_{2m})\right\| \ \leqslant\ \mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right).
\end{aligned}
\tag{4.16}
$$

Taking $k = M$ in (4.9), we get

$$
\begin{aligned}
J_M &= vec(I_{2p})^T \widehat{C}\left[\widehat{Q}^{-1}\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\right]^M \widehat{Q}^{-1} \\
&\quad \left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}) \\
&= vec(I_{2p})^T \widehat{C}\widehat{Q}^{-1}\Bigg[\left(\widehat{N}\widehat{Q}^{-1}\right)^M \widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^0 \\
&\quad + \left(\widehat{N}\widehat{Q}^{-1}\right)^{M-1}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^1 \\
&\quad + \left(\widehat{N}\widehat{Q}^{-1}\right)^{M-2}\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^2+\ldots \\
&\quad + \left(\widehat{N}\widehat{Q}^{-1}\right)^0\widehat{\widehat{F}}\widehat{Q}^{-1}\left(\left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\widehat{Q}^{-1}\right)^M\Bigg] \\
&\quad \left(I_{2n}-\widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}vec(I_{2m}).
\end{aligned}
\tag{4.17}
$$

As earlier, while assigning that $\left\|\widehat{Q}^{-1}\right\| < 1$ and $\left\|\widehat{\widehat{F}}\right\| < 1$, bounding the right hand side of the above equation we get

$$\begin{aligned}\left|J_M\right| &\leqslant \|vec(I_{2p})^T\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{N}\right\|^M \left\|\widehat{Q}^{-1}\right\|^M \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \\ &\left[\left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}\right) + \ldots + \left(\frac{1}{1-\left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\|}\right)^{M+1}\right] \\ &\left\|\widehat{B}\right\| \|vec(I_{2m})\| \\ &\leqslant \mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right).\end{aligned} \tag{4.18}$$

Substituting (4.12), (4.14), (4.16) and (4.18) in (4.9), we get

$$\begin{aligned}\left\|\zeta^{M^{err}}\right\|_{H_2}^2 = \left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}^2 &= J_0 + J_1 + J_2 + \ldots + J_M \\ &\leqslant \mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right).\end{aligned} \tag{4.19}$$

This result is independent of M, except that it should not be infinity. We know

$$\widehat{\widehat{F}} = \left(I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n}\right) \quad \text{or} \quad \left\|\widehat{\widehat{F}}\right\| = \left\|\left(I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n}\right)\right\|.$$

Now using the triangle inequality property (i.e., $\|X+Y\| \leqslant \|X\| + \|Y\|$) we get

$$\left\|\widehat{\widehat{F}}\right\| \leqslant \left\|\left(I_{2n} \otimes \widehat{F}\right)\right\| + \left\|\left(\widehat{F} \otimes I_{2n}\right)\right\|.$$

In the above, if we use the Kronecker product property ($\|X \otimes Y\| = \|X\| \|Y\|$)[35, 34], then

$$\begin{aligned}\left\|\widehat{\widehat{F}}\right\| &\leqslant \|I_{2n}\| \left\|\widehat{F}\right\| + \left\|\widehat{F}\right\| \|I_{2n}\| \\ &\leqslant \mathcal{O}\left(\left\|\widehat{F}\right\|\right).\end{aligned}$$

We also know $\widehat{F} = \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix}$. Thus, using the matrix norm property [38] in the above we get

$$\mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right) \leqslant \mathcal{O}\left(\|F\|\right). \tag{4.20}$$

Substituting (4.20) in (4.19) we get

$$\left\|\zeta^{M^{err}}\right\|_{H_2}^2 = \left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}^2 \leqslant \mathcal{O}\left(\|F\|\right).$$

Following theorem summarizes this.

Theorem 8. *Let F be the constant perturbation introduced in A, $\widehat{F} = \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix}$, $\widehat{\widehat{F}} = \left(I_{2n} \otimes \widehat{F} + \widehat{F} \otimes I_{2n}\right)$, and $\widehat{Q} = \left(-\begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}\right)$. If $\left\|\widehat{\widehat{F}}\right\| < 1$, $\widehat{Q}$ is invertible and $\left\|\widehat{Q}^{-1}\right\| < 1$, then TBIRKA satisfies the second condition of backward stability with respect to the inexact linear solves, i.e.,* (2.11).

4.2 Subsystem Approach

Since, as mentioned earlier subsystem approach works only for the SISO systems. Hence, in the following discussion in this section $B = b \in \mathbb{R}^{n\times 1}$, $C = c \in \mathbb{R}^{1\times n}$, and we only have one bilinear matrix $N \in \mathbb{R}^{n\times n}$ in (1.1), (1.2), and (1.3)[1]. Thus, in (4.4), for the original truncated bilinear dynamical system ζ^M, the k^{th} order transfer function,

$$H_k\left(s_1,\ \ldots,\ s_k\right) = c\left(s_k I - A\right)^{-1} N\left(s_{k-1} I - A\right)^{-1}\ \ldots N\left(s_1 I - A\right)^{-1} b. \tag{4.21}$$

Similarly, in (4.5), for the perturbed truncated bilinear dynamical system $\widetilde{\zeta}^M$, the k^{th} order transfer function,

$$\begin{aligned} \widetilde{H}_k\left(s_1,\ \ldots,\ s_k\right) = c\left(s_k I - (A+F)\right)^{-1} \\ N\left(s_{k-1} I - (A+F)\right)^{-1}\ \ldots N\left(s_1 I - (A+F)\right)^{-1} b. \end{aligned} \tag{4.22}$$

To prove the condition (4.7), we first abstract out the term containing the perturbation F from the normed difference between the two corresponding transfer functions (of the original system and the perturbed system) in Lemma 1. Next, in Lemma 2, for $k = 2$, we show that the norm of this term is order of the norm of F. Finally, we generalize the result of Lemma 2 in Lemma 3 (from $k = 2$ to any general k) by using induction.

Lemma 1. *Let the original bilinear dynamical system be defined as in* (4.4) *and the perturbed bilinear dynamical system be defined as in* (4.5)*. Then,*

$$\left\|H_k\left(s_1,\ \ldots,\ s_k\right)-\widetilde{H}_k\left(s_1,\ \ldots,\ s_k\right)\right\|_{H_2}^2 \leqslant \left\|c\mathcal{K}^{-1}\left(s_k\right)\right\|_{H_2}^2 \left\|\mathcal{K}^{-1}\left(s_{k-1}\right)\right\|_{H_2}^2 \ldots \left\|\mathcal{K}^{-1}\left(s_1\right)\right\|_{H_2}^2 \left\|U(s_1,\ \ldots,\ s_k)\right\|_{H_\infty}^2 \left\|\mathcal{K}^{-1}\left(s_1\right)b\right\|_{H_\infty}^2,$$

where $\mathcal{K}\left(s_i\right)=\left(s_iI_n-A\right)$ *for* $i=1,\ \ldots,\ k$*, and*

$$\begin{aligned} U(s_1,\ldots,\ s_k) = \mathcal{K}\left(s_1\right)\ldots\mathcal{K}\left(s_{k-1}\right)\Bigg(&N\mathcal{K}^{-1}\left(s_{k-1}\right)\ldots N\mathcal{K}^{-1}\left(s_2\right)N \\ &-\left(I_n-F\mathcal{K}^{-1}\left(s_k\right)\right)^{-1} \\ &N\mathcal{K}^{-1}\left(s_{k-1}\right)\left(I_n-F\mathcal{K}^{-1}\left(s_{k-1}\right)\right)^{-1}\ldots N\mathcal{K}^{-1}\left(s_2\right)\left(I_n-F\mathcal{K}^{-1}\left(s_2\right)\right)^{-1} \\ &N\left(I_n-\mathcal{K}^{-1}\left(s_1\right)F\right)^{-1}\Bigg). \end{aligned} \tag{4.23}$$

Proof. Using the definition of H_2-norm (1.6), we get

$$\begin{aligned} &\left\|H_k\left(s_1,\ \ldots,\ s_k\right)-\widetilde{H}_k\left(s_1,\ \ldots,\ s_k\right)\right\|_{H_2}^2 \\ &=\left(\frac{1}{2\pi}\right)^k \underset{m\to\infty}{Lim}\int_{-m}^{m}\ldots\int_{-m}^{m}\Big\|c\mathcal{K}^{-1}\left(i\omega_k\right)N\mathcal{K}^{-1}\left(i\omega_{k-1}\right)\ldots N\mathcal{K}^{-1}\left(i\omega_1\right)b \\ &\quad-c\left(\mathcal{K}\left(i\omega_k\right)-F\right)^{-1}N\left(\mathcal{K}\left(i\omega_{k-1}\right)-F\right)^{-1}\ldots N\left(\mathcal{K}\left(i\omega_2\right)-F\right)^{-1} \\ &\qquad N\left(\mathcal{K}\left(i\omega_1\right)-F\right)^{-1}b\Big\|_F^2\, d\omega_1\ldots d\omega_k \\ &=\left(\frac{1}{2\pi}\right)^k \underset{m\to\infty}{Lim}\int_{-m}^{m}\ldots\int_{-m}^{m}\Bigg\|c\mathcal{K}^{-1}\left(i\omega_k\right)\Bigg(N\mathcal{K}^{-1}\left(i\omega_{k-1}\right)\ldots N\mathcal{K}^{-1}\left(i\omega_2\right)N \\ &\quad-\left(I_n-F\mathcal{K}^{-1}\left(i\omega_k\right)\right)^{-1}N\mathcal{K}^{-1}\left(i\omega_{k-1}\right)\left(I_n-F\mathcal{K}^{-1}\left(i\omega_{k-1}\right)\right)^{-1}\ldots \\ &\qquad N\mathcal{K}^{-1}\left(i\omega_2\right)\left(I_n-F\mathcal{K}^{-1}\left(i\omega_2\right)\right)^{-1}N\left(I_n-\mathcal{K}^{-1}\left(i\omega_1\right)F\right)^{-1}\Bigg)\mathcal{K}^{-1}\left(i\omega_1\right)b\Bigg\|_F^2 \\ &\qquad d\omega_1\ldots d\omega_k \end{aligned}$$

$$= \left(\frac{1}{2\pi}\right)^k \underset{m\to\infty}{Lim} \int_{-m}^{m} \dots \int_{-m}^{m} \left\| c\mathcal{K}^{-1}(i\omega_k)\,\mathcal{K}^{-1}(i\omega_{k-1}) \dots \mathcal{K}^{-1}(i\omega_1) \right.$$
$$\mathcal{K}(i\omega_1) \dots \mathcal{K}(i\omega_{k-1}) \left(N\mathcal{K}^{-1}(i\omega_{k-1}) \dots N\mathcal{K}^{-1}(i\omega_2)\, N \right.$$
$$- \left(I_n - F\mathcal{K}^{-1}(i\omega_k)\right)^{-1} N\mathcal{K}^{-1}(i\omega_{k-1}) \left(I_n - F\mathcal{K}^{-1}(i\omega_{k-1})\right)^{-1} \dots$$
$$\left. N\mathcal{K}^{-1}(i\omega_2) \left(I_n - F\mathcal{K}^{-1}(i\omega_2)\right)^{-1} N \left(I_n - \mathcal{K}^{-1}(i\omega_1)\, F\right)^{-1} \right) \mathcal{K}^{-1}(i\omega_1)\, b \Big\|_F^2$$
$$d\omega_1 \dots d\omega_k.$$

Using $U(s_1,\ \dots,\ s_k)$ given by (4.23), $\|XYZ\|_F \leqslant \|X\|_F \|YZ\|_F$, $\|YZ\|_F \leqslant \|Y\|_F \|Z\|_2$, and comparison integral inequality[1] [33] for any matrices $X,\ Y$, and Z, in the above equation, we have

$$\left\| H_k(s_1,\ \dots,\ s_k) - \widetilde{H}_k(s_1,\ \dots,\ s_k) \right\|_{H_2}^2 \leqslant \left(\frac{1}{2\pi}\right)^k \underset{m\to\infty}{Lim} \int_{-m}^{m} \dots \int_{-m}^{m} \left\| c\mathcal{K}^{-1}(i\omega_k) \right\|_F^2$$
$$\left\| \mathcal{K}^{-1}(i\omega_{k-1}) \right\|_F^2 \dots \left\| \mathcal{K}^{-1}(i\omega_1) \right\|_F^2 \left\| U(i\omega_1,\ \dots,\ i\omega_k) \right\|_2^2 \left\| \mathcal{K}^{-1}(i\omega_1)\, b \right\|_2^2 d\omega_1 \dots d\omega_k. \tag{4.24}$$

From the mean value theorem of integration [33] we know

$$\int_{-m}^{m} \int_{-m}^{m} f(i\omega_2) g(i\omega_1, i\omega_2)\, h(i\omega_1)\, d\omega_1 d\omega_2$$
$$= \int_{-m}^{m} f(i\omega_2) \left(\int_{-m}^{m} g(i\omega_1, i\omega_2)\, h(i\omega_1)\, d\omega_1 \right) d\omega_2$$
$$\leqslant \int_{-m}^{m} f(i\omega_2) \left(\underset{\mathrm{c}\in\mathbb{R}}{max} \left(g(i\mathrm{c}, i\omega_2)\right) \int_{-m}^{m} h(i\omega_1)\, d\omega_1 \right) d\omega_2$$
$$\leqslant \underset{\mathrm{c,d}\in\mathbb{R}}{max} (g(i\mathrm{c}, i\mathrm{d})) \int_{-m}^{m} f(i\omega_2)\, d\omega_2 \int_{-m}^{m} h(i\omega_1)\, d\omega_1.$$

[1] This inequality says if $f(x)$ and $g(x)$ are integrable over $[a, b]$ and $f(x) \leqslant g(x)$, then $\int_a^b f(x)\, dx \leqslant \int_a^b g(x)\, dx$. Note that although we have improper integrals here, this inequality still holds because of the earlier assumption that such integrals give a finite value.

Using this property in (4.24) we get[2]

$$\begin{aligned}
&\left\|H_k\left(s_1,\ \ldots,\ s_k\right)-\widetilde{H}_k\left(s_1,\ \ldots,\ s_k\right)\right\|_{H_2}^2 \\
&\qquad\leqslant\left(\frac{1}{2\pi}\right)^k \underset{m\to\infty}{Lim}\int_{-m}^{m}\ldots\int_{-m}^{m}\left\|c\mathcal{K}^{-1}\left(i\omega_k\right)\right\|_F^2\left\|\mathcal{K}^{-1}\left(i\omega_{k-1}\right)\right\|_F^2\ldots\left\|\mathcal{K}^{-1}\left(i\omega_1\right)\right\|_F^2 \\
&\qquad\qquad d\omega_1\ldots d\omega_k \underset{\omega_1,\ \ldots,\ \omega_k\in\mathbb{R}}{max}\left\|U\left(i\omega_1,\ \ldots,\ i\omega_k\right)\right\|_2^2\ \underset{\omega_1\in\mathbb{R}}{max}\left\|\mathcal{K}^{-1}\left(i\omega_1\right)b\right\|_2^2 \\
&\qquad\leqslant\left\|c\mathcal{K}^{-1}\left(s_k\right)\right\|_{H_2}^2\left\|\mathcal{K}^{-1}\left(s_{k-1}\right)\right\|_{H_2}^2\ldots\left\|\mathcal{K}^{-1}\left(s_1\right)\right\|_{H_2}^2 \\
&\qquad\qquad \left\|U\left(s_1,\ \ldots,\ s_k\right)\right\|_{H_\infty}^2\left\|\mathcal{K}^{-1}\left(s_1\right)b\right\|_{H_\infty}^2.
\end{aligned}$$

□

Lemma 2. *Let $\|F\|_2 < 1$, where F is the perturbation introduced in the A matrix of the input dynamical system. Also, let $\left\|\mathcal{K}^{-1}\left(s_i\right)\right\|_{H_\infty} < 1$ for $i = 1$ and 2, where $\mathcal{K}\left(s_i\right) = \left(s_i I_n - A\right)$ with I_n being the identity matrix. Then,*

$$\left\|U_2\right\|_{H_\infty} \propto \mathcal{O}\left(\|F\|_2\right).$$

where $U_2 = U\left(s_1,\ s_2\right)$ from (4.23).

Proof. Substituting $k = 2$ in (4.23), we get

$$U_2 = \mathcal{K}\left(s_1\right)\left(N - \left(I_n - F\mathcal{K}^{-1}\left(s_2\right)\right)^{-1} N\left(I_n - \mathcal{K}^{-1}\left(s_1\right)F\right)^{-1}\right).$$

If $\left\|F\mathcal{K}^{-1}\left(s_2\right)\right\|_{H_\infty} < 1$ and $\left\|\mathcal{K}^{-1}\left(s_1\right)F\right\|_{H_\infty} < 1$, then by the Neumann series, we get[3]

$$\begin{aligned}
U_2 &= \mathcal{K}\left(s_1\right)\left(N - \left(I_n + F\mathcal{K}^{-1}\left(s_2\right) + \left(F\mathcal{K}^{-1}\left(s_2\right)\right)^2 + \cdots\right)N\right. \\
&\qquad \left.\left(I_n + \mathcal{K}^{-1}\left(s_1\right)F + \left(\mathcal{K}^{-1}\left(s_1\right)F\right)^2 + \cdots\right)\right) \\
&= \mathcal{K}\left(s_1\right)\left(N - N - N\mathcal{K}^{-1}\left(s_1\right)F\left(I_n + \mathcal{K}^{-1}\left(s_1\right)F + \cdots\right)\right. \\
&\qquad \left. - F\mathcal{K}^{-1}\left(s_2\right)\left(I_n + F\mathcal{K}^{-1}\left(s_2\right) + \cdots\right)N\left(I_n + \mathcal{K}^{-1}\left(s_1\right)F + \left(\mathcal{K}^{-1}\left(s_1\right)F\right)^2 + \cdots\right)\right)
\end{aligned}$$

[2]As mentioned in Footnote 1, the improper integrals here do not affect application of this mean value theorem because all such integrals are assumed to give a finite value.

[3]From [38, page 527], we know $\left(I - A\right)^{-1} = \sum_{k=0}^{\infty} A^k$ when $\|A\| < 1$ for any matrix norm. Here, for the first inequality we have $\left\|F\mathcal{K}^{-1}\left(s_2\right)\right\|_{H_\infty} < 1$ or $\underset{\omega_2\in\mathbb{R}}{max}\left\|F\mathcal{K}^{-1}\left(i\omega_2\right)\right\|_2 < 1$, and hence, the applicable matrix norm is 2−norm. Similarly for the second inequality.

$$
\begin{aligned}
&= \mathcal{K}(s_1)\Big(- N\mathcal{K}^{-1}(s_1) F \left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1} \\
&\quad - F\mathcal{K}^{-1}(s_2) \left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} N \left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1} \Big) \\
&= \mathcal{K}(s_1)\Big(- N\mathcal{K}^{-1}(s_1) F - F\mathcal{K}^{-1}(s_2) \left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} N\Big) \left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1}.
\end{aligned}
$$

Taking $H_\infty-$norm on both sides, and using $\|XY\|_2 \leqslant \|X\|_2 \|Y\|_2$ and $\|X+Y\|_2 \leqslant \|X\|_2 + \|Y\|_2$, for any two matrices X and Y, we get

$$
\begin{aligned}
\|U_2\|_{H_\infty} \leqslant \underset{\omega_1,\omega_2\in\mathbb{R}}{max} \Bigg(& \|\mathcal{K}(i\omega_1)\|_2 \Bigg(\|N\|_2 \left\|\mathcal{K}^{-1}(i\omega_1)\right\|_2 \|F\|_2 + \|F\|_2 \left\|\mathcal{K}^{-1}(i\omega_2)\right\|_2 \\
& \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_2)\right)^{-1}\right\|_2 \|N\|_2 \Bigg) \left\|\left(I_n - \mathcal{K}^{-1}(i\omega_1) F\right)^{-1}\right\|_2 \Bigg) \\
\leqslant & \|\mathcal{K}(s_1)\|_{H_\infty} \|N\|_2 \|F\|_2 \Bigg(\left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty} + \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} \\
& \underset{\omega_2\in\mathbb{R}}{max} \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_2)\right)^{-1}\right\|_2 \Bigg) \underset{\omega_1\in\mathbb{R}}{max} \left\|\left(I_n - \mathcal{K}^{-1}(i\omega_1) F\right)^{-1}\right\|_2.
\end{aligned}
\tag{4.25}
$$

Technically by definition of the $H_\infty-$norm and how $\mathcal{K}(s)$ is defined in our hypotheses, $\|\mathcal{K}(s_1)\|_{H_\infty} = \|\mathcal{K}(s_2)\|_{H_\infty} = \|\mathcal{K}(s)\|_{H_\infty}$, however, for sake of exposition, we keep them separate. Similarly for the $H_\infty-$norm of inverses of $\mathcal{K}(s_1)$ and $\mathcal{K}(s_2)$.

To abstract $\|F\|_2$ out from the above inequality, let us look at $\underset{\omega_2\in\mathbb{R}}{max} \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_2)\right)^{-1}\right\|_2$ separately. Recall, while applying Neumann series we assumed that $\|F\mathcal{K}^{-1}(s_2)\|_{H_\infty} < 1$ or $\underset{\omega_2\in\mathbb{R}}{max} \|F\mathcal{K}^{-1}(i\omega_2)\|_2 < 1$. Since the maximum of such a norm is less than one, we have for all $\omega_2 \in \mathbb{R}$, $\|F\mathcal{K}^{-1}(i\omega_2)\|_2 < 1$. Using this along with Lemma 2.3.3 from [27][4] in the above expression, we get

$$
\begin{aligned}
\underset{\omega_2\in\mathbb{R}}{max} \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_2)\right)^{-1}\right\|_2 &\leqslant \underset{\omega_2\in\mathbb{R}}{max} \frac{1}{1 - \|F\mathcal{K}^{-1}(i\omega_2)\|_2} \\
&\leqslant \frac{1}{1 - \underset{\omega_2\in\mathbb{R}}{max} \|F\mathcal{K}^{-1}(i\omega_2)\|_2} \\
&\leqslant \frac{1}{1 - \|F\mathcal{K}^{-1}(s_2)\|_{H_\infty}}.
\end{aligned}
\tag{4.26}
$$

[4]If $F \in \mathbb{R}^{n\times n}$ and $\|F\|_p < 1$, then $I - F$ is nonsingular and $(I-F)^{-1} = \sum_{k=0}^{\infty} F^k$ with $\left\|(I-F)^{-1}\right\|_p \leqslant \frac{1}{1-\|F\|_p}$.

If we assume $\|F\|_2 < 1$ and $\|\mathcal{K}^{-1}(s_2)\|_{H_\infty} < 1$ (as in our hypotheses), then using earlier used matrix norm properties, we get

$$\begin{aligned}\left\|F\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} = \underset{\omega_2\in\mathbb{R}}{max}\left\|F\mathcal{K}^{-1}(i\omega_2)\right\|_2 &\leqslant \|F\|_2 \underset{\omega_2\in\mathbb{R}}{max}\left\|\mathcal{K}^{-1}(i\omega_2)\right\|_2 \\ &\leqslant \|F\|_2 \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} \\ &\leqslant 1,\end{aligned}$$

as assumed for applying Neumann series earlier as well as Lemma 2.3.3 from [27] above. Thus, no extra assumptions beyond those in hypotheses are needed. Further, we also get

$$1 - \|F\|_2 \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} \leqslant 1 - \left\|F\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} \qquad \text{or} \tag{4.27}$$

$$\frac{1}{1 - \left\|F\mathcal{K}^{-1}(s_2)\right\|_{H_\infty}} \leqslant \frac{1}{1 - \|F\|_2 \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty}}. \tag{4.28}$$

Similarly, by assuming $\|F\|_2 < 1$ and $\|\mathcal{K}^{-1}(s_1)\|_{H_\infty} < 1$ (as in our hypotheses), we can bound the last term of (4.25) as follows:

$$\underset{\omega_1\in\mathbb{R}}{max}\left\|\left(I_n - \mathcal{K}^{-1}(i\omega_1)F\right)^{-1}\right\|_2 \leqslant \frac{1}{1 - \left\|\mathcal{K}^{-1}(s_1)F\right\|_{H_\infty}} \qquad \text{and} \tag{4.29}$$

$$\frac{1}{1 - \left\|\mathcal{K}^{-1}(s_1)F\right\|_{H_\infty}} \leqslant \frac{1}{1 - \left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty}\|F\|_2}. \tag{4.30}$$

Substituting (4.26)-(4.28) and (4.29)-(4.30) in (4.25), we get

$$\begin{aligned}\|U_2\|_{H_\infty} \leqslant \|\mathcal{K}(s_1)\|_{H_\infty} \|N\|_2 \|F\|_2 &\left[\left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty} + \frac{\left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty}}{1 - \|F\|_2 \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty}}\right] \\ &\left(\frac{1}{1 - \left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty}\|F\|_2}\right).\end{aligned}$$

From the above inequality it is clear that if $\|F\|_2 \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} < 1$ and $\left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty}\|F\|_2 < 1$, which are true from our hypotheses, then

$$\|U_2\|_{H_\infty} = \mathcal{O}(\|F\|_2).$$

□

Lemma 3. *Let $\|F\|_2 < 1$, where F is the perturbation introduced in the A matrix of the input dynamical system. Also, let $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty} < 1$ for $i = 1,\ 2,\ \ldots,\ k$, where $\mathcal{K}(s_i) = (s_i I_n - A)$ with I_n being the identity matrix. Then,*

$$\|U_k\|_{H_\infty} \propto \mathcal{O}\left(\|F\|_2\right),$$

where $U_k = U(s_1,\ \ldots,\ s_k)$ from (4.23).

Proof. We prove this by mathematical induction.

Base Case :

$k = 1$ is the linear system case already proved in [10] (see below theorem).

Theorem 9. *[10] Let F be the constant perturbation introduced in the A matrix of the input dynamical system. If $\|\mathcal{K}^{-1}(s)\|_{H_\infty} < 1$ and $\|F\|_2 < 1$, then*

$$\left\|H(s) - \widetilde{H}(s)\right\|_{H_2} \leqslant \frac{\|c\mathcal{K}^{-1}(s)\|_{H_2}\,\|\mathcal{K}^{-1}(s)\,b\|_{H_\infty}}{1 - \|\mathcal{K}^{-1}(s)\|_{H_\infty}\,\|F\|_2}\,\|F\|_2,$$

where $\mathcal{K}(s) = (sI_n - A)$ for $s \in \{s_1,\ \ldots,\ s_k\}$.

$k = 2$ has been proved above (Lemma 2).

Induction Hypothesis :

From (4.23), we know for $k = L$

$$\begin{aligned} U_L = \mathcal{K}(s_1)\ldots\mathcal{K}(s_{L-1})\Bigg(&N\mathcal{K}^{-1}(s_{L-1})\ldots N\mathcal{K}^{-1}(s_2)\,N - \left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \\ &N\mathcal{K}^{-1}(s_{L-1})\left(I_n - F\mathcal{K}^{-1}(s_{L-1})\right)^{-1}\ldots N\mathcal{K}^{-1}(s_2)\left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\ &N\left(I_n - \mathcal{K}^{-1}(s_1)\,F\right)^{-1}\Bigg). \end{aligned} \tag{4.31}$$

Let $\|U_L\|_{H_\infty} = \mathcal{O}\left(\|F\|_2\right)$.

Induction Step :

We show the above for $k = L + 1$. Again, from (4.23), we know

$$U_{L+1} = \mathcal{K}(s_1) \dots \mathcal{K}(s_L) \Bigg(N\mathcal{K}^{-1}(s_L) \dots N\mathcal{K}^{-1}(s_2) N - \left(I_n - F\mathcal{K}^{-1}(s_{L+1})\right)^{-1} \\ N\mathcal{K}^{-1}(s_L) \left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \dots N\mathcal{K}^{-1}(s_2) \left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\ N\left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1} \Bigg).$$

We first write U_{L+1} in terms of U_L. Using our hypotheses, we have $\|F\mathcal{K}^{-1}(s_{L+1})\|_{H_\infty} < \|F\|_2 \|\mathcal{K}^{-1}(s_{L+1})\|_{H_\infty} < 1$, and hence, applying Neumann series above, we get

$$\begin{aligned} U_{L+1} &= \mathcal{K}(s_1) \dots \mathcal{K}(s_L) \Bigg(N\mathcal{K}^{-1}(s_L) \dots N\mathcal{K}^{-1}(s_2) N \\ &\quad - \left(I_n + F\mathcal{K}^{-1}(s_{L+1}) + \left(F\mathcal{K}^{-1}(s_{L+1})\right)^2 + \cdots\right) \\ &\quad N\mathcal{K}^{-1}(s_L) \left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \dots N\mathcal{K}^{-1}(s_2) \left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\ &\quad N\left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1} \Bigg) \\ &= \mathcal{K}(s_1) \dots \mathcal{K}(s_L) \Bigg(N\mathcal{K}^{-1}(s_L) \dots N\mathcal{K}^{-1}(s_2) N \\ &\quad - N\mathcal{K}^{-1}(s_L) \left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \dots N\mathcal{K}^{-1}(s_2) \left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\ &\quad N\left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1} \\ &\quad - F\mathcal{K}^{-1}(s_{L+1}) \left(I_n - F\mathcal{K}^{-1}(s_{L+1})\right)^{-1} N\mathcal{K}^{-1}(s_L) \left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \dots \\ &\quad N\mathcal{K}^{-1}(s_2) \left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} N\left(I_n - \mathcal{K}^{-1}(s_1) F\right)^{-1} \Bigg). \end{aligned}$$

In the above equation, taking $N\mathcal{K}^{-1}(s_L)$ common from the first two terms of the

bigger bracket, we have

$$
\begin{aligned}
&= \mathcal{K}(s_1)\ldots\mathcal{K}(s_L)\Bigg(N\mathcal{K}^{-1}(s_L)\Bigg(N\mathcal{K}^{-1}(s_{L-1})\ldots N\mathcal{K}^{-1}(s_2)\,N \\
&\quad -\left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \\
&\qquad N\mathcal{K}^{-1}(s_{L-1})\left(I_n - F\mathcal{K}^{-1}(s_{L-1})\right)^{-1}\ldots N\mathcal{K}^{-1}(s_2)\left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\
&\qquad N\left(I_n - \mathcal{K}^{-1}(s_1)\,F\right)^{-1}\Bigg) \\
&\quad - F\mathcal{K}^{-1}(s_{L+1})\left(I_n - F\mathcal{K}^{-1}(s_{L+1})\right)^{-1} N\mathcal{K}^{-1}(s_L)\left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1}\ldots \\
&\qquad N\mathcal{K}^{-1}(s_2)\left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} N\left(I_n - \mathcal{K}^{-1}(s_1)\,F\right)^{-1}\Bigg).
\end{aligned}
\tag{4.32}
$$

Now we look at expression of U_L. Multiplying $\mathcal{K}^{-1}(s_{L-1})\ldots\mathcal{K}^{-1}(s_1)$ on both the sides of (4.31) from left, we get

$$
\begin{aligned}
\mathcal{K}^{-1}(s_{L-1})\ldots\mathcal{K}^{-1}(s_1)\,U_L = \Bigg(&N\mathcal{K}^{-1}(s_{L-1})\ldots N\mathcal{K}^{-1}(s_2)\,N - \left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1} \\
&N\mathcal{K}^{-1}(s_{L-1})\left(I_n - F\mathcal{K}^{-1}(s_{L-1})\right)^{-1}\ldots N\mathcal{K}^{-1}(s_2)\left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\
&N\left(I_n - \mathcal{K}^{-1}(s_1)\,F\right)^{-1}\Bigg).
\end{aligned}
\tag{4.33}
$$

Substituting (4.33) in (4.32), we get

$$
\begin{aligned}
U_{L+1} = \mathcal{K}(s_1)\ldots\mathcal{K}(s_L)\Bigg(&N\mathcal{K}^{-1}(s_L)\Bigg(\mathcal{K}^{-1}(s_{L-1})\ldots\mathcal{K}^{-1}(s_1)\,U_L\Bigg) \\
&- F\mathcal{K}^{-1}(s_{L+1})\left(I_n - F\mathcal{K}^{-1}(s_{L+1})\right)^{-1} \\
&\quad N\mathcal{K}^{-1}(s_L)\left(I_n - F\mathcal{K}^{-1}(s_L)\right)^{-1}\ldots N\mathcal{K}^{-1}(s_2)\left(I_n - F\mathcal{K}^{-1}(s_2)\right)^{-1} \\
&\quad N\left(I_n - \mathcal{K}^{-1}(s_1)\,F\right)^{-1}\Bigg).
\end{aligned}
$$

Taking $H_\infty-$norm on both sides, and as earlier, using the norm inequality properties

in the above equation, we get

$$\begin{aligned}\|U_{L+1}\|_{H_\infty} \leqslant \max_{\omega_1,\ \ldots,\ \omega_{L+1}\in\mathbb{R}} \Bigg[& \|\mathcal{K}(i\omega_1)\|_2 \ldots \|\mathcal{K}(i\omega_L)\|_2 \Bigg(\|N\|_2 \left\|\mathcal{K}^{-1}(i\omega_L)\right\|_2 \ldots \left\|\mathcal{K}^{-1}(i\omega_1)\right\|_2 \\ & \|U(i\omega_1,\ \ldots,\ i\omega_L)\|_2 + \|F\|_2 \left\|\mathcal{K}^{-1}(i\omega_{L+1})\right\|_2 \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_{L+1})\right)^{-1}\right\|_2 \\ & \|N\|_2 \left\|\mathcal{K}^{-1}(i\omega_L)\right\|_2 \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_L)\right)^{-1}\right\|_2 \ldots \\ & \|N\|_2 \left\|\mathcal{K}^{-1}(i\omega_2)\right\|_2 \left\|\left(I_n - F\mathcal{K}^{-1}(i\omega_2)\right)^{-1}\right\|_2 \\ & \|N\|_2 \left\|\left(I_n - \mathcal{K}^{-1}(i\omega_1)\,F\right)^{-1}\right\|_2 \Bigg) \Bigg].\end{aligned}$$

Similar to (4.26) and (4.28), here also, using Lemma 2.3.3 from [27] we get

$$\begin{aligned}\|U_{L+1}\|_{H_\infty} \leqslant & \|\mathcal{K}(s_1)\|_{H_\infty} \ldots \|\mathcal{K}(s_L)\|_{H_\infty} \|N\|_2 \left\|\mathcal{K}^{-1}(s_L)\right\|_{H_\infty} \ldots \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty} \\ & \Bigg[\left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty} \|U_L\|_{H_\infty} + \\ & \frac{\|N\|_2^{L-1} \left\|\mathcal{K}^{-1}(s_{L+1})\right\|_{H_\infty}}{\left(1 - \|F\|_2 \left\|\mathcal{K}^{-1}(s_{L+1})\right\|_{H_\infty}\right) \ldots \left(1 - \|F\|_2 \left\|\mathcal{K}^{-1}(s_2)\right\|_{H_\infty}\right)} \\ & \cdot \frac{\|F\|_2}{1 - \left\|\mathcal{K}^{-1}(s_1)\right\|_{H_\infty} \|F\|_2} \Bigg].\end{aligned}$$

From induction hypothesis we know $\|U_L\|_{H_\infty} \propto \mathcal{O}(\|F\|_2)$. Using this we get

$$\|U_{L+1}\|_{H_\infty} \propto \mathcal{O}(\|F\|_2).$$

□

Theorem 10. *If hypotheses of Lemmas 1 and 3 holds, then*

$$\left\|H_k(s_1,\ \ldots,\ s_k) - \widetilde{H}_k(s_1,\ \ldots,\ s_k)\right\|_{H_2}^2 = \mathcal{O}\left(\|F\|_2^2\right)$$

or

TBIRKA satisfies the second condition of backward stability with respect to inexact linear solves.

Proof. Directly follows from combining the results of Lemmas 1 and 3. □

Corollary 2. *Assuming the hypotheses of Theorem 7, and either of Theorem 8 or Theorem 10 are satisfied, then TBIRKA is backward stable with respect to the inexact linear solves.*

In the next section, we analyze all the involved matrices.

4.3 Invertibility of Involved Matrices

Similar to previous chapter, here also we have assumed invertibility of seven matrices. Most of these invertibility assumptions directly come from the control system theory as well as the model reduction theory of bilinear systems. We have also assumed invertibility of few newly proposed matrices. In this section, we summarize/ analyze all these assumptions in the order of appearance of the corresponding matrix in this chapter. We first summarize the invertibility assumptions from literature.

(a) In the H_2-norm definition of a truncated bilinear dynamical system (1.9), we assume that $(-A \otimes I_n - I_n \otimes A)$ is invertible. This is a standard definition. Please see Lemma 5.1 of [26].

(b) Here again we assume invertibility of $\left(\widetilde{W}_r^T \widetilde{V}_r\right)$. Directly coming from literature. This is easy to enforce and come from TBIRKA. Please see Algorithm 4.3 of [24] or Algorithm 2 of [26] .

(c) In (2.4), we assume the middle term, i.e.,

$$\left(- \begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_r \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_r \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \check{A} \end{bmatrix} \right)$$

is invertible. This comes from the H_2-norm of the error system $(\zeta^M - \zeta_r^M)$. Please see Theorem 5.1 of [26].

(d) We assume invertibility of $(-\Lambda \otimes I_n - I_r \otimes A)$ in Algorithm 2.2. This again comes from TBIRKA. Please see Algorithm 4.3 of [24] or Algorithm 2 of [26].

(e) Again, as earlier, here also we assume invertibility of $(s_k I_n - A)$ and $(s_k I_n - (A + F))$ in (4.21) and (4.22), respectively. These come from the transfer function definitions. Please see Section 2 of [26] and Theorem 4.1 of [10], respectively.

During the backward stability analysis of TBIRKA, we assume invertibility of some newly proposed matrices. Next, we analyze these matrices. Note that below, we discuss the matrix in (a) before the matrix in (b) although the latter appears first in this chapter. This is done for ease of exposition.

(a) We assume invertibility of $\widehat{Q}$ given in (4.8). Also listed below for easy access.

$$\widehat{Q} = -\begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}.$$

This is one of the most important assumption in obtaining a backward stable TBIRKA (see Corollary 2). Hence, here we relate this invertibility assumption with the underlying bilinear dynamical system. If we define $A_2 = \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}$, $I_{2n} = \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix}$, and $\widehat{Q} = Q_1 \otimes Q_2$, where $Q_1, Q_2 \in \mathbb{R}^{2n \times 2n}$ are any two matrices, then $\widehat{Q}$ can be rewritten as

$$\begin{aligned}
-A_2 \otimes I_{2n} - I_{2n} \otimes A_2 &= Q_1 \otimes Q_2 \qquad \text{or} \\
-\left(A_2 \otimes I_{2n}\right) vec(I_{2n}) - \left(I_{2n} \otimes A_2\right) vec(I_{2n}) &= \left(Q_1 \otimes Q_2\right) vec(I_{2n}) \quad \text{or} \\
-A_2^T - A_2 &= Q_2 Q_1^T \qquad \text{or} \\
-\begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}^T - \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} &= Q_2 Q_1^T \qquad \text{or} \\
\begin{bmatrix} -A^T - A & 0 \\ 0 & -A^T - A \end{bmatrix} &= Q_2 Q_1^T.
\end{aligned}$$

If $\left(-A^T - A\right)$ is invertible, then Q_1 and Q_2 are invertible. This implies that $\widehat{Q} = (Q_1 \otimes Q_2)$ is invertible. As in the case of BIRKA stability (see (b) on page 36-37), $\left(-A^T - A\right)$ is directly related to the underlying Lyapunov equation.

(b) In (4.6) and (4.9), we assume invertibility of

$$\left(- \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A+F \end{bmatrix} \right)$$

and $\left(\widehat{Q} - \widehat{\widehat{F}} \right)$, respectively, both of which represent the same matrix (i.e., $\widehat{Q}$ with perturbation). This matrix is invertible if $\Big(- (A+F)^T - (A+F) \Big)$ is invertible.

Next, we look at the conditioning of the problem and the perturbation expression, leading to the accuracy of the reduced system.

4.4 Accuracy of the Reduced System

Next, as in the previous chapter, we compute the accuracy for the reduced system obtained after using inexact TBIRKA. Assume that TBIRKA satisfies the hypotheses of above Corollary 2, i.e., TBIRKA is backward stable with respect to the inexact linear solves. Then, from Theorem 3 we get

$$\begin{aligned} \frac{\left\| g\left(\zeta^M\right) - \widetilde{g}\left(\zeta^M\right) \right\|_{H_2}}{\left\| g\left(\zeta^M\right) \right\|_{H_2}} &= \mathcal{O}\left(\kappa(\zeta^M) \, \|F\| \right), \quad \text{or} \\ \frac{\left\| \zeta_r^M - \widetilde{\zeta}_r^M \right\|_{H_2}}{\left\| \zeta_r^M \right\|_{H_2}} &= \mathcal{O}\left(\kappa(\zeta^M) \, \|F\| \right), \end{aligned} \tag{4.34}$$

where, as earlier, g denotes exact TBIRKA, $\widetilde{g}$ denotes inexact TBIRKA, ζ^M is the original full model, $\kappa(\zeta^M)$ is the condition number of ζ^M (discussed below), and F is the perturbation in ζ^M. Also, $g\left(\zeta^M\right) = \zeta_r^M$, and $\widetilde{g}\left(\zeta^M\right) = \widetilde{\zeta}_r^M$.

Thus, the accuracy of the reduced system is dependent upon the condition number of the problem and the perturbation. First, we compute the conditioning of the original bilinear system with respect to computing the H_2-norm of the error system $\zeta^{M^{err}} = \zeta^M - \widetilde{\zeta}^M$. This is easy to compute and gives a good approximation to the conditioning of the original bilinear system with respect to computing the H_2-norm of the error system $\zeta_r^M - \widetilde{\zeta}_r^M$ (as needed in (4.34)). Similar analysis has been done in [21].

Recall the error expression as defined in (4.9) or

$$\begin{aligned}\left\|\zeta^{M^{err}}\right\|_{H_2}^2 = \left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}^2 = vec(I_{2p})^T \widehat{C} \sum_{k=0}^{M}\left[\widehat{Q}^{-1}\left(I_{2n} - \widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{N}\right]^k \widehat{Q}^{-1} \\ \left(I_{2n} - \widehat{\widehat{F}}\widehat{Q}^{-1}\right)^{-1}\widehat{B}\ vec(I_{2m}), \\ = J_0 + J_1 + J_2 + \ldots + J_M,\end{aligned}$$

where J_0, J_1, J_2, and J_M are defined in (4.10), (4.13), (4.15), and (4.17), respectively. Thus, the error expression above can be bounded as follow:

$$\left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2} \leqslant |J_0| + |J_1| + |J_2| + \ldots + |J_M|\,. \tag{4.35}$$

Recall that J_0, J_1, J_2, and J_M have already been bounded in (4.12), (4.14), (4.16), and (4.18), respectively. For stability, in Corollary 2, we have assumed $\left\|\widehat{\widehat{F}}\right\| < 1$, which gives

$$\frac{1}{1 - \left\|\widehat{\widehat{F}}\right\|\left\|\widehat{Q}^{-1}\right\|} \leqslant \frac{1}{1 - \left\|\widehat{Q}^{-1}\right\|}.$$

Using this new bound in (4.12), (4.14), (4.16), and (4.18) we get

$$\begin{aligned}\left|J_k\right| \leqslant \left\|vec(I_{2p})^T\right\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\widehat{N}\right\|^k \left\|\widehat{Q}^{-1}\right\|^k \left\|\widehat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \\ \left[\left(\frac{1}{1 - \left\|\widehat{Q}^{-1}\right\|}\right) + \ldots + \left(\frac{1}{1 - \left\|\widehat{Q}^{-1}\right\|}\right)^{k+1}\right] \left\|\widehat{B}\right\| \left\|vec(I_{2m})\right\|,\end{aligned}$$

for $k = 0, \ldots, M$. Using the above in (4.35) we get

$$\begin{aligned}
\left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2} \leqslant \|vec(I_{2p})^T\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\hat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \left(\frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}\right) \\
\left[1 + \left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\| \left(1 + \frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}\right)\right. \\
+ \left\|\widehat{N}\right\|^2 \left\|\widehat{Q}^{-1}\right\|^2 \left(1 + \frac{1}{1-\left\|\widehat{Q}^{-1}\right\|} + \left(\frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}\right)^2\right) \\
+ \ldots \\
\left. + \left\|\widehat{N}\right\|^M \left\|\widehat{Q}^{-1}\right\|^M \left(1 + \frac{1}{1-\left\|\widehat{Q}^{-1}\right\|} + \ldots + \left(\frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}\right)^M\right)\right] \\
\left\|\widehat{B}\right\| \|vec(I_{2m})\| .
\end{aligned}$$

We know that, if S is an arithmetic progression of the form $1 + a(1+x) + a^2(1+x+x^2) + \ldots + a^m(1+x+x^2+\ldots+x^m)$, then $S = \frac{1}{1-x}\left[\frac{a^{m+1}-1}{a-1} - \frac{x\left((ax)^{m+1}-1\right)}{ax-1}\right]$. Using this property in the above inequality, we get

$$\begin{aligned}
\left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2} \leqslant \|vec(I_{2p})^T\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left\|\hat{\widehat{F}}\right\| \left\|\widehat{Q}^{-1}\right\| \left(\frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}\right) \\
\left(\frac{1}{1-\frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}}\right) \left[\frac{\left\|\widehat{N}\right\|^{M+1} \left\|\widehat{Q}^{-1}\right\|^{M+1} - 1}{\left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\| - 1}\right. \\
\left. - \left(\frac{1}{1-\left\|\widehat{Q}^{-1}\right\|}\right) \frac{\frac{\left\|\widehat{N}\right\|^{M+1} \left\|\widehat{Q}^{-1}\right\|^{M+1}}{\left(1-\left\|\widehat{Q}^{-1}\right\|\right)^{M+1}} - 1}{\frac{\left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\|}{\left(1-\left\|\widehat{Q}^{-1}\right\|\right)} - 1}\right] \\
\left\|\widehat{B}\right\| \|vec(I_{2m})\| \quad \text{or}
\end{aligned}$$

$$\frac{\left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}}{\|\zeta^M\|_{H_2}} \leqslant \|vec(I_{2p})^T\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left[\frac{\left(1 - \left\|\widehat{Q}^{-1}\right\|\right)^{M+1} - \left\|\widehat{N}\right\|^{M+1} \left\|\widehat{Q}^{-1}\right\|^{M+1}}{\left(1 - \left\|\widehat{Q}^{-1}\right\|\right)^{M+1} \left(1 - \left\|\widehat{Q}^{-1}\right\| - \left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\|\right)} - \frac{1 - \left\|\widehat{N}\right\|^{M+1} \left\|\widehat{Q}^{-1}\right\|^{M+1}}{1 - \left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\|}\right] \left\|\widehat{B}\right\| \|vec(I_{2m})\| \frac{\|A\|}{\|\zeta_M\|_{H_2}} \frac{\left\|\widehat{\widehat{F}}\right\|}{\|A\|}.$$

Since, the condition number $k(\zeta^M)$ gives the relative change in the output (in our case this is $\frac{\left\|\zeta^M - \widetilde{\zeta}^M\right\|_{H_2}}{\|\zeta^M\|_{H_2}}$) with respect to relative change in the input (in our case this is $\frac{\|F\|}{\|A\|}$ because we are perturbing the matrix A only), using the above inequality and the fact that $\mathcal{O}\left(\left\|\widehat{\widehat{F}}\right\|\right) \leqslant \mathcal{O}\left(\|F\|\right)$ (from proof of Theorem 8) we have

$$k(\zeta^M) = \|vec(I_{2p})^T\| \left\|\widehat{C}\right\| \left\|\widehat{Q}^{-1}\right\| \left[\frac{\left(1 - \left\|\widehat{Q}^{-1}\right\|\right)^{M+1} - \left\|\widehat{N}\right\|^{M+1} \left\|\widehat{Q}^{-1}\right\|^{M+1}}{\left(1 - \left\|\widehat{Q}^{-1}\right\|\right)^{M+1} \left(1 - \left\|\widehat{Q}^{-1}\right\| - \left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\|\right)} - \frac{1 - \left\|\widehat{N}\right\|^{M+1} \left\|\widehat{Q}^{-1}\right\|^{M+1}}{1 - \left\|\widehat{N}\right\| \left\|\widehat{Q}^{-1}\right\|}\right] \left\|\widehat{B}\right\| \|vec(I_{2m})\| \frac{\|A\|}{\|\zeta_M\|_{H_2}}. \tag{4.36}$$

In the experimental results section, we show that condition numbers of the problems under consideration are fairly small[2]. This implies that our problems are well conditioned with respect to computing the H_2-norm of the error system $\zeta^{M^{err}}$.

Second, we determine the upper bound on the perturbation F with respect to the residuals of inexact linear solves defined in Theorem 7, simultaneously. That is,

$$R_{B_k} = F\widetilde{V}_k \qquad \text{and} \qquad R_{C_k}^T = \widetilde{W}_k^T F \qquad \text{for } k = 1, \ \ldots, \ M.$$

If we revisit the TBIRKA algorithm (Algorithm 2.2), we ideally want to perform stability analysis at the end of line 3d. (where the intermediate projection matrices are summed to obtain the reduced model). However, until now, we have performed

stability analysis at the end of line 3b. and the end of the every step of the `for` loop of line 3c. We have done this for two reasons. First, the two analysis are equivalent mathematically, and second, the latter is more easy to implement than the former. When we derive the expression for perturbation, the former analysis gives a single expression for perturbation, which we need and is not possible in the latter case. Hence, in the rest of this section, we refer to end of line 3d. for our analysis. That is,

$$R_B = F\widetilde{V} \qquad \text{and} \qquad R_C^T = \widetilde{W}^T F \tag{4.37}$$

where $R_B = \sum_{k=1}^{M} R_{B_k}$, $R_C = \sum_{k=1}^{M} R_{C_k}$, $\widetilde{V} = \sum_{k=1}^{M} V_k$ and $\widetilde{W} = \sum_{k=1}^{M} W_k$.

From the assumption of Theorem 7, for satisfying the first condition of backward stability in TBIRKA we need to use the iterative solver based upon the framework given by (4.1), i.e.,

$$\begin{bmatrix} V_1^T \\ V_2^T \\ \vdots \\ V_M^T \end{bmatrix} \begin{bmatrix} R_{C_1} & R_{C_2} & \cdots & R_{C_M} \end{bmatrix} = 0 \qquad \text{and}$$

$$\begin{bmatrix} W_1^T \\ W_2^T \\ \vdots \\ W_M^T \end{bmatrix} \begin{bmatrix} R_{B_1} & R_{B_2} & \cdots & R_{B_M} \end{bmatrix} = 0.$$

This implies

$$\widetilde{W} \perp [R_{B_1},\ R_{B_2},\ \ldots,\ R_{B_M}] \quad \text{and} \quad \widetilde{V} \perp [R_{C_1},\ R_{C_2},\ \ldots,\ R_{C_M}]. \tag{4.38}$$

If we define

$$F = R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1} \widetilde{W}^T + \widetilde{V} \left(\widetilde{W}^T\widetilde{V}\right)^{-1} R_C^T, \tag{4.39}$$

then using (4.38), (4.37) is satisfied. Also, as earlier discussed $\left(\widetilde{W}^T\widetilde{V}\right)$ is assumed to be invertible. The following theorem gives a bound on this perturbation F. This theorem is similar to Theorem 5 from [21].

Theorem 11. *Let R_{B_k}, R_{C_k}, $\widetilde{W}_k$ and $\widetilde{V}_k$ are defined in* (4.1). *F be defined as in* (4.39). *Also, $R_B = \sum_{k=1}^{M} R_{B_k}$, $R_C = \sum_{k=1}^{M} R_{C_k}$, $\widetilde{V} = \sum_{k=1}^{M} V_k$, $\widetilde{W} = \sum_{k=1}^{M} W_k$ and assume $\left(\widetilde{W}^T\widetilde{V}\right)$ is nonsingular. Then, the perturbation F satisfies.*

$$\|F\|_2 \leqslant \|F\|_F \leqslant \sqrt{r}\Big\{\max_i \|R_{B_i}\| \left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\| + \max_i \|R_{C_i}\| \left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|\Big\}.$$

Proof. Note that

$$\begin{aligned} F &= R_B\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T + \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T. \\ \|F\|_F &= \left\|R_B\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T + \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T\right\|_F \\ \|F\|_F &\leqslant \left\|R_B\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|_F + \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_C^T\right\|_F. \end{aligned}$$

Now taking the first term from above expression as

$$\begin{aligned} \left\|R_B\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|_F &\leqslant \|R_B\|_F \left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\| \\ &\leqslant \sqrt{r}\left(\max_i \|R_{B_i}\|\right)\left\|\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T\right\|. \end{aligned}$$

Similarly taking the second term as

$$\begin{aligned} \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}R_c^T\right\|_F &\leqslant \left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\| \|R_c\|_F \\ &\leqslant \sqrt{r}\left(\max_i \|R_{C_i}\|\right)\left\|\widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1}\right\|. \end{aligned}$$

So, finally the expression $\|F\|_F$ given as

$$\|F\|_2 \leqslant \|F\|_F \leqslant \sqrt{r}\Big\{\max_i \|R_{B_i}\| \left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\| + \max_i \|R_{C_i}\| \left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|\Big\}.$$

□

From the above expression of $\|F\|$, we see that the norm of the perturbation is proportional to the norm of the two residuals sum obtained while solving the linear systems arises at step 3b. and 3c. in TBIRKA. The norm of perturbation is also proportional to the norm of two other quantities $\left\|(\widetilde{W}^T\widetilde{V})^{-1}\widetilde{W}^T\right\|$ and $\left\|\widetilde{V}(\widetilde{W}^T\widetilde{V})^{-1}\right\|$. These two quantities are very less dependent on accuracy of the linear systems we solve.

Also, they are not sensitive to different initializations of TBIRKA as well as different reduced system sizes. This behavior is similar to the related quantities obtained in the stability analysis of BIRKA [21].

From (4.34), $\left\|\zeta_r^M - \widetilde{\zeta}_r^M\right\|$ is proportional to the conditioning of the problem $\left(\kappa(\zeta^M)\right)$ and $\|F\|$. The problem is usually well conditioned. From Theorem 11, $\|F\|$ is directly proportional to $\|R_B\|$ and $\|R_C\|$, where R_B and R_C are the sum of residuals in a TBIRKA iteration. Thus, we get a more accurate reduced system as we iteratively solve the linear systems more accurately arising in TBIRKA. This result is very useful in deciding the stopping tolerance for the linear solves. We support this with experimental results in the next section.

4.5 Numerical Experiments

Here, we first revisit the constraints imposed while satisfying the first and second conditions of stability. Satisfying the first condition for a backward stable TBIRKA requires using a Petrov-Galerkin based iterative solver and achieving some extra-orthogonality conditions during the linear solves (see Corollary 2 or Theorem 7).

As mentioned earlier, we use BiCG as the underlying iterative solver because it is based upon the Petrov-Galerkin framework, however, we do not attempt to satisfy the extra-orthogonalities mentioned above for simplicity. We do this during PMOR in the next chapter where the MOR algorithm is simpler than TBIRKA, and hence, satisfying extra-orthogonalities is not hard.

Although these orthogonalities are not satisfied by TBIRKA, it turns to be backward stable experimentally (we get a more accurate reduced model as we solve the linear systems more accurately). This is because these are sufficiency conditions not necessary.

We have seen two ways of satisfying the second condition of stability for TBIRKA with respect to inexact linear solves. Now, we discuss how to practically use those results. In the complete system approach, we have two constraints; $\|\widehat{Q}\| < 1$ and $\|\widehat{\widehat{F}}\| < 1$ (see Corollary 2 and Theorem 8). Here, $\widehat{Q}$ depends only upon the input

dynamical system, and hence, is a rigid constraint. $\widehat{\widehat{F}}$ depends on the subspaces $\widetilde{V}$ and $\widetilde{W}$ and also on the residuals R_B and R_C. As in the case of BIRKA, here also, $\widehat{\widehat{F}}$ is less sensitive to $\widetilde{V}$ and $\widetilde{W}$ and more to the residuals of the linear system that we are solving. Hence, if we need to check that for a given input model would we get a more accurate reduced model by solving the linear systems to a much smaller stopping tolerance (say machine precision), then we only check $\widehat{\widehat{F}} < 1$ for a large stopping tolerance (say 10^{-2}), which guarantees satisfying this constraint for the smaller stopping tolerance (machine precision as above). This saves the effort in solving the linear systems to a smaller stopping tolerance if TBIRKA is unstable for that particular input model and that small stopping tolerance.

In the subsystem approach, the two constraints are $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty} < 1$ and $\|F\| < 1$ (see Corollary 2 and Theorem 10). For $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty} < 1$ as well we check only for a large stopping tolerance and this guarantees that this constraint holds for a smaller stopping tolerance. This is true because by changing the stopping tolerance, the interpolation points do not vary much. We demonstrate this experimentally in our subsequent subsections. The constraint $\|F\| < 1$ can be satisfied in the same way as $\|\widehat{\widehat{F}}\| < 1$ in the complete system approach.

We perform experiments to show three different cases for satisfying the second condition of stability for TBIRKA. First, when the constraints of both the approaches (i.e., complete system and subsystem) are satisfied. Second, when only the constraints of the complete system approach are satisfied but not of the subsystem approach. Finally, the third, when the constraints of the subsystem approach are satisfied but not of the complete system approach. These experiments are performed on a flow model [18]. This model gives us a SISO bilinear dynamical system, which has been already defined in Section 3.3.1.

4.5.1 Constraints of Both Approaches Satisfied

Here, we take N = 10, L = 1 and $v = 0.1$ that gives us a SISO bilinear dynamical system of size 110 [12, 24]. We truncate the Volterra series up to 4 terms, i.e., $M = 4$,

based upon the values in [24] and [26]. We initialize the input reduced system in TBIRKA by random matrices such that initially the first set of constraints of the subsystem approach are satisfied (i.e., $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty} < 1$). The stopping tolerance for TBIRKA is taken as 10^{-6}, and we reduce this model to size 6. Both these values are again chosen based upon the values in [12] and [24]. This leads to solving the linear systems of size 110×110. While using BiCG, we use two different stopping tolerances (10^{-6} and 10^{-10}). Ideally, we should obtain a more accurate reduced model when using the smaller BiCG tolerance.

TBIRKA Iteration	$\|\widehat{\widehat{F}}\|$	$\|F\|$	$\|\mathcal{K}^{-1}(s_i)\|_{H_\infty}$
1	8.8738×10^{-3}	6.1926×10^{-3}	5.56314×10^{-1}
2	3.0384×10^{-2}	2.1509×10^{-2}	5.41302×10^{-1}
3	2.5004×10^{-2}	1.7639×10^{-2}	5.41260×10^{-1}
4	2.9488×10^{-2}	2.0791×10^{-2}	5.40870×10^{-1}
5	2.8530×10^{-2}	2.0134×10^{-2}	5.40833×10^{-1}
6	2.9275×10^{-2}	2.0655×10^{-2}	5.40793×10^{-1}
7	2.9161×10^{-2}	2.0575×10^{-2}	5.40784×10^{-1}
8	2.9293×10^{-2}	2.0668×10^{-2}	5.40778×10^{-1}
9	2.9277×10^{-2}	2.0656×10^{-2}	5.40776×10^{-1}
10	2.9302×10^{-2}	2.0674×10^{-2}	5.40774×10^{-1}
11	2.9300×10^{-2}	2.0672×10^{-2}	5.40774×10^{-1}
12	2.9305×10^{-2}	2.0676×10^{-2}	5.40773×10^{-1}
13	2.9305×10^{-2}	2.0676×10^{-2}	5.40773×10^{-1}
14	2.9306×10^{-2}	2.0676×10^{-2}	5.40773×10^{-1}

Table 4.1: Second condition constraint values for the complete system and the subsystem approaches when using BiCG stopping tolerance of 10^{-6}.

First, we look at the constraints of the second condition of the complete system approach (as discussed above). $\widehat{Q}$ is invertible here. We also have $\left\|\widehat{Q}^{-1}\right\|$ less than one

TBIRKA Iteration	$\|\hat{\hat{F}}\|$	$\|F\|$	$\|\mathcal{K}^{-1}(s_i)\|_{H_\infty}$
1	4.9676×10^{-8}	3.5217×10^{-8}	5.56314×10^{-1}
2	1.1830×10^{-7}	8.2901×10^{-8}	5.41302×10^{-1}
3	4.3547×10^{-7}	2.7917×10^{-7}	5.41260×10^{-1}
4	1.9356×10^{-7}	1.3649×10^{-7}	5.40870×10^{-1}
5	2.0118×10^{-7}	1.4096×10^{-7}	5.40833×10^{-1}
6	2.0720×10^{-7}	1.4472×10^{-7}	5.40793×10^{-1}
7	2.0654×10^{-7}	1.4428×10^{-7}	5.40785×10^{-1}
8	2.0616×10^{-7}	1.4400×10^{-7}	5.40778×10^{-1}
9	2.0601×10^{-7}	1.4390×10^{-7}	5.40776×10^{-1}
10	2.0593×10^{-7}	1.4384×10^{-7}	5.40774×10^{-1}
11	2.0590×10^{-7}	1.4382×10^{-7}	5.40774×10^{-1}
12	2.0588×10^{-7}	1.4381×10^{-7}	5.40773×10^{-1}
13	2.0587×10^{-7}	1.4380×10^{-7}	5.40773×10^{-1}
14	2.0587×10^{-7}	1.4380×10^{-7}	5.40773×10^{-1}

Table 4.2: Second condition constraint values for the complete system and the subsystem approaches when using BiCG stopping tolerance of 10^{-10}.

(i.e., 5.5716×10^{-1}). Finally, $\left\|\hat{\hat{F}}\right\|$, at the end of the first TBIRKA step, for the BiCG stopping tolerances of 10^{-6} and 10^{-10} is 8.8738×10^{-3} and 4.9676×10^{-8}, respectively, both of which are also less than one. These values are less than one at the end of all the other TBIRKA steps as well. This data is given in two tables (Table 4.1 and 4.2; see the second columns there).

Next, we look at the constraints of the second condition of the subsystem approach (as discussed above). $\mathcal{K}(s_i)$ is invertible here. Also, as earlier, due to the particular initialization of TBIRKA, $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty}$, at the end of the first TBIRKA step, for BiCG tolerances of 10^{-6} and 10^{-10} both is less than one (5.5631×10^{-1}). Finally, $\|F\|$, at the end of the first TBIRKA step, for the BiCG stopping tolerance of 10^{-6}

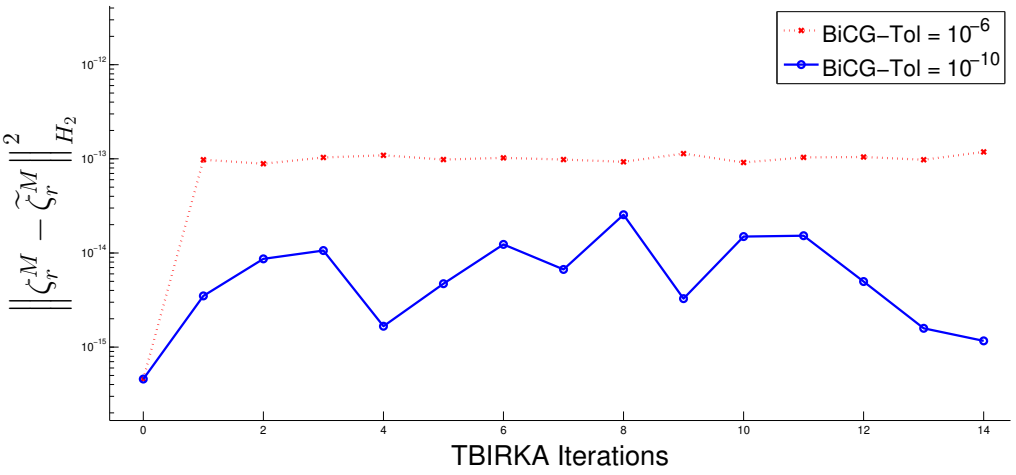

Figure 4.1: Accuracy of the reduced system plotted at each TBIRKA iteration for the two different stopping tolerances in BiCG; Flow model of size 110 (satisfying the constraints in both the complete and the subsystem approach). Here, the x-axis is in the linear scale and the y-axis is in the log scale.

and 10^{-10} is 6.1927×10^{-3} and 3.5218×10^{-8}, respectively, both of which are also less than one. These values are less than one at the end of all the other TBIRKA steps as well (see the third and the fourth columns of Table 4.1 and Table 4.2). The condition number for our input model, as defined in (4.36), is 2.4752×10^{-2}. This shows that the flow model is well-conditioned.

The linear systems arising during the model reduction process are ill-conditioned. Hence, we use a preconditioned BiCG here. The preconditioner that we use is incomplete LU [22]. The drop tolerance in the preconditioner is taken as 10^{-3} based upon the range given in [22].

The accuracy result is given in Figure 4.1. Here, we have the accuracy of the reduced system $\left(\left\|\zeta_r^M - \tilde{\zeta}_r^M\right\|_{H_2}\right)$ on the y-axis in log scale and the TBIRKA iterations on the x-axis in linear scale. From Figure 4.1, it is again evident that we get a more accurate reduced model as we solve the linear systems more accurately (solid line is below the dotted one at all TBIRKA steps).

4.5.2 Constraints of only the Complete System Approach Satisfied

For this experiments, we take the original system same as in the previous subsection (Section 4.5.1) (i.e., N = 10, L = 1 and $v = 0.1$). The only difference is the initialization of the input reduced system. Here, we take the initial input random matrices such that at the start the constraints of subsystem approach are not satisfied (i.e., $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty} \geqslant 1$). This violation is carried forward in the next few TBIRKA steps as well. Again, the stopping tolerance for TBIRKA is taken as 10^{-6}, and we reduce this model to size 6. Both these values are again chosen based upon the values in [12] and [24]. Also, while using BiCG we use two different stopping tolerances (10^{-6} and 10^{-10}).

Next, we look at the constraints of the second condition of the complete system approach. As we know, the value of $\widehat{Q}$ depends only upon the input model, which is the same as that of the previous subsection, thus $\widehat{Q}$ is invertible and less than one. The value of $\left\|\widehat{\widehat{F}}\right\|$, at the end of the first TBIRKA step, for the BiCG stopping tolerances of 10^{-6} and 10^{-10} is 6.3434×10^{-2} and 1.4573×10^{-6}, respectively, both of which are also less than one. These values are less than one at the end of all the other TBIRKA steps as well. Please look at Table 4.3. As earlier subsection, the flow model is well-conditioned.

Here, although the subsystem approach constraints are violated and the complete system approach constraints are satisfied, we still achieve a stable TBIRKA. This is because, as earlier, both of these approaches provide sufficiency conditions for stability, and only one needs to be satisfied. The accuracy result is given in Figure 4.2. Here, we have accuracy of the reduced system $\left(\left\|\zeta_r^M - \widetilde{\zeta}_r^M\right\|_{H_2}\right)$ on the y-axis in the log scale and the TBIRKA iterations on the x-axis in the linear scale. From Figure 4.2, we do not observe any significant difference in the values of $\left(\left\|\zeta_r^M - \widetilde{\zeta}_r^M\right\|_{H_2}\right)$ for the two BiCG tolerances at the starting TBIRKA iterations. The dotted line, which corresponds to the BiCG stopping tolerance 10^{-6} and the solid line, which corresponds to the BiCG stopping tolerance 10^{-10} almost coincide. TBIRKA gets more consistent as it

converges to the ideal interpolation points. Hence, towards the end of the TBIRKA iterations (iteration 14 to iteration 17), the solid line should be below the dotted line. It is again evident that we get a more accurate reduced model as we solve the linear systems more accurately (solid line is below the dotted one at all TBIRKA steps).

TBIRKA	$\|\widehat{\widehat{F}}\|$	
Iteration	BiCG-Tol of 10^{-6}	BiCG-Tol of 10^{-10}
1	6.3434×10^{-2}	1.4573×10^{-6}
2	5.5772×10^{-2}	1.7022×10^{-7}
3	1.7442×10^{-2}	2.2269×10^{-7}
4	2.4148×10^{-2}	1.8257×10^{-7}
5	2.6761×10^{-2}	1.8346×10^{-7}
6	2.7373×10^{-2}	3.5462×10^{-7}
7	2.8725×10^{-2}	1.9356×10^{-7}
8	2.8844×10^{-2}	2.0118×10^{-7}
9	2.9122×10^{-2}	2.0740×10^{-7}
10	2.9209×10^{-2}	2.0659×10^{-7}
11	2.9264×10^{-2}	2.0621×10^{-7}
12	2.9285×10^{-2}	2.0603×10^{-7}
13	2.9297×10^{-2}	2.0594×10^{-7}
14	2.9301×10^{-2}	2.0590×10^{-7}
15	2.9304×10^{-2}	2.0588×10^{-7}
16	2.9305×10^{-2}	2.0587×10^{-7}
17	2.9306×10^{-2}	2.0587×10^{-7}

Table 4.3: Second condition constraint values for the complete system approach when using BiCG stopping tolerances of 10^{-6} and 10^{-10}.

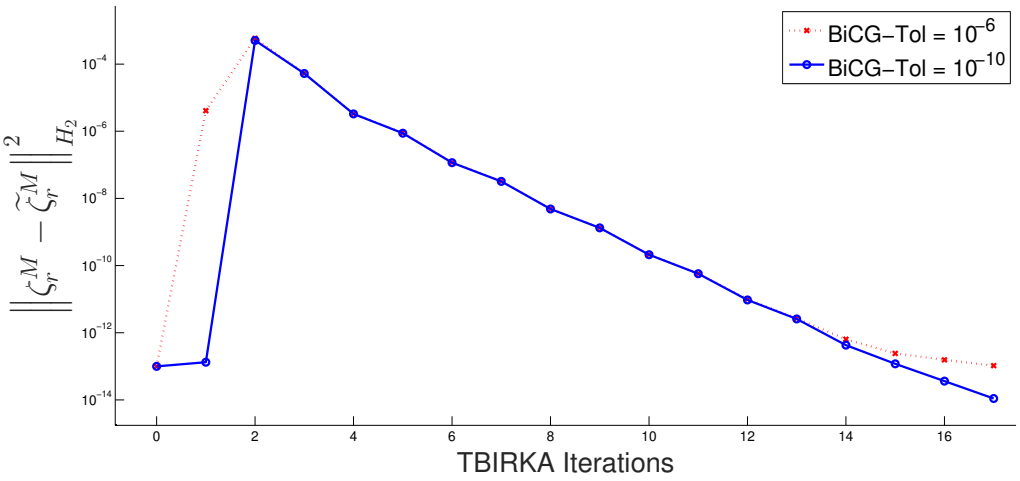

Figure 4.2: Accuracy of the reduced system plotted at each TBIRKA iteration for the two different stopping tolerances in BiCG; Flow model of size 110 (satisfying the constraints of the complete system approach but not of the subsystem approach). Here, the x-axis is in the linear scale and the y-axis is in the log scale.

4.5.3 Constraints of only the Subsystem Approach Satisfied

For our last experiment, we take N = 10, L = 1 and $v = 0.065$ that gives us a SISO bilinear dynamical system of size 110. These parameters are deliberately chosen different than the earlier two experiments because this leads to $\left\|\widehat{Q}^{-1}\right\|$ being greater than one (1.0221×10^0). Thus, the first constraint of the complete system approach is directly violated.

Again, we truncate the Volterra series up to 4 terms, i.e. $M = 4$, as discussed in Section 4.5.1, which has been taken from [24] and [26]. We initialize the input reduced system in TBIRKA by random matrices such that initially, the first set of constraints of the subsystem approach are satisfied (i.e., $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty} < 1$). The stopping tolerance for TBIRKA is taken as 10^{-6}, and we reduce this model to size 6, as earlier based upon the values in [12] and [24]. This leads to solving the linear systems of size 110×110. While using BiCG, we use two different stopping tolerances (10^{-4} and 10^{-8}). This choice is slightly different than the earlier two experiments so as to demonstrate our main conjecture (while satisfying the constraints imposed by

TBIRKA	BiCG-Tol of 10^{-4}		BiCG-Tol of 10^{-8}	
Iteration	$\|\mathcal{K}^{-1}(s_i)\|_{H_\infty}$	$\|F\|$	$\|\mathcal{K}^{-1}(s_i)\|_{H_\infty}$	$\|F\|$
1	9.39899×10^{-1}	1.4028×10^{-2}	9.39899×10^{-1}	2.0097×10^{-5}
2	8.96287×10^{-1}	1.4298×10^{-2}	8.96279×10^{-1}	1.3542×10^{-5}
3	8.99243×10^{-1}	2.0362×10^{-2}	8.99242×10^{-1}	1.6097×10^{-5}
4	9.00299×10^{-1}	2.0751×10^{-2}	9.00313×10^{-1}	1.5380×10^{-5}
5	9.00012×10^{-1}	2.1881×10^{-2}	9.00012×10^{-1}	1.5632×10^{-5}
6	9.00205×10^{-1}	2.1054×10^{-2}	9.00220×10^{-1}	1.5496×10^{-5}
7	9.00128×10^{-1}	2.1343×10^{-2}	9.00130×10^{-1}	1.5561×10^{-5}
8	9.00161×10^{-1}	2.1327×10^{-2}	9.00169×10^{-1}	1.5540×10^{-5}
9	9.00139×10^{-1}	2.1340×10^{-2}	9.00147×10^{-1}	1.5554×10^{-5}
10	9.00147×10^{-1}	2.1337×10^{-2}	9.00155×10^{-1}	1.5549×10^{-5}
11	9.00142×10^{-1}	2.1340×10^{-2}	9.00150×10^{-1}	1.5552×10^{-5}
12	9.00144×10^{-1}	2.1339×10^{-2}	9.00151×10^{-1}	1.5551×10^{-5}
13	9.00142×10^{-1}	2.1340×10^{-2}	9.00150×10^{-1}	1.5551×10^{-5}
14	9.00143×10^{-1}	2.1339×10^{-2}	9.00151×10^{-1}	1.5551×10^{-5}
15	9.00143×10^{-1}	2.1340×10^{-2}	9.00150×10^{-1}	1.5551×10^{-5}
16	9.00143×10^{-1}	2.1340×10^{-2}	9.00150×10^{-1}	1.5551×10^{-5}

Table 4.4: Second condition constraint values for the subsystem approach when using BiCG tolerances of 10^{-4} and 10^{-8}.

the subsystem approach, as we solve the linear systems more accurately, we get a more accurate reduced system).

Next, we look at the constraints of the second condition of the subsystem approach. $\mathcal{K}(s_i)$ is invertible here. Also, as earlier, due to the particular initialization of TBIRKA, $\|\mathcal{K}^{-1}(s_i)\|_{H_\infty}$, at the end of the first TBIRKA step, for BiCG tolerances of 10^{-4} and 10^{-8} both is less than one (9.3989×10^{-1}). Also, $\|F\|$, at the end of the first TBIRKA step, for the BiCG stopping tolerances of 10^{-4} and 10^{-8} is 1.4028×10^{-2} and 2.0098×10^{-5}, respectively, both of which are also less than one. These values are

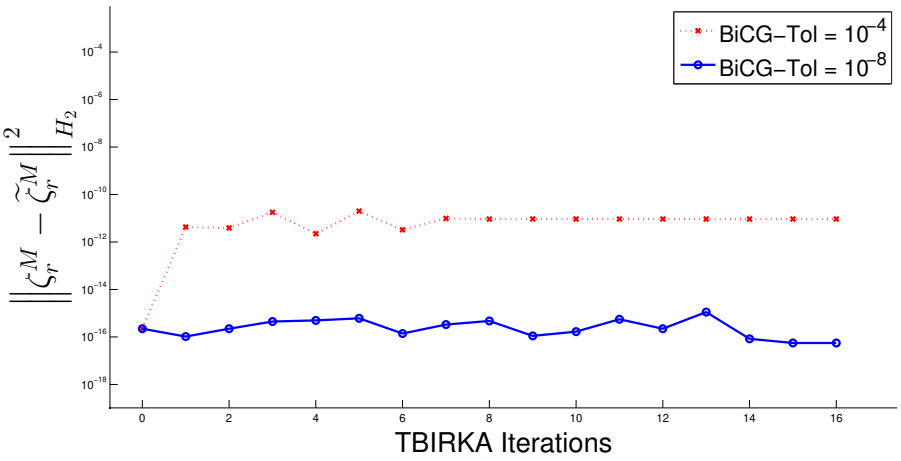

Figure 4.3: Accuracy of the reduced system plotted at each TBIRKA iteration for the two different stopping tolerances in BiCG; Flow model of size 110 (satisfying the constraints in the subsystem approach but not in the complete system approach). Here, the x-axis is in the linear scale and the y-axis is in the log scale.

less than one at the end of all the other TBIRKA steps as well (see Table 4.4). The condition number for our problem, as defined in (4.36), is 1.8275×10^{-1}. This shows that the flow model is well-conditioned.

Here, although the constraints for the complete system approach are violated and the constraints for the subsystem approach are satisfied, we still achieve a stable TBIRKA. This because, as earlier, both of these approaches provide sufficiency conditions for stability, and only one needs to be satisfied.

The accuracy result for this is given in Figure 4.3. Again, we have accuracy of the reduced system $\left(\left\|\zeta_r^M - \widetilde{\zeta}_r^M\right\|_{H_2}\right)$ on the y-axis in the log scale and the TBIRKA iterations on the x-axis in the linear scale. From this figure, it is again evident that we get a more accurate reduced model as we solve the linear systems more accurately (solid line is below the dotted one at all the TBIRKA steps).

CHAPTER 5

STABILITY ANALYSIS IN PMOR

While performing exact MOR of *non-parametric linear* dynamical systems, the projection matrices at the one iterative step of the MOR algorithm (IRKA [30]) have the form as follows:

$$\begin{aligned} V &= \left[(\sigma_1 E - A)^{-1} B \mathbb{r}_1, \ \ldots, \ (\sigma_r E - A)^{-1} B \mathbb{r}_r \right] \quad \text{and} \\ W &= \left[(\sigma_1 E - A)^{-T} C^T \mathbb{l}_1, \ \ldots, \ (\sigma_r E - A)^{-T} C^T \mathbb{l}_r \right], \end{aligned} \tag{5.1}$$

where σ_i with $i = 1, \ \ldots, \ r$ denote the shifts where interpolation is performed; r is the size to which we want to reduce the input dynamical system; and $\mathbb{l}_i$ and $\mathbb{r}_i$ are the left and the right tangent directions, respectively, which are used during interpolation.

As mentioned earlier, we focus on IPMOR [9] for MOR of *parametric linear* dynamical systems. Using inexact linear solves in building the projection matrices in IPMOR, we get[1]

$$\begin{aligned} \widetilde{V} &= \left[\widetilde{V}_1 (p^1) \ \ \cdots \ \ \widetilde{V}_K (p^1) \ \ \cdots\cdots \ \ \widetilde{V}_1 (p^L) \ \ \cdots \ \ \widetilde{V}_K (p^L) \right] \quad \text{and} \\ \widetilde{W} &= \left[\widetilde{W}_1 (p^1) \ \ \cdots \ \ \widetilde{W}_K (p^1) \ \ \cdots\cdots \ \ \widetilde{W}_1 (p^L) \ \ \cdots \ \ \widetilde{W}_K (p^L) \right] \end{aligned} \tag{5.2}$$

[1]IPMOR is not an iterative algorithm. Hence, the reduced system is obtained in one step. This is unlike all the earlier algorithms.

with

$$\left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)\widetilde{V}_i\left(p^j\right) = B\left(p^j\right)\mathbb{r}_{ij} + R_{B_i}\left(p^j\right) \quad \text{and} \tag{5.3}$$

$$\left(\sigma_i E\left(p^j\right) - A\left(p^j\right)\right)^T\widetilde{W}_i\left(p^j\right) = C^T\left(p^j\right)\mathbb{l}_{ij} + R_{C_i}\left(p^j\right), \tag{5.4}$$

where as earlier, σ_i with $i = 1, \ldots, K$ denote the shifts; $p^j \in \mathbb{R}^v$ with $j = 1, \ldots, L$ denote the set of parameters; $\mathbb{l}_{ij}$ and $\mathbb{r}_{ij}$ denote the left and the right tangent directions, respectively; and $R_{B_i}\left(p^j\right)$ and $R_{C_i}\left(p^j\right)$ denote the residuals. Thus, the reduced model obtained by the the inexact IPMOR algorithm can be connected to the original full model by using the Petrov-Galerkin projection as

$$\begin{aligned}&\widetilde{E}_r\left(p\right) = \widetilde{W}^T E\left(p\right)\widetilde{V}, \quad \widetilde{A}_r\left(p\right) = \widetilde{W}^T A\left(p\right)\widetilde{V}, \quad \widetilde{B}_r\left(p\right) = \widetilde{W}^T B\left(p\right), \quad \text{and}\\ &\widetilde{C}_r\left(p\right) = C\left(p\right)\widetilde{V},\end{aligned} \tag{5.5}$$

where $p \in \{p^1, \ldots, p^L\}$.

For backward stability, next we need to apply the exact IPMOR algorithm on a perturbed full model. Let F be the perturbation in $A\left(p\right)$ only[2], i.e., the perturbed system matrices be denoted as follows:

$$\widetilde{E}\left(p\right) = E\left(p\right), \quad \widetilde{A}\left(p\right) = A\left(p\right) + F, \quad \widetilde{B}\left(p\right) = B\left(p\right), \quad \widetilde{C}\left(p\right) = C\left(p\right).$$

Here, applying exact IPMOR algorithm on this perturbed system leads to a system of linear equations given by

$$\left(\sigma_i E\left(p^j\right) - \left(A\left(p^j\right) + F\right)\right)\widetilde{V}_i\left(p^j\right) = B\left(p^j\right)\mathbb{r}_{ij}, \quad \text{and} \tag{5.6}$$

$$\left(\sigma_i E\left(p^j\right) - \left(A\left(p^j\right) + F\right)\right)^T\widetilde{W}_i\left(p^j\right) = C^T\left(p^j\right)\mathbb{l}_{ij}. \tag{5.7}$$

To be able to satisfy the first condition of stability (see (2.10)), where the inexact IPMOR applied onto the original dynamical system is equal to the exact IPMOR applied onto the perturbed dynamical system, we have to use the same $\widetilde{V}$ and $\widetilde{W}$ here. Thus, the reduced model obtained by exact IPMOR algorithm can be connected to the perturbed full model by using the Petrov-Galerkin projection as

$$\widehat{E}_r\left(p\right) = \widetilde{E}_r\left(p\right), \quad \widehat{A}_r\left(p\right) = \widetilde{A}_r\left(p\right) + \widetilde{W}^T F\widetilde{V}, \quad \widehat{B}_r\left(p\right) = \widetilde{B}_r\left(p\right), \qquad \widehat{C}_r\left(p\right) = \widetilde{C}_r\left(p\right),$$

[2]The derivations that we do next, can be easily done if we consider perturbations in $E\left(p\right)$, $B\left(p\right)$, and $C\left(p\right)$ individually as well [10, 21].

where $\widetilde{V}$ and $\widetilde{W}$, as earlier, are given by (5.2); $p \in \{p^1, \ldots, p^L\}$; and $\widetilde{E}_r(p)$, $\widetilde{A}_r(p)$, $\widetilde{B}_r(p)$, and $\widetilde{C}_r(p)$ are given by (5.5). Thus, from the above equation, if $\widetilde{W}^T F \widetilde{V} = 0$, then $\widehat{A}_r(p) = \widetilde{A}_r(p)$ and we satisfy the first condition of stability.

From (5.3) – (5.6) and (5.4) – (5.7), we get

$$\begin{aligned} R_{B_i}\left(p^j\right) = F\widetilde{V}_i\left(p^j\right) \quad \text{and} \quad R_{C_i}\left(p^j\right)^T = \widetilde{W}_i\left(p^j\right)^T F \quad \forall\ j = 1, \ldots, L \quad \text{or} \\ R_B = F\widetilde{V} \quad \text{and} \quad R_C^T = \widetilde{W}^T F, \end{aligned} \tag{5.8}$$

where

$$\begin{aligned} R_B &= \left[R_{B_1}(p^1) \quad \cdots \quad R_{B_K}(p^1) \quad \cdots\cdots \quad R_{B_1}\left(p^L\right) \quad \cdots \quad R_{B_K}\left(p^L\right)\right], \\ R_C &= \left[R_{C_1}(p^1) \quad \cdots \quad R_{C_K}(p^1) \quad \cdots\cdots \quad R_{C_1}\left(p^L\right) \quad \cdots \quad R_{C_K}\left(p^L\right)\right]. \end{aligned} \tag{5.9}$$

In (5.8), if we multiply $\widetilde{W}^T$ from left in the first equation and $\widetilde{V}$ from right in the second equation, then we get

$$\widetilde{W}^T R_B = \widetilde{W}^T F \widetilde{V} \quad \text{and} \quad R_C^T \widetilde{V} = \widetilde{W}^T F \widetilde{V}. \tag{5.10}$$

Similar to the non-parametric case, we need $\widetilde{W}^T F \widetilde{V} = 0$ for $\widehat{A}_r(p) = \widetilde{A}_r(p)$. This can be achieved if

$$\text{either} \quad \widetilde{W}^T R_B = 0 \quad \text{or} \quad R_C^T \widetilde{V} = 0. \tag{5.11}$$

The first equation of (5.11) is satisfied if

$$\begin{bmatrix} \widetilde{W}_1(p^1)^T \\ \vdots \\ \widetilde{W}_K(p^1)^T \\ \vdots \\ \vdots \\ \widetilde{W}_1\left(p^L\right)^T \\ \vdots \\ \widetilde{W}_K\left(p^L\right)^T \end{bmatrix} \left[R_{B_1}(p^1) \quad \cdots \quad R_{B_K}(p^1) \quad \cdots\cdots \quad R_{B_1}\left(p^L\right) \quad \cdots \quad R_{B_K}\left(p^L\right)\right] = 0. \tag{5.12}$$

The second equation of (5.11) is satisfied if

$$\begin{bmatrix} R_{C_1}\left(p^1\right)^T \\ \vdots \\ R_{C_K}\left(p^1\right)^T \\ \vdots \\ \vdots \\ R_{C_1}\left(p^L\right)^T \\ \vdots \\ R_{C_K}\left(p^L\right)^T \end{bmatrix} \begin{bmatrix} \widetilde{V}_1\left(p^1\right) & \cdots & \widetilde{V}_K\left(p^1\right) & \cdots\cdots & \widetilde{V}_1\left(p^L\right) & \cdots & \widetilde{V}_K\left(p^L\right) \end{bmatrix} = 0. \quad (5.13)$$

Theorem 12. *Let the inexact linear solves in IPMOR, that is* (5.3) *and* (5.4)*, be solved while satisfying* (5.12) *and* (5.13)*. Then, IPMOR satisfies the first condition of backward stability with respect to these inexact linear solves, i.e.,* (2.10)*.*

Next, satisfying the second condition of stability in IPMOR (Theorem 13 below) leads to constraints of the same form as that for IRKA (Theorem 4.3 from [10]) with the difference that system matrices here are dependent on the parameters, which was not the case earlier.

Theorem 13. *Let F be the constant perturbation introduced in $A(p)$. If $\left\|\mathbb{K}^{-1}(s;\ p)\right\|_{H_\infty} < 1$ and $\|F\| < 1$, then*

$$\left\|H(s;\ p) - \widetilde{H}(s;\ p)\right\|_{H_2} \leqslant \frac{\left\|C(p)\,\mathbb{K}^{-1}(s;\ p)\right\|_{H_2} \left\|\mathbb{K}^{-1}(s;\ p)\,B(p)\right\|_{H_\infty}}{1 - \left\|\mathbb{K}^{-1}(s;\ p)\right\|_{H_\infty} \|F\|} \|F\|, \quad (5.14)$$

where $\mathbb{K}(s;\ p) = (sI_n - A(p))$, $s \in \{s_1,\ \ldots,\ s_K\}$, and $p \in \{p_1,\ \ldots,\ p_L\}$. That is, IPMOR satisfies the second condition of backward stability with respect to the inexact linear solves (5.3) *and* (5.4)*, i.e.,* (2.11)*.*

Proof. Similar to Theorem 4.3 in [10]. □

Corollary 3. *Assuming the hypotheses of Theorem 12 and Theorem 13 are satisfied, then IPMOR algorithm is backward stable with respect to the inexact linear solves.*

Next, we discuss how the conditions for stability (given by Corollary 3 or Theorem 12-13) can be easily satisfied. Satisfying (5.14) is not hard and we support this in the results section later.

Satisfying (5.12) and (5.13) is similar to satisfying (4.1) in the previous chapter (Stability of TBIRKA)[3]. As mentioned earlier, this requires more work than in IRKA (reducing first-order *non-parametric linear* dynamical systems in [10]) as well as in BIRKA (reducing first-order *non-parametric bilinear* dynamical systems in Chapter 3 and [21]).

Also, as discussed in the previous chapter, this can be easily achieved by using the framework proposed while reducing second-order *non-parametric linear* dynamical systems in [46] (AIRGA algorithm). Hence, in the next section (Section 5.1) we propose this new framework and also highlight differences between ours and the framework of [46].

5.1 Satisfying Extra-Orthogonality for Stability

We divide satisfying extra orthogonalities problem into three parts; diagonal matrix part, upper triangular matrix part, and lower triangular matrix part. For making the diagonal part zero of (5.12) and (5.13), we need

$$\begin{aligned} \widetilde{W}_i\left(p^j\right) &\perp R_{B_i}\left(p^j\right) \qquad \text{for } i = 1,\ \ldots,\ K \text{ and } j = 1,\ \ldots, L; \quad \text{and} \\ \widetilde{V}_i\left(p^j\right) &\perp R_{C_i}\left(p^j\right) \qquad \text{for } i = 1,\ \ldots,\ K \text{ and } j = 1,\ \ldots, L, \end{aligned} \tag{5.15}$$

Next, we look at the upper triangular and lower triangular parts of the matrices in (5.12) and (5.13). Since, the arguments for (5.12) exactly carry to (5.13), we focus on the former only to avoid repetition. We need the orthogonalities below for ensuring that the upper and lower triangular part of the matrix is zero in (5.12) and (5.12), respectively.

[3]Now onwards, we would only talk about (5.12) and (5.13) since all results easily carryover to TBIRKA.

$$
\begin{cases}
\left[\widetilde{W}_1\left(p^1\right)\right] \perp R_{B_2}\left(p^1\right) \\
\left[\widetilde{W}_1\left(p^1\right) \quad \widetilde{W}_2\left(p^1\right)\right] \perp R_{B_3}\left(p^1\right) \\
\qquad \vdots \\
\left[\widetilde{W}_1\left(p^1\right) \quad \cdots \quad \widetilde{W}_{K-1}\left(p^1\right)\right] \perp R_{B_K}\left(p^1\right) \\
\qquad \vdots \\
\qquad \vdots \\
\begin{cases}
\left[\widetilde{W}_1\left(p^1\right) \quad \cdots \quad \widetilde{W}_K\left(p^1\right) \quad \cdots\cdots \quad \widetilde{W}_1\left(p^{L-1}\right) \quad \cdots \quad \widetilde{W}_K\left(p^{L-1}\right)\right] \perp R_{B_1}\left(p^L\right) \\
\left[\widetilde{W}_1\left(p^1\right) \quad \cdots \quad \widetilde{W}_K\left(p^1\right) \quad \cdots\cdots \quad \widetilde{W}_1\left(p^{L-1}\right) \quad \cdots \quad \widetilde{W}_K\left(p^{L-1}\right) \quad \widetilde{W}_1\left(p^L\right)\right] \perp R_{B_2}\left(p^L\right) \\
\qquad \vdots \\
\left[\widetilde{W}_1\left(p^1\right) \quad \cdots \quad \widetilde{W}_K\left(p^1\right) \quad \cdots\cdots \quad \widetilde{W}_1\left(p^L\right) \quad \cdots \quad \widetilde{W}_{K-1}\left(p^L\right)\right] \perp R_{B_K}\left(p^L\right).
\end{cases}
\end{cases}
\tag{5.16}
$$

$$
\begin{cases}
\widetilde{W}_2\left(p^1\right) \quad \perp \left[R_{B_1}\left(p^1\right)\right] \\
\widetilde{W}_3\left(p^1\right) \quad \perp \left[R_{B_1}\left(p^1\right) \quad R_{B_2}\left(p^1\right)\right] \\
\qquad \vdots \\
\widetilde{W}_K\left(p^1\right) \quad \perp \left[R_{B_1}\left(p^1\right) \quad \cdots \quad R_{B_{K-1}}\left(p^1\right)\right] \\
\qquad \vdots \\
\qquad \vdots \\
\begin{cases}
\widetilde{W}_1\left(p^L\right) \quad \perp \left[R_{B_1}\left(p^1\right) \quad \cdots \quad R_{B_K}\left(p^1\right) \quad \cdots\cdots \quad R_{B_1}\left(p^{L-1}\right) \cdots \quad R_{B_K}\left(p^{L-1}\right)\right] \\
\widetilde{W}_2\left(p^L\right) \quad \perp \left[R_{B_1}\left(p^1\right) \quad \cdots \quad R_{B_K}\left(p^1\right) \quad \cdots\cdots \quad R_{B_1}\left(p^{L-1}\right) \cdots \quad R_{B_K}\left(p^{L-1}\right) \quad R_{B_1}\left(p^L\right)\right] \\
\qquad \vdots \\
\widetilde{W}_K\left(p^L\right) \quad \perp \left[R_{B_1}\left(p^1\right) \quad \cdots \quad R_{B_K}\left(p^1\right) \quad \cdots\cdots \quad R_{B_1}\left(p^L\right) \cdots \quad R_{B_{K-1}}\left(p^L\right)\right].
\end{cases}
\end{cases}
\tag{5.17}
$$

Next, we describe the choice of the linear solver that would satisfy the above three types of orthogonalities. *First*, (5.15) can be easily satisfied if we use a Petrov-Galerkin based iterative solver (as earlier, BiCG). This is the same as done for IRKA and BIRKA stability analyses (in [10] and Chapter 3 - [21], respectively). In AIRGA stability analysis (in [46]), authors use a Ritz-Galerkin based iterative solver (CG).

Second, to satisfy 5.16 and 5.17, we adapt BiCG (discussed below), and to do this with no code changes as well as cheaply, we propose a *new* variant of Recycling BiCG [3, 4] (in the following subsection; Section 5.2). This was not needed for IRKA or BIRKA stability analysis. However, AIRGA stability analysis required doing similar derivations with two notable differences from the work here;

1. they proposed CG as the underlying iterative solver, and hence used off-the-shelf Recycling CG [39], and

2. the number of orthogonalities to be satisfied there were much lesser (due to absence of parameters) leading to ease in making the iterative solver converge cheaply.

Next, we adapt the two components of the BiCG algorithm to satisfy the above discussed orthogonalities.

5.1.1 Adapted Bi-Lanczos

Assume we are trying to solve the dual linear systems of the form

$$\mathcal{A}\mathbf{x} = \mathbf{b} \qquad \text{and} \qquad \mathcal{A}^T\mathbf{y} = \mathbf{c}, \tag{5.18}$$

where $\mathcal{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b},\ \mathbf{c} \in \mathbb{R}^{n\times 1}$. We refer $\mathcal{A}\mathbf{x} = \mathbf{b}$ as the primary system and $\mathcal{A}^T\mathbf{y} = \mathbf{c}$ as the dual system. Let $\mathbf{x}_0$ be the initial solution vector with $\mathbf{r}_0 = \mathbf{b} - \mathcal{A}\mathbf{x}_0$ as the corresponding residual for the primary system, and $\mathbf{y}_0$ be the initial solution vector with $\widetilde{\mathbf{r}}_0 = \mathbf{c} - \mathcal{A}^T\mathbf{y}_0$ as the corresponding residual for the dual system. The bi-Lanczos algorithm computes good bases of the generated Krylov subspaces involving

$\mathcal{A}$ and $\mathbf{r}_0$ for the primary system, and $\mathcal{A}^T$ and $\widetilde{\mathbf{r}}_0$ for the dual system as follows:

$$\begin{aligned}
&\mathbf{v}_{q+1} \in \mathcal{K}^q\left(\mathcal{A},\ \mathbf{r}_0\right) = span\{\mathbf{r}_0,\ \mathcal{A}\mathbf{r}_0,\ \ldots,\ \mathcal{A}^q\mathbf{r}_0\} \quad s.t. \quad \mathbf{v}_{q+1} \perp \left[\mathbf{w}_1\ \ldots\ \mathbf{w}_q\right] \quad \text{and} \\
&\mathbf{w}_{q+1} \in \mathcal{K}^q\left(\mathcal{A}^T,\ \widetilde{\mathbf{r}}_0\right) = span\{\widetilde{\mathbf{r}}_0,\ \mathcal{A}^T\widetilde{\mathbf{r}}_0,\ \ldots,\ \mathcal{A}^{Tq}\widetilde{\mathbf{r}}_0\} \quad s.t. \quad \mathbf{w}_{q+1} \perp \left[\mathbf{v}_1\ \ldots\ \mathbf{v}_q\right],
\end{aligned} \tag{5.19}$$

where $\mathbf{v}_{q+1}$ and $\mathbf{w}_{q+1}$ are the Lanczos vectors of the respective systems, of (5.18), at the $(q+1)^{th}$ iterative step. Also, $\mathbf{v}_1 = \dfrac{\mathbf{r}_0}{\|\mathbf{r}_0\|}$ and $\mathbf{w}_1 = \dfrac{\widetilde{\mathbf{r}}_0}{\|\widetilde{\mathbf{r}}_0\|}$[4].

Assuming, we are carrying some residual $\widetilde{\mathbf{r}}$, which we need to make orthogonal to the final solution of the primary system of (5.18), and similarly we are carrying some residual $\mathbf{r}$, which we need to make orthogonal to the final solution of the dual system of (5.18). Then, the adapted bi-Lanczos algorithm above would consist of the following steps:

$$\begin{aligned}
&\mathbf{v}_{q+1} \in \mathcal{K}^q\left(\mathcal{A},\ \mathbf{r}_0\right) \quad s.t. \quad \mathbf{v}_{q+1} \perp \left[\mathbf{w}_1\ \ldots\ \mathbf{w}_q\ \widetilde{\mathbf{r}}\right] \quad \text{and} \\
&\mathbf{w}_{q+1} \in \mathcal{K}^q\left(\mathcal{A}^T,\ \widetilde{\mathbf{r}}_0\right) \quad s.t. \quad \mathbf{w}_{q+1} \perp \left[\mathbf{v}_1\ \ldots\ \mathbf{v}_q\ \mathbf{r}\right].
\end{aligned} \tag{5.20}$$

Next, we generalize the above adapted bi-Lanczos algorithm for a series of dual linear systems

$$\mathcal{A}_i\mathbf{x}_i = \mathbf{b}_i \qquad \text{and} \qquad \mathcal{A}_i^T\mathbf{y}_i = \mathbf{c}_i, \tag{5.21}$$

for $i = 1,\ \ldots, \mathcal{L}$. Note that we solve these linear systems inexactly and make the solutions of one set of linear systems orthogonal to the residuals obtained from solving **all** the previous sets of linear systems. This mimics the behavior of satisfying the orthogonality conditions given by (5.17).

[4]Here, the first equation of (5.19) is implemented using

$$\mathbf{v}_{q+1} = A\mathbf{v}_q - c_1\mathbf{v}_1 - c_2\mathbf{v}_2 - \ldots - c_{q-1}\mathbf{v}_{q-1} - c_q\mathbf{v}_q.$$

Finally, the orthogonality conditions of the first equation of (5.19) gives us $c_1,\ c_2,\ \ldots,\ c_q$. Similarly, the second equation of (5.19) is implemented using

$$\mathbf{w}_{q+1} = A^T\mathbf{w}_q - \widetilde{c}_1\mathbf{w}_1 - \widetilde{c}_2\mathbf{w}_2 - \ldots - \widetilde{c}_{q-1}\mathbf{w}_{q-1} - \widetilde{c}_q\mathbf{w}_q.$$

Here, $\widetilde{c}_1,\ \widetilde{c}_2,\ \ldots,\ \widetilde{c}_q$ are similarly obtained. For a complete derivation of this, please see [49].

Using an iterative method to solve the first set of equations of (5.21), i.e., for $i = 1$, implies we eventually solve the following equations:

$$\mathcal{A}_1 \dot{\mathbf{x}}_1 = \mathbf{b}_1 + \mathbf{r}_1 \qquad \text{and} \qquad \mathcal{A}_1^T \dot{\mathbf{y}}_1 = \mathbf{c}_1 + \widetilde{\mathbf{r}}_1, \tag{5.22}$$

where $\dot{\mathbf{x}}_1$ and $\dot{\mathbf{y}}_1$ are the final solution vectors; and $\mathbf{r}_1$ and $\widetilde{\mathbf{r}}_1$ are the final residuals of the primary and the dual systems, respectively.

Next, while solving the second set of equations of (5.21), i.e., for $i = 2$,

$$\mathcal{A}_2 \mathbf{x}_2 = \mathbf{b}_2 \qquad \text{and} \qquad \mathcal{A}_2^T \mathbf{y}_2 = \mathbf{c}_2, \tag{5.23}$$

we need a good basis of the two generated Krylov subspaces such that the solution of the primary system of (5.23) is orthogonal to $\widetilde{\mathbf{r}}_1$, and the solution of the dual system of (5.23) is orthogonal to $\mathbf{r}_1$. Hence, here, the adapted bi-Lanczos algorithm would consist of the following procedure:

$$\begin{aligned} &(\mathbf{v_2})_{q+1} \in \mathcal{K}^q\left(\mathcal{A}_2,\ (\mathbf{r}_2)_0\right) \quad s.t. \quad (\mathbf{v}_2)_{q+1} \perp \Big[(\mathbf{w}_2)_1\ \ldots\ (\mathbf{w}_2)_q\ \widetilde{\mathbf{r}}_1\Big] \quad \text{and} \\ &(\mathbf{w_2})_{q+1} \in \mathcal{K}^q\left(\mathcal{A}_2^T,\ (\widetilde{\mathbf{r}}_2)_0\right) \quad s.t. \quad (\mathbf{w}_2)_{q+1} \perp \Big[(\mathbf{v}_2)_1\ \ldots\ (\mathbf{v}_2)_q\ \mathbf{r}_1\Big], \end{aligned} \tag{5.24}$$

where $(\mathbf{v}_2)_{q+1}$ and $(\mathbf{w}_2)_{q+1}$ are the Lanczos vectors of the respective systems of (5.23) at the $(q+1)^{th}$ iterative step such that $(\mathbf{v}_2)_1 = \dfrac{(\mathbf{r}_2)_0}{\|\,(\mathbf{r}_2)_0\,\|}$ and $(\mathbf{w}_2)_1 = \dfrac{(\widetilde{\mathbf{r}}_2)_0}{\|\,(\widetilde{\mathbf{r}}_2)_0\,\|}$; and $(\mathbf{r}_2)_0$ and $(\widetilde{\mathbf{r}}_2)_0$ are the initial residuals of the respective systems of (5.23).

We repeat this procedures for $i = 3,\ \ldots,\ \mathcal{L}$ in (5.21). To summarize, while solving the last set of equations of (5.21), i.e., for $i = \mathcal{L}$

$$\mathcal{A}_{\mathcal{L}} \mathbf{x}_{\mathcal{L}} = \mathbf{b}_{\mathcal{L}} \qquad \text{and} \qquad \mathcal{A}_{\mathcal{L}}^T \mathbf{y}_{\mathcal{L}} = \mathbf{c}_{\mathcal{L}}, \tag{5.25}$$

we need a good basis of the two generated Krylov subspaces such that the solution of the primary system of (5.25) is orthogonal to the residuals $\widetilde{\mathbf{r}}_1$, $\widetilde{\mathbf{r}}_2$, ..., and $\widetilde{\mathbf{r}}_{\mathcal{L}-1}$ coming from the previously solved dual linear systems; and the solution of the dual system of (5.25) is orthogonal to the residuals $\mathbf{r}_1$, $\mathbf{r}_2$, ..., and $\mathbf{r}_{\mathcal{L}-1}$ coming from the previously solved primary systems. Hence, here, the adapted bi-Lanczos algorithm would consist of the following procedure:

$$\begin{aligned} &(\mathbf{v}_{\mathcal{L}})_{q+1} \in \mathcal{K}^q\left(\mathcal{A}_{\mathcal{L}},\ (\mathbf{r}_{\mathcal{L}})_0\right) \quad s.t. \quad (\mathbf{v}_{\mathcal{L}})_{q+1} \perp \Big[(\mathbf{w}_{\mathcal{L}})_1\ \ldots\ (\mathbf{w}_{\mathcal{L}})_q\ \widetilde{\mathbf{r}}_1\ \widetilde{\mathbf{r}}_2\ \ldots\ \widetilde{\mathbf{r}}_{\mathcal{L}-1}\Big] \quad \text{and} \\ &(\mathbf{w}_{\mathcal{L}})_{q+1} \in \mathcal{K}^q\left(\mathcal{A}_{\mathcal{L}}^T,\ (\widetilde{\mathbf{r}}_{\mathcal{L}})_0\right) \quad s.t. \quad (\mathbf{w}_{\mathcal{L}})_{q+1} \perp \Big[(\mathbf{v}_{\mathcal{L}})_1\ \ldots\ (\mathbf{v}_{\mathcal{L}})_q\ \mathbf{r}_1\ \mathbf{r}_2\ \ldots\ \mathbf{r}_{\mathcal{L}-1}\Big], \end{aligned} \tag{5.26}$$

where $(\mathbf{v}_{\mathcal{L}})_{q+1}$ and $(\mathbf{w}_{\mathcal{L}})_{q+1}$ are the Lanczos vectors of the respective systems of (5.25) at the $(q+1)^{th}$ iterative step such that $(\mathbf{v}_{\mathcal{L}})_1 = \dfrac{(\mathbf{r}_{\mathcal{L}})_0}{\|\,(\mathbf{r}_{\mathcal{L}})_0\,\|}$ and $(\mathbf{w}_{\mathcal{L}})_1 = \dfrac{(\widetilde{\mathbf{r}}_{\mathcal{L}})_0}{\|\,(\widetilde{\mathbf{r}}_{\mathcal{L}})_0\,\|}$; and $(\mathbf{r}_{\mathcal{L}})_0$ and $(\widetilde{\mathbf{r}}_{\mathcal{L}})_0$ are the initial residuals of the respective systems of (5.25).

5.1.2 Adapted Petrov-Galerkin

If we are trying to solve the linear systems given in (5.18) by the BiCG method, then (5.19) gives good bases of the two generated Krylov subspaces. The solution updates here are given as

$$\mathbf{x}_q = \mathbf{x}_0 + \mathcal{V}_q z_q \qquad \text{and} \qquad \mathbf{y}_q = \mathbf{y}_0 + \mathcal{W}_q \widetilde{z}_q, \tag{5.27}$$

where $\mathcal{V}_q = [\mathbf{v}_1 \ \ldots \mathbf{v}_q]$ and $\mathcal{W}_q = [\mathbf{w}_1 \ \ldots \ \mathbf{w}_q]$ are the basis defined by the bi-Lanczos process in (5.19). In BiCG, these z_q and $\widetilde{z}_q$ are defined by a Petrov-Galerkin projection

$$\mathbf{r}_q \perp \mathcal{W}_q \quad \text{and} \quad \widetilde{\mathbf{r}}_q \perp \mathcal{V}_q,$$
$$\text{where} \quad \mathbf{r}_q = \mathbf{r}_0 - \mathcal{A}\mathcal{V}_q z_q \text{ and } \widetilde{\mathbf{r}}_q = \widetilde{\mathbf{r}}_0 - \mathcal{A}^T \mathcal{W}_q \widetilde{z}_q.$$

Assume we are carrying some solution vector $\widehat{\mathbf{y}}$, which we need to make orthogonal to the final residual of the primary linear system in (5.18). In the similar manner, assume we are also carrying some solution vector $\widehat{\mathbf{x}}$, which we need to make orthogonal to the final residual of the dual linear system in (5.18). Then, the adapted Petrov-Galerkin process would consist of the following procedure:

$$\mathbf{r}_q \perp [\mathcal{W}_q \ \widehat{\mathbf{y}}] \quad \text{and} \quad \widetilde{\mathbf{r}}_q \perp [\mathcal{V}_q \ \widehat{\mathbf{x}}]\,. \tag{5.28}$$

Again, we generalize the above Petrov-Galerkin process for a series of dual linear systems, defined in (5.21). As earlier, we solve these linear systems inexactly and make the residuals of the one set of linear systems orthogonal to the solution vector obtained from solving **all** the previous sets of linear systems. This mimics the behavior of satisfying the orthogonality condition given by (5.16).

Assume that after solving the first set of equations of (5.21), i.e., for $i = 1$ (5.22), we obtain $\dot{\mathbf{x}}_1$ and $\dot{\mathbf{y}}_1$ as the final solution vectors of the two respective systems.

Next, for $i = 2$ in (5.21), lets look at the second set of linear systems defined by (5.23). Here, we need to make the final residual of the primary system (i.e., $\mathbf{r}_2$) orthogonal to $\mathbf{y}_1$ and the final residual of the dual system (i.e., $\widetilde{\mathbf{r}}_2$) orthogonal to $\mathbf{x}_1$. Hence, here, the adapted Petrov-Galerkin process is defined as

$$(\mathbf{r}_2)_q \perp \Big[(\mathbf{w}_2)_1 \ \ldots \ (\mathbf{w}_2)_q \ \dot{\mathbf{y}}_1\Big] \quad \text{and} \quad (\widetilde{\mathbf{r}}_2)_q \perp \Big[(\mathbf{v}_2)_1 \ \ldots \ (\mathbf{w}_2)_q \ \dot{\mathbf{x}}_1\Big], \tag{5.29}$$

where $(\mathbf{r}_2)_q$ and $(\widetilde{\mathbf{r}}_2)_q$ are the residuals of the respective systems of (5.23) at the q^{th} iterative step. Note that $\mathbf{r}_2$ and $\widetilde{\mathbf{r}}_2$ are the final residuals of the respective systems of (5.23), respectively (at convergence of BiCG).

Similarly, we repeat this process for $i = 3, \ \ldots, \ \mathcal{L}$ in (5.21). Thus, for $i = \mathcal{L}$, lets look at the last set of linear systems defined by (5.25). Here, we need to make the final residual of the primary system (i.e., $\mathbf{r}_{\mathcal{L}}$) orthogonal to the solutions $\dot{\mathbf{y}}_1$, $\dot{\mathbf{y}}_2$, ..., and $\dot{\mathbf{y}}_{\mathcal{L}-1}$ coming from the previously solved dual systems; and the final residual of the dual system (i.e., $\widetilde{\mathbf{r}}_{\mathcal{L}}$) orthogonal to the solutions $\dot{\mathbf{x}}_1$, $\dot{\mathbf{x}}_2$, ..., and $\dot{\mathbf{x}}_{\mathcal{L}-1}$ coming from the previously solved primary systems. Hence, here, the adapted Petrov-Galerkin process is as

$$\begin{aligned} &(\mathbf{r}_{\mathcal{L}})_q \perp \Big[(\mathbf{w}_{\mathcal{L}})_1 \ \ldots \ (\mathbf{w}_{\mathcal{L}})_q \ \dot{\mathbf{y}}_1 \ \dot{\mathbf{y}}_2 \ \ldots \ \dot{\mathbf{y}}_{\mathcal{L}-1}\Big] \quad \text{and} \\ &(\widetilde{\mathbf{r}}_{\mathcal{L}})_q \perp \Big[(\mathbf{v}_{\mathcal{L}})_1 \ \ldots \ (\mathbf{v}_{\mathcal{L}})_q \ \dot{\mathbf{x}}_1 \ \dot{\mathbf{x}}_2 \ \ldots \ \dot{\mathbf{x}}_{\mathcal{L}-1}\Big], \end{aligned} \tag{5.30}$$

where $(\mathbf{r}_{\mathcal{L}})_q$ and $(\widetilde{\mathbf{r}}_{\mathcal{L}})_q$ are the residuals of the respective systems of (5.25) at the q^{th} iterative step. Note that $\mathbf{r}_{\mathcal{L}}$ and $\widetilde{\mathbf{r}}_{\mathcal{L}}$ are the final residuals of the respective systems in (5.25) (at convergence of BiCG). Next, we look at changes to RBiCG.

5.2 Changes to RBiCG and Building Recycle Spaces

Developing the BiCG algorithm that is based upon the adapted bi-Lanczos process and the adapted Petrov-Galerkin projection, discussed in the previous section, is feasible, however, this requires too many code change with the additional drawback that new BiCG's efficient version may not exist. Also, as the number of linear systems to

be solved increases, the number of orthogonalization to be done also increase linearly. As discussed earlier, using Recycling BiCG (RBiCG) [3, 4] helps alleviate both these problems. Hence, in the following subsections, we describe the changes to be done with RBiCG code and also show how the choice of the recycle space helps in easily as well as efficiently achieving the desired orthogonalities.

5.2.1 Changes for implementing the Adapted Bi-Lanczos Process

For solving the linear systems in (5.18) using the BiCG method, we need good bases of the generated Krylov subspaces. These bases are computed by the bi-Lanczos algorithm as given by the relation in (5.19). One can also write these bi-Lanczos relations as a pair of 3-term recurrences in the matrix form as [43]

$$\begin{aligned} \mathcal{A}\mathcal{V}_q &= \mathcal{V}_{q+1}\underline{\mathrm{T}}_q \qquad \text{and} \qquad \mathcal{A}^T\mathcal{W}_q = \mathcal{W}_{q+1}\underline{\widetilde{\mathrm{T}}}_q, \\ &\text{such that} \qquad \mathcal{V}_q \perp_b \mathcal{W}_q, \end{aligned}$$

where $\underline{\mathrm{T}}_q$ and $\underline{\widetilde{\mathrm{T}}}_q$ are tridiagonal matrices of size $(q+1) \times q$. Also, $\mathcal{V}_q = [\mathbf{v}_1 \ \ldots \ \mathbf{v}_q]$, $\mathcal{W}_q = [\mathbf{w}_1 \ \ldots \ \mathbf{w}_q]$, and $\perp_b$ denotes the bi-orthogonality.

In [2], authors have proposed a recycling variant of BiCG, called Recycling BiCG (RBiCG), where the solutions for the two systems of (5.18) are searched in augmented Krylov subspaces $[\mathcal{V}_q \ U]$ and $\left[\mathcal{W}_q \ \widetilde{U}\right]$, respectively, with $U = [u_1, \ \ldots, \ u_\tau]$ and $\widetilde{U} = [\widetilde{u}_1, \ \ldots, \ \widetilde{u}_\tau]$ being two input recycle spaces such that following bi-orthogonality is achieved:

$$[\mathcal{V}_q \ C] \perp_b \left[\mathcal{W}_q \ \widetilde{C}\right], \tag{5.31}$$

where $C = AU$ and $\widetilde{C} = A^T\widetilde{U}$.

For us, we still need to search in the augmented Krylov subspaces $[\mathcal{V}_q \ U]$ and $\left[\mathcal{W}_q \ \widetilde{U}\right]$, however, we need to implement the following bi-orthogonality here

$$[\mathcal{V}_q \ U] \perp_b \left[\mathcal{W}_q \ \widetilde{U}\right]. \tag{5.32}$$

This leads to new augmented bi-Lanczos relations given by

$$\begin{aligned} \left(I - U\widehat{U}^T\right)\mathcal{A}\mathcal{V}_q &= \mathcal{V}_{q+1}\underline{\mathrm{T}}_q \quad \text{and} \\ \left(I - \widetilde{U}\breve{U}^T\right)\mathcal{A}^T\mathcal{W}_q &= \mathcal{W}_{q+1}\underline{\widetilde{\mathrm{T}}}_q, \end{aligned} \tag{5.33}$$

where $\widehat{U} = \left[\frac{\widetilde{u}_1}{u_1^T\widetilde{u}_1}, \ \ldots, \ \frac{\widetilde{u}_\tau}{u_\tau^T\widetilde{u}_\tau}\right]$ and $\breve{U} = \left[\frac{u_1}{\widetilde{u}_1^T u_1}, \ \ldots, \ \frac{u_\tau}{\widetilde{u}_\tau^T u_\tau}\right]$. If $D = \widetilde{U}^T U$ or $D = \begin{bmatrix} \widetilde{u}_1^T u_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \widetilde{u}_\tau^T u_\tau \end{bmatrix}$, which is usually enforced from the input U and $\widetilde{U}$ (see [2]), then $\widehat{U} = \widetilde{U}D^{-T}$ and $\breve{U} = UD^{-1}$.

5.2.2 Changes for implementing the Adapted Petrov-Galerkin Process

In the standard BiCG algorithm, for solving the linear systems in (5.18), if $\mathcal{V}_q$ and $\mathcal{W}_q$ define the good basis of the respective columns of Krylov subspaces, then the solution at q^{th} iteration of BiCG is given by (5.27), i.e.,

$$\mathbf{x}_q = \mathbf{x}_0 + \mathcal{V}_q z_q \qquad \text{and} \qquad \mathbf{y}_q = \mathbf{y}_0 + \mathcal{W}_q \widetilde{z}_q,$$

where, as earlier, $\mathbf{x}_0$ and $\mathbf{y}_0$, are the initial guess for the respective linear systems of (5.18).

In the standard RBiCG [2], as discussed above, the solution iterates have the form

$$\mathbf{x}_q = \mathbf{x}_0 + \mathcal{V}_q z_q + U\widehat{z}_q \qquad \text{and} \qquad \mathbf{y}_q = \mathbf{y}_0 + \mathcal{W}_q \widetilde{z}_q + \widetilde{U}\breve{z}_q, \tag{5.34}$$

where z_q, $\widehat{z}_q$, $\widetilde{z}_q$, and $\breve{z}_q$ are determined by a Petrov-Galerkin projection. This is same for us. If at the q^{th} iterative step, $\mathbf{r}_q$ and $\widetilde{\mathbf{r}}_q$ are the residuals of the primary and dual systems, respectively, then the chosen Petrov-Galerkin projection give

$$\begin{aligned} \mathbf{r}_q &= \mathbf{r}_0 - \mathcal{A}\mathcal{V}_q z_q - \mathcal{A}U\widehat{z}_q \perp \left[\mathcal{W}_q \ \widetilde{C}\right] \quad \text{and} \\ \widetilde{\mathbf{r}}_q &= \widetilde{\mathbf{r}}_0 - \mathcal{A}^T\mathcal{W}_q\widetilde{z}_q - \mathcal{A}^T\widetilde{U}\breve{z}_q \perp \left[\mathcal{V}_q \ C\right], \end{aligned}$$

where, as earlier, $\mathbf{r}_0$ and $\widetilde{\mathbf{r}}_0$ are the initial residuals of the respective linear systems of (5.18).

For us, these projections do not suffice (because we need to eliminate use of C and $\widetilde{C}$, completely). Hence, we use the Petrov-Galerkin projection as follows:

$$\mathbf{r}_q = \mathbf{r}_0 - \mathcal{A}\mathcal{V}_q z_q - \mathcal{A}U\widehat{z}_q \perp \left[\mathcal{W}_q\ \widetilde{U}\right] \quad \text{and} \tag{5.35}$$

$$\widetilde{\mathbf{r}}_q = \widetilde{\mathbf{r}}_0 - \mathcal{A}^T\mathcal{W}_q\widetilde{z}_q - \mathcal{A}^T\widetilde{U}\widetilde{z}_q \perp \left[\mathcal{V}_q\ U\right]. \tag{5.36}$$

Using (5.35) for the primary system analysis, we get

$$\begin{bmatrix} \mathcal{W}_q^T \\ \widetilde{U}^T \end{bmatrix} \left[\mathbf{r}_0 - \mathcal{A}\mathcal{V}_q z_q - \mathcal{A}U\widehat{z}_q\right] = 0. \tag{5.37}$$

Let $\xi = \left\| \left(I - U\widehat{U}^T\right)\mathbf{r}_0 \right\|_2$ and $\mathbf{v}_1 = \frac{\left(I-U\widehat{U}^T\right)\mathbf{r}_0}{\xi}$, where $\mathbf{v}_1$ is the first Lanczos vector with respect to primary system. Then, $\mathbf{r}_0$ can be re-written as

$$\mathbf{r}_0 = U\widehat{U}^T\mathbf{r}_0 + r_0 - U\widehat{U}^T\mathbf{r}_0$$

$$\mathbf{r}_0 = U\widehat{U}^T\mathbf{r}_0 + \xi\mathbf{v}_1$$

$$\mathbf{r}_0 = U\widehat{U}^T\mathbf{r}_0 + \xi\mathcal{V}_{q+1}e_1,$$

where e_1 is the first column of the identity matrix. The above equation can be written in the matrix form as

$$\mathbf{r}_0 = \begin{bmatrix} U & \mathcal{V}_{q+1} \end{bmatrix} \begin{bmatrix} \widehat{U}^T\mathbf{r}_0 \\ \xi e_1 \end{bmatrix}. \tag{5.38}$$

Let $\mathcal{A}U = U\mathcal{M}$, where M is the unknown matrix, then

$$\widetilde{U}^T\mathcal{A}U = \widetilde{U}^TU\mathcal{M} \qquad \text{or}$$

$$\mathcal{M} = \left(\widetilde{U}^TU\right)^{-1}\widetilde{U}^T\mathcal{A}U = D^{-1}\widetilde{U}^T\mathcal{A}U.$$

with D defined in the previous section.

Now, let us look at the term

$$\mathcal{A}\mathcal{V}_q z_q + \mathcal{A}U\widehat{z}_q = \begin{bmatrix} U\widehat{U}^T\mathcal{A}\mathcal{V}_q + \mathcal{V}_{q+1}\underline{\mathrm{T}}_q & \mathcal{A}U \end{bmatrix} \begin{bmatrix} z_q \\ \widehat{z}_q \end{bmatrix} \cdots \text{(Using (5.33))}$$

$$= \begin{bmatrix} U\widehat{U}^T\mathcal{A}\mathcal{V}_q + \mathcal{V}_{q+1}\underline{\mathrm{T}}_q & U\mathcal{M} \end{bmatrix} \begin{bmatrix} z_q \\ \widehat{z}_q \end{bmatrix} \cdots \text{(Using the assumption } AU = UM\text{)}$$

$$= \begin{bmatrix} U & \mathcal{V}_{q+1} \end{bmatrix} \begin{bmatrix} \widehat{U}^T\mathcal{A}\mathcal{V}_q & \mathcal{M} \\ \underline{\mathrm{T}}_q & 0 \end{bmatrix} \begin{bmatrix} z_q \\ \widehat{z}_q \end{bmatrix}. \tag{5.39}$$

Putting the values from (5.38) and (5.39) to (5.37), we get

$$\begin{bmatrix} \mathcal{W}_q^T \\ \widetilde{U}^T \end{bmatrix} \left[\begin{bmatrix} U & \mathcal{V}_{q+1} \end{bmatrix} \begin{bmatrix} \widehat{U}^T \mathbf{r}_0 \\ \xi e_1 \end{bmatrix} - \begin{bmatrix} U & \mathcal{V}_{q+1} \end{bmatrix} \begin{bmatrix} \widehat{U}^T \mathcal{A} \mathcal{V}_q & \mathcal{M} \\ \underline{\mathrm{T}}_q & 0 \end{bmatrix} \begin{bmatrix} z_q \\ \widehat{z}_q \end{bmatrix} \right] = 0 \quad \text{or}$$

$$\begin{bmatrix} \mathcal{W}_q^T \\ \widetilde{U}^T \end{bmatrix} \begin{bmatrix} U & \mathcal{V}_{q+1} \end{bmatrix} \left[\begin{bmatrix} \widehat{U}^T \mathbf{r}_0 \\ \xi e_1 \end{bmatrix} - \begin{bmatrix} \widehat{U}^T \mathcal{A} \mathcal{V}_q & \mathcal{M} \\ \underline{\mathrm{T}}_q & 0 \end{bmatrix} \begin{bmatrix} z_q \\ \widehat{z}_q \end{bmatrix} \right] = 0. \tag{5.40}$$

Using the bi-orthogonality condition (5.32) in the above expression, we get[5]

$$\begin{bmatrix} \widehat{U}^T \mathbf{r}_0 \\ \xi e_1 \end{bmatrix} - \begin{bmatrix} \widehat{U}^T \mathcal{A} \mathcal{V}_q & \mathcal{M} \\ T_q & 0 \end{bmatrix} \begin{bmatrix} z_q \\ \widehat{z}_q \end{bmatrix} = 0. \tag{5.41}$$

Thus, we can find the values of z_q and $\widehat{z}_q$ from the above expression as

$$\begin{aligned} z_q &= \xi T_q^{-1} e_1, \\ \widehat{z}_q &= \mathcal{M}^{-1} \left(\widehat{U}^T \mathbf{r}_0 - \widehat{U}^T \mathcal{A} \mathcal{V}_q z_q \right). \end{aligned}$$

Substituting the value of z_q and $\widehat{z}_q$ in the first equation of (5.34), we get the updated solution of the primary system as

$$\mathbf{x}_q = \mathbf{x}_0 + U \mathcal{M}^{-1} \widehat{U}^T \mathbf{r}_0 + \left(I - U \mathcal{M}^{-1} \widehat{U}^T \mathcal{A} \right) \mathcal{V}_q \xi T_q^{-1} e_1. \tag{5.42}$$

Similarly, using (5.36) for the dual system analysis, let $\widetilde{\xi} = \left\| \left(I - \widetilde{U} \breve{U}^T \right) \widetilde{\mathbf{r}}_0 \right\|_2$ and $\mathbf{w}_1 = \frac{\left(I - \widetilde{U} \breve{U}^T \right) \widetilde{\mathbf{r}}_0}{\widetilde{\xi}}$, where $\mathbf{w}_1$ is the first Lanczos vector with respect to the dual system. Also, let $\mathcal{A}^T \widetilde{U} = \widetilde{U} \widetilde{\mathcal{M}}$, where $\widetilde{\mathcal{M}} = D^{-T} U^T \mathcal{A}^T \widetilde{U}$, then we get the updated solution of the dual system as

$$\mathbf{y}_q = \mathbf{y}_0 + \widetilde{U} \widetilde{\mathcal{M}}^{-1} \breve{U}^T \widetilde{\mathbf{r}}_0 + \left(I - \widetilde{U} \widetilde{\mathcal{M}}^{-1} \breve{U}^T \mathcal{A}^T \right) \mathcal{W}_q \widetilde{\xi} \widetilde{T}_q^{-1} e_1. \tag{5.43}$$

Note that the solution updates of this new RBiCG (5.42)-(5.43) require that M and $\widetilde{M}$ be invertible. This is usually not a problem as seen by numerical experiments.

[5]Note that the dimension of e_1 in (5.41) is one less than that of e_1 in (5.40), although both denote the first canonical vector.

5.2.3 Building Recycle Subspaces

Assume that we want to solve the linear systems in (5.3) and (5.4). Also, assume that the recycle spaces are of the form of $span\{U_{ij}\}$ and $span\{\widetilde{U}_{ij}\}$ for the primary and dual systems, respectively, where columns of $U_{ij},\ \widetilde{U}_{ij} \in \mathbb{R}^{n\times(2\times(i-1)\times j)}$ are linearly independent. In our case, the recycle spaces are defined as below.

$*$ For the first linear system (recall (5.3) and (5.4), where $i = 1$ and $j = 1$):

$$U_{11} = [\ ] \text{ and } \widetilde{U}_{11} = [\ ].$$

$*$ For the second linear system (recall (5.3) and (5.4), where $i = 2$ and $j = 1$):

$$U_{21} = [R_{C_1}(p^1)\ \ \widetilde{W}_1(p^1)] \quad \text{and} \quad \widetilde{U}_{21} = [R_{B_1}(p^1)\ \ \widetilde{V}_1(p^1)].$$

$*$ Similarly, for the last linear system (recall (5.3) and (5.4), where $i = K$ and $j = L$):

$$\begin{aligned} U_{KL} =&[R_{C_1}(p^1)\ \ldots\ R_{C_K}(p^1)\ \ldots\ R_{C_1}(p^L)\ \ldots\ R_{C_{K-1}}(p^L)\\ &\widetilde{W}_1(p^1)\ \ldots\ \widetilde{W}_K(p^1)\ \ldots\ \widetilde{W}_1(p^L)\ \ldots\ \widetilde{W}_{K-1}(p^L)] \quad \text{and}\\ \widetilde{U}_{KL} =&[R_{B_1}(p^1)\ \ldots\ R_{B_K}(p^1)\ \ldots\ R_{B_1}(p^L)\ \ldots\ R_{B_{K-1}}(p^L)\\ &\widetilde{V}_1(p^1)\ \ldots\ \widetilde{V}_K(p^1)\ \ldots\ \widetilde{V}_1(p^L)\ \ldots\ \widetilde{V}_{K-1}(p^L)]. \end{aligned}$$

As mentioned earlier, in some cases, this choice of the recycle spaces can actually accelerate the convergence of the linear system under consideration. In cases, when these recycle spaces deteriorate the convergence, this behavior can be bounded.

As done for the previous two chapters, before discussing results, we first compute the expression for accuracy of the reduced system in the next section (Section 5.3.)

5.3 Computing Accuracy

From Theorem 3, we know that if IPMOR is backward stable, then the accuracy of the reduced system is

$$\frac{\left\|H_r(s;\ p) - \widetilde{H}_r(s;\ p)\right\|_{H_2}}{\|H_r(s;\ p)\|_{H_2}} = \mathcal{O}\left(\mathcal{K}\left(H(s;\ p)\right)\cdot\|F\|\right), \tag{5.44}$$

where, as earlier, $H(s;\ p) = C(p)(sE(p) - A(p))^{-1}B(p)$, $\widetilde{H}(s;\ p) = C(p)(sE(p) - (A(p) + F))^{-1}B(p)$, $\mathcal{K}\left(H(s;\ p)\right)$ is the condition number of $H(s;\ p)$, and F is the

perturbation in the input dynamical system. Thus, the accuracy of the reduced system is dependent on the condition number of the problem and the perturbation in the system.

As discussed in the earlier two chapters, the condition number of the input dynamical system with respect to computing $\left\|H_r(s;\ p) - \widetilde{H}_r(s;\ p)\right\|_{H_2}$, can be approximated well by the condition number of the input dynamical system with respect to computing $\left\|H(s;\ p) - \widetilde{H}(s;\ p)\right\|_{H_2}$. Thus, from Theorem 13, we have

$$\left\|H(s;\ p) - \widetilde{H}(s;\ p)\right\|_{H_2} \leqslant \frac{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\|_{H_2}\,\|\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_\infty}}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}\,\|F\|}\,\|F\|. \tag{5.45}$$

Let $\|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty} < 1$ and $\|F\| < 1$ (already assume in Corollary 3 and Theorem 13), then we have $\|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}\,\|F\| < 1$, and hence,

$$\frac{1}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}\,\|F\|} < \frac{1}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}}.$$

Substituting the above in (5.45) we get

$$\left\|H(s;\ p) - \widetilde{H}(s;\ p)\right\|_{H_2} \leqslant \frac{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\|_{H_2}\,\|\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_\infty}}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}}\,\|F\| \quad \text{or}$$

$$\frac{\left\|H(s;\ p) - \widetilde{H}(s;\ p)\right\|_{H_2}}{\|H(s;\ p)\|_{H_2}} \leqslant \frac{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\|_{H_2}\,\|\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_\infty}}{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_2}} \cdot \frac{1}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}} \cdot \|F\| \quad \text{or}$$

$$\frac{\left\|H(s;\ p) - \widetilde{H}(s;\ p)\right\|_{H_2}}{\|H(s;\ p)\|_{H_2}} \leqslant \frac{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\|_{H_2}\,\|\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_\infty}}{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_2}} \cdot \frac{\|A(s;\ p)\|}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}} \cdot \frac{\|F\|}{\|A(s;\ p)\|}.$$

Thus, from the above inequality we get that the condition number of the input dynamical system is

$$\mathcal{K}(H(s;\ p)) = \frac{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\|_{H_2}\,\|\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_\infty}}{\|C(p)\,\mathbb{K}^{-1}(s;\ p)\,B(p)\|_{H_2}} \cdot \frac{\|A(s;\ p)\|}{1 - \|\mathbb{K}^{-1}(s;\ p)\|_{H_\infty}}. \tag{5.46}$$

Usually, the condition numbers of the problems under consideration are fairly small[2]. Also, we assume invertibility of $\mathbb{K}(s;\ p)$ or $(sI_n - A(p))$ in our analysis. This comes from the transfer function definitions [9].

Hence, next, we relate perturbation F with the residual R_B and R_C defined in (5.9). Rewriting (5.8), we have

$$R_B = F\widetilde{V} \qquad \text{and} \qquad R_C^T = \widetilde{W}^T F.$$

From the backward stability assumption given in Theorem 12, collectively we can write the above equations as

$$F = R_B \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T + \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1} R_C^T, \tag{5.47}$$

assuming $\widetilde{W}^T\widetilde{V}$ is invertible[6]. The following theorem gives a upper bound on perturbation F.

Theorem 14. *Let for L different parameters and K different shifts; $\widetilde{V}$ and $\widetilde{W}$, be given by* (5.2)*; R_B and R_C be given as in* (5.9)*; and F be given as in* (5.47)*. Assume, $\left(\widetilde{W}^T\widetilde{V}\right)$ is invertible. Then, the perturbation F satisfies*

$$\|F\| \leqslant \sqrt{K \times L}\left\{ \max_i \|R_B(:,i)\| \left\| \left(\widetilde{W}^T\widetilde{V}\right)^{-1}\widetilde{W}^T \right\| + \max_i \|R_C(:,i)\| \left\| \widetilde{V}\left(\widetilde{W}^T\widetilde{V}\right)^{-1} \right\| \right\}.$$

Proof. Similar to Theorem 6 in Chapter 3 or Theorem 5 in [21]. □

Thus, by using condition number expression from (5.46) and perturbation upper bound from Theorem 14 into accuracy expression (5.44), we get that for a well-conditioned input dynamical system, as we solve the linear systems more accurately in the backward stable IPMOR, we get a more accurate reduced system. We support this with experiments as well in the next section (Section 5.4).

Thus, as in the previous two chapters, this is the main outcome of using a backward stable model reduction algorithm, which gives the end user flexibility in deciding how accurately to solve the linear systems to get a sufficiently accurate reduced system.

[6]This is usually easily achieved as has been shown for non-parametric linear case [10, Section 4.1] and non-parametric bilinear case [21, Section 4.1 and Chapter 3, Section 3.2.1].

5.4 Numerical Experiments

We perform preliminary experiment on the FOM model [32, 16, 40]. This model consists of a parametric linear dynamical system of size $n = 1006$, as

$$E\dot{x}(t) = A(p)\,x(t) + Bu(t),$$
$$y(t) = Cx(t),$$

where $E = I_{n\times n}$, $A(p) = diag(A_1(p),\ A_2,\ A_3, A_4)$, and
$B^T = C = \left[\underbrace{10\ \ldots\ 10}_{6}\ \ \underbrace{1\ \ldots\ 1}_{1000}\right]$ with

$$A_1(p) = \begin{bmatrix} -1 & p \\ -p & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -1 & 200 \\ -200 & 1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} -1 & 400 \\ -400 & 1 \end{bmatrix}, \quad \text{and}$$
$$A_4 = -diag(1,\ \ldots, 1000).$$

For our experiments, we take two interpolation points, i.e., $K = 2$, such that $\sigma_i \in [0.99,\ 1]$ and two parameters, i.e., $L = 2$, such that $p^j \in [99.99,\ 100]$. All these values are chosen based upon similar values in [16]. Thus, the size of the reduced system obtained is 4.

This leads to solving linear systems of size 1006×1006. As earlier, here also, for solving the linear systems while computing V and W by a direct method (exact IPMOR), we use a backslash in Matlab. As discussed earlier, we use the RBiCG variant along with the recycle spaces as proposed in the previous section. While using RBiCG, we use two different stopping tolerances (10^{-2} and 10^{-3}). These choice of stopping tolerances ensures that RBiCG takes same number of steps for convergence so that we can compare the two cases. Ideally, we should obtain a more accurate reduced model when using the smaller RBiCG tolerance.

We implement our codes in MATLAB (2015a), and test on a machine with the following configuration: Intel Xeon(R) CPU E5-1620 V3 @ 3.50 GHz., frequency 1200 MHz., 8 CPU, 64 GB RAM.

First, let us look at the assumptions for backward stability of IPMOR (see Corollary 3 and Theorem 13). $\mathbb{K}(s,\ p)$ is invertible here. We also have $\|\mathbb{K}^{-1}(s,\ p)\|_{H_\infty}$ less

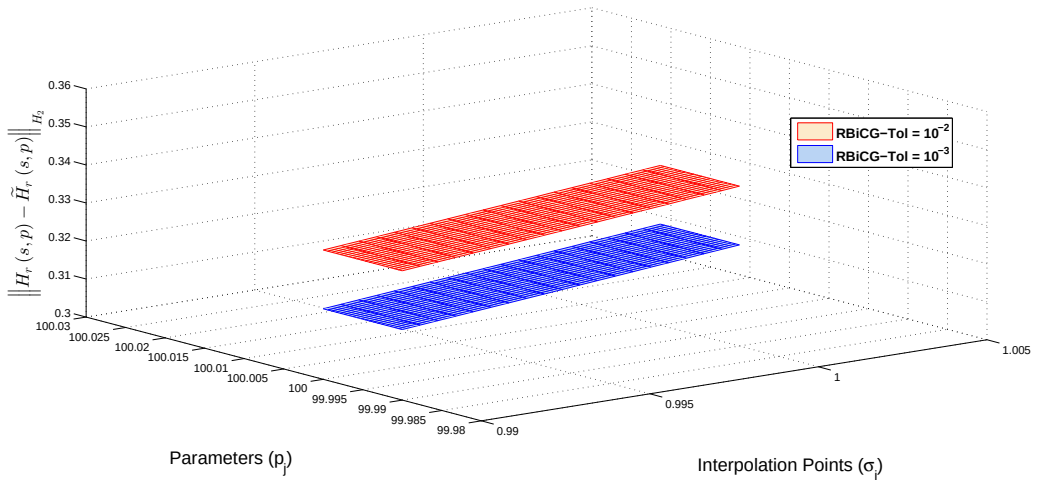

Figure 5.1: Accuracy of the reduced system plotted with respect to interpolation points and parameters for the two different stopping tolerances in RBiCG; FOM model of size 1006.

than one (i.e., 5.0251×10^{-1}). Finally, $\|F\|$, for the RBiCG stopping tolerances of 10^{-2} and 10^{-3} is 7.0474×10^{-1} and 1.9037×10^{-1}, respectively, both of which are also less than one. Note that this is a single step algorithm so we do not iterate. The condition number for our problem, as defined in (5.46), is 2.5024×10^{-1}. This shows that the FOM model is well-conditioned.

The accuracy result for this is given in Figure 5.1. Here, we have accuracy of the reduced system $\left(\left\| H_r\left(s,\ p\right) - \widetilde{H}_r\left(s,\ p\right) \right\|_{H_2} \right)$ on the z-axis, interpolation points (i.e., σ_i) on x-axis, and parameters (i.e., p^j) on the y-axis. From Figure 5.1, it is again evident that we get a more accurate reduced model as we solve the linear systems more accurately (blue surface is below to the red surface).

Finally, we support our final claim that the way required orthogonalities are achieved (see Section 5.2.3), it often does not deteriorate the convergence of our linear solves. Thus, we solve all linear systems arising in the IPMOR algorithm with BiCG as well RBiCG. Table 5.1 gives the iteration count of the two solvers. It is evident that using the recycle spaces as formulated in Section 5.2.3, accelerates the convergence of the solver (savings of about 70% to 73%).

It is important to note that by using a recycle space we are doing extra work in terms of more number of inner products. Hence, the savings in time would be less

Linear System in IPMOR	Stopping tolerance 10^{-2}		Stopping tolerance 10^{-3}	
	BiCG Iteration Count	RBiCG Iteration Count	BiCG Iteration Count	RBiCG Iteration Count
1	62	62	94	94
2	62	1	94	8
3	62	1	94	1
4	62	3	94	8
Total	**248**	**67**	**376**	**111**

Table 5.1: Convergence analysis of BiCG and RBiCG at two different stopping tolerances; FOM Model.

than the savings in iteration count.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this dissertation, we perform stability analysis of MOR algorithms for reducing first order *non-parametric/ parametric* and *linear/ bilinear* dynamical systems with respect to inexact linear solves. Since MOR algorithms for reducing *non-parametric linear* dynamical systems have been studied earlier, we focus on MOR algorithms for reducing *non-parametric bilinear* (summarized in the following two headings) and *parametric linear* (summarized in the last heading) dynamical system. Study of MOR algorithms for reducing *parametric bilinear* dynamical systems forms part of our future work.

Stability Analysis of BIRKA

BIRKA [12], (which is a standard algorithm for reducing non-parametric bilinear dynamical systems), provides a locally H_2-optimal reduced model. The most expensive part of BIRKA is finding solutions of large linear systems of equations. Iterative algorithms are a method of choice for such systems but they find solutions only up to a certain tolerance. Hence, we show that BIRKA is backward stable with respect to these inexact linear solves under some mild assumptions. We also analyze the accuracy of the inexact reduced system obtained from a backward stable BIRKA. We support all our results with numerical experiments.

The first assumption is that $\widehat{Q}$ in (3.14) is invertible. In Section 3.2.1, we have given

a better characterization of this invertibility assumption (in terms of the underlying Lyapunov equation). However, this requires further analysis and forms ours first future work.

The second and the third assumptions involve bounding $\left\|\widehat{Q}^{-1}\right\|$ and $\left\|\widehat{\widehat{F}}\right\|$ (given after (3.19)) by one. Although for both our experimental models we have shown that these assumptions are easily satisfied, they may not always hold. $\widehat{Q}$ is dependent on the input dynamical system and $\widehat{\widehat{F}}$ on the stopping tolerance of our underlying linear solver. Hence, the second future work here involves identifying the categories of bilinear dynamical systems and the range of linear solver stopping tolerances when these would be true.

While computing the accuracy, we have given an expression for the condition number of the bilinear system with respect to computing the H_2-norm of the error between the perturbed model and the original model. This condition number is an approximation to the condition we want to compute. That is, the condition number of the bilinear system with respect to computing the H_2-norm of the error between the inexact reduced model and the original model. This forms our third future work.

Stability Analysis of Other Efficient Algorithms for Bilinear MOR

Here, we extend the stability analysis done for BIRKA in the previous chapter to other cheaper and efficient algorithms for bilinear MOR. This includes TBIRKA [24, 26], balanced truncation based [13], Gramian based [50], moment-matching based [8], and implicit Volterra series based [1]. Specifically we work with TBIRKA, as it forms the base of all such efficient algorithms. In TBIRKA, fulfilling the first condition for stability leads to constraints on the iterative linear solver, which are similar to those obtained during BIRKA's stability analysis. The second condition for TBIRKA can be satisfied by two different approaches, complete system approach and subsystem approach. The complete system approach works for both SISO and MIMO cases, but the subsystem approach works only for SISO case. However, both have an advantage because they are sufficiency conditions and depending upon the input dynamical system, one may be more easier to satisfy then other one.

The stability analysis as done for BIRKA and TBIRKA here, all give us sufficiency

conditions for a stable underlying MOR algorithm. Hence, the first future work here is to derive the necessary conditions for the same. In recent years, there have been a lot of efforts in performing data-driven MOR algorithm (specially using Loewner framework [7]). The second future work here is to apply this stability analysis to such classes of algorithms as well. Finally, the third future work is to extend this stability analysis to the cases when instead of a dynamical system, the underlying differential equation is studied [31].

Stability Analysis in PMOR

We study stability of a interpolatory MOR algorithm for parametric linear dynamical systems with respect to inexact linear solves, that is, the IPMOR algorithm [9]. This analysis is easily extendible to other MOR algorithms for such systems. Besides deriving the two conditions for stability, accuracy expression, and subsequent experimentation, our novel contribution here has been achieving extra-orthogonalities for stability without any code changes to the underlying iterative solver as well as doing all this cheaply. As a outcome of this research, we also develop a new variant of the Recycling BiCG algorithm [3, 4].

Here, the first future work involves more rigorous experimentation with a larger problem. Also, since IPMOR is a single iteration algorithm, it would be good to extend this stability analysis to other more optimal PMOR algorithms, e.g., piecewise H_2-optimal PMOR algorithm [9] that iterate to the ideal interpolation points. This forms the second future work. Finally, as a third future work, we also plan to generalize this theory to other MOR algorithms for parametric dynamical systems (second/ third-orders and bilinear/ nonlinear terms).

BIBLIOGRAPHY

[1] Ahmad, M. I., Baur, U., and Benner, P. (2017). Implicit volterra series interpolation for model reduction of bilinear systems. *Journal of Computational and Applied Mathematics*, 316:15–28.

[2] Ahuja, K. (2009). Recycling Bi-Lanczos Algorithms: BiCG, CGS, and BiCGSTAB. Master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

[3] Ahuja, K. (2011). *Recycling Krylov Subspaces and Preconditioners*. PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

[4] Ahuja, K., de Sturler, E., Gugercin, S., and Chang, E. R. (2012). Recycling BiCG with an application to model reduction. *SIAM Journal on Scientific Computing*, 34(4):A1925–A1949.

[5] Antoulas, A. C. (2005). *Approximation of Large-Scale Dynamical Systems*. SIAM Advances in Design and Control, Philadelphia, PA, USA.

[6] Antoulas, A. C., Beattie, C. A., and Gugercin, S. (2010). Interpolatory model reduction of large-scale dynamical systems. In Mohammadpour, J. and Grigoriadis, K. M., editors, *Efficient Modeling and Control of Large-Scale Systems*, pages 3–58. Springer, Berlin/ Heidelberg, Germany.

[7] Antoulas, A. C., Gosea, I. V., and Ionita, A. C. (2016). Model reduction of bilinear systems in the Loewner framework. *SIAM Journal on Scientific Computing*, 38(5):B889–B916.

[8] Bai, Z. and Skoogh, D. (2006). A projection method for model reduction of bilinear dynamical systems. *Linear Algebra and its Applications*, 415(2–3):406–425.

[9] Baur, U., Beattie, C., Benner, P., and Gugercin, S. (2011). Interpolatory projection methods for parameterized model reduction. *SIAM Journal on Scientific Computing*, 33(5):2489–2518.

[10] Beattie, C., Gugercin, S., and Wyatt, S. (2012). Inexact solves in interpolatory model reduction. *Linear Algebra and its Applications*, 436(8):2916–2943.

[11] Beattie, C. A. and Gugercin, S. (2006). Inexact solves in Krylov-based model reduction. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 3405–3411.

[12] Benner, P. and Breiten, T. (2012). Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. *SIAM Journal on Matrix Analysis and Applications*, 33(3):859–885.

[13] Benner, P. and Damm, T. (2011). Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM Journal on Control and Optimization*, 49(2):686–711.

[14] Benner, P., Grundel, S., and Hornung, N. (2015a). Parametric model order reduction with a small $\mathcal{H}_2$-error using radial basis functions. *Advances in Computational Mathematics*, 41(5):1231–1253.

[15] Benner, P., Gugercin, S., and Willcox, K. (2015b). A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531.

[16] Benner, P., Mehrmann, V., and C. Sorensen, D. (2005). *Dimension Reduction of Large-Scale Systems*, volume 45 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin/ Heidelberg, Germany.

[17] Breiten, T. (2013). *Interpolation Methods for Model Reduction of Large-Scale Dynamical Systems*. PhD thesis, Otto-Von-Guericke University of Magdeburg, Magdeburg, Germany.

[18] Breiten, T. and Damm, T. (2010). Krylov subspace methods for model order reduction of bilinear control systems. *Systems & Control Letters*, 59(8):443–450.

[19] Bunse-Gerstner, A., Kubalińska, D., Vossen, G., and Wilczek, D. (2010). h_2-norm optimal model reduction for large scale discrete dynamical MIMO systems. *Journal of Computational and Applied Mathematics*, 233(5):1202–1216.

[20] Carracedo Rodriguez, A., Gugercin, S., and Borggaard, J. (2018). Interpolatory model reduction of parameterized bilinear dynamical systems. *Advances in Computational Mathematics*, 44(6):1887–1916.

[21] Choudhary, R. and Ahuja, K. (2018). Stability analysis of bilinear iterative rational Krylov algorithm. *Linear Algebra and its Applications*, 538:56–88.

[22] Chow, E. and Saad, Y. (1998). Approximate inverse preconditioners via sparse-sparse iterations. *SIAM Journal on Scientific Computing*, 19(3):995–1023.

[23] Demmel, J. W. (1997). *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, USA.

[24] Flagg, G. M. (2012). *Interpolation Methods for the Model Reduction of Bilinear Systems*. PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

[25] Flagg, G. M., Beattie, C., and Gugercin, S. (2012). Convergence of the iterative rational Krylov algorithm. *Systems & Control Letters*, 61(6):688 – 691.

[26] Flagg, G. M. and Gugercin, S. (2015). Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(2):549–579.

[27] Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*, volume 3. Johns Hopkins University Press.

[28] Grimme, E. J. (1997). *Krylov Projection Methods for Model Reduction*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

[29] Gugercin, S. (2003). *Projection Methods for Model Reduction of Large-Scale Dynamical Systems*. PhD thesis, Rice University, Houston, TX, USA.

[30] Gugercin, S., Antoulas, A. C., and Beattie, C. (2008). $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. *SIAM Journal on Matrix Analysis and Applications*, 30(2):609–638.

[31] Han, T. and Han, Y. (2013). Numerical solution for super large scale systems. *IEEE Access*, 1:537–544.

[32] Ionita, A. and Antoulas, A. (2014). Data-driven parametrized model reduction in the Loewner framework. *SIAM Journal on Scientific Computing*, 36(3):A984–A1007.

[33] Jeffrey, A. (2003). *Handbook of Mathematical Formulas and Integrals*. Academic Press, Cambridge, MA, USA.

[34] Lancaster, P. and Farahat, H. K. (1972). Norms on direct sums and tensor products. *Mathematics of Computation*, 26(118):401–414.

[35] Laub, A. J. (2004). *Matrix Analysis for Scientists And Engineers*. SIAM, Philadelphia, PA, USA.

[36] Lou, D. and Weiland, S. (2018). Parametric model order reduction for large-scale and complex thermal systems. In *2018 European Control Conference (ECC)*, pages 2593–2598.

[37] Lu, D., Su, Y., and Bai, Z. (2016). Stability analysis of the two-level orthogonal Arnoldi procedure. *SIAM Journal on Matrix Analysis and Applications*, 37(1):195–214.

[38] Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, PA, USA.

[39] Parks, M. L., De Sturler, E., Mackey, G., Johnson, D. D., and Maiti, S. (2006). Recycling Krylov subspaces for sequences of linear systems. *SIAM Journal on Scientific Computing*, 28(5):1651–1674.

[40] Penzl, T. (2006). Algorithms for model reduction of large dynamical systems. *Linear Algebra and its Application*, 415(2–3):322–343.

[41] Philips, J. R. (2003). Projection-based approaches for model reduction of weakly nonlinear, time-varying systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 22(02):171–187.

[42] Rugh, W. J. (1981). *Nonlinear System Theory: The Volterra/Wiener Approach*. Johns Hopkins Series in Information Sciences and Systems. Johns Hopkins University Press, Baltimore, Maryland, USA.

[43] Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, USA, second edition.

[44] Sandberg, H. (2012). Parameterized model order reduction using extended balanced truncation. In *51st IEEE Conference on Decision and Control (CDC)*, pages 4291–4296. IEEE.

[45] Schilders, W. H., Van der Vorst, H. A., and Rommes, J. (2008). *Model Order Reduction: Theory, Research Aspects and Applications*, volume 13. Springer, Berlin/ Heidelberg, Germany.

[46] Singh, N. P. and Ahuja, K. (2018). Stability analysis of inexact solves in moment matching based model reduction. *Preprint arXiv:1803.09283*.

[47] Stykel, T. (2002). *Analysis and Numerical Solution of Generalized Lyapunov Equations*. PhD thesis, Institut für Mathematik, Technische Universität, Berlin, Berlin, Germany.

[48] Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*. SIAM, Philadelphia, PA, USA.

[49] Van der Vorst, H. A. (2003). *Iterative Krylov Methods for Large Linear Systems*, volume 13. Cambridge University Press.

[50] Xu, K. L., Jiang, Y. L., and Yang, Z. X. (2017). H_2 optimal model order reduction by two-sided technique on Grassmann manifold via the cross-Gramian of bilinear systems. *International Journal of Control*, 90(3):616–626.

[51] Zhang, L. and Lam, J. (2002). On H_2 model reduction of bilinear systems. *Automatica*, 38(2):205–216.