

Issues in Interpolatory Model Reduction: Inexact Solves,  
Second-order Systems and DAEs

Sarah Wyatt

Dissertation submitted to the Faculty of the  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Mathematics

Serkan Gugercin, Chair

Christoper Beattie

Jeff Borggaard

Eric de Sturler

May 1, 2012

Blacksburg, Virginia

Keywords: Krylov reduction, Inexact Solves, Second-order Systems, DAEs

Copyright 2012, Sarah Wyatt

# Issues in Interpolatory Model Reduction: Inexact Solves, Second-order Systems and DAEs

Sarah Wyatt

(ABSTRACT)

Dynamical systems are mathematical models characterized by a set of differential or difference equations. Model reduction aims to replace the original system with a reduced system of significantly smaller dimension that still describes the important dynamics of the large-scale model. Interpolatory model reduction methods define a reduced model that interpolates the full model at selected interpolation points. The reduced model may be obtained through a Krylov reduction process or by using the Iterative Rational Krylov Algorithm (IRKA), which iterates this Krylov reduction process to obtain an optimal  $\mathcal{H}_2$  reduced model.

This dissertation studies interpolatory model reduction for first-order descriptor systems, second-order systems, and DAEs. The main computational cost of interpolatory model reduction is the associated linear systems. Especially in the large-scale setting, inexact solves become desirable if not necessary. With the introduction of inexact solutions, however, exact interpolation no longer holds. While the effect of this loss of interpolation has previously been studied, we extend the discussion to the preconditioned case. Then we utilize IRKA's convergence behavior to develop preconditioner updates.

We also consider the interpolatory framework for DAEs and second-order systems. While

interpolation results still hold, the singularity associated with the DAE often results in unbounded model reduction errors. Therefore, we present a theorem that guarantees interpolation and a bounded model reduction error. Since this theorem relies on expensive projectors, we demonstrate how interpolation can be achieved without explicitly computing the projectors for index-1 and Hessenberg index-2 DAEs. Finally, we study reduction techniques for second-order systems. Many of the existing methods for second-order systems rely on the model's associated first-order system, which results in computations of a  $2n$  system. As a result, we present an IRKA framework for the reduction of second-order systems that does not involve the associated  $2n$  system. The resulting algorithm is shown to be effective for several dynamical systems.

# Acknowledgments

I would like to express my gratitude to Dr. Gugercin, who has been an absolutely phenomenal advisor. Beginning with my first question, Dr. Gugercin has exemplified the qualities of patience, intellectual generosity and superb communication of his vast knowledge. I would especially like to thank him for his time, understanding and care during every step of this process. Also, I would like to thank Dr. de Sturler, Dr. Beattie, Dr. Zietsman, Dr. Borggaard, and Dr. Drmac, who have shown me the beauty of numerical analysis. In addition, I am very grateful to my high school and college freshmen instructors, especially Miss Benson, Mrs. Hines, and Ms. Anderson; without these teachers, I would never have studied math beyond calculus. I would like to thank my students at New River Community College, Radford University and Virginia Tech for daily reminding me of the other purpose for my graduate work.

Furthermore, I would like to thank my friends and family. Many thanks especially to Shannon and Lynn for their encouragement and support. A very sincere thanks to Nikolas, who, among so many other gifts, built my computer, showed me the efficient way, challenged the veracity of my assumptions, and taught me to enjoy my life. Finally and most importantly,

I would like to give my most heartfelt thank you to my parents, Leslie, Sean, Will and Sabrina for loving and supporting me in every way possible. I would especially like to thank Nikolas and my family for their compassion during and after the fall semester of 2008 and for imbuing my life with love and happiness.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Linear Dynamical Systems . . . . .	1
1.2	Notation: . . . . .	2
1.3	Model Reduction . . . . .	2
1.4	Model Reduction by Projection and Interpolation . . . . .	4
1.5	$\mathcal{H}_2$ and $\mathcal{H}_\infty$ Norms . . . . .	8
1.6	Interpolation using $\mathcal{H}_2$ Optimality Conditions . . . . .	10
1.7	Optimal $\mathcal{H}_2$ Model Reduction . . . . .	13
1.8	Dissertation Goals and Organization . . . . .	14
<b>2</b>	<b>Preconditioned Iterative Solves in Model Reduction</b>	<b>17</b>
2.1	Preconditioning Techniques . . . . .	19

2.2	Upper Bounds for Subspace Angles . . . . .	22
2.2.1	Numerical Example of the Subspace Angles . . . . .	28
2.3	Pointwise Error . . . . .	29
2.4	Backward Error . . . . .	38
2.4.1	Backward Error for Left and Right Preconditioning . . . . .	40
2.4.2	Backward Error for Split Preconditioning . . . . .	46
2.5	Properties of the Backward Error Term: $\mathbf{F}_{2r}$ . . . . .	50
<b>3</b>	<b>Preconditioner Updates</b>	<b>54</b>
3.1	Preconditioner Updates . . . . .	56
3.2	Sparse Approximate Inverse (SAI) Updates . . . . .	57
3.2.1	Sparse Approximate Inverse (SAI) Updates in IRKA . . . . .	57
3.3	Numerical Results for SAI Updates . . . . .	59
3.3.1	Models Studied . . . . .	59
3.4	Effect of the SAI Update . . . . .	61
3.5	Using $\ \Delta\mathbf{EM}_i\ $ to Update . . . . .	68
3.6	Bellavia et al. Updates . . . . .	73
3.7	Numerical Results for the Bellavia et al. Update . . . . .	80

3.8	Effect of Initial Preconditioner . . . . .	86
3.8.1	Effect of Preconditioner Accuracy . . . . .	91
3.8.2	Effect of IRKA[ <b>Mid</b> , <b>f</b> ] and IRKA[ <b>G</b> , <b>f</b> ] on SAI Updates . . . . .	91
3.8.3	Effect of IRKA[ <b>Mid</b> , <b>f</b> ] and IRKA[ <b>G</b> , <b>f</b> ] on Bellavia et al. Updates . . . . .	92
<b>4</b>	<b>Interpolatory Methods for DAEs</b>	<b>98</b>
4.1	Interpolatory Model Reduction of DAEs . . . . .	99
4.2	Index-1 DAEs . . . . .	111
4.3	Numerical Results for Index-1 DAEs . . . . .	117
4.3.1	TL1 Model . . . . .	117
4.3.2	TL2 Model . . . . .	118
4.4	Hessenberg Index-2 DAEs . . . . .	119
4.4.1	Related Computational Issues to the Reduction of Hessenberg Index-2 DAEs . . . . .	132
4.4.2	IRKA for Hessenberg Index-2 DAEs . . . . .	137
4.4.3	$\mathbf{B}_2 \neq \mathbf{0}$ Case . . . . .	140
4.5	Numerical Results for Hessenberg Index-2 DAEs . . . . .	142
4.5.1	Problem 1 . . . . .	142



4.5.2	Problem 2 . . . . .	143
<b>5</b>	<b>Model Reduction of Second-order Systems</b>	<b>149</b>
5.1	Second-Order Systems . . . . .	150
5.2	Balanced Truncation Methods for Second-order Systems . . . . .	154
5.3	An IRKA framework for Second-order Systems . . . . .	159
5.3.1	SOR-IRKA . . . . .	162
5.3.2	SO-IRKA . . . . .	164
5.4	Numerical Results for the Effect of the Shift Reduction Step . . . . .	169
5.4.1	Building Model . . . . .	170
5.4.2	Beam Model . . . . .	171
5.4.3	12a Model . . . . .	173
5.5	Comparison with Balanced Truncation Methods . . . . .	174
5.5.1	Building Model . . . . .	175
5.5.2	Beam Model . . . . .	175
5.5.3	12a Model . . . . .	176
5.5.4	Butterfly Gyro Model . . . . .	178

<b>6 Conclusion</b>	<b>181</b>
<b>Bibliography</b>	<b>184</b>

# List of Figures

2.1	$\frac{\ \mathbf{F}_{2r}\ }{\ \mathbf{A}\ }$ when BiCG fails to converge . . . . .	50
2.2	$\frac{\ \mathbf{F}_{2r}\ }{\ \mathbf{A}\ }$ when BiCG converges . . . . .	51
2.3	$\frac{\ \mathbf{F}_{2r}\ }{\ \mathbf{A}\ }$ in IRKA: <i>Good Shift</i> . . . . .	52
2.4	$\frac{\ \mathbf{F}_{2r}\ }{\ \mathbf{A}\ }$ in IRKA: <i>Poor Shift</i> . . . . .	53
3.1	IRKA Shift Evolution for $\sigma_3$ of the Rail 1357 Model using BiCG . . . . .	55
3.2	GMRES Iterations (Rail Model) . . . . .	69
3.3	GMRES Iterations (CD Model) . . . . .	70
3.4	GMRES Iterations (1r Model) . . . . .	71
4.1	MNA Sigma Plot . . . . .	101
4.2	TL1 Model: Amplitude Bode Plots of $\mathbf{G}(s)$ and $\mathbf{G}_r(s)$ . . . . .	118
4.3	TL2 Model: Amplitude Bode Plots of $\mathbf{G}(s)$ and $\mathbf{G}_r(s)$ . . . . .	119

4.4	Oseen Equation: Problem 1, Frequency Response . . . . .	143
4.5	Oseen Equation: Problem 1, Time Domain Response . . . . .	144
4.6	Oseen Equation: Problem 2, Time Domain Response, First Output . . . . .	145
4.7	Oseen Equation: Problem 2, Time Domain Response, First Output: Balanced Truncation and One Step Interpolation . . . . .	146
4.8	Oseen Equation: Problem 2, Time Domain Response, First Output: Balanced Truncation and IRKA . . . . .	147
4.9	Oseen Equation: Problem 2, Time Domain Response, Second Output . . . . .	147
4.10	Oseen Equation: Problem 2, Frequency Domain Response . . . . .	148
5.1	Building Model Bode Plot ( $V \neq W$ ) . . . . .	170
5.2	Beam Model Bode Plot ( $V = W$ ) . . . . .	172
5.3	12a Model Bode Plot ( $V = W$ ) . . . . .	174
5.4	Building Model Bode Plot . . . . .	176
5.5	Beam Model Bode Plot . . . . .	177
5.6	12a Model Bode Plot . . . . .	178
5.7	Butterfly Gyro Model Bode Plot . . . . .	179

# List of Tables

1.1	Notation . . . . .	2
2.1	Rail 1357; $r = 6$ ; BiCG with ILU Left Preconditioning; $\sin(\Theta(\mathbf{V}_r, \widehat{\mathbf{V}}_r))$ . . .	29
3.1	Shift Information for the Rail Model . . . . .	60
3.2	Shift Information for the CD Model . . . . .	61
3.3	Shift Information for the 1r Model . . . . .	62
3.4	Total GMRES Iterations for $\mathbf{V}_r$ in IRKA . . . . .	68
3.5	$\frac{ \sigma_p - \sigma_j }{ \sigma_p }$ . . . . .	72
3.6	$F_G$ and $\ \mathbf{R}_i + \Delta \mathbf{E} \mathbf{M}_i\ _F$ (Rail Model) . . . . .	73
3.7	$F_G$ and $\ \mathbf{R}_i + \Delta \mathbf{E} \mathbf{M}_i\ _F$ (CD Model) . . . . .	74
3.8	$F_G$ and $\ \mathbf{R}_i + \Delta \mathbf{E} \mathbf{M}_i\ _F$ (1r Model) . . . . .	75
3.9	Factor of Additional GMRES Iterations and Preconditioners . . . . .	76

3.10	Instances of Abandoned Update . . . . .	81
3.11	Total GMRES Iterations for $\mathbf{V}_r$ . . . . .	84
3.12	Preconditioner Update Bounds . . . . .	85
3.13	$\ \mathbf{R}_k \mathbf{P}_k\ $ and Associated Terms for the Last IRKA Iteration of the Rail Model	86
3.14	$\ \mathbf{R}_k \mathbf{P}_k\ $ and Associated Terms for the Last IRKA Iteration of the CD Model	94
3.15	$\ \mathbf{R}_k \mathbf{P}_k\ $ and Associated Terms for the Last IRKA Iteration of the 1r Model	95
3.16	Rail Model: Total GMRES Iterations for $\mathbf{V}_r$ . . . . .	95
3.17	Rail Model: Incomplete LU Decompositions Computed for $\mathbf{V}_r$ . . . . .	96
3.18	CD Model: Total GMRES Iterations for $\mathbf{V}_r$ . . . . .	96
3.19	CD Model: Incomplete LU Decompositions Computed for $\mathbf{V}_r$ . . . . .	96
3.20	1r Model: Total GMRES Iterations for $\mathbf{V}_r$ . . . . .	96
3.21	1r Model: Incomplete LU Decompositions Computed for $\mathbf{V}_r$ . . . . .	97
4.1	Model Reduction Errors for the TL1 Model . . . . .	118
4.2	Model Reduction Errors for the TL2 Model . . . . .	119
4.3	Oseen Equations: Problem 1 . . . . .	143
4.4	Oseen Equations: Problem 2 . . . . .	145
5.1	Building Model Errors, reducing from $2r$ to $r$ , $\mathbf{V}_r \neq \mathbf{W}_r$ . . . . .	171

5.2	Beam Model Errors, reducing from $2r$ to $r$ , $\mathbf{V}_r = \mathbf{W}_r$ . . . . .	172
5.3	12a Model Errors, reducing from $2r$ to $r$ , $\mathbf{V}_r = \mathbf{W}_r$ . . . . .	173
5.4	Building Model Errors . . . . .	175
5.5	Beam Model Errors . . . . .	176
5.6	12a Model Errors . . . . .	177
5.7	Shift Iteration for the Butterfly Gyro Model . . . . .	180

# Chapter 1

## Introduction

### 1.1 Linear Dynamical Systems

Dynamical systems are mathematical models characterized by a set of differential or difference equations which capture the behavior of natural and artificial processes. Today's problems often lead to an almost insatiable demand for more precision, requiring a myriad of equations to describe the system. Oftentimes, these dynamical systems may involve thousands and even millions of equations. Although these complex models may capture the overall dynamics of the system more accurately, limited computational resources, inaccuracy, and ill-conditioning often result in these large models being computationally cumbersome or even intractable to use in a practical setting. Therefore, the original system is replaced with a reduced system described by a smaller set of equations. The aim of model reduction is to



obtain this reduced system, which will still delineate the important intricacies of the original system and yet be feasible to use in practice.

## 1.2 Notation:

Table 1.1: Notation

$\mathbb{R}^{m \times n}$	set of real matrices of size $m$ by $n$
$\mathbb{C}^{m \times n}$	set of complex matrices of size $m$ by $n$
$s$	a complex number in $\mathbb{C}$
$\mathbf{A}^T$	transpose of $\mathbf{A}$
	regardless of whether $\mathbf{A}$ is real or complex
$\text{span}\{\mathbf{x}_1, \dots, \mathbf{x}_l\}$	span of the vectors $\mathbf{x}_1, \dots, \mathbf{x}_l$
$\ \mathbf{A}\ , \ \mathbf{A}\ _2$	2-induced norm of $\mathbf{A}$
$\ \mathbf{A}\ _F$	Frobenius norm of $\mathbf{A}$
$\mathbf{I}$	Identity matrix of appropriate size
$i$	$\sqrt{-1}$
$\ \mathbf{H}\ _{\mathcal{H}_\infty}$	$\mathcal{H}_\infty$ norm of $\mathbf{H}(s)$
$\ \mathbf{H}\ _{\mathcal{H}_2}$	$\mathcal{H}_2$ norm of $\mathbf{H}(s)$
$\text{Null}(\mathbf{M})$	Null space of $\mathbf{M}$
$\text{Ran}(\mathbf{M})$	Range of $\mathbf{M}$

## 1.3 Model Reduction

We focus primarily on linear multiple-input/multiple-output (MIMO) systems and follow the exposition of [7]. For an input,  $\mathbf{u}(t)$ , and output,  $\mathbf{y}(t)$ , let  $\tilde{\mathbf{u}}(s)$  and  $\tilde{\mathbf{y}}(s)$  denote the system inputs and outputs, respectively, in the Laplace transform domain. Then the state-space

form of the linear MIMO system is given as

$$\text{Find } \tilde{\mathbf{v}}(s) \text{ such that } \mathbf{K}(s)\tilde{\mathbf{v}}(s) = \mathbf{B}(s)\tilde{\mathbf{u}}(s), \text{ then } \tilde{\mathbf{y}}(s) := \mathbf{C}(s)\tilde{\mathbf{v}}(s) \quad (1.3.1)$$

where the matrices  $\mathbf{K}(s) \in \mathbb{C}^{n \times n}$ ,  $\mathbf{C}(s) \in \mathbb{C}^{p \times n}$ ,  $\mathbf{B}(s) \in \mathbb{C}^{n \times m}$  are analytic in the right-half plane and  $\mathbf{K}(s)$  is of full rank throughout the right-half plane as well. Using (1.3.1) to solve for  $\tilde{\mathbf{y}}(s)$  leads to

$$\tilde{\mathbf{y}}(s) = \mathbf{C}(s)\mathbf{K}(s)^{-1}\mathbf{B}(s)\tilde{\mathbf{u}}(s) = \mathbf{H}(s)\tilde{\mathbf{u}}(s), \quad (1.3.2)$$

implying that the transfer function of (1.3.1) is given as

$$\mathbf{H}(s) = \mathbf{C}(s)\mathbf{K}(s)^{-1}\mathbf{B}(s). \quad (1.3.3)$$

We refer to (1.3.3) as the *generalized coprime realization*. One of the benefits of this framework is its versatility since MIMO systems with various structures, such as second-order systems or parametric models, can be described by (1.3.3). For example, a first-order descriptor system,  $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ , with constant matrices  $\mathbf{E}$ ,  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times m}$ , and  $\mathbf{C} \in \mathbb{R}^{p \times n}$  is described in this framework by taking  $\mathbf{C}(s) = \mathbf{C}$ ,  $\mathbf{B}(s) = \mathbf{B}$ , and  $\mathbf{K}(s) = s\mathbf{E} - \mathbf{A}$ .

The dimension of the associated state space is defined to be equal to that of the dimension of  $\mathbf{K}(s)$ . Since  $\mathbf{K}(s)$  is an  $n \times n$  matrix, the dimension of  $\mathbf{H}(s)$  is  $n$ . Since oftentimes  $n$  is so large it renders the model too cumbersome for efficient simulation or control applications,

the goal of model reduction is to obtain a dynamical system with a state space form as

$$\text{Find } \tilde{\mathbf{v}}(s) \text{ such that } \mathbf{K}_r(s)\tilde{\mathbf{v}}(s) = \mathbf{B}_r(s)\tilde{\mathbf{u}}(s), \text{ then } \tilde{\mathbf{y}}_r(s) := \mathbf{C}_r(s)\tilde{\mathbf{v}}(s) \quad (1.3.4)$$

where  $\tilde{\mathbf{y}}_r(s)$  is the reduced output,  $\mathbf{K}_r(s) \in \mathbb{C}^{r \times r}$ ,  $\mathbf{C}_r(s) \in \mathbb{C}^{p \times r}$ ,  $\mathbf{B}_r(s) \in \mathbb{C}^{r \times m}$  and  $r \ll n$ .

The transfer function of the reduced system is then

$$\mathbf{H}_r(s) = \mathbf{C}_r(s)\mathbf{K}_r(s)^{-1}\mathbf{B}_r(s). \quad (1.3.5)$$

Since the goal is ultimately to use  $\mathbf{H}_r(s)$  as a surrogate for  $\mathbf{H}(s)$ , the reduced model needs to satisfy additional criteria. First,  $\mathbf{H}_r(s)$  must be obtained in a computationally feasible and efficient manner even for large-scale dynamical systems. Secondly, system properties and structure present in the full-order model ideally should also be represented in the reduced model. Finally and perhaps most importantly,  $\mathbf{H}_r(s)$  needs to capture the input and output relationship of the original system. For more details, see [2].

## 1.4 Model Reduction by Projection and Interpolation

In this section, we describe the Petrov-Galerkin approximation process that leads to a reduced-order system,  $\mathbf{H}_r(s)$ . Assume  $\mathcal{V}_r$  and  $\mathcal{W}_r$  are the right and left modeling  $r$ -dimensional subspaces of  $\mathbb{R}^n$  with  $\mathcal{V}_r \cap \mathcal{W}_r^\perp = \{0\}$ . For inputs,  $\tilde{\mathbf{u}}(s)$ , the reduced output,  $\tilde{\mathbf{y}}_r(s)$ , is then

defined as

$$\text{Find } \tilde{\mathbf{v}}(s) \in \mathcal{V}_r \text{ such that } \mathbf{W}_r^T (\mathcal{K}(s) \tilde{\mathbf{v}}(s) - \mathcal{B}(s) \tilde{\mathbf{u}}(s)) = 0 \quad (1.4.1)$$

$$\text{then } \tilde{\mathbf{y}}_r(s) := \mathcal{C}(s) \tilde{\mathbf{v}}(s). \quad (1.4.2)$$

In the Laplace domain, this projection process results in a reduced transfer function

$$\mathcal{H}_r(s) = \mathcal{C}_r(s) \mathcal{K}_r(s)^{-1} \mathcal{B}_r(s) \quad (1.4.3)$$

where

$$\begin{aligned} \mathcal{K}_r(s) &= \mathbf{W}_r^T \mathcal{K}(s) \mathbf{V}_r \in \mathbb{C}^{r \times r}, & \mathcal{B}_r(s) &= \mathbf{W}_r^T \mathcal{B}(s) \in \mathbb{C}^{p \times r}, \\ \text{and } \mathcal{C}_r(s) &= \mathcal{C}(s) \mathbf{V}_r \in \mathbb{C}^{r \times m}. \end{aligned} \quad (1.4.4)$$

For interpolatory model reduction,  $\mathbf{V}_r$  and  $\mathbf{W}_r$  are defined to enforce certain interpolation conditions. For the SISO case ( $m = 1, p = 1$ ), for example, we aim to match certain moments. The  $k^{\text{th}}$  moment of  $\mathcal{H}(s)$  at a point  $\sigma_i \in \mathbb{C}$  is defined as the  $k^{\text{th}}$  derivative of the transfer function,  $\mathcal{H}(s)$ , evaluated at  $\sigma_i$ . The aim of model reduction by moment matching is for the reduced model,  $\mathcal{H}_r(s)$ , to interpolate  $\mathcal{H}(s)$  and a certain number of its derivatives at selected interpolation points or shifts, which will be denoted by  $\sigma_k$ . To achieve Hermite

interpolation, for example, our goal is to find a reduced-order model,  $\mathcal{H}_r(s)$ , such that

$$\mathcal{H}_r(\sigma_k) = \mathcal{H}(\sigma_k) \quad \text{and} \quad \mathcal{H}'_r(\sigma_k) = \mathcal{H}'(\sigma_k) \quad \text{for} \quad k = 1, \dots, r.$$

In the general MIMO case, the multiple-input and multiple-output structure implies that the moment matching requirement is too restrictive; instead, we aim to construct reduced-order models that tangentially interpolate the full-order model. Given a set of interpolation points  $\{\sigma_i\}_{i=1}^r, \{\mu_i\}_{i=1}^r \subset \mathbb{C}$  and sets of right-tangential directions,  $\{\mathbf{b}_i\}_{i=1}^r \subset \mathbb{C}^m$ , and left-tangential directions,  $\{\mathbf{c}_i\}_{i=1}^r \subset \mathbb{C}^p$ , we say  $\mathcal{H}_r(s)$  tangentially interpolates  $\mathcal{H}(s)$  in the following sense:

$$\mathcal{H}(\sigma_j)\mathbf{b}_j = \mathcal{H}_r(\sigma_j)\mathbf{b}_j \quad \mathbf{c}_i^T \mathcal{H}(\mu_i) = \mathbf{c}_i^T \mathcal{H}_r(\mu_i) \quad \text{for } i, j = 1, \dots, r. \quad (1.4.5)$$

We say that  $\mathcal{H}_r(s)$  bitangentially interpolates  $\mathcal{H}(s)$  at  $\mu_k$  provided

$$\mathbf{c}_k^T \mathcal{H}'(\mu_k) \mathbf{b}_k = \mathbf{c}_k^T \mathcal{H}'_r(\mu_k) \mathbf{b}_k. \quad (1.4.6)$$

Naturally, we desire for the process with which we achieve the interpolation to be numerically robust. Previous research has proven that the computation of moments is extremely ill-conditioned [39]. Fortunately, the following theorem elucidates a way to achieve (1.4.5) and (1.4.6) without explicit computation of the interpolated quantities.

**Theorem 1.1.** [3] Suppose that  $\mathbf{B}(s)$ ,  $\mathbf{C}(s)$ , and  $\mathbf{K}(s)$  are analytic at  $\sigma \in \mathbb{C}$  and  $\mu \in \mathbb{C}$ .

Also, let  $\mathbf{K}(\sigma)$ ,  $\mathbf{K}(\mu)$ ,  $\mathbf{K}_r(\sigma) = \mathbf{W}_r^T \mathbf{K}(\sigma) \mathbf{V}_r$ , and  $\mathbf{K}_r(\mu) = \mathbf{W}_r^T \mathbf{K}(\mu) \mathbf{V}_r$  have full rank. Let nonnegative integers  $M$  and  $N$  be given as well as nontrivial vectors,  $\mathbf{b} \in \mathbb{R}^m$  and  $\mathbf{c} \in \mathbb{R}^p$ .

Let the quantity  $\mathbf{H}^{(m)}(\sigma)$  denote the  $m^{\text{th}}$  derivative of  $\mathbf{H}(s)$  with respect to  $s$  evaluated at  $s = \sigma$  and  $\mathcal{D}_\sigma^l f$  denote the  $l^{\text{th}}$  derivative of the univariate function  $f(s)$  evaluated at  $s = \sigma$ .

a) If  $\mathcal{D}_\sigma^i [\mathbf{K}(s)^{-1} \mathbf{B}(s)] \mathbf{b} \in \text{Ran}(\mathbf{V}_r)$  for  $i = 0, \dots, N$ , then  $\mathbf{H}^{(l)}(\sigma) \mathbf{b} = \mathbf{H}_r^{(l)}(\sigma) \mathbf{b}$  for

$l = 0, \dots, N$ .

b) If  $(\mathbf{c}^T \mathcal{D}_\mu^j [\mathbf{C}(s) \mathbf{K}(s)^{-1}])^T \in \text{Ran}(\mathbf{W}_r)$  for  $j = 0, \dots, M$ , then  $\mathbf{c}^T \mathbf{H}^{(l)}(\mu) = \mathbf{c}^T \mathbf{H}_r^{(l)}(\mu)$  for

$l = 0, \dots, M$ .

c) If (a) and (b) hold with  $\sigma = \mu$ , then  $\mathbf{c}^T \mathbf{H}^{(l)}(\sigma) \mathbf{b} = \mathbf{c}^T \mathbf{H}_r^{(l)}(\sigma) \mathbf{b}$  for  $l = 0, \dots, M + N + 1$ .

To implement this theorem so that Hermite interpolation is achieved, for example, the matrices are constructed as:

$$\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r] = [\mathbf{K}(\sigma_1)^{-1} \mathbf{B}(\sigma_1) \mathbf{b}_1, \dots, \mathbf{K}(\sigma_r)^{-1} \mathbf{B}(\sigma_r) \mathbf{b}_r], \quad (1.4.7)$$

$$\mathbf{W}_r^T = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_r^T \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1^T \mathbf{C}(\sigma_1) \mathbf{K}(\sigma_1)^{-1} \\ \vdots \\ \mathbf{c}_r^T \mathbf{C}(\sigma_r) \mathbf{K}(\sigma_r)^{-1} \end{bmatrix}. \quad (1.4.8)$$

Deflating  $\mathbf{V}_r$  and  $\mathbf{W}_r$  if necessary, we will assume  $\mathbf{V}_r$  and  $\mathbf{W}_r$  to be full-rank. The reduced-order model  $\mathbf{H}_r(s) = \mathbf{C}_r(s) \mathbf{K}_r(s)^{-1} \mathbf{B}_r(s)$  is then defined by (1.4.4) using  $\mathbf{V}_r$  and  $\mathbf{W}_r$  as defined in (1.4.7) - (1.4.8). From Theorem 1.1,  $\mathbf{H}_r(s)$  tangentially interpolates  $\mathbf{H}(s)$  as

defined in (1.4.5) and (1.4.6).

## 1.5 $\mathcal{H}_2$ and $\mathcal{H}_\infty$ Norms

Throughout this dissertation, we will use the  $\mathcal{H}_2$  and  $\mathcal{H}_\infty$  norms. The  $\mathcal{H}_2$  norm is defined as

$$\|\mathfrak{H}\|_{\mathcal{H}_2} := \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathfrak{H}(i\omega)\|_F^2 d\omega \right)^{1/2}. \quad (1.5.1)$$

For more details, see [2].

The  $\mathcal{H}_2$  norm is one way to evaluate the performance of the reduced-order model. In order for  $\mathfrak{H}_r(s)$  to be of practical use, we desire  $\mathfrak{H}_r(s)$  to capture the relationship between the input and output of the system, namely we want  $\max_{t>0} \|\mathbf{y}(t) - \mathbf{y}_r(t)\|_\infty$  to be uniformly small over all inputs  $\mathbf{u}(t)$ . As shown in [47], assuming  $\mathbf{u}(t)$  is such that  $\int_0^\infty \|\mathbf{u}(t)\|_2^2 dt \leq 1$ , then

$$\max_{t>0} \|\mathbf{y}(t) - \mathbf{y}_r(t)\|_\infty \leq \|\mathfrak{H} - \mathfrak{H}_r\|_{\mathcal{H}_2} \quad (1.5.2)$$

with equality holding for the SISO case. Therefore, to minimize  $\max_{t>0} \|\mathbf{y}(t) - \mathbf{y}_r(t)\|_\infty$ , we want  $\|\mathfrak{H} - \mathfrak{H}_r\|_{\mathcal{H}_2}$  to be minimized. Assuming that  $\mathfrak{H}(s)$  is a stable dynamical system, the

$\mathcal{H}_\infty$  norm is defined as

$$\|\mathfrak{H}\|_{\mathcal{H}_\infty} := \sup_{\omega} \|\mathfrak{H}(i\omega)\|_2.$$

For first-order descriptor systems,  $\mathfrak{H}(s) = s\mathbf{E} - \mathbf{A}$ , such that  $\mathbf{E}$  is singular, 0 must be a nondefective eigenvalue of  $\mathbf{E}$  so that  $\mathfrak{H}(s)$  remains bounded at  $\infty$ . One advantage of the  $\mathcal{H}_\infty$  norm is its ability to capture the physical properties of the system. If we assume that  $\|\mathfrak{H} - \mathfrak{H}_r\|_{\mathcal{H}_\infty} \leq \alpha$  where  $\alpha$  is a positive scalar, then

$$\|\mathfrak{H} - \mathfrak{H}_r\|_{\mathcal{H}_\infty} = \sup_{\mathbf{u} \neq 0} \frac{\|\mathbf{y} - \mathbf{y}_r\|_{\mathcal{L}_2}}{\|\mathbf{u}\|_{\mathcal{L}_2}} \leq \alpha.$$

For SIMO ( $m = 1$ ) and MISO ( $p = 1$ ) systems, a similar relationship also holds in the  $\mathcal{H}_2$  norm:

$$\|\mathfrak{H} - \mathfrak{H}_r\|_{\mathcal{H}_2} = \sup_{\mathbf{u} \neq 0} \frac{\|\mathbf{y} - \mathbf{y}_r\|_{\mathcal{L}_\infty}}{\|\mathbf{u}\|_{\mathcal{L}_2}}.$$

In this way, the  $\mathcal{H}_\infty$  and  $\mathcal{H}_2$  norms describe the input and output relationship of the error system. While both norms provide metrics for the fidelity of the reduced models, we will be especially interested in the  $\mathcal{H}_2$  norm and the associated optimal  $\mathcal{H}_2$  model reduction problem as discussed in the next section.



## 1.6 Interpolation using $\mathcal{H}_2$ Optimality Conditions

For the optimal  $\mathcal{H}_2$  model reduction problem, we consider the case when  $\mathcal{K}(s) = s\mathbf{E} - \mathbf{A}$  in the generalized coprime realization, implying that we aim to reduce a full-order system of the form:

$$\mathbf{H}(s) : \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t), \end{cases} \quad (1.6.1)$$

where  $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times m}$ , and  $\mathbf{C} \in \mathbb{R}^{p \times n}$ . Given a system  $\mathbf{H}(s)$  as in (1.6.1), the aim of optimal  $\mathcal{H}_2$  model reduction is to find a reduced-order system,  $\mathbf{H}_r(s)$ , such that

$$\mathbf{H}_r(s) = \min_{\substack{\deg(\hat{\mathbf{H}}_r)=r \\ \hat{\mathbf{H}}: \text{stable}}} \|\mathbf{H}(s) - \hat{\mathbf{H}}_r(s)\|_{\mathcal{H}_2}. \quad (1.6.2)$$

The significance of constructing a reduced-order model  $\mathbf{H}_r(s)$  which satisfies (1.6.2) follows from (1.5.2); by finding a reduced-order system that minimizes the  $\mathcal{H}_2$  error, the maximum difference between the outputs  $\mathbf{y}(t)$  and  $\mathbf{y}_r(t)$  is as small as possible. This is an extremely important feature for our reduced-order model to possess since the reduced-order model needs to capture the relationship between the input and output of the original model. To obtain the model that satisfies (1.6.2), we first present two results for SISO systems without proof. The proofs may be found in [45], [47] and [62].

**Theorem 1.2.** [45] *Given the full-order SISO model  $H(s)$  and a reduced-order model  $H_r(s)$ , let  $\lambda_i$  and  $\hat{\lambda}_j$  be the poles of  $H(s)$  and  $H_r(s)$ , respectively. Suppose that the poles of  $H_r(s)$*

are distinct. Let  $\Phi_i$  and  $\hat{\Phi}_j$  denote the residues of the transfer functions  $H(s)$  and  $H_r(s)$  at their poles  $\lambda_i$  and  $\hat{\lambda}_j$ , respectively:  $\Phi_i = \text{res}[H(s), \lambda_i]$ ,  $i = 1, \dots, n$  and  $\hat{\Phi}_j = \text{res}[H_r(s), \hat{\lambda}_j]$ ,  $j = 1, \dots, r$ . The  $\mathcal{H}_2$  norm of the error system is given by

$$\|H(s) - H_r(s)\|_{\mathcal{H}_2}^2 = \sum_{i=1}^n \Phi_i (H(-\lambda_i) - H_r(-\lambda_i)) + \sum_{j=1}^r \hat{\Phi}_j (H_r(-\hat{\lambda}_j) - H(-\hat{\lambda}_j)).$$

Theorem 1.2 describes the relationship between the  $\mathcal{H}_2$  error and the poles of both  $H(s)$  and  $H_r(s)$ . To minimize  $\|H(s) - H_r(s)\|_{\mathcal{H}_2}$ , we want  $H_r(s)$  to match  $H(s)$  at both the reflected poles of  $H(s)$  and at the mirror images of its own poles. While Gugercin and Antoulas in [46] illustrated the benefits of choosing the interpolation points,  $\sigma_i$ , to be the mirror images of the poles of  $H(s)$  associated with the larger residuals, [47] proves that the second term of the sum is actually more important. In fact, [47] shows that the optimal selection of interpolation points,  $\sigma_i$ , is  $\sigma_i = -\hat{\lambda}_i$ . Therefore, Theorem 1.2 shows the important connection between the poles of the full and reduced-order models. The next theorem also reflects the pivotal importance of the reduced-order model's poles.

**Theorem 1.3.** *Meier-Luenberger [62] Let  $H(s)$  be the full-order SISO system and  $H_r(s)$  be a minimizer for  $\|H(s) - H_r(s)\|_{\mathcal{H}_2}$  with the simple poles of  $H_r(s)$  denoted by  $\hat{\lambda}_k$ . Then*

$$H(-\hat{\lambda}_k) = H_r(-\hat{\lambda}_k) \quad \text{and} \quad H'(-\hat{\lambda}_k) = H_r'(-\hat{\lambda}_k) \quad \text{for} \quad k = 1, \dots, r.$$

The Meier-Luenberger conditions as stated in Theorem 1.3 provide the first-order necessary

conditions for  $\mathcal{H}_2$  optimality, namely for a reduced-order model to satisfy (1.6.2), Hermite interpolation of  $H(s)$  must occur at the mirror images of the poles of  $H_r(s)$ . For the MIMO case, first-order necessary conditions have been derived as discussed in [47], [76], and [25]. Below we state a theorem for the first-order necessary conditions in the MIMO case as presented and proved in [47].

**Theorem 1.4.** [47] *Suppose  $\mathbf{H}(s)$  and  $\mathbf{H}_r(s)$  are real stable dynamical systems. Let*

$$\mathbf{H}_r(s) = \sum_{i=1}^r \frac{1}{s - \hat{\lambda}_i} \mathbf{c}_i \mathbf{b}_i^T$$

where  $\{\hat{\lambda}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i \mathbf{b}_i^T\}_{i=1}^r$  are the simple poles and residues of  $\mathbf{H}_r(s)$ , respectively. Furthermore, if  $\mathbf{H}_r(s)$  satisfies

$$\mathbf{H}_r(s) = \min_{\substack{\deg(\hat{\mathbf{H}}_r)=r \\ \hat{\mathbf{H}}: \text{stable}}} \|\mathbf{H}(s) - \hat{\mathbf{H}}_r(s)\|_{\mathcal{H}_2}, \quad (1.6.3)$$

then for  $i = 1, 2, \dots, r$

- 1)  $\mathbf{H}(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{H}_r(-\hat{\lambda}_i) \mathbf{b}_i$
- 2)  $\mathbf{c}_i^T \mathbf{H}(-\hat{\lambda}_i) = \mathbf{c}_i^T \mathbf{H}_r(-\hat{\lambda}_i)$
- 3)  $\mathbf{c}_i^T \mathbf{H}'(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(-\hat{\lambda}_i) \mathbf{b}_i.$

## 1.7 Optimal $\mathcal{H}_2$ Model Reduction

A plethora of research has shown that finding a global minimizer for the  $\mathcal{H}_2$  error is an extremely arduous process. While the existence of a global minimizer for the SISO case is guaranteed, a similar guarantee has yet to be proven for the MIMO case [47]. As a result, the common method is to focus on fulfilling only the first-order conditions for  $\mathcal{H}_2$  optimality. Unfortunately, most of these methods rely on the computation of dense matrix operations. See [81], [70], [23], [62], [53], and [80] for more details. Especially for large-scale dynamical systems, the expensive computations involved often make these methods not practical. To circumvent these issues, Gugercin et al. in [47] proposed the “Iterative Rational Krylov Algorithm” (IRKA), which employs iterative rational Krylov steps such that upon convergence the first-order necessary conditions are satisfied. The key feature of IRKA is its ability to satisfy the first-order necessary conditions without explicitly computing solutions of the expensive Lyapunov equations. IRKA iterates the Krylov reduction process and assigns  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  as the new interpolation points until the iteration converges to the optimal shift selection as defined by the first-order necessary conditions. For more details about the theoretical motivation behind the algorithm, see [47].

### Algorithm 1.7.1. [47] IRKA for MIMO $\mathcal{H}_2$ Optimal Tangential Interpolation

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2.  $\mathbf{V}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_r ]$ .

$$3. \mathbf{W}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_r ].$$

4. *while (not converged)*

$$(a) \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r.$$

(b) Compute  $\mathbf{Y}^T \mathbf{A}_r \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_r \mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_r - \mathbf{A}_r$ .

$$(c) \sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r) \text{ for } i = 1, \dots, r, \mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r, \text{ and } \mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i.$$

$$(d) \mathbf{V}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_r ].$$

$$(e) \mathbf{W}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_r ].$$

$$5. \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r.$$

Since the eigenvalue problem is of dimension  $r$ , the main cost of IRKA is the solution of the  $2r$  linear systems at each iteration. Especially in a large-scale setting, inexact solves will need to be employed in solving these systems. Using inexact solves creates new concerns since exact Hermite interpolation of  $\mathbf{H}(s)$  will no longer hold.

## 1.8 Dissertation Goals and Organization

The main aim of this dissertation is to contribute to the study of interpolatory model reduction. Chapter 2 continues the discussion of [7] and extends the results to the cases when left, right or split preconditioning are implemented in the inexact solve. While the upper bounds

for the subspace angles and pointwise error are similar for all preconditioning techniques, the backward error result for the unpreconditioned case does not trivially extend to the preconditioned case. Instead, the results of Chapter 2 prove that for left and right preconditioning, a backward error result similar to [7] does not hold unless unique orthogonality conditions, which are not readily available in iterative methods, are imposed. In addition, this chapter proves that the backward error result requiring a Petrov-Galerkin framework holds when split preconditioning is employed. Due to the importance of preconditioning, Chapter 3<sup>1</sup> develops preconditioning techniques that utilize the convergence of the shifts in the IRKA iteration. Two update methods are studied, namely sparse approximate inverse preconditioners (SAI) and a preconditioner update technique as proposed in [11]. We consider the theoretical properties of these updates as well as presenting a numerical study of the updates applied to the reduction of three dynamical systems.

In the remaining chapters, we focus on implementing IRKA for different types of dynamical systems, namely DAEs and second-order systems. Chapter 4 considers the reduction of DAEs, where the singular  $\mathbf{E}$  matrix may potentially cause unbounded model reduction errors when existing interpolatory methods are employed. To remedy this, we present a new interpolation result and an algorithm for the reduction of DAEs. While this algorithm proves to be effective, it depends on projectors that are computationally expensive. As a result, the remainder of the chapter is devoted to theoretically and numerically illustrating how the explicit computation of the projectors can be circumvented for index-1 and Hessenberg index-2 DAEs. Finally, Chapter 5 presents an algorithm for the reduction of second-order systems

---

<sup>1</sup>Chapter 3 is the result of collaboration with Dr. Eric de Sturler.

using the IRKA framework. The resulting algorithm is not just a trivial extension of IRKA as there are several implementation issues introduced with the second-order structure. We consider these issues and conclude with a numerical study of four models, which illustrates the competitiveness of the proposed algorithm in comparison to existing methods. Therefore, this dissertation offers a significant contribution to the study of first-order descriptor systems, second-order systems and DAEs in the interpolatory framework.

## Chapter 2

# Preconditioned Iterative Solves in Model Reduction

For interpolatory model reduction methods, the main cost emanates from the construction of the matrices  $\mathbf{V}_r$  and  $\mathbf{W}_r$ , which requires the solution of  $2r$  systems. Although interpolatory methods assume that the systems are solved directly, the need for more accuracy often augments the dimension of the dynamical system to the point where direct solves become computationally infeasible. Since  $\mathcal{K}(s)$  is typically sparse, this is an ideal setting in which to employ iterative methods, such as GMRES and BiCG. For more details about these solvers, see [69], [32], [41], [5], and [74]. However, the introduction of inexact solves implies exact interpolation of the full-order model no longer holds. Hence, we wish to quantify the effect of these inexact solves on the overall model reduction procedure. Suppose that  $\hat{\mathbf{v}}_j$  and  $\hat{\mathbf{w}}_i$



are approximate solutions to the linear systems

$$\mathcal{K}(\sigma_j)\mathbf{v}_j = \mathcal{B}(\sigma_j)\mathbf{b}_j \quad (2.0.1)$$

$$\mathcal{K}(\mu_i)^T \mathbf{w}_i = \mathcal{C}(\mu_i)^T \mathbf{c}_i \quad (2.0.2)$$

with corresponding residuals  $\boldsymbol{\eta}_j$  and  $\boldsymbol{\xi}_i$  defined as

$$\boldsymbol{\eta}_j = \mathcal{K}(\sigma_j)\widehat{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \quad \text{and} \quad \boldsymbol{\xi}_i = \mathcal{K}(\mu_i)^T \widehat{\mathbf{w}}_i - \mathcal{C}(\mu_i)^T \mathbf{c}_i. \quad (2.0.3)$$

We denote the resulting inexact matrices by

$$\widehat{\mathbf{V}}_r = [\widehat{\mathbf{v}}_1, \dots, \widehat{\mathbf{v}}_r] \quad \text{and} \quad \widehat{\mathbf{W}}_r^T = \begin{bmatrix} \widehat{\mathbf{w}}_1^T \\ \vdots \\ \widehat{\mathbf{w}}_r^T \end{bmatrix}. \quad (2.0.4)$$

Then the inexact reduced-order model is given by

$$\widehat{\mathcal{H}}_r(s) = \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s),$$

where

$$\widehat{\mathcal{K}}_r(s) = \widehat{\mathbf{W}}_r^T \mathcal{K}(s) \widehat{\mathbf{V}}_r, \quad \widehat{\mathcal{B}}_r(s) = \widehat{\mathbf{W}}_r^T \mathcal{B}(s), \quad \text{and} \quad \widehat{\mathcal{C}}_r(s) = \mathcal{C}(s) \widehat{\mathbf{V}}_r. \quad (2.0.5)$$

In [7], the effect of unpreconditioned inexact solves in this framework was examined. The aim of our work is to extend the results of [7] to the preconditioned case. In Section 2.1, we define the notation for the different types of preconditioning techniques considered, namely left, right and split preconditioning. Using this notation, we closely follow the exposition of [7] to extend the upper bounds for the subspace angle and pointwise errors to the preconditioned case in Section 2.2 and Section 2.3. While the pointwise and subspace angle upper bounds are similar to those as in the unpreconditioned case, the backward error result of [7] does not trivially extend to the preconditioned case. As a result, the remainder of the chapter is devoted to stating and proving the orthogonality conditions required for the backward error results in the left, right and split preconditioning cases. For left and right preconditioning, a backward error result similar to the one obtained in [7] requires a special orthogonality condition, which can not be easily implemented in existing iterative solve methods. However, we will conclude the chapter by proving that a Petrov-Galerkin framework combined with a split preconditioner provides a similar backward error result as shown in [7].

## 2.1 Preconditioning Techniques

To improve the convergence of the linear solve, a preconditioner is often used. The aim of preconditioning a linear system,  $\mathbf{M}\mathbf{g} = \mathbf{h}$ , is to find a matrix  $\mathbf{P}$  such that the preconditioned system has superior convergence properties, namely the eigenvalues of the preconditioned system are clustered, ideally away from the origin. In solving the linear system, there are

three common types of preconditioning, namely left preconditioning, right preconditioning and split preconditioning. With left preconditioning, we compute a preconditioner  $\mathbf{P}$  and then solve the preconditioned system  $\mathbf{P}^{-1}\mathbf{M}\mathbf{g} = \mathbf{P}^{-1}\mathbf{h}$ . For right preconditioning, we solve  $\mathbf{M}\mathbf{P}^{-1}\mathbf{y} = \mathbf{h}$ , and then the solution to the unpreconditioned linear system is given by  $\mathbf{g} = \mathbf{P}^{-1}\mathbf{y}$ . Finally, split preconditioning can be employed; for example, if  $\mathbf{P} = \mathbf{L}\mathbf{U}$ , then we solve  $\mathbf{L}^{-1}\mathbf{M}\mathbf{U}^{-1}\mathbf{u} = \mathbf{L}^{-1}\mathbf{h}$  and the solution to the unpreconditioned system is given by  $\mathbf{g} = \mathbf{U}^{-1}\mathbf{u}$ .

Applying split preconditioning in the context of interpolatory model reduction, we let

$\mathbf{N} = \mathbf{L}\mathbf{U}$  be a preconditioner and let  $\hat{\mathbf{v}}_j = \mathbf{U}_j^{-1}\hat{\mathbf{u}}_j$  and  $\hat{\mathbf{w}}_j = \mathbf{L}_j^{-T}\hat{\mathbf{z}}_j$  be the inexact solutions for the split preconditioned systems,

$$\mathbf{L}_j^{-1}\mathcal{K}(\sigma_j)\mathbf{U}_j^{-1}\mathbf{u}_j = \mathbf{L}_j^{-1}\mathcal{B}(\sigma_j)\mathbf{b}_j \quad (2.1.1)$$

$$\mathbf{U}_j^{-T}\mathcal{K}(\sigma_j)^T\mathbf{L}_j^{-T}\mathbf{z}_j = \mathbf{U}_j^{-T}\mathcal{C}(\sigma_j)^T\mathbf{c}_j \quad (2.1.2)$$

with associated residuals

$$\boldsymbol{\eta}_j = \mathbf{L}_j^{-1}\mathcal{K}(\sigma_j)\mathbf{U}_j^{-1}\hat{\mathbf{u}}_j - \mathbf{L}_j^{-1}\mathcal{B}(\sigma_j)\mathbf{b}_j \quad (2.1.3)$$

$$\boldsymbol{\xi}_j = \mathbf{U}_j^{-T}\mathcal{K}(\sigma_j)^T\mathbf{L}_j^{-T}\hat{\mathbf{z}}_j - \mathbf{U}_j^{-T}\mathcal{C}(\sigma_j)^T\mathbf{c}_j. \quad (2.1.4)$$

If left preconditioning is applied to (2.0.1) and (2.0.2), then we define  $\widehat{\mathbf{v}}_j$  and  $\widehat{\mathbf{w}}_j$  to be the inexact solutions for the left preconditioned systems

$$\mathbf{N}_j^{-1}\mathcal{K}(\sigma_j)\mathbf{v}_j = \mathbf{N}_j^{-1}\mathcal{B}(\sigma_j)\mathbf{b}_j \quad (2.1.5)$$

$$\mathbf{N}_j^{-T}\mathcal{K}(\sigma_j)^T\mathbf{w}_j = \mathbf{N}_j^{-T}\mathcal{C}(\sigma_j)^T\mathbf{c}_j \quad (2.1.6)$$

with associated residuals

$$\boldsymbol{\eta}_j = \mathbf{N}_j^{-1}\mathcal{K}(\sigma_j)\widehat{\mathbf{v}}_j - \mathbf{N}_j^{-1}\mathcal{B}(\sigma_j)\mathbf{b}_j \quad (2.1.7)$$

$$\boldsymbol{\xi}_j = \mathbf{N}_j^{-T}\mathcal{K}(\sigma_j)^T\widehat{\mathbf{w}}_j - \mathbf{N}_j^{-T}\mathcal{C}(\sigma_j)^T\mathbf{c}_j. \quad (2.1.8)$$

If right preconditioning is applied, then we let  $\widehat{\mathbf{v}}_j$  and  $\widehat{\mathbf{w}}_j$  be the inexact solutions for the right preconditioned systems

$$\mathcal{K}(\sigma_j)\mathbf{R}^{-1}\mathbf{y}_j = \mathcal{B}(\sigma_j)\mathbf{b}_j \quad \text{where} \quad \mathbf{v}_j = \mathbf{R}^{-1}\mathbf{y}_j \quad (2.1.9)$$

$$\mathcal{K}(\sigma_j)^T\mathbf{R}^{-T}\mathbf{z}_j = \mathcal{C}(\sigma_j)^T\mathbf{c}_j \quad \text{where} \quad \mathbf{w}_j = \mathbf{R}^{-T}\mathbf{z}_j \quad (2.1.10)$$

with associated residuals

$$\boldsymbol{\eta}_j = \mathcal{K}(\sigma_j)\mathbf{R}^{-1}\widehat{\mathbf{y}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \quad (2.1.11)$$

$$\boldsymbol{\xi}_j = \mathcal{K}(\sigma_j)^T\mathbf{R}^{-T}\widehat{\mathbf{z}}_j - \mathcal{C}(\sigma_j)^T\mathbf{c}_j. \quad (2.1.12)$$

For all preconditioning techniques, we denote the resulting inexact matrices by

$$\widehat{\mathbf{V}}_r = [\widehat{\mathbf{v}}_1, \dots, \widehat{\mathbf{v}}_r] \quad \text{and} \quad \widehat{\mathbf{W}}_r^T = \begin{bmatrix} \widehat{\mathbf{w}}_1^T \\ \vdots \\ \widehat{\mathbf{w}}_r^T \end{bmatrix}, \quad (2.1.13)$$

and the inexact reduced-order model as

$$\widehat{\mathcal{H}}_r(s) = \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s).$$

## 2.2 Upper Bounds for Subspace Angles

Since the range of the matrices in (1.4.7) - (1.4.8) or (2.1.13) ultimately determines the reduced-order model, we are interested in establishing the relationship between the inexact and exact spaces. In general, if  $\mathcal{X}$  and  $\mathcal{Y}$  are subspaces of  $\mathbb{C}^n$ , then the angle between subspaces  $\Theta(\mathcal{X}, \mathcal{Y})$  is defined as

$$\sup_{\mathbf{x} \in \mathcal{X}} \inf_{\mathbf{y} \in \mathcal{Y}} \frac{\|\mathbf{y} - \mathbf{x}\|}{\|\mathbf{x}\|} = \sin(\Theta(\mathcal{X}, \mathcal{Y})).$$

For the unpreconditioned case, an upper bound for the subspace angles,  $\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r)$  and  $\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r)$ , where  $\mathcal{V}_r, \widehat{\mathcal{V}}_r, \mathcal{W}_r$ , and  $\widehat{\mathcal{W}}_r$  are the subspaces associated with the matrices  $\mathbf{V}_r, \widehat{\mathbf{V}}_r, \mathbf{W}_r, \widehat{\mathbf{W}}_r$ , was given in [7]. The next three theorems illustrate that the result of [7] can be extended to the preconditioned case. We begin by presenting the theorem and proof for the case of split preconditioning using the same notation and reasoning as in [7]. The results and proofs associated with left and right preconditioning follow similarly; hence, we only state the result.

**Theorem 2.1.** *Let the columns of  $\mathbf{V}_r, \widehat{\mathbf{V}}_r, \mathbf{W}_r$  and  $\widehat{\mathbf{W}}_r$  be exact and approximate solutions to (2.1.1) and (2.1.2). Suppose approximate solutions are computed to a relative tolerance of  $\varepsilon$  using split preconditioning, so that the associated residuals,  $\boldsymbol{\eta}_i$  and  $\boldsymbol{\xi}_i$  as defined in (2.1.3) and (2.1.4), satisfy  $\|\boldsymbol{\eta}_i\| \leq \varepsilon \|\mathbf{L}_i^{-1} \mathbf{B}(\sigma_i) \mathbf{b}_i\|$  and  $\|\boldsymbol{\xi}_i\| \leq \varepsilon \|\mathbf{U}_i^{-T} \mathbf{C}(\sigma_i)^T \mathbf{c}_i\|$ . Denoting the associated subspaces as  $\mathcal{V}_r, \mathcal{W}_r, \widehat{\mathcal{V}}_r$ , and  $\widehat{\mathcal{W}}_r$ , then*

$$\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r \mathbf{D}_u)} \quad (2.2.1)$$

$$\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r \mathbf{D}_z)} \quad (2.2.2)$$

where  $\mathbf{D}_u = \text{diag}((\|\mathcal{K}(\sigma_1)^{-1} \mathbf{L}_1\| \|\mathbf{L}_1^{-1} \mathbf{B}(\sigma_1) \mathbf{b}_1\|)^{-1}, \dots, (\|\mathcal{K}(\sigma_r)^{-1} \mathbf{L}_r\| \|\mathbf{L}_r^{-1} \mathbf{B}(\sigma_r) \mathbf{b}_r\|)^{-1})$ ,

$\mathbf{D}_z = \text{diag}((\|\mathcal{K}(\sigma_1)^{-T}\mathbf{U}_1^T\| \|\mathbf{U}_1^{-T}\mathbf{C}(\sigma_1)^T\mathbf{c}_1\|)^{-1}, \dots, (\|\mathcal{K}(\sigma_r)^{-T}\mathbf{U}_r^T\| \|\mathbf{U}_r^{-T}\mathbf{C}(\sigma_r)^T\mathbf{c}_r\|)^{-1})$ , and

$$\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r \widetilde{\mathbf{D}}_u)} \max_i \|\mathbf{U}_i^{-1}\| \kappa_2(\mathbf{L}_i^{-1} \mathcal{K}(\sigma_i) \mathbf{U}_i^{-1}, \widehat{\mathbf{u}}_i) \quad (2.2.3)$$

$$\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r \widetilde{\mathbf{D}}_z)} \max_i \|\mathbf{L}_i^{-T}\| \kappa_2(\mathbf{U}^{-T} \mathcal{K}(\sigma_i)^T \mathbf{L}^{-T}, \widehat{\mathbf{z}}_i) \quad (2.2.4)$$

where  $\widetilde{\mathbf{D}}_u = \text{diag}(1/\|\widehat{\mathbf{u}}_1\|, \dots, 1/\|\widehat{\mathbf{u}}_r\|)$ ,  $\widetilde{\mathbf{D}}_z = \text{diag}(1/\|\widehat{\mathbf{z}}_1\|, \dots, 1/\|\widehat{\mathbf{z}}_r\|)$ ,

the quantities  $\kappa_2(\mathbf{L}_i^{-1} \mathcal{K}(\sigma_i) \mathbf{U}_i^{-1}, \widehat{\mathbf{u}}_i) = \frac{\|(\mathbf{U}_i \mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i\| \|\mathbf{L}_i^{-1} \mathcal{B}(\sigma_i) \mathbf{b}_i\|}{\|\widehat{\mathbf{u}}_i\|}$  and

$\kappa_2(\mathbf{U}_i^{-T} \mathcal{K}(\sigma_i)^T \mathbf{L}_i^{-T}, \widehat{\mathbf{z}}_i) = \frac{\|(\mathbf{L}_i^T \mathcal{K}(\sigma_i)^{-T} \mathbf{U}_i^T\| \|\mathbf{U}_i^{-T} \mathbf{C}(\sigma_i)^T \mathbf{c}_i\|}{\|\widehat{\mathbf{z}}_i\|}$  are the condition numbers of the  $i^{\text{th}}$

linear system, and  $\varsigma_{\min}(\mathbf{M})$  is the smallest singular value of the matrix  $\mathbf{M}$ .

*Proof.* Since the proofs are similar, we will only prove (2.2.1) and (2.2.3) by following an analogous argument as seen in [7].

Write  $\widehat{\mathbf{V}}_r = \mathbf{V}_r + \mathbf{E}$  with  $\mathbf{E} = [\mathcal{K}(\sigma_1)^{-1} \mathbf{L}_1 \boldsymbol{\eta}_1, \dots, \mathcal{K}(\sigma_r)^{-1} \mathbf{L}_r \boldsymbol{\eta}_r]$ . Then

$$\begin{aligned} \sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) &= \max_{\widehat{\mathbf{v}} \in \widehat{\mathcal{V}}_r} \min_{\mathbf{v} \in \mathcal{V}_r} \frac{\|\mathbf{v} - \widehat{\mathbf{v}}\|}{\|\widehat{\mathbf{v}}\|} \\ &= \max_{x_i} \min_{z_i} \frac{\|\sum_{i=1}^r z_i \mathcal{K}(\sigma_i)^{-1} \mathcal{B}(\sigma_i) \mathbf{b}_i - \sum_{i=1}^r x_i \widehat{\mathbf{v}}_i\|}{\|\sum_{i=1}^r x_i \widehat{\mathbf{v}}_i\|} \\ &= \max_{x_i} \min_{z_i} \frac{\|\sum_{i=1}^r (z_i - x_i) \mathcal{K}(\sigma_i)^{-1} \mathcal{B}(\sigma_i) \mathbf{b}_i - x_i \mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i \boldsymbol{\eta}_i\|}{\|\sum_{i=1}^r x_i \widehat{\mathbf{v}}_i\|} \\ &\leq \max_{x_i} \frac{\|\sum_{i=1}^r x_i \mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i \boldsymbol{\eta}_i\|}{\|\sum_{i=1}^r x_i \widehat{\mathbf{v}}_i\|} = \max_{\mathbf{x}} \frac{\|\mathbf{E} \mathbf{x}\|}{\|\widehat{\mathbf{V}}_r \mathbf{x}\|} = \max_{\mathbf{x}} \frac{\|\mathbf{E} \mathbf{D} \mathbf{x}\|}{\|\widehat{\mathbf{V}}_r \mathbf{D} \mathbf{x}\|} \end{aligned}$$

where  $\mathbf{D} = \text{diag}(d_1, \dots, d_r)$  is a positive definite diagonal matrix. We may bound the

numerator as follows:

$$\begin{aligned} \|\mathbf{ED}\mathbf{x}\| &\leq \|\mathbf{ED}\| \|\mathbf{x}\| \leq \sqrt{r} \|\mathbf{x}\| \max_i (d_i \|\mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i \boldsymbol{\eta}_i\|) \\ &\leq \sqrt{r} \|\mathbf{x}\| \max_i (d_i \|\mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i\| \|\boldsymbol{\eta}_i\|). \end{aligned}$$

This gives

$$\sin \Theta(\widehat{\mathbf{V}}_r, \mathbf{V}_r) \leq \sqrt{r} \frac{\max_i (d_i \|\mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i\| \|\boldsymbol{\eta}_i\|)}{\min_{\mathbf{x}} (\|\widehat{\mathbf{V}}_r \mathbf{D} \mathbf{x}\| / \|\mathbf{x}\|)} = \sqrt{r} \frac{\max_i (d_i \|\mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i\| \|\boldsymbol{\eta}_i\|)}{\zeta_{\min}(\widehat{\mathbf{V}}_r \mathbf{D})}. \quad (2.2.5)$$

While this bound holds for any  $d_i$ , we cite the *Column Equilibration Theorem* of van der Sluis [75] to argue that the optimal choice of the diagonal constant will satisfy

$$d_i \|\mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i\| \|\boldsymbol{\eta}_i\| = C$$

where the constant  $C$  is independent of  $i = 1, \dots, r$ . Since the residuals satisfy  $\|\boldsymbol{\eta}_i\| \approx \varepsilon \|\mathbf{L}_i^{-1} \mathcal{B}(\sigma_i) \mathbf{b}_i\|$ , we take  $C = \varepsilon$  and  $d_i = (\|\mathcal{K}(\sigma_i)^{-1} \mathbf{L}_i\| \|\mathbf{L}_i^{-1} \mathcal{B}(\sigma_i) \mathbf{b}_i\|)^{-1}$  to achieve the best bound. This leads to (2.2.1). To obtain a more computationally feasible upper bound, we may define the diagonal matrix  $\mathbf{D}$  to be  $\widetilde{\mathbf{D}}_u = \text{diag}(1/\|\widehat{\mathbf{u}}_1\|, \dots, 1/\|\widehat{\mathbf{u}}_r\|)$ , which leads to (2.2.3).  $\square$

**Theorem 2.2.** *Let the columns of  $\mathbf{V}_r$ ,  $\widehat{\mathbf{V}}_r$ ,  $\mathbf{W}_r$  and  $\widehat{\mathbf{W}}_r$  be exact and approximate solutions to (2.1.5) and (2.1.6). Suppose approximate solutions are computed to a relative tolerance of  $\varepsilon$ , so that the associated residuals,  $\boldsymbol{\eta}_i$  and  $\boldsymbol{\xi}_i$  as defined in (2.1.7) and (2.1.8), satisfy*



$\|\boldsymbol{\eta}_i\| \leq \varepsilon \|\mathbf{N}_i^{-1} \mathbf{B}(\sigma_i) \mathbf{b}_i\|$  and  $\|\boldsymbol{\xi}_i\| \leq \varepsilon \|\mathbf{N}_i^{-T} \mathbf{C}(\sigma_i)^T \mathbf{c}_i\|$ . Denoting the associated subspaces as  $\mathcal{V}_r, \mathcal{W}_r, \widehat{\mathcal{V}}_r$ , and  $\widehat{\mathcal{W}}_r$ , then

$$\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r \mathbf{D}_v)} \quad (2.2.6)$$

$$\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r \mathbf{D}_w)} \quad (2.2.7)$$

where  $\mathbf{D}_v = \text{diag}((\|\mathcal{K}(\sigma_1)^{-1} \mathbf{N}_1\| \|\mathbf{N}_1^{-1} \mathbf{B}(\sigma_1) \mathbf{b}_1\|)^{-1}, \dots, (\|\mathcal{K}(\sigma_r)^{-1} \mathbf{N}_r\| \|\mathbf{N}_r^{-1} \mathbf{B}(\sigma_r) \mathbf{b}_r\|)^{-1})$ ,

$\mathbf{D}_w = \text{diag}((\|\mathcal{K}(\sigma_1)^{-T} \mathbf{N}_1^T\| \|\mathbf{N}_1^{-T} \mathbf{C}(\sigma_1)^T \mathbf{c}_1\|)^{-1}, \dots, (\|\mathcal{K}(\sigma_r)^{-T} \mathbf{N}_r^T\| \|\mathbf{N}_r^{-T} \mathbf{C}(\sigma_r)^T \mathbf{c}_r\|)^{-1})$ , and

$$\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r \widetilde{\mathbf{D}}_v)} \max_i \kappa_2(\mathbf{N}_i^{-1} \mathcal{K}(\sigma_i), \widehat{\mathbf{v}}_i) \quad (2.2.8)$$

$$\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r \widetilde{\mathbf{D}}_w)} \max_i \kappa_2(\mathbf{N}_i^{-T} \mathcal{K}(\sigma_i)^T, \widehat{\mathbf{w}}_i) \quad (2.2.9)$$

where  $\widetilde{\mathbf{D}}_v = \text{diag}(1/\|\widehat{\mathbf{v}}_1\|, \dots, 1/\|\widehat{\mathbf{v}}_r\|)$ ,  $\widetilde{\mathbf{D}}_w = \text{diag}(1/\|\widehat{\mathbf{w}}_1\|, \dots, 1/\|\widehat{\mathbf{w}}_r\|)$ ,

the quantities  $\kappa_2(\mathbf{N}_i^{-1} \mathcal{K}(\sigma_i), \widehat{\mathbf{v}}_i) = \frac{\|\mathcal{K}(\sigma_i)^{-1} \mathbf{N}_i\| \|\mathbf{N}_i^{-1} \mathbf{B}(\sigma_i) \mathbf{b}_i\|}{\|\widehat{\mathbf{v}}_i\|}$  and

$\kappa_2(\mathbf{N}_i^{-T} \mathcal{K}(\sigma_i)^T, \widehat{\mathbf{w}}_i) = \frac{\|\mathcal{K}(\sigma_i)^{-T} \mathbf{N}_i^T\| \|\mathbf{N}_i^{-T} \mathbf{C}(\sigma_i)^T \mathbf{c}_i\|}{\|\widehat{\mathbf{w}}_i\|}$  are the condition numbers of the  $i^{\text{th}}$  linear

system, and  $\varsigma_{\min}(\mathbf{M})$  is the smallest singular value of the matrix  $\mathbf{M}$ .

*Proof.* The proof follows in a manner similar to the split preconditioned case once we write

$$\widehat{\mathbf{V}}_r = \mathbf{V}_r + \mathbf{E} \text{ with } \mathbf{E} = [\mathcal{K}(\sigma_1)^{-1} \mathbf{N}_1 \boldsymbol{\eta}_1, \dots, \mathcal{K}(\sigma_r)^{-1} \mathbf{N}_r \boldsymbol{\eta}_r]. \quad \square$$

**Theorem 2.3.** *Let the columns of  $\mathbf{V}_r$ ,  $\widehat{\mathbf{V}}_r$ ,  $\mathbf{W}_r$  and  $\widehat{\mathbf{W}}_r$  be exact and approximate solutions to the right preconditioned systems (2.1.9) and (2.1.10). Suppose approximate solutions are computed to a relative tolerance of  $\varepsilon$ , so that the associated residuals,  $\boldsymbol{\eta}_i$  and  $\boldsymbol{\xi}_i$  as defined in (2.1.11) and (2.1.12), satisfy  $\|\boldsymbol{\eta}_i\| \leq \varepsilon \|\mathcal{B}(\sigma_i)\mathbf{b}_i\|$  and  $\|\boldsymbol{\xi}_i\| \leq \varepsilon \|\mathcal{C}(\sigma_i)^T \mathbf{c}_i\|$ . Denoting the associated subspaces as  $\mathcal{V}_r, \mathcal{W}_r, \widehat{\mathcal{V}}_r$ , and  $\widehat{\mathcal{W}}_r$ , then*

$$\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r \mathbf{D}_v)} \quad (2.2.10)$$

$$\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r \mathbf{D}_w)} \quad (2.2.11)$$

where  $\mathbf{D}_v = \text{diag}((\|\mathcal{K}(\sigma_1)^{-1}\|\|\mathcal{B}(\sigma_1)\mathbf{b}_1\|)^{-1}, \dots, (\|\mathcal{K}(\sigma_r)^{-1}\|\|\mathcal{B}(\sigma_r)\mathbf{b}_r\|)^{-1})$ ,

$\mathbf{D}_w = \text{diag}((\|\mathcal{K}(\sigma_1)^{-T}\|\|\mathcal{C}(\sigma_1)^T \mathbf{c}_1\|)^{-1}, \dots, (\|\mathcal{K}(\sigma_r)^{-T}\|\|\mathcal{C}(\sigma_r)^T \mathbf{c}_r\|)^{-1})$ , and

$$\sin \Theta(\widehat{\mathcal{V}}_r, \mathcal{V}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r \widetilde{\mathbf{D}}_v)} \max_i \kappa_2(\mathcal{K}(\sigma_i), \widehat{\mathbf{v}}_i) \quad (2.2.12)$$

$$\sin \Theta(\widehat{\mathcal{W}}_r, \mathcal{W}_r) \leq \frac{\varepsilon \sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r \widetilde{\mathbf{D}}_w)} \max_i \kappa_2(\mathcal{K}(\sigma_i)^T, \widehat{\mathbf{w}}_i) \quad (2.2.13)$$

where  $\widetilde{\mathbf{D}}_v = \text{diag}(1/\|\widehat{\mathbf{v}}_1\|, \dots, 1/\|\widehat{\mathbf{v}}_r\|)$ ,  $\widetilde{\mathbf{D}}_w = \text{diag}(1/\|\widehat{\mathbf{w}}_1\|, \dots, 1/\|\widehat{\mathbf{w}}_r\|)$ ,

the quantities  $\kappa_2(\mathcal{K}(\sigma_i), \widehat{\mathbf{v}}_i) = \frac{\|\mathcal{K}(\sigma_i)^{-1}\|\|\mathcal{B}(\sigma_i)\mathbf{b}_i\|}{\|\widehat{\mathbf{v}}_i\|}$  and

$\kappa_2(\mathcal{K}(\sigma_i)^T, \widehat{\mathbf{w}}_i) = \frac{\|\mathcal{K}(\sigma_i)^{-T}\|\|\mathcal{C}(\sigma_i)^T \mathbf{c}_i\|}{\|\widehat{\mathbf{w}}_i\|}$  are the condition numbers of the  $i^{\text{th}}$  linear system,

and  $\varsigma_{\min}(\mathbf{M})$  is the smallest singular value of the matrix  $\mathbf{M}$ .

*Proof.* After writing  $\widehat{\mathbf{V}}_r = \mathbf{V}_r + \mathbf{E}$  with  $\mathbf{E} = [\mathfrak{K}(\sigma_1)^{-1}\boldsymbol{\eta}_1, \dots, \mathfrak{K}(\sigma_r)^{-1}\boldsymbol{\eta}_r]$ , the proof is similar to the split preconditioning case.  $\square$

### 2.2.1 Numerical Example of the Subspace Angles

These bounds illustrate that the stopping tolerance and the conditioning of the linear systems are two important factors involved in the angle between the inexact and exact subspaces. For all preconditioning techniques, the upper bound suggests that the difference between the inexact and exact subspaces will decrease by a factor of  $\varepsilon$  as the stopping tolerance decreases. Furthermore, the conditioning of the linear systems will also impact the decay in the upper bound as the tolerance decreases. This behavior is observed in our numerical data. For example, we computed the subspace angles associated with one-step of interpolatory model reduction for the Rail Model. The Rail Model emerges from a semi-discretized heat transfer problem for the optimal cooling of steel profiles. After a finite element discretization, we obtain a descriptor system of the form

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t)$$

where  $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n \times n}$  and  $\mathbf{B} \in \mathbb{R}^{n \times 7}, \mathbf{C} \in \mathbb{R}^{6 \times n}$  and  $n$  depends on the mesh width of the discretization. For more details, see [12] and [13]. We only present our data for the SISO system of order  $n = 1357$  that relates the sixth input to the second output of the system.

In Table 2.1, *good shift* and *poor shift* refer to shift selections that result in well-conditioned and poorly-conditioned linear systems, respectively. In the *good shift* case, we observe the  $\varepsilon$  decay as anticipated. These results are similar to those observed in the unpreconditioned case, implying that the preconditioned linear systems do not necessarily change the reduced-order model despite improving the convergence of the iterative solves.

Table 2.1: Rail 1357;  $r = 6$ ; BiCG with ILU Left Preconditioning;  $\sin(\Theta(\mathbf{V}_r, \widehat{\mathbf{V}}_r))$

Tolerance	Good shift	Poor shift
$1 \times 10^{-1}$	$2.72 \times 10^{-1}$	$1.00 \times 10^0$
$1 \times 10^{-2}$	$5.15 \times 10^{-2}$	$9.94 \times 10^{-1}$
$1 \times 10^{-3}$	$2.34 \times 10^{-3}$	$9.91 \times 10^{-1}$
$1 \times 10^{-4}$	$5.34 \times 10^{-4}$	$9.96 \times 10^{-1}$
$1 \times 10^{-5}$	$1.31 \times 10^{-5}$	$9.94 \times 10^{-1}$
$1 \times 10^{-6}$	$1.93 \times 10^{-6}$	$9.90 \times 10^{-1}$
$1 \times 10^{-7}$	$6.78 \times 10^{-7}$	$4.89 \times 10^{-1}$
$1 \times 10^{-8}$	$7.04 \times 10^{-9}$	$7.17 \times 10^{-1}$
$1 \times 10^{-9}$	$3.92 \times 10^{-9}$	$1.69 \times 10^{-1}$
$1 \times 10^{-10}$	$3.45 \times 10^{-10}$	$9.97 \times 10^{-3}$

## 2.3 Pointwise Error

Another main concern is to quantify the perturbation error introduced with inexact solves. In [7], upper bounds for the pointwise error are proven, and we use the notation and reasoning of [7] to extend these results to the preconditioned case. We consider perturbations in both

$\mathbf{B}$  and  $\mathbf{C}$  by defining the perturbed transfer functions,

$$\mathcal{H}_{\delta\mathbf{B}}(s) = \mathbf{C}(s)\mathcal{K}(s)^{-1}(\mathbf{B}(s) + \delta\mathbf{B}) \quad \text{and} \quad \mathcal{H}_{\delta\mathbf{C}}(s) = (\mathbf{C}(s) + \delta\mathbf{C})\mathcal{K}(s)^{-1}\mathbf{B}(s)$$

and define the following condition numbers associated with perturbations  $\delta\mathbf{B}$  and  $\delta\mathbf{C}$  at

$s = \sigma$  :

$$\begin{aligned} \text{cond}_{\mathbf{B}}(\mathcal{H}(\sigma)) &= \frac{\|\mathbf{C}(\sigma)\mathcal{K}(\sigma)^{-1}\| \|\mathbf{B}(\sigma)\|}{\|\mathcal{H}(\sigma)\|} \\ \text{cond}_{\mathbf{C}}(\mathcal{H}(\sigma)) &= \frac{\|\mathbf{C}(\sigma)\| \|\mathcal{K}(\sigma)^{-1}\mathbf{B}(\sigma)\|}{\|\mathcal{H}(\sigma)\|}. \end{aligned}$$

Furthermore, we define for values of  $s$  such that  $\mathcal{K}_r(s)$  and  $\widehat{\mathcal{K}}_r(s)$  are nonsingular the following functions:

$$\begin{aligned} \mathcal{P}_r(s) &= \mathcal{K}(s)\mathbf{V}_r\mathcal{K}_r(s)^{-1}\mathbf{W}_r^T, & \mathcal{Q}_r(s) &= \mathbf{V}_r\mathcal{K}_r(s)^{-1}\mathbf{W}_r^T\mathcal{K}(s), \\ \widehat{\mathcal{P}}_r(s) &= \mathcal{K}(s)\widehat{\mathbf{V}}_r\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T, & \widehat{\mathcal{Q}}_r(s) &= \widehat{\mathbf{V}}_r\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T\mathcal{K}(s). \end{aligned} \quad (2.3.1)$$

We note that  $\mathcal{P}_r(s)$ ,  $\mathcal{Q}_r(s)$ ,  $\widehat{\mathcal{P}}_r(s)$ , and  $\widehat{\mathcal{Q}}_r(s)$  are skew projectors and differentiable with respect to  $s$  with

$$\begin{aligned} \widehat{\mathcal{P}}_r'(s) &= \left(\mathbf{I} - \widehat{\mathcal{P}}_r\right) \mathcal{K}'(s)\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r, & \widehat{\mathcal{Q}}_r'(s) &= \widehat{\mathcal{Q}}_r\mathcal{K}(s)^{-1}\mathcal{K}'(s) \left(\mathbf{I} - \widehat{\mathcal{Q}}_r\right) \\ \mathcal{P}_r'(s) &= \left(\mathbf{I} - \mathcal{P}_r\right) \mathcal{K}'(s)\mathcal{K}(s)^{-1}\mathcal{P}_r, & \mathcal{Q}_r'(s) &= \mathcal{Q}_r\mathcal{K}(s)^{-1}\mathcal{K}'(s) \left(\mathbf{I} - \mathcal{Q}_r\right). \end{aligned} \quad (2.3.2)$$

The importance of these projectors becomes evident once we note that the pointwise transfer function error is expressed in terms of the projectors as follows:

$$\begin{aligned}
\mathcal{H}(s) - \widehat{\mathcal{H}}_r(s) &= \mathbf{C}(s)\mathcal{K}(s)^{-1} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(s) \right) \mathcal{B}(s) \\
&= \mathbf{C}(s) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(s) \right) \mathcal{K}(s)^{-1} \mathcal{B}(s) \\
&= \mathbf{C}(s) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(s) \right) \mathcal{K}(s)^{-1} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(s) \right) \mathcal{B}(s).
\end{aligned}$$

Moreover, the error in the transfer function's first derivative is expressed in terms of these projectors as

$$\begin{aligned}
\mathcal{H}'(s) - \widehat{\mathcal{H}}_r'(s) &= \frac{d}{ds} [\mathbf{C}(s)\mathcal{K}(s)^{-1}] \left( \mathbf{I} - \widehat{\mathcal{P}}_r(s) \right) \mathcal{B}(s) \\
&\quad + \mathbf{C}(s) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(s) \right) \frac{d}{ds} [\mathcal{K}(s)^{-1} \mathcal{B}(s)] \\
&\quad - \mathbf{C}(s) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(s) \right) \frac{d}{ds} [\mathcal{K}(s)^{-1}] \left( \mathbf{I} - \widehat{\mathcal{P}}_r(s) \right) \mathcal{B}(s).
\end{aligned} \tag{2.3.3}$$

In order to bound these errors, we define the following subspaces:

$$\begin{aligned}
\mathfrak{P}_r(s) &= \text{Ran } \mathcal{P}_r(s) = \text{Ran } \mathcal{K}(s)\mathbf{V}_r, & \mathfrak{Q}_r(s) &= \text{Ker } \left( \mathbf{W}_r^T \mathcal{K}(s) \right)^\perp, \\
\widehat{\mathfrak{P}}_r(s) &= \text{Ran } \widehat{\mathcal{P}}_r(s) = \text{Ran } \mathcal{K}(s)\widehat{\mathbf{V}}_r, & \widehat{\mathfrak{Q}}_r(s) &= \text{Ker } \left( \widehat{\mathbf{W}}_r^T \mathcal{K}(s) \right)^\perp, \\
\mathfrak{B}_m(s) &= \text{Ran } \mathcal{K}(s)^{-1} \mathcal{B}(s), & \mathfrak{C}_p(s) &= \text{Ker } \left( \mathbf{C}(s)\mathcal{K}(s)^{-1} \right)^\perp,
\end{aligned}$$

and note the following equalities:

$$\|\widehat{\mathcal{P}}_r(s)\| = \|\mathbf{I} - \widehat{\mathcal{P}}_r(s)\| = \frac{1}{\cos \Theta(\widehat{\mathfrak{P}}_r(s), \widehat{\mathcal{W}}_r)} \quad (2.3.4)$$

$$\|\widehat{\mathcal{Q}}_r(s)\| = \|\mathbf{I} - \widehat{\mathcal{Q}}_r(s)\| = \frac{1}{\cos \Theta(\widehat{\mathfrak{Q}}_r(s), \widehat{\mathcal{V}}_r)}. \quad (2.3.5)$$

Using these quantities, the next three theorems extend the results of [7] to the preconditioned case. It is important to emphasize that these results are very similar to the unpreconditioned case as shown in [7] once the appropriate residuals are included.

**Theorem 2.4.** *Given the full-order model  $\mathcal{H}(s) = \mathbf{C}(s)\mathcal{K}(s)^{-1}\mathbf{B}(s)$ , the interpolation points  $\{\sigma_j\}, \{\mu_i\} \in \mathbb{C}$  and the corresponding tangential directions,  $\{\mathbf{b}_i\} \in \mathbb{C}^m$  and  $\{\mathbf{c}_i\} \in \mathbb{C}^p$ , let the inexact interpolatory reduced model  $\widehat{\mathcal{H}}_r(s) = \widehat{\mathbf{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{B}}_r(s)$  be constructed using split preconditioning so that (2.1.1) - (2.1.4) hold. Then the pointwise tangential interpolation error is*

$$\frac{\|\widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j - \mathcal{H}(\sigma_j)\mathbf{b}_j\|}{\|\mathcal{H}(\sigma_j)\mathbf{b}_j\|} \leq \text{cond}_{\mathbf{B}}(\mathcal{H}(\sigma_j)\mathbf{b}_j) \frac{\sin \Theta(\mathbf{c}_p(\sigma_j), \widehat{\mathcal{W}}_r)}{\cos \Theta(\widehat{\mathfrak{P}}_r(\sigma_j), \widehat{\mathcal{W}}_r)} \frac{\|\mathbf{L}_j\boldsymbol{\eta}_j\|}{\|\mathbf{B}(\sigma_j)\mathbf{b}_j\|} \quad (2.3.6)$$

$$\frac{\|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T \mathcal{H}(\mu_i)\|}{\|\mathbf{c}_i^T \mathcal{H}(\mu_i)\|} \leq \text{cond}_{\mathbf{C}}(\mathbf{c}_i^T \mathcal{H}(\mu_i)) \frac{\sin \Theta(\mathfrak{B}_m(\mu_i), \widehat{\mathcal{V}}_r)}{\cos \Theta(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)} \frac{\|\mathbf{U}_i^T \boldsymbol{\xi}_i\|}{\|\mathbf{c}_i^T \mathbf{C}(\mu_i)\|}. \quad (2.3.7)$$

If  $\mu_i = \sigma_i$  then,

$$|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T \mathcal{H}(\mu_i)\mathbf{b}_i| \leq \frac{\|\mathcal{K}(\mu_i)^{-1}\| \|\mathbf{L}_i\boldsymbol{\eta}_i\| \|\mathbf{U}_i^T \boldsymbol{\xi}_i\|}{\max\left(\cos \Theta(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r), \cos \Theta(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)\right)}, \quad (2.3.8)$$

and

$$\begin{aligned}
|\mathbf{c}_i^T \mathcal{H}'(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}'_r(\mu_i) \mathbf{b}_i| &\leq M \left( \frac{\|\mathbf{L}_i \boldsymbol{\eta}_i\|}{\cos \Theta(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r)} + \frac{\|\mathbf{U}_i^T \boldsymbol{\xi}_i\|}{\cos \Theta(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)} \right. \\
&\quad \left. + \frac{\|\mathbf{L}_i \boldsymbol{\eta}_i\|}{\cos \Theta(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r)} \frac{\|\mathbf{U}_i^T \boldsymbol{\xi}_i\|}{\cos \Theta(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)} \right) \quad (2.3.9)
\end{aligned}$$

with  $M = \max\left(\left\|\frac{d}{ds} [\mathbf{c}_i^T \mathbf{C} \mathcal{K}^{-1}] \Big|_{\mu_i}\right\|, \left\|\frac{d}{ds} [\mathcal{K}^{-1} \mathbf{B} \mathbf{b}_i] \Big|_{\mu_i}\right\|, \left\|\frac{d}{ds} [\mathcal{K}^{-1}] \Big|_{\mu_i}\right\|\right)$ .

*Proof.* Using (2.1.3), we have that  $\widehat{\mathbf{v}}_j = \mathcal{K}(\sigma_j)^{-1}(\mathcal{B}(\sigma_j) \mathbf{b}_j + \mathbf{L}_j \boldsymbol{\eta}_j)$ , implying that  $\mathcal{K}(\sigma_j) \widehat{\mathbf{v}}_j = \mathcal{B}(\sigma_j) \mathbf{b}_j + \mathbf{L}_j \boldsymbol{\eta}_j \in \widehat{\mathfrak{P}}_r(\sigma_j)$ . Since  $\widehat{\mathfrak{P}}_r$  is a skew projector,  $(\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) (\mathcal{B}(\sigma_j) \mathbf{b}_j + \mathbf{L}_j \boldsymbol{\eta}_j) = \mathbf{0}$  or equivalently

$$(\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) \mathcal{B}(\sigma_j) \mathbf{b}_j = -(\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) \mathbf{L}_j \boldsymbol{\eta}_j. \quad (2.3.10)$$

We then define  $\widehat{\Pi}$  to be the orthogonal projector onto  $\widehat{\mathcal{W}}_r = \text{Ker}(\widehat{\mathfrak{P}}_r(s))^\perp$ , implying

$\mathbf{I} - \widehat{\mathfrak{P}}_r(s) = (\mathbf{I} - \widehat{\Pi}) (\mathbf{I} - \widehat{\mathfrak{P}}_r(s))$ . Also, define  $\Gamma$  to be an orthogonal projector onto  $\mathfrak{C}_p(\sigma_j)$ .

This implies that  $\text{Ran}(\mathbf{I} - \Gamma) = \text{Ker}(\mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1})$ , and so  $\mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1} = \mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1} \Gamma$ .

Using these observations, we can express the pointwise error as follows:

$$\begin{aligned}
\widehat{\mathcal{H}}_r(\sigma_j) \mathbf{b}_j - \mathcal{H}(\sigma_j) \mathbf{b}_j &= -\mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1} (\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) \mathcal{B}(\sigma_j) \mathbf{b}_j \\
&= \mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1} (\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) \mathbf{L}_j \boldsymbol{\eta}_j \quad (2.3.11)
\end{aligned}$$

$$= \mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1} (\mathbf{I} - \widehat{\Pi}) (\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) \mathbf{L}_j \boldsymbol{\eta}_j \quad (2.3.12)$$

$$= \mathbf{C}(\sigma_j) \mathcal{K}(\sigma_j)^{-1} \Gamma (\mathbf{I} - \widehat{\Pi}) (\mathbf{I} - \widehat{\mathfrak{P}}_r(\sigma_j)) \mathbf{L}_j \boldsymbol{\eta}_j. \quad (2.3.13)$$



Taking norms results in (2.3.6):

$$\begin{aligned} \|\widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j - \mathcal{H}(\sigma_j)\mathbf{b}_j\| &\leq \|(\mathbf{I} - \widehat{\Pi}) \Gamma (\mathbf{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1})^T\| \cdot \|\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\| \cdot \|\mathbf{L}_j\boldsymbol{\eta}_j\| \\ &\leq \|\mathbf{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\| \cdot \frac{\sin \Theta(\mathbf{C}_p(\sigma_j), \widehat{\mathcal{W}}_r)}{\cos \Theta(\widehat{\mathcal{P}}_r(\sigma_j), \widehat{\mathcal{W}}_r)} \cdot \|\mathbf{L}_j\boldsymbol{\eta}_j\|. \end{aligned}$$

To prove (2.3.7), we follow a similar reasoning using

$$\mathbf{c}_i^T \mathbf{C}(\mu_i) (\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)) = -\boldsymbol{\xi}_i^T \mathbf{U}_i (\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)) \quad (2.3.14)$$

and defining the orthogonal projector,  $\widehat{\boldsymbol{\Xi}}$ , onto  $\widehat{\mathcal{V}}_r = \text{Ran}(\widehat{\mathcal{Q}}_r(s))$ , which gives

$\mathbf{I} - \widehat{\mathcal{Q}}_r(s) = (\mathbf{I} - \widehat{\mathcal{Q}}_r(s)) (\mathbf{I} - \widehat{\boldsymbol{\Xi}})$ . This yields

$$\begin{aligned} \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T \mathcal{H}(\mu_i) &= \mathbf{c}_i^T \mathbf{C}(\mu_i) (\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)) \mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i) \\ &\leq \boldsymbol{\xi}_i^T \mathbf{U}_i (\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)) (\mathbf{I} - \widehat{\boldsymbol{\Xi}}) \mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i). \end{aligned}$$

Taking norms, we then have

$$\begin{aligned} \|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T \mathcal{H}(\mu_i)\| &\leq \|\mathbf{U}_i^T \boldsymbol{\xi}_i\| \cdot \|\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)\| \cdot \|(\mathbf{I} - \widehat{\boldsymbol{\Xi}}) \mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i)\| \\ &\leq \|\mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i)\| \cdot \frac{\sin \Theta(\mathcal{B}_m(\mu_i), \widehat{\mathcal{V}}_r)}{\cos \Theta(\widehat{\mathcal{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)} \cdot \|\mathbf{U}_i^T \boldsymbol{\xi}_i\|. \end{aligned}$$

If  $\mu_i = \sigma_i$ , then

$$\begin{aligned} \mathbf{c}_i^T \mathcal{H}(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) \mathbf{b}_i &= \mathbf{c}_i^T \mathbf{C}(\mu_i) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i) \right) \mathcal{K}(\mu_i)^{-1} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i) \right) \mathcal{B}(\mu_i) \mathbf{b}_i \\ &= \boldsymbol{\xi}_i^T \mathbf{U}_i \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i) \right) \mathcal{K}(\mu_i)^{-1} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i) \right) \mathbf{L}_i \boldsymbol{\eta}_i \\ &= \begin{cases} \boldsymbol{\xi}_i^T \mathbf{U}_i \mathcal{K}(\mu_i)^{-1} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i) \right) \mathbf{L}_i \boldsymbol{\eta}_i, & \text{or} \\ \boldsymbol{\xi}_i^T \mathbf{U}_i \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i) \right) \mathcal{K}(\mu_i)^{-1} \mathbf{L}_i \boldsymbol{\eta}_i, \end{cases} \end{aligned}$$

resulting in the following:

$$|\mathbf{c}_i^T \mathcal{H}(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) \mathbf{b}_i| \leq \|\mathbf{U}_i^T \boldsymbol{\xi}_i\| \cdot \|\mathbf{L}_i \boldsymbol{\eta}_i\| \cdot \|\mathcal{K}(\mu_i)^{-1}\| \cdot \|\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\|$$

and

$$|\mathbf{c}_i^T \mathcal{H}(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) \mathbf{b}_i| \leq \|\mathbf{U}_i^T \boldsymbol{\xi}_i\| \cdot \|\mathbf{L}_i \boldsymbol{\eta}_i\| \cdot \|\mathcal{K}(\mu_i)^{-1}\| \cdot \|\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)\|.$$

Combining these bounds, we obtain (2.3.8). The last inequality comes from (2.3.3) with

$s = \mu_i$ :

$$\begin{aligned} \mathbf{c}_i^T \mathcal{H}'(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r'(\mu_i) \mathbf{b}_i &= \frac{d}{ds} [\mathbf{c}_i^T \mathbf{C} \mathcal{K}^{-1}]|_{\mu_i} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i) \right) \mathcal{B}(\mu_i) \mathbf{b}_i \\ &\quad + \mathbf{c}_i^T \mathbf{C}(\mu_i) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i) \right) \frac{d}{ds} [\mathcal{K}^{-1} \mathcal{B} \mathbf{b}_i]|_{\mu_i} \\ &\quad - \mathbf{c}_i^T \mathbf{C}(\mu_i) \left( \mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i) \right) \frac{d}{ds} [\mathcal{K}^{-1}]|_{\mu_i} \left( \mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i) \right) \mathcal{B}(\mu_i) \mathbf{b}_i. \end{aligned}$$

Using the Cauchy-Schwarz inequality along with (2.3.10) and (2.3.14), the conclusion readily

follows once we note that

$$\begin{aligned}
& \left| \mathbf{c}_i^T \mathcal{H}'(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r'(\mu_i) \mathbf{b}_i \right| \leq \left| \frac{d}{ds} [\mathbf{c}_i^T \mathbf{C} \mathcal{K}^{-1}] \Big|_{\mu_i} (\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)) \mathbf{L}_i \boldsymbol{\eta}_i \right| \\
& \quad + \left| \boldsymbol{\xi}_i^T \mathbf{U}_i (\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)) \frac{d}{ds} [\mathcal{K}^{-1} \mathcal{B} \mathbf{b}_i] \Big|_{\mu_i} \right| \\
& \quad + \left| \boldsymbol{\xi}_i^T \mathbf{U}_i (\mathbf{I} - \widehat{\mathcal{Q}}_r(\mu_i)) \frac{d}{ds} [\mathcal{K}^{-1}] \Big|_{\mu_i} (\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)) \mathbf{L}_i \boldsymbol{\eta}_i \right| \\
& \leq \left\| \frac{d}{ds} [\mathbf{c}_i^T \mathbf{C} \mathcal{K}^{-1}] \Big|_{\mu_i} \right\| \cdot \frac{\|\mathbf{L}_i \boldsymbol{\eta}_i\|}{\cos \Theta(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r)} \\
& \quad + \frac{\|\mathbf{U}_i^T \boldsymbol{\xi}_i\|}{\cos \Theta(\widehat{\mathcal{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)} \cdot \left\| \frac{d}{ds} [\mathcal{K}^{-1} \mathcal{B} \mathbf{b}_i] \Big|_{\mu_i} \right\| \\
& \quad + \left\| \frac{d}{ds} [\mathcal{K}^{-1}] \Big|_{\mu_i} \right\| \cdot \frac{\|\mathbf{L}_i \boldsymbol{\eta}_i\|}{\cos \Theta(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r)} \frac{\|\mathbf{U}_i^T \boldsymbol{\xi}_i\|}{\cos \Theta(\widehat{\mathcal{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)}.
\end{aligned}$$

□

**Theorem 2.5.** *Given the full-order model  $\mathcal{H}(s) = \mathbf{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)$ , the interpolation points  $\{\sigma_j\}, \{\mu_i\} \in \mathbb{C}$  and the corresponding tangential directions,  $\{\mathbf{b}_i\} \in \mathbb{C}^m$  and  $\{\mathbf{c}_i\} \in \mathbb{C}^p$ , let the inexact interpolatory reduced model  $\widehat{\mathcal{H}}_r(s) = \widehat{\mathbf{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s)$  be constructed using left preconditioning so that (2.1.5) - (2.1.8) hold. Then the pointwise tangential interpolation error is*

$$\frac{\|\widehat{\mathcal{H}}_r(\sigma_j) \mathbf{b}_j - \mathcal{H}(\sigma_j) \mathbf{b}_j\|}{\|\mathcal{H}(\sigma_j) \mathbf{b}_j\|} \leq \text{cond}_{\mathcal{B}}(\mathcal{H}(\sigma_j) \mathbf{b}_j) \frac{\sin \Theta(\mathbf{C}_p(\sigma_j), \widehat{\mathcal{W}}_r)}{\cos \Theta(\widehat{\mathfrak{P}}_r(\sigma_j), \widehat{\mathcal{W}}_r)} \frac{\|\mathbf{N}_j \boldsymbol{\eta}_j\|}{\|\mathcal{B}(\sigma_j) \mathbf{b}_j\|} \quad (2.3.15)$$

$$\frac{\|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T \mathcal{H}(\mu_i)\|}{\|\mathbf{c}_i^T \mathcal{H}(\mu_i)\|} \leq \text{cond}_{\mathbf{C}}(\mathbf{c}_i^T \mathcal{H}(\mu_i)) \frac{\sin \Theta(\mathcal{B}_m(\mu_i), \widehat{\mathcal{V}}_r)}{\cos \Theta(\widehat{\mathcal{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r)} \frac{\|\mathbf{N}_i^T \boldsymbol{\xi}_i\|}{\|\mathbf{c}_i^T \mathbf{C}(\mu_i)\|}. \quad (2.3.16)$$

If  $\mu_i = \sigma_i$  then,

$$|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \mathcal{H}(\mu_i) \mathbf{b}_i| \leq \frac{\|\mathcal{K}(\mu_i)^{-1}\| \|\mathbf{N}_i \boldsymbol{\eta}_i\| \|\mathbf{N}_i^T \boldsymbol{\xi}_i\|}{\max\left(\cos \Theta\left(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r\right), \cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)\right)}, \quad (2.3.17)$$

and

$$\begin{aligned} |\mathbf{c}_i^T \mathcal{H}'(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r'(\mu_i) \mathbf{b}_i| \leq M & \left( \frac{\|\mathbf{N}_i \boldsymbol{\eta}_i\|}{\cos \Theta\left(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r\right)} + \frac{\|\mathbf{N}_i^T \boldsymbol{\xi}_i\|}{\cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)} \right. \\ & \left. + \frac{\|\mathbf{N}_i \boldsymbol{\eta}_i\|}{\cos \Theta\left(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r\right)} \frac{\|\mathbf{N}_i^T \boldsymbol{\xi}_i\|}{\cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)} \right) \end{aligned} \quad (2.3.18)$$

with  $M = \max\left(\left\| \frac{d}{ds} [\mathbf{c}_i^T \mathbf{C} \mathcal{K}^{-1}] \Big|_{\mu_i} \right\|, \left\| \frac{d}{ds} [\mathcal{K}^{-1} \mathbf{B} \mathbf{b}_i] \Big|_{\mu_i} \right\|, \left\| \frac{d}{ds} [\mathcal{K}^{-1}] \Big|_{\mu_i} \right\| \right)$ .

**Theorem 2.6.** *Given the full-order model  $\mathcal{H}(s) = \mathbf{C}(s) \mathcal{K}(s)^{-1} \mathbf{B}(s)$ , the interpolation points  $\{\sigma_j\}, \{\mu_i\} \in \mathbb{C}$  and the corresponding tangential directions,  $\{\mathbf{b}_i\} \in \mathbb{C}^m$  and  $\{\mathbf{c}_i\} \in \mathbb{C}^p$ , let the inexact interpolatory reduced model  $\widehat{\mathcal{H}}_r(s) = \widehat{\mathbf{C}}_r(s) \widehat{\mathcal{K}}_r(s)^{-1} \widehat{\mathbf{B}}_r(s)$  be constructed using right preconditioning so that (2.1.9) - (2.1.12). Then the pointwise tangential interpolation error is*

$$\frac{\|\widehat{\mathcal{H}}_r(\sigma_j) \mathbf{b}_j - \mathcal{H}(\sigma_j) \mathbf{b}_j\|}{\|\mathcal{H}(\sigma_j) \mathbf{b}_j\|} \leq \text{cond}_{\mathbf{B}}(\mathcal{H}(\sigma_j) \mathbf{b}_j) \frac{\sin \Theta\left(\mathbf{c}_p(\sigma_j), \widehat{\mathcal{W}}_r\right)}{\cos \Theta\left(\widehat{\mathfrak{P}}_r(\sigma_j), \widehat{\mathcal{W}}_r\right)} \frac{\|\boldsymbol{\eta}_j\|}{\|\mathbf{B}(\sigma_j) \mathbf{b}_j\|} \quad (2.3.19)$$

$$\frac{\|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T \mathcal{H}(\mu_i)\|}{\|\mathbf{c}_i^T \mathcal{H}(\mu_i)\|} \leq \text{cond}_{\mathbf{C}}(\mathbf{c}_i^T \mathcal{H}(\mu_i)) \frac{\sin \Theta\left(\mathbf{B}_m(\mu_i), \widehat{\mathcal{V}}_r\right)}{\cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)} \frac{\|\boldsymbol{\xi}_i\|}{\|\mathbf{c}_i^T \mathbf{C}(\mu_i)\|}. \quad (2.3.20)$$

If  $\mu_i = \sigma_i$  then,

$$|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \mathcal{H}(\mu_i) \mathbf{b}_i| \leq \frac{\|\mathcal{K}(\mu_i)^{-1}\| \|\boldsymbol{\eta}_i\| \|\boldsymbol{\xi}_i\|}{\max\left(\cos \Theta\left(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r\right), \cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)\right)} \quad (2.3.21)$$

and

$$\begin{aligned} |\mathbf{c}_i^T \mathcal{H}'(\mu_i) \mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}'_r(\mu_i) \mathbf{b}_i| \leq M & \left( \frac{\|\boldsymbol{\eta}_i\|}{\cos \Theta\left(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r\right)} + \frac{\|\boldsymbol{\xi}_i\|}{\cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)} \right. \\ & \left. + \frac{\|\boldsymbol{\eta}_i\|}{\cos \Theta\left(\widehat{\mathfrak{P}}_r(\mu_i), \widehat{\mathcal{W}}_r\right)} \frac{\|\boldsymbol{\xi}_i\|}{\cos \Theta\left(\widehat{\mathfrak{Q}}_r(\mu_i), \widehat{\mathcal{V}}_r\right)} \right) \end{aligned} \quad (2.3.22)$$

with  $M = \max\left(\left\|\frac{d}{ds} [\mathbf{c}_i^T \mathbf{C} \mathcal{K}^{-1}] \Big|_{\mu_i}\right\|, \left\|\frac{d}{ds} [\mathcal{K}^{-1} \mathbf{B} \mathbf{b}_i] \Big|_{\mu_i}\right\|, \left\|\frac{d}{ds} [\mathcal{K}^{-1}] \Big|_{\mu_i}\right\|\right).$

## 2.4 Backward Error

In this section, we examine the backward error emanating from the inexact solves. In solving the linear systems, we have the standard backward error result that  $\widehat{\mathbf{v}}_j$  and  $\widehat{\mathbf{w}}_j$  are exact solutions to systems with perturbed righthand sides, namely

$$\widehat{\mathbf{v}}_j = \mathcal{K}(\sigma_j)^{-1}(\boldsymbol{\eta}_j + \mathbf{B}(\sigma_j) \mathbf{b}_j) \quad \text{and} \quad \widehat{\mathbf{w}}_j = \mathcal{K}(\sigma_j)^{-T}(\boldsymbol{\xi}_j + \mathbf{C}(\sigma_j)^T \mathbf{c}_j).$$

As a result, the inexact reduced model,  $\widehat{\mathcal{H}}_r(s) = \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s)$ , exactly interpolates a perturbed full-order model:

$$\widehat{\mathcal{H}}(s) = (\boldsymbol{\xi}_j^T + \mathbf{c}_j^T \mathcal{C}(s))\mathcal{K}(s)^{-1}(\mathcal{B}(s)\mathbf{b}_j + \boldsymbol{\eta}_j)$$

at  $s = \sigma_j$ . It is important to note that  $\widehat{\mathcal{H}}_r(s)$  will interpolate a different  $\widehat{\mathcal{H}}(s)$  for each  $\sigma_1, \dots, \sigma_r$ . Due to this shift dependence, the authors of [7] suggest imposing a Petrov-Galerkin framework so that the backward error no longer depends on the particular shift selection. We follow the discussion of [7] to define the Petrov-Galerkin structure as follows: let  $\mathcal{P}_k$  and  $\mathcal{Q}_k$  be subspaces of  $\mathbb{C}^n$  with  $\mathcal{P}_k^\perp \cap \mathcal{Q}_k = \{0\}$ . Let  $\tilde{\mathbf{v}}_j \in \mathcal{P}_k$  and  $\tilde{\mathbf{w}}_j \in \mathcal{Q}_k$  be the inexact solutions of  $\mathcal{K}(\sigma_j)\mathbf{v}_j = \mathcal{B}(\sigma_j)\mathbf{b}_j$  and  $\mathcal{K}(\sigma_j)^T\mathbf{w}_j = \mathcal{C}(\sigma_j)^T\mathbf{c}_j$ , respectively. Then, the Petrov-Galerkin structure gives that

$$\mathcal{K}(\sigma_j)\tilde{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \in \mathcal{Q}_k^\perp \quad \text{and} \quad \mathcal{K}(\sigma_j)^T\tilde{\mathbf{w}}_j - \mathcal{C}(\sigma_j)^T\mathbf{c}_j \in \mathcal{P}_k^\perp. \quad (2.4.1)$$

Assuming the Petrov-Galerkin framework is present in the inexact solver, the authors of [7] prove that the computed inexact reduced-order model,  $\tilde{\mathcal{H}}_r(s) = \tilde{\mathcal{C}}_r(s)\tilde{\mathcal{K}}_r(s)^{-1}\tilde{\mathcal{B}}_r(s)$ , obtained via the Petrov-Galerkin process exactly tangentially interpolates the perturbed full-order model,  $\tilde{\mathcal{H}}(s) = \mathcal{C}(s)(\mathcal{K}(s) + \mathbf{Z})^{-1}\mathcal{B}(s)$ , at each  $\sigma_i$  where  $\mathbf{Z}$  is a rank  $2r$  perturbation matrix. Since  $\tilde{\mathcal{H}}_r(s)$  interpolates  $\tilde{\mathcal{H}}(s)$  for all of the interpolation points, this result is significant as it provides a more attractive backward error result. Moreover, it suggests that Petrov-Galerkin solvers, such as BiCG, have a distinct advantage over other solvers, such

as GMRES, when employed in the interpolatory model reduction setting. Of course, BiCG is notorious for its disadvantages of serious breakdowns and erratic convergence issues. See [69], [41], [5], and [74] for more details. In fact, for many models considered in the numerical studies for this dissertation, BiCG only converged if the system was preconditioned, implying that a preconditioned BiCG algorithm is not only desirable but oftentimes necessary. In the preconditioning process, we want to maintain a similar backward error result as in [7]. The next two theorems delineate an important deviation from the unpreconditioned case as they prove that the backward error result of [7] does not extend trivially when left and right preconditioning are present in the inexact solve. Instead of requiring a Petrov-Galerkin framework, which is readily available in such solvers as BiCG, the left and right preconditioned cases require orthogonality conditions that are not easily implemented numerically. Fortunately, the last theorem proves that split preconditioning provides a similar backward error result to the unpreconditioned case, namely, split preconditioning requires a Petrov-Galerkin framework as shown in Theorem 2.9. Therefore, these findings suggest a noteworthy distinction between preconditioning techniques with respect to the model reduction backward error.

### 2.4.1 Backward Error for Left and Right Preconditioning

The next two theorems state the conditions required for the computed inexact reduced-order model to exactly tangentially interpolate a perturbed full-order model for all of the interpolation points. Unfortunately, these conditions for left and right preconditioning require

rather awkward orthogonality requirements. Let  $\tilde{\boldsymbol{\eta}}_j$  and  $\tilde{\boldsymbol{\xi}}_j$  be the residuals associated with the solutions  $\tilde{\mathbf{v}}_j$  and  $\tilde{\mathbf{w}}_j$  obtained through a preconditioned inexact solve. For the case of left preconditioning, the necessary orthogonality condition for the backward error result is

$$\mathbf{N}_j \tilde{\mathbf{v}}_j \perp \tilde{\boldsymbol{\xi}}_j \quad \text{and} \quad \mathbf{N}_j^T \tilde{\mathbf{w}}_j \perp \tilde{\boldsymbol{\eta}}_j. \quad (2.4.2)$$

Meanwhile, the orthogonality requirement for the right preconditioned systems is

$$\mathbf{R}_j^{-1} \tilde{\mathbf{y}}_j \perp \tilde{\boldsymbol{\xi}}_j \quad \text{and} \quad \mathbf{R}_j^T \tilde{\mathbf{z}}_j \perp \tilde{\boldsymbol{\eta}}_j \quad (2.4.3)$$

where  $\tilde{\mathbf{v}}_j = \mathbf{R}_j^{-1} \tilde{\mathbf{y}}_j$  and  $\tilde{\mathbf{w}}_j = \mathbf{R}_j^{-T} \tilde{\mathbf{z}}_j$ . Conditions (2.4.2) and (2.4.3) are somewhat troublesome as they are not readily available in the implementation of inexact solvers. Therefore, these next two theorems are significant as they illustrate that the backward error result for the left and right preconditioning cases is not a trivial extension of the unpreconditioned case.

**Theorem 2.7.** *Given  $\mathcal{H}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)$ ,  $r$  interpolation points  $\{\sigma_j\}_{j=1}^r$ , and the tangential directions  $\{\mathbf{b}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i\}_{i=1}^r$ , let the inexact solutions  $\tilde{\mathbf{v}}_j$  for  $\mathcal{K}(\sigma_j)^{-1}\mathcal{B}(\sigma_j)\mathbf{b}_j$  and  $\tilde{\mathbf{w}}_j$  for  $\mathcal{K}(\sigma_j)^{-T}\mathcal{C}(\sigma_j)^T\mathbf{c}_j$  be obtained using a left preconditioner,  $\mathbf{N}_j$ . Let  $\tilde{\mathbf{V}}_r$  and  $\tilde{\mathbf{W}}_r$  denote the corresponding inexact interpolatory bases; i.e.*

$$\tilde{\mathbf{V}}_r = [ \tilde{\mathbf{v}}_1, \quad \cdots, \quad \tilde{\mathbf{v}}_r ] \quad \text{and} \quad \tilde{\mathbf{W}}_r = [ \tilde{\mathbf{w}}_1, \quad \cdots, \quad \tilde{\mathbf{w}}_r ]. \quad (2.4.4)$$



Define the residuals

$$\tilde{\boldsymbol{\eta}}_j = \mathbf{N}_j^{-1} \mathcal{K}(\sigma_j) \tilde{\mathbf{v}}_j - \mathbf{N}_j^{-1} \mathcal{B}(\sigma_j) \mathbf{b}_j, \quad \tilde{\boldsymbol{\xi}}_j = \mathbf{N}_j^{-T} \mathcal{K}(\sigma_j)^T \tilde{\mathbf{w}}_j - \mathbf{N}_j^{-T} \mathcal{C}(\sigma_j) \mathbf{c}_j, \quad (2.4.5)$$

where the residuals and inexact solutions satisfy the following orthogonality conditions:

$$\mathbf{N}_j \tilde{\mathbf{v}}_j \perp \tilde{\boldsymbol{\xi}}_j \quad \text{and} \quad \mathbf{N}_j^T \tilde{\mathbf{w}}_j \perp \tilde{\boldsymbol{\eta}}_j. \quad (2.4.6)$$

Also, define the residual matrices

$$\mathbf{R}_b = [\mathbf{N}_1 \tilde{\boldsymbol{\eta}}_1, \mathbf{N}_2 \tilde{\boldsymbol{\eta}}_2, \dots, \mathbf{N}_r \tilde{\boldsymbol{\eta}}_r], \quad \mathbf{R}_c = [\mathbf{N}_1^T \tilde{\boldsymbol{\xi}}_1, \mathbf{N}_2^T \tilde{\boldsymbol{\xi}}_2, \dots, \mathbf{N}_r^T \tilde{\boldsymbol{\xi}}_r], \quad (2.4.7)$$

and the rank- $2r$  matrix

$$\mathbf{F}_{2r} = \mathbf{R}_b (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \tilde{\mathbf{W}}_r^T + \tilde{\mathbf{V}}_r (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \mathbf{R}_c^T. \quad (2.4.8)$$

Also, let  $\tilde{\mathcal{H}}_r(s) = \tilde{\mathcal{C}}_r(s) \tilde{\mathcal{K}}_r(s)^{-1} \tilde{\mathcal{B}}_r(s)$  denote the computed inexact reduced model where

$$\tilde{\mathcal{K}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{K}(s) \tilde{\mathbf{V}}_r, \quad \tilde{\mathcal{B}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{B}(s), \quad \text{and} \quad \tilde{\mathcal{C}}_r(s) = \mathcal{C}(s) \tilde{\mathbf{V}}_r. \quad (2.4.9)$$

Then,  $\tilde{\mathcal{H}}_r(s)$  exactly tangentially interpolates the perturbed full-order model

$$\tilde{\mathcal{H}}_r(s) = \mathcal{C}(s) (\mathcal{K}(s) - \mathbf{F}_{2r})^{-1} \mathcal{B}(s), \quad (2.4.10)$$

for each  $\sigma_i$ , namely, for  $i = 1, \dots, r$ ,

$$\tilde{\mathcal{H}}(\sigma_i)\mathbf{b}_i = \tilde{\mathcal{H}}_r(\sigma_i)\mathbf{b}_i, \quad \mathbf{c}_i^T \tilde{\mathcal{H}}(\sigma_i) = \mathbf{c}_i^T \tilde{\mathcal{H}}_r(\sigma_i), \quad \text{and} \quad \mathbf{c}_i^T \tilde{\mathcal{H}}'(\sigma_i)\mathbf{b}_i = \mathbf{c}_i^T \tilde{\mathcal{H}}_r'(\sigma_i)\mathbf{b}_i.$$

*Proof.* In order for the inexact reduced-order model,  $\tilde{\mathcal{H}}_r(s)$ , to exactly tangentially interpolate  $\tilde{\mathcal{H}}_r(s)$ , we need

$$(\mathcal{K}(\sigma_i) - \mathbf{F}_{2r}) \tilde{\mathbf{v}}_i = \mathcal{B}(\sigma_i)\mathbf{b}_i \quad \text{and} \quad \tilde{\mathbf{w}}_i^T (\mathcal{K}(\sigma_i) - \mathbf{F}_{2r}) = \mathbf{c}_i^T \mathcal{C}(\sigma_i) \quad \text{for } i = 1, \dots, r,$$

or equivalently

$$\mathbf{F}_{2r} \tilde{\mathbf{v}}_j = \mathcal{K}(\sigma_j) \tilde{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \quad \text{and} \quad \tilde{\mathbf{w}}_j^T \mathbf{F}_{2r} = \tilde{\mathbf{w}}_j^T \mathcal{K}(\sigma_j) - \mathbf{c}_j^T \mathcal{C}(\sigma_j).$$

From (2.4.5), we have

$$\mathbf{N}_j \tilde{\boldsymbol{\eta}}_j = \mathcal{K}(\sigma_j) \tilde{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \quad \text{and} \quad \tilde{\boldsymbol{\xi}}_j^T \mathbf{N}_j = \tilde{\mathbf{w}}_j^T \mathcal{K}(\sigma_j) - \mathbf{c}_j^T \mathcal{C}(\sigma_j).$$

Therefore,  $\mathbf{F}_{2r}$  must satisfy

$$\mathbf{F}_{2r} \tilde{\mathbf{v}}_j = \mathbf{N}_j \tilde{\boldsymbol{\eta}}_j \quad \text{and} \quad \tilde{\mathbf{w}}_j^T \mathbf{F}_{2r} = \tilde{\boldsymbol{\xi}}_j^T \mathbf{N}_j.$$

In matrix form, these conditions are equivalent to  $\mathbf{F}_{2r} \tilde{\mathbf{V}}_r = \mathbf{R}_b$  and  $\tilde{\mathbf{W}}_r^T \mathbf{F}_{2r} = \mathbf{R}_c^T$ . The orthogonality assumptions in (2.4.6) guarantee that  $\tilde{\mathbf{W}}_r^T \mathbf{R}_b = \mathbf{0}$  and  $\mathbf{R}_c^T \tilde{\mathbf{V}}_r = \mathbf{0}$ . Note also

that

$$\tilde{\mathcal{K}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{K}(s) \tilde{\mathbf{V}}_r = \tilde{\mathbf{W}}_r^T (\mathcal{K}(s) - \mathbf{F}_{2r}) \tilde{\mathbf{V}}_r$$

since the orthogonality assumptions in (2.4.6) imply  $\tilde{\mathbf{W}}_r^T \mathbf{F}_{2r} \tilde{\mathbf{V}}_r = \mathbf{0}$ .  $\square$

It is important to note that the proof of this theorem is similar to the unpreconditioned case as presented in [7]. The important deviation from the unpreconditioned case, however, lies in how the residual matrices,  $\mathbf{R}_b$  and  $\mathbf{R}_c$ , and orthogonality conditions must be defined in order for the following to hold:

$$\tilde{\mathbf{W}}_r^T \mathbf{R}_b = \mathbf{0}, \quad \mathbf{R}_c^T \tilde{\mathbf{V}}_r = \mathbf{0}, \quad \text{and} \quad \tilde{\mathbf{W}}_r^T \mathbf{F}_{2r} \tilde{\mathbf{V}}_r = \mathbf{0} \quad (2.4.11)$$

in the left preconditioned case. Without these equalities holding, the computed reduced-order model will not exactly tangentially interpolate a perturbed dynamical system for all interpolation points. In the next theorem, the residual matrices,  $\mathbf{R}_b$  and  $\mathbf{R}_c$ , along with the orthogonality condition are modified to ensure that the equations in (2.4.11) are satisfied. The proof is similar to the previous argument used for the left preconditioned case, and so we only state the theorem.

**Theorem 2.8.** *Given  $\mathcal{H}(s) = \mathbf{C}(s)\mathcal{K}(s)^{-1}\mathbf{B}(s)$ ,  $r$  interpolation points  $\{\sigma_j\}_{j=1}^r$ , and the tangential directions  $\{\mathbf{b}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i\}_{i=1}^r$ , let the inexact solutions  $\tilde{\mathbf{v}}_j$  for  $\mathcal{K}(\sigma_j)^{-1}\mathbf{B}(\sigma_j)\mathbf{b}_j$  and  $\tilde{\mathbf{w}}_j$  for  $\mathcal{K}(\sigma_j)^{-T}\mathbf{C}(\sigma_j)^T\mathbf{c}_j$  be obtained using a right preconditioner,  $\mathbf{R}_j$ . Let  $\tilde{\mathbf{V}}_r$  and  $\tilde{\mathbf{W}}_r$*

denote the corresponding inexact interpolatory bases; i.e.

$$\tilde{\mathbf{V}}_r = [ \tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_r ] \quad \text{and} \quad \tilde{\mathbf{W}}_r = [ \tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_r ]. \quad (2.4.12)$$

Let the residuals be defined as in (2.1.11) and (2.1.12), where the residuals and inexact solutions satisfy the following orthogonality conditions:

$$\mathbf{R}_j^{-1} \tilde{\mathbf{y}}_j \perp \tilde{\boldsymbol{\xi}}_j \quad \text{and} \quad \mathbf{R}_j^{-T} \tilde{\mathbf{z}}_j \perp \tilde{\boldsymbol{\eta}}_j.$$

Also, define the residual matrices

$$\mathbf{R}_b = [ \tilde{\boldsymbol{\eta}}_1, \tilde{\boldsymbol{\eta}}_2, \dots, \tilde{\boldsymbol{\eta}}_r ], \quad \mathbf{R}_c = [ \tilde{\boldsymbol{\xi}}_1, \tilde{\boldsymbol{\xi}}_2, \dots, \tilde{\boldsymbol{\xi}}_r ], \quad (2.4.13)$$

and the rank- $2r$  matrix

$$\mathbf{F}_{2r} = \mathbf{R}_b (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \tilde{\mathbf{W}}_r^T + \tilde{\mathbf{V}}_r (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \mathbf{R}_c^T. \quad (2.4.14)$$

Also, let  $\tilde{\mathcal{H}}_r(s) = \tilde{\mathcal{C}}_r(s) \tilde{\mathcal{K}}_r(s)^{-1} \tilde{\mathcal{B}}_r(s)$  denote the computed inexact reduced model where

$$\tilde{\mathcal{K}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{K}(s) \tilde{\mathbf{V}}_r, \quad \tilde{\mathcal{B}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{B}(s), \quad \text{and} \quad \tilde{\mathcal{C}}_r(s) = \mathcal{C}(s) \tilde{\mathbf{V}}_r. \quad (2.4.15)$$

Then,  $\tilde{\mathcal{H}}_r(s)$  exactly tangentially interpolates the perturbed full-order model

$$\tilde{\mathcal{H}}_r(s) = \mathbf{C}(s)(\mathcal{K}(s) - \mathbf{F}_{2r})^{-1}\mathbf{B}(s), \quad (2.4.16)$$

for each  $\sigma_i$ , namely, for  $i = 1, \dots, r$ ,

$$\tilde{\mathcal{H}}(\sigma_i)\mathbf{b}_i = \tilde{\mathcal{H}}_r(\sigma_i)\mathbf{b}_i, \quad \mathbf{c}_i^T \tilde{\mathcal{H}}(\sigma_i) = \mathbf{c}_i^T \tilde{\mathcal{H}}_r(\sigma_i), \quad \text{and} \quad \mathbf{c}_i^T \tilde{\mathcal{H}}'(\sigma_i)\mathbf{b}_i = \mathbf{c}_i^T \tilde{\mathcal{H}}_r'(\sigma_i)\mathbf{b}_i.$$

## 2.4.2 Backward Error for Split Preconditioning

Perhaps the most important contribution of this chapter is the next theorem, which states and proves the presence of a backward error result when split preconditioning is employed in an inexact solver with an embedded Petrov-Galerkin framework. The proof of this result follows a similar argument as employed for left and right preconditioning; however, the residual matrices,  $\mathbf{R}_b$  and  $\mathbf{R}_c$ , are defined for split preconditioning so that only the Petrov-Galerkin framework is required. As a result, the absence of the awkward orthogonality conditions implies that split preconditioning offers an advantage in the interpolatory model reduction framework.

**Theorem 2.9.** *Given a full-order model  $\mathcal{H}(s) = \mathbf{C}(s)\mathcal{K}(s)^{-1}\mathbf{B}(s)$ , interpolation points  $\{\sigma_j\}_{j=1}^r$ , and tangential directions  $\{\mathbf{b}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i\}_{i=1}^r$ , let the inexact solutions  $\tilde{\mathbf{v}}_j$  for  $\mathcal{K}(\sigma_j)^{-1}\mathbf{B}(\sigma_j)\mathbf{b}_j$  and  $\tilde{\mathbf{w}}_j$  for  $\mathcal{K}(\sigma_j)^{-T}\mathbf{C}(\sigma_j)^T\mathbf{c}_j$  be obtained using split preconditioning in a Petrov-Galerkin framework as in (2.4.1). Let  $\tilde{\mathbf{V}}_r$  and  $\tilde{\mathbf{W}}_r$  denote the corresponding inexact*

interpolatory bases:

$$\tilde{\mathbf{V}}_r = [ \tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_r ] \quad \text{and} \quad \tilde{\mathbf{W}}_r = [ \tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_r ]. \quad (2.4.17)$$

Define the residuals

$$\tilde{\boldsymbol{\eta}}_j = \mathbf{L}_j^{-1} \mathcal{K}(\sigma_j) \mathbf{U}_j^{-1} \tilde{\mathbf{u}}_j - \mathbf{L}_j^{-1} \mathcal{B}(\sigma_j) \mathbf{b}_j, \quad \tilde{\boldsymbol{\xi}}_j = \mathbf{U}_j^{-T} \mathcal{K}(\sigma_j)^T \mathbf{L}_j^{-T} \tilde{\mathbf{z}}_j - \mathbf{U}_j^{-T} \mathcal{C}(\sigma_j)^T \mathbf{c}_j, \quad (2.4.18)$$

residual matrices

$$\mathbf{R}_b = [ \mathbf{L}_1 \tilde{\boldsymbol{\eta}}_1, \mathbf{L}_2 \tilde{\boldsymbol{\eta}}_2, \dots, \mathbf{L}_r \tilde{\boldsymbol{\eta}}_r ], \quad \mathbf{R}_c = [ \mathbf{U}_1^T \tilde{\boldsymbol{\xi}}_1, \mathbf{U}_2^T \tilde{\boldsymbol{\xi}}_2, \dots, \mathbf{U}_r^T \tilde{\boldsymbol{\xi}}_r ], \quad (2.4.19)$$

and the rank- $2r$  matrix

$$\mathbf{F}_{2r} = \mathbf{R}_b (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \tilde{\mathbf{W}}_r^T + \tilde{\mathbf{V}}_r (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \mathbf{R}_c^T. \quad (2.4.20)$$

Let  $\tilde{\mathcal{H}}_r(s) = \tilde{\mathcal{C}}_r(s) \tilde{\mathcal{K}}_r(s)^{-1} \tilde{\mathcal{B}}_r(s)$  denote the computed inexact reduced model via the Petrov-Galerkin process where

$$\tilde{\mathcal{K}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{K}(s) \tilde{\mathbf{V}}_r, \quad \tilde{\mathcal{B}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{B}(s), \quad \text{and} \quad \tilde{\mathcal{C}}_r(s) = \mathcal{C}(s) \tilde{\mathbf{V}}_r. \quad (2.4.21)$$

Then,  $\tilde{\mathcal{H}}_r(s)$  exactly tangentially interpolates the perturbed full-order model

$$\tilde{\mathcal{H}}(s) = \mathbf{C}(s)(\mathcal{K}(s) - \mathbf{F}_{2r})^{-1}\mathbf{B}(s) \quad (2.4.22)$$

at each  $\sigma_i$ , namely:

$$\tilde{\mathcal{H}}(\sigma_i)\mathbf{b}_i = \tilde{\mathcal{H}}_r(\sigma_i)\mathbf{b}_i, \quad \mathbf{c}_i^T \tilde{\mathcal{H}}(\sigma_i) = \mathbf{c}_i^T \tilde{\mathcal{H}}_r(\sigma_i),$$

$$\text{and } \mathbf{c}_i^T \tilde{\mathcal{H}}'(\sigma_i)\mathbf{b}_i = \mathbf{c}_i^T \tilde{\mathcal{H}}_r'(\sigma_i)\mathbf{b}_i \quad \text{for each } i = 1, \dots, r.$$

*Proof.* In order for the inexact reduced-order model,  $\tilde{\mathcal{H}}_r(s)$ , to exactly tangentially interpolate  $\tilde{\mathcal{H}}(s)$ , we need

$$(\mathcal{K}(\sigma_i) - \mathbf{F}_{2r})\tilde{\mathbf{v}}_i = \mathbf{B}(\sigma_i)\mathbf{b}_i \quad \text{and} \quad \tilde{\mathbf{w}}_i^T (\mathcal{K}(\sigma_i) - \mathbf{F}_{2r}) = \mathbf{c}_i^T \mathbf{C}(\sigma_i) \quad \text{for } i = 1, \dots, r,$$

or equivalently

$$\mathbf{F}_{2r}\tilde{\mathbf{v}}_j = \mathcal{K}(\sigma_j)\tilde{\mathbf{v}}_j - \mathbf{B}(\sigma_j)\mathbf{b}_j \quad \text{and} \quad \tilde{\mathbf{w}}_j^T \mathbf{F}_{2r} = \tilde{\mathbf{w}}_j^T \mathcal{K}(\sigma_j) - \mathbf{c}_j^T \mathbf{C}(\sigma_j).$$

From (2.4.18), we have

$$\mathbf{L}_j \tilde{\boldsymbol{\eta}}_j = \mathcal{K}(\sigma_j)\tilde{\mathbf{v}}_j - \mathbf{B}(\sigma_j)\mathbf{b}_j \quad \text{and} \quad \tilde{\boldsymbol{\xi}}_j^T \mathbf{U}_j = \tilde{\mathbf{w}}_j^T \mathcal{K}(\sigma_j) - \mathbf{c}_j^T \mathbf{C}(\sigma_j).$$

Therefore,  $\mathbf{F}_{2r}$  must satisfy

$$\mathbf{F}_{2r} \tilde{\mathbf{v}}_j = \mathbf{L}_j \tilde{\boldsymbol{\eta}}_j \quad \text{and} \quad \tilde{\mathbf{w}}_j^T \mathbf{F}_{2r} = \tilde{\boldsymbol{\xi}}_j^T \mathbf{U}_j,$$

which is equivalent to  $\mathbf{F}_{2r} \tilde{\mathbf{V}}_r = \mathbf{R}_b$  and  $\tilde{\mathbf{W}}_r^T \mathbf{F}_{2r} = \mathbf{R}_c^T$ . By the orthogonality assumptions, we have  $\tilde{\mathbf{W}}_r^T \mathbf{R}_b = \mathbf{0}$  and  $\mathbf{R}_c^T \tilde{\mathbf{V}}_r = \mathbf{0}$ . Note also that

$$\tilde{\mathcal{K}}_r(s) = \tilde{\mathbf{W}}_r^T \mathcal{K}(s) \tilde{\mathbf{V}}_r = \tilde{\mathbf{W}}_r^T (\mathcal{K}(s) - \mathbf{F}_{2r}) \tilde{\mathbf{V}}_r$$

since the orthogonality assumptions give  $\tilde{\mathbf{W}}_r^T \mathbf{F}_{2r} \tilde{\mathbf{V}}_r = \mathbf{0}$ . □

*Remark:* Since a similar backward error result as in [7] holds, other results from [7] can be extended to the case of split preconditioning. For example, an analogous result to Theorem 4.2 of [7] holds in the split preconditioned case, namely the rank  $2r$  perturbation term can be bounded as

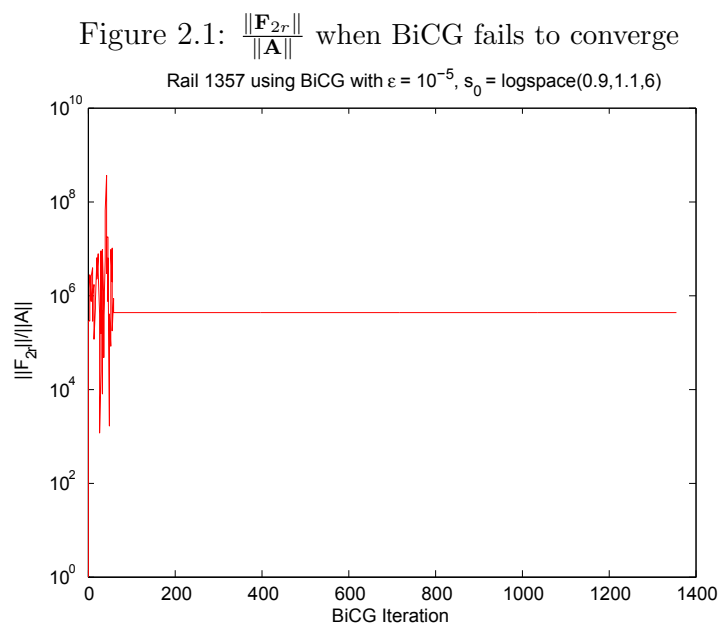
$$\|\mathbf{F}_{2r}\|_F \leq \sqrt{r} \|\tilde{\phi}_r\| \left( \max_i \|\mathbf{L}_i\| \frac{\|\boldsymbol{\eta}_i\|}{\|\tilde{\mathbf{v}}_i\|} \varsigma_{\min}(\tilde{\mathbf{V}}_r \mathbf{D}_v)^{-1} + \max_i \|\mathbf{U}_i\| \frac{\|\boldsymbol{\xi}_i\|}{\|\tilde{\mathbf{w}}_i\|} \varsigma_{\min}(\tilde{\mathbf{W}}_r \mathbf{D}_w)^{-1} \right),$$

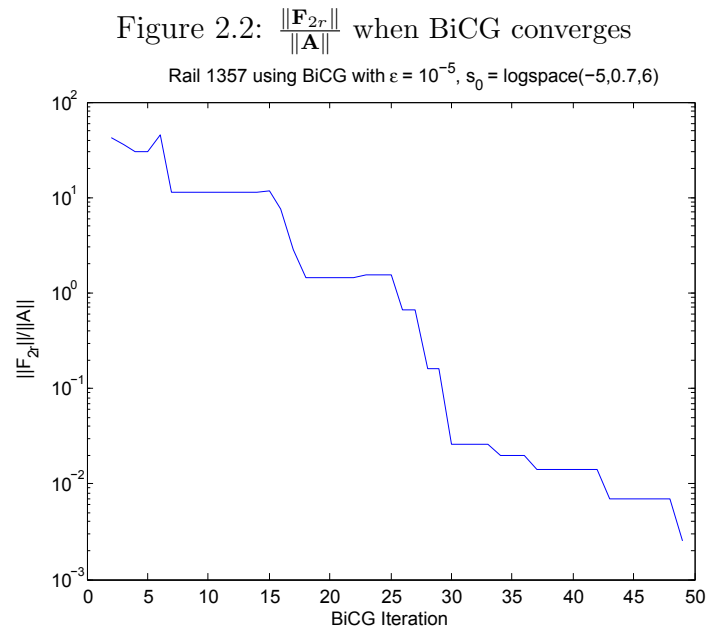
where  $\tilde{\phi}_r = \tilde{\mathbf{V}}_r (\tilde{\mathbf{W}}_r^T \tilde{\mathbf{V}}_r)^{-1} \tilde{\mathbf{W}}_r$ . Perhaps most importantly, the presence of a backward error result for the case of split preconditioning allows us to extend the results of [7] to a backward error result for the IRKA algorithm, namely the computed reduced-order model obtained in a Petrov-Galerkin framework via split preconditioning exactly tangentially interpolates a nearby full-order dynamical system,  $\tilde{\mathcal{H}}(s) = \mathbf{C}(s\mathbf{E} - (\mathbf{A} - \mathbf{F}_{2r}))^{-1} \mathbf{B}$ .



## 2.5 Properties of the Backward Error Term: $\mathbf{F}_{2r}$

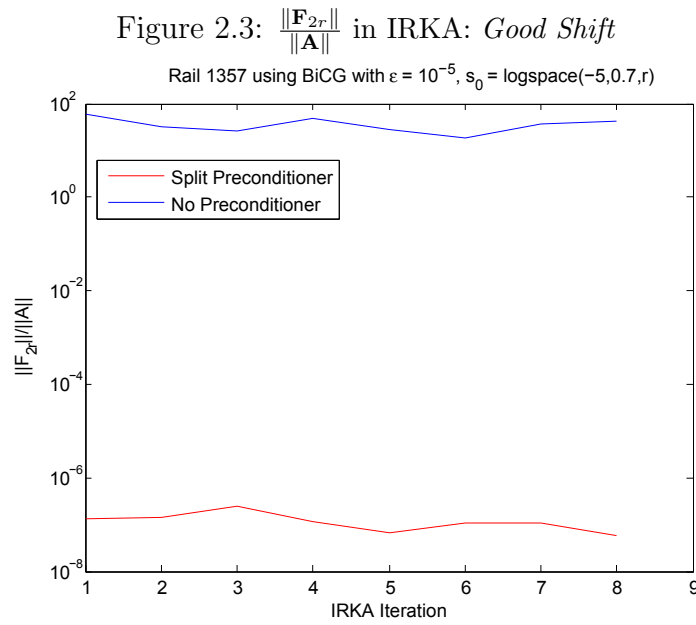
In this section, we study the backward error term by considering the Rail Model of dimension  $n = 1357$ , which was reduced to dimension  $r = 6$ , as discussed in Section 2.2.1. As mentioned previously, one of the key issues with BiCG is that it oftentimes fails to converge. Due to the presence of  $\mathbf{R}_b$  and  $\mathbf{R}_c$  in the  $\mathbf{F}_{2r}$  term, it follows that the size of  $\|\mathbf{F}_{2r}\|$  would decrease as the residuals decrease in the convergence of the iterative solve. We observe this expected decay in our numerical results. In Figure 2.1, BiCG experiences a near breakdown due to a poor initial guess, and we observe a stagnation of the  $\frac{\|\mathbf{F}_{2r}\|}{\|\mathbf{A}\|}$  term. Meanwhile, in Figure 2.2 a nice decay is shown as BiCG converges. It is important to emphasize that this data is representative of the results found when applying BiCG to several other models in addition to the Rail Model, namely the convergence of BiCG coincided with a decline in the  $\frac{\|\mathbf{F}_{2r}\|}{\|\mathbf{A}\|}$  quantity.

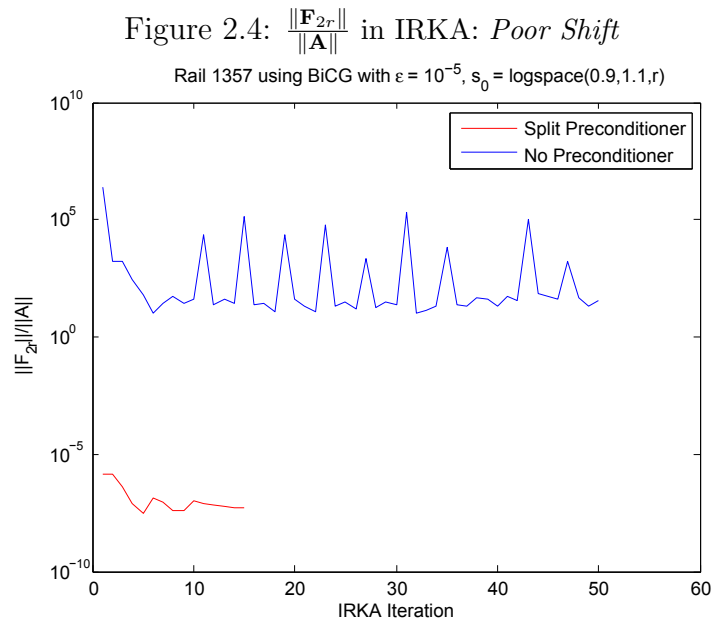




In Figure 2.3 and Figure 2.4, we present the  $\frac{\|\mathbf{F}_{2r}\|}{\|\mathbf{A}\|}$  quantity throughout the IRKA iteration for the unpreconditioned and split preconditioned cases. For a *good shift selection* as displayed in Figure 2.3, both unpreconditioned and preconditioned IRKA iterations converged in eight iterations. However, there is a noticeable difference of several orders in the  $\frac{\|\mathbf{F}_{2r}\|}{\|\mathbf{A}\|}$  term between the two cases. In Figure 2.4, the distinction between the preconditioned and unpreconditioned case is even more significant. With the *poor shift selection*, unpreconditioned IRKA failed to converge in fifty iterations. However, once the systems were preconditioned using an incomplete LU with a drop tolerance of 0.5, IRKA converged in fewer than twenty iterations. This convergence behavior is captured by the  $\|\mathbf{F}_{2r}\|$  term. As shown in Figure 2.3 and Figure 2.4, the  $\frac{\|\mathbf{F}_{2r}\|}{\|\mathbf{A}\|}$  term oscillated for the unpreconditioned case while remained smoother and smaller for the preconditioned case. This suggests that preconditioning improves the convergence of BiCG, and thereby the behavior of the  $\|\mathbf{F}_{2r}\|$  term. It is important

to emphasize that these results are very representative of those observed with other models and further indicate the importance of preconditioning.





# Chapter 3

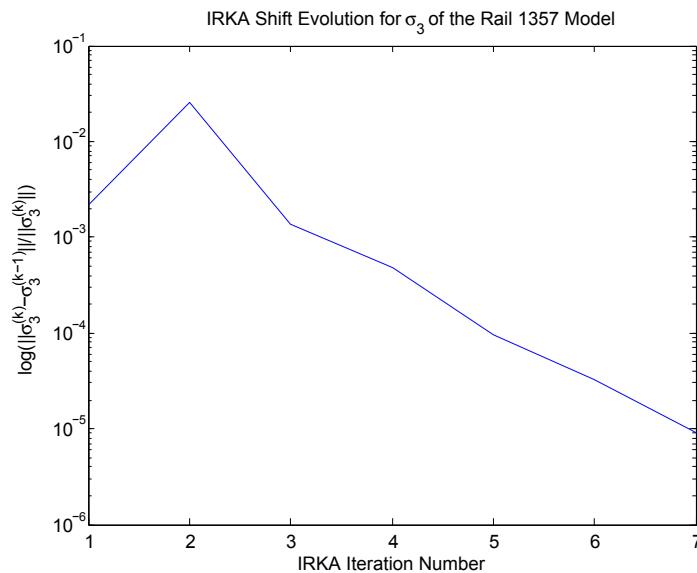
## Preconditioner Updates

**Collaboration Statement:** *This chapter is the result of collaboration with Dr. de Sturler.*

In this chapter, we design preconditioning techniques specifically for IRKA. As Chapter 2 illustrated, there are several advantages associated with Petrov-Galerkin solvers, namely the backward error result and the simultaneous computation of  $\mathbf{v}_i$  and  $\mathbf{w}_i$ . However, oftentimes the main drawbacks of Petrov-Galerkin solvers, namely serious breakdowns and erratic convergence, are only removed when the system is preconditioned. Even when the solver does not experience convergence difficulties, preconditioning for generic iterative solvers often becomes desirable to improve the speed of convergence. While the preconditioner improves convergence, it also adds to the computational cost of the algorithm. Since IRKA requires  $2r$  linear systems at each step, the cost associated with preconditioning every linear system becomes significant, especially for large systems. Fortunately, the convergence behavior of IRKA suggests that computing a new preconditioner for every linear system may be unnec-

essary. In Figure 3.1, for example, the relative difference from iteration  $k$  to iteration  $k - 1$  of IRKA for the Rail Model's third shift is given, which is representative of the behavior observed for the other shifts. Figure 3.1 displays that the shifts converge as IRKA converges, implying that a preconditioner from iteration  $k - 1$  of IRKA may also be appropriate for iteration  $k$  of IRKA. Moreover, the convergence of shifts implies that an update of an initial

Figure 3.1: IRKA Shift Evolution for  $\sigma_3$  of the Rail 1357 Model using BiCG



incomplete LU decomposition may only be required rather than completely computing a new incomplete LU decomposition. While improving the convergence of a sequence of systems has been studied, for example as in [65], the main aim of this chapter is to develop preconditioner update methods specific to the systems required by IRKA. To do so, we study two types of preconditioner updates, namely sparse approximate inverse updates (SAI) and an approach suggested in Bellavia et al. [11]. The beginning of the chapter is devoted to the first approach. In Section 3.2, we provide a brief overview of SAI updates and define

the associated problem in the context of interpolatory model reduction. We then present numerical results for three models in Section 3.3 and discuss how to efficiently update the preconditioner based on the difference of the shifts, the matrix  $\mathbf{E}$ , and the preconditioner update. The latter part of the chapter is devoted to the Bellavia et al. update [11]. While this update was originally proposed for the inverse of the Jacobian arising in the Newton iteration, the method and its associated theoretical results can be extended to interpolatory model reduction. One of the issues of the Bellavia et al. update is that oftentimes it is abandoned due to near singularity in the update; hence, we conclude the chapter with a discussion of the importance of strategically recomputing the incomplete LU decomposition during IRKA.

### 3.1 Preconditioner Updates

Preconditioners for a matrix  $\mathbf{A}$  are categorized as either explicit or implicit. An implicit preconditioner requires the solution of a linear system within each step of the iterative solve while explicit preconditioners rely on a known approximation for  $\mathbf{A}^{-1}$  and only require a matrix-vector product. Since applying an incomplete LU decomposition requires the solution of two sparse triangular solves, incomplete LU decompositions belong to the latter category. While the incomplete LU preconditioner has proven effective, IRKA requires  $2r$  linear systems and therefore up to  $2r$  preconditioners at every iteration. To mitigate these costs, we consider an explicit preconditioner as an update to an initial incomplete LU pre-

conditioner. For previous works discussing the update of a preconditioner, see [20], [37], [38], [1] and [78].

## 3.2 Sparse Approximate Inverse (SAI) Updates

The explicit preconditioner we consider is a sparse approximate inverse. In constructing a preconditioner  $\mathbf{P}$  for a matrix  $\mathbf{A}$ , we would like  $\mathbf{PA} \approx \mathbf{I}$  (for left preconditioning) and  $\mathbf{AP} \approx \mathbf{I}$  (for right preconditioning). SAI preconditioners minimize the associated error norm  $\|\mathbf{I} - \mathbf{PA}\|$  or  $\|\mathbf{I} - \mathbf{AP}\|$  for a given sparsity pattern. By choosing the Frobenius norm, we have

$$\|\mathbf{I} - \mathbf{AP}\|_F^2 = \sum_{i=1}^n \|\mathbf{e}_i - \mathbf{A}\mathbf{p}_i\|_2^2.$$

Hence, the minimization problem reduces to a linear least squares problem for each row or column of  $\mathbf{A}$  with the number of unknowns being equal to the number of nonzeros allowed in the row or column. As discussed in [28], [40], [58], and [60], the least squares problems may be solved with iterative or direct methods. For more details about these preconditioners, see [16], [14], [28], [30], [34], [40], [42], [43] and [59], and the references therein.

### 3.2.1 Sparse Approximate Inverse (SAI) Updates in IRKA

In this section, we define a SAI update for the sequence of linear systems that arise in interpolatory model reduction.



To do so, let  $\mathbf{K}_0 = \sigma_0 \mathbf{E} - \mathbf{A}$  and  $\mathbf{K}_k = \sigma_k \mathbf{E} - \mathbf{A}$  denote coefficient matrices. Let  $\mathbf{P}_0$  be a very good preconditioner for  $\mathbf{K}_0$ . In defining a cheap preconditioner update,  $\mathbf{P}_k$ , for  $\mathbf{K}_k$ , we would like  $\mathbf{K}_0 \mathbf{P}_0 \approx \mathbf{K}_k \mathbf{P}_k$ . Since we assume that  $\mathbf{P}_0$  is a good preconditioner,  $\mathbf{K}_0 \mathbf{P}_0 \approx \mathbf{K}_k \mathbf{P}_k$  implies that  $\mathbf{P}_k$  will be a good preconditioner for  $\mathbf{K}_k$  [1]. To this end, we express  $\mathbf{K}_k$  in terms of  $\mathbf{K}_0$ :

$$\mathbf{K}_k = \sigma_k \mathbf{E} - \mathbf{A} = \mathbf{K}_0 + (\sigma_k - \sigma_0) \mathbf{E} = \mathbf{K}_0 (\mathbf{I} + (\sigma_k - \sigma_0) \mathbf{K}_0^{-1} \mathbf{E}).$$

Then

$$\mathbf{K}_0 \mathbf{P}_0 = \mathbf{K}_0 (\mathbf{I} + (\sigma_k - \sigma_0) \mathbf{K}_0^{-1} \mathbf{E}) (\mathbf{I} + (\sigma_k - \sigma_0) \mathbf{K}_0^{-1} \mathbf{E})^{-1} \mathbf{P}_0 = \mathbf{K}_k \mathbf{P}_k$$

where  $\mathbf{K}_k = \mathbf{K}_0 (\mathbf{I} + (\sigma_k - \sigma_0) \mathbf{K}_0^{-1} \mathbf{E})$  and  $\mathbf{P}_k = (\mathbf{I} + (\sigma_k - \sigma_0) \mathbf{K}_0^{-1} \mathbf{E})^{-1} \mathbf{P}_0$ .

Define  $\mathbf{M}_k \approx (\mathbf{I} + (\sigma_k - \sigma_0) \mathbf{K}_0^{-1} \mathbf{E})^{-1}$ , then  $\mathbf{K}_0 \mathbf{P}_0 \approx \mathbf{K}_k \mathbf{M}_k \mathbf{P}_0$ , implying  $\mathbf{K}_0 \approx \mathbf{K}_k \mathbf{M}_k$ . At this point, we employ the motivation for the SAI computation where  $\mathbf{A} \mathbf{P} \approx \mathbf{I}$  is solved through minimizing  $\|\mathbf{I} - \mathbf{A} \mathbf{P}\|_F$ . Applying this reasoning to  $\mathbf{K}_0 \approx \mathbf{K}_k \mathbf{M}_k$  suggests a preconditioner update that satisfies

$$\min \|\mathbf{K}_k \mathbf{M}_k - \mathbf{K}_0\|_F^2 = \min \sum_{i=1}^n \|\mathbf{K}_k \mathbf{M}_k^i - \mathbf{K}_0^i\|_2^2. \quad (3.2.1)$$

where  $\mathbf{K}_0^i$  and  $\mathbf{M}_k^i$  denote the  $i^{\text{th}}$  columns of  $\mathbf{K}_0$  and  $\mathbf{M}_k$ , respectively. Although iterative methods are available to solve the least squares problems, we assume a direct solve is used.

Since  $\mathbf{K}_0\mathbf{P}_0 \approx \mathbf{K}_k\mathbf{M}_k\mathbf{P}_0$ , the system  $\mathbf{K}_k\mathbf{x} = \mathbf{b}$  is right preconditioned with  $\mathbf{P}_0$  and a SAI update as

$$\mathbf{K}_k\mathbf{M}_k\mathbf{P}_0\mathbf{x} = \mathbf{b}.$$

The idea of defining a preconditioner based on the Frobenius norm minimization has been suggested previously in several papers; see for example [28], [52], and [26]. In these papers, however, the goal is to improve an existing preconditioner whereas (3.2.1) aims to define a preconditioner for a new matrix.

### 3.3 Numerical Results for SAI Updates

#### 3.3.1 Models Studied

To study the SAI update's effectiveness, we considered three models, namely the Rail, CD and 1r models. While these models are small, they serve as proof of concepts examples. The Rail Model obtained from a heat transfer problem was previously discussed in Section 2.2.1. To initialize IRKA, we used six logarithmically spaced points in the interval  $[10^{-5}, 10^{0.7}]$ . IRKA converged with a  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  error of  $1.03 \times 10^{-3}$ , and a final shift set as shown in Table 3.1. Table 3.1 also provides  $\|\sigma^{(1)}\mathbf{E} - \mathbf{A}\|$  and  $\|\sigma^{(f)}\mathbf{E} - \mathbf{A}\|$  where  $\sigma^{(1)}$  and  $\sigma^{(f)}$  are the shifts from the first and final IRKA iterations, respectively. Among the studied models, the Rail Model involves the smallest  $\|\sigma^{(f)}\mathbf{E} - \mathbf{A}\|$  and  $\|\sigma^{(1)}\mathbf{E} - \mathbf{A}\|$  terms. While the optimal  $\mathcal{H}_2$

Table 3.1: Shift Information for the Rail Model

Final Shifts	$\ \sigma^{(1)}\mathbf{E} - \mathbf{A}\ $	$\ \sigma^{(f)}\mathbf{E} - \mathbf{A}\ $
$1.88 \times 10^{-5}$	$4.80 \times 10^{-5}$	$4.80 \times 10^{-5}$
$2.92 \times 10^{-4}$	$4.80 \times 10^{-5}$	$4.80 \times 10^{-5}$
$4.47 \times 10^{-3}$	$4.83 \times 10^{-5}$	$4.91 \times 10^{-5}$
$5.04 \times 10^{-2}$	$5.63 \times 10^{-5}$	$6.65 \times 10^{-5}$
$2.38 \times 10^{-1}$	$3.24 \times 10^{-4}$	$2.17 \times 10^{-4}$
$1.07 \times 10^0$	$4.35 \times 10^{-3}$	$9.40 \times 10^{-4}$

points for the Rail Model are real, the CD Model has complex poles. The CD Model is a SISO model of dimension 120 that describes the dynamics emerging from the lens actuator and the radial arm position of a CD player. Without assuming prior knowledge, we initialized with real shifts, namely forty points logarithmically spaced in the interval  $[10^0, 10^4]$ . The resulting  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  error is  $3.78 \times 10^{-6}$ . The final shifts as well as the norms of the coefficient matrices corresponding to the first and last IRKA iterations are given in Table 3.2. Finally, we studied the 1r Model. The 1r Model describes component 1r of the International Space Station using 270 states, 3 inputs and 3 outputs, and we reduced only the SISO model relating the first input to the first output. The shift initialization was 24 logarithmically spaced points in the interval  $[10^{-3}, 10^1]$ , and the resulting final shifts are given in Table 3.3. Also, in Table 3.3, we present  $\|\sigma^{(1)}\mathbf{E} - \mathbf{A}\|$  and  $\|\sigma^{(f)}\mathbf{E} - \mathbf{A}\|$  for the first and last IRKA iterations. The  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  error is  $1.75 \times 10^{-2}$ .

Table 3.2: Shift Information for the CD Model

Final Shifts	$\ \sigma^{(1)}\mathbf{E} - \mathbf{A}\ $	$\ \sigma^{(f)}\mathbf{E} - \mathbf{A}\ $
$2.24 \times 10^{-1} \pm 2.25 \times 10^1 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$4.60 \times 10^0 \pm 4.76 \times 10^1 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$6.36 \times 10^0 \pm 6.43 \times 10^1 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$7.27 \times 10^1 \pm 7.37 \times 10^1 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$7.54 \times 10^0 \pm 7.38 \times 10^1 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$7.83 \times 10^0 \pm 7.77 \times 10^1 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$1.97 \times 10^1 \pm 1.96 \times 10^2 i$	$4.33 \times 10^4$	$4.33 \times 10^4$
$2.18 \times 10^2 \pm 4.28 \times 10^2 i$	$4.33 \times 10^4$	$4.34 \times 10^4$
$1.22 \times 10^1 \pm 3.06 \times 10^2 i$	$4.33 \times 10^4$	$4.34 \times 10^4$
$2.75 \times 10^2 \pm 4.40 \times 10^2 i$	$4.33 \times 10^4$	$4.38 \times 10^4$
$1.16 \times 10^1 \pm 5.81 \times 10^2 i$	$4.33 \times 10^4$	$4.39 \times 10^4$
$1.27 \times 10^1 \pm 6.35 \times 10^2 i$	$4.33 \times 10^4$	$4.40 \times 10^4$
$1.32 \times 10^1 \pm 6.60 \times 10^2 i$	$4.33 \times 10^4$	$4.40 \times 10^4$
$5.30 \times 10^1 \pm 3.82 \times 10^3 i$	$4.33 \times 10^4$	$4.71 \times 10^4$
$7.63 \times 10^1 \pm 4.14 \times 10^3 i$	$4.33 \times 10^4$	$4.75 \times 10^4$
$2.10 \times 10^2 \pm 5.20 \times 10^3 i$	$4.33 \times 10^4$	$4.85 \times 10^4$
$1.62 \times 10^2 \pm 1.04 \times 10^4 i$	$4.34 \times 10^4$	$5.37 \times 10^4$
$1.81 \times 10^2 \pm 1.08 \times 10^4 i$	$4.34 \times 10^4$	$5.42 \times 10^4$
$1.20 \times 10^2 \pm 1.21 \times 10^4 i$	$4.35 \times 10^4$	$5.54 \times 10^4$
$5.12 \times 10^2 \pm 2.55 \times 10^4 i$	$4.38 \times 10^4$	$6.88 \times 10^4$
$4.33 \times 10^2 \pm 4.33 \times 10^4 i$	$4.46 \times 10^4$	$8.66 \times 10^4$

### 3.4 Effect of the SAI Update

In this section, we are interested in applying the SAI update proposed in Section 3.2.1 to answer two questions. First, we want to determine if the SAI update is competitive compared with recomputing a new incomplete LU decomposition. Also, we would like to study the effect of applying only one SAI update in the IRKA iteration. To do so, we consider the IRKA algorithm with three different types of preconditioning methods. For all of the algorithms, we assume that GMRES is the iterative solver and that a direct solve is used for the least

Table 3.3: Shift Information for the 1r Model

Final Shifts	$\ \sigma^{(1)}\mathbf{E} - \mathbf{A}\ $	$\ \sigma^{(f)}\mathbf{E} - \mathbf{A}\ $
$3.87 \times 10^{-3} \pm 7.75 \times 10^{-1}i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$9.96 \times 10^{-3} \pm 1.99 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$1.18 \times 10^{-2} \pm 2.29 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$1.56 \times 10^{-2} \pm 2.48 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$1.53 \times 10^{-2} \pm 2.56 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$1.95 \times 10^{-2} \pm 3.91 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$2.85 \times 10^{-2} \pm 5.62 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$4.25 \times 10^{-2} \pm 7.92 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$4.76 \times 10^{-2} \pm 9.22 \times 10^0i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$1.74 \times 10^{-1} \pm 3.49 \times 10^1i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$1.89 \times 10^{-1} \pm 3.79 \times 10^1i$	$3.76 \times 10^3$	$3.76 \times 10^3$
$2.76 \times 10^{-1} \pm 4.79 \times 10^1i$	$3.76 \times 10^3$	$3.76 \times 10^3$

squares problems associated with the SAI update. The sparsity pattern of the SAI update is taken to be that of  $\mathbf{A}^2$ . To compare methods, we will only use the convergence of GMRES. Ideally, our comparison would also include timings since the start-up costs vary between the update methods. However, due to the subtleties of MATLAB's compiled and interpreted code, we only present our comparison in terms of GMRES iterations. In Algorithm 3.4.1, IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ] is presented; this algorithm computes an incomplete LU decomposition for each  $\sigma_i\mathbf{E} - \mathbf{A}$  with a drop tolerance of  $10^{-8}$ . Especially since the drop tolerance is so small, this results in a computationally intensive algorithm that is expected to give convergence roughly similar to using an exact solve. Therefore, IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ] serves as a lower bound benchmark for the convergence of GMRES rather than as a practical method to be implemented in the large-scale setting.

**Algorithm 3.4.1. IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]: IRKA with Incomplete LU Decompositions**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .

2. Compute  $\mathbf{L}_i\mathbf{U}_i \approx (\sigma_i\mathbf{E} - \mathbf{A})$  for  $i = 1, \dots, r$ .

3. For  $i = 1, \dots, r$ , solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{U}_i^{-1}\mathbf{L}_i^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and  $(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{L}_i^{-T}\mathbf{U}_i^{-T}\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$ .

4.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  and  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .

5. while (not converged)

(a)  $\mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C}\mathbf{V}_r$ .

(b) Compute  $\mathbf{Y}^T\mathbf{A}_r\mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T\mathbf{E}_r\mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda\mathbf{E}_r - \mathbf{A}_r$ .

(c)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T\mathbf{Y}^T\mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r\mathbf{X}\mathbf{e}_i$ .

(d) Compute  $\mathbf{L}_i\mathbf{U}_i \approx (\sigma_i\mathbf{E} - \mathbf{A})$  for  $i = 1, \dots, r$ .

(e) For  $i = 1, \dots, r$ , solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{U}_i^{-1}\mathbf{L}_i^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and

$$(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{L}_i^{-T}\mathbf{U}_i^{-T}\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i.$$

(f)  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  and  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .

6.  $\mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C}\mathbf{V}_r$ .

We also considered IRKA with the SAI update implemented at every step after the first shift of the first IRKA iteration. The implementation details are given in IRKA[ $\mathbf{M}_i$ ] (Algorithm

3.4.2). We assume that the shifts are ordered from smallest to largest, so that the first shift corresponds to the smallest shift. With this algorithm, only one incomplete LU decomposition for the smallest shift is computed. The reasoning for computing the incomplete LU for the smallest shift is that we expect this shift to correspond to the hardest system to solve, implying this system benefits the most from a very good preconditioner. The remaining systems are solved using the SAI updates,  $\mathbf{M}_i$  and  $\mathbf{N}_i$ . If  $k$  is the number of IRKA iterations, this implies that  $2rnk - n$  least squares problems are required in addition to one incomplete LU decomposition. The work required for the least squares problems is independent of  $n$ , implying that the cost for the least squares problems remains linear in  $n$ .

**Algorithm 3.4.2. IRKA[ $\mathbf{M}_i$ ]: IRKA with  $\mathbf{M}_i$  SAI Updates**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2. Compute  $\mathbf{L}_1\mathbf{U}_1 \approx (\sigma_1\mathbf{E} - \mathbf{A})$ .
3. For  $i = 2, \dots, r$ , compute SAI updates,  $\mathbf{M}_i$  and  $\mathbf{N}_i$ , by solving
 
$$\min \|(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{M}_i - (\sigma_1\mathbf{E} - \mathbf{A})\|_F \text{ and } \min \|(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{N}_i - (\sigma_1\mathbf{E} - \mathbf{A})^T\|_F.$$
4. Solve  $(\sigma_1\mathbf{E} - \mathbf{A})\mathbf{U}_1^{-1}\mathbf{L}_1^{-1}\mathbf{v}_1 = \mathbf{B}\mathbf{b}_1$  and  $(\sigma_1\mathbf{E} - \mathbf{A})^T\mathbf{L}_1^{-T}\mathbf{U}_1^{-T}\mathbf{w}_1 = \mathbf{C}^T\mathbf{c}_1$ .
5. For  $i = 2, \dots, r$ , solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{M}_i\mathbf{U}_1^{-1}\mathbf{L}_1^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and
 
$$(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{N}_i\mathbf{L}_1^{-T}\mathbf{U}_1^{-T}\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i.$$
6.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  and  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .

7. *while (not converged)*

(a)  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ .

(b) Compute  $\mathbf{Y}^T \mathbf{A}_r \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_r \mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_r - \mathbf{A}_r$ .

(c)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ .

(d) For  $i = 1, \dots, r$ , compute SAI updates,  $\mathbf{M}_i$  and  $\mathbf{N}_i$ , by solving

$$\min \|(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{M}_i - (\sigma_1 \mathbf{E} - \mathbf{A})\|_F \text{ and } \min \|(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{N}_i - (\sigma_1 \mathbf{E} - \mathbf{A})^T\|_F.$$

(e) For  $i = 1, \dots, r$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{M}_i \mathbf{U}_1^{-1} \mathbf{L}_1^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and

$$(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{N}_i \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i.$$

(f)  $\mathbf{V}_r = [ \mathbf{v}_1, \dots, \mathbf{v}_r ]$  and  $\mathbf{W}_r = [ \mathbf{w}_1, \dots, \mathbf{w}_r ]$ .

8.  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ .

Due to the large number of least squares problems associated with IRKA[ $\mathbf{M}_i$ ], we also implemented IRKA[ $\mathbf{M}_{\frac{r}{2}+1}$ ] (Algorithm 3.4.3). In terms of the computations required in constructing the preconditioners, IRKA[ $\mathbf{M}_{\frac{r}{2}+1}$ ] involves substantially less work as it computes only one incomplete LU decomposition and one SAI update. For the first IRKA step, the incomplete LU is computed as  $\mathbf{L}_1 \mathbf{U}_1 \approx \sigma_1 \mathbf{E} - \mathbf{A}$  and an SAI update is computed for  $\sigma_{\frac{r}{2}+1} \mathbf{E} - \mathbf{A}$ . The first  $\frac{r}{2}$  systems are then preconditioned using  $\mathbf{L}_1 \mathbf{U}_1$  while the remaining systems are preconditioned using the SAI update. Throughout the remainder of the IRKA iteration, the incomplete LU and the SAI update are reused for all remaining systems in a similar manner.



**Algorithm 3.4.3. IRKA[ $M_{\frac{r}{2}+1}$ ]: IRKA with  $M_{\frac{r}{2}+1}$  SAI Update**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2. Compute  $\mathbf{L}_1 \mathbf{U}_1 \approx (\sigma_1 \mathbf{E} - \mathbf{A})$ .
3. Compute SAI updates,  $M_{\frac{r}{2}+1}$  and  $N_{\frac{r}{2}+1}$  by solving
 
$$\min \|(\sigma_{\frac{r}{2}+1} \mathbf{E} - \mathbf{A}) M_{\frac{r}{2}+1} - (\sigma_1 \mathbf{E} - \mathbf{A})\|_F \text{ and } \min \|(\sigma_{\frac{r}{2}+1} \mathbf{E} - \mathbf{A})^T N_{\frac{r}{2}+1} - (\sigma_1 \mathbf{E} - \mathbf{A})^T\|_F.$$
4. For  $i = 1, \dots, \frac{r}{2}$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_1^{-1} \mathbf{L}_1^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
5. For  $i = \frac{r}{2} + 1, \dots, r$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) M_{\frac{r}{2}+1} \mathbf{U}_1^{-1} \mathbf{L}_1^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T N_{\frac{r}{2}+1} \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
6.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  and  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .
7. while (not converged)
  - (a)  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ .
  - (b) Compute  $\mathbf{Y}^T \mathbf{A}_r \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_r \mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_r - \mathbf{A}_r$ .
  - (c)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ .
  - (d) For  $i = 1, \dots, \frac{r}{2}$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_1^{-1} \mathbf{L}_1^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .

(e) For  $i = \frac{r}{2} + 1, \dots, r$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A})\mathbf{M}_{\frac{r}{2}+1}\mathbf{U}_1^{-1}\mathbf{L}_1^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$

and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{N}_{\frac{r}{2}+1} \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .

(f)  $\mathbf{V}_r = [ \mathbf{v}_1, \dots, \mathbf{v}_r ]$  and  $\mathbf{W}_r = [ \mathbf{w}_1, \dots, \mathbf{w}_r ]$ .

8.  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ .

In Table 3.4, we present the total number of GMRES iterations for the computation of  $\mathbf{V}_r$  throughout the IRKA iteration using the proposed preconditioners. The results are representative of the behavior also observed in the computation of  $\mathbf{W}_r$ . Even though IRKA[ $\mathbf{M}_i$ ] for the Rail Model resulted in over five times as many GMRES iterations, IRKA[ $\mathbf{M}_i$ ] applied to the CD and 1r models yielded exactly the same number of GMRES iterations as IRKA[ $\mathbf{L}_i \mathbf{U}_i$ ]. This suggests that IRKA[ $\mathbf{M}_i$ ] is competitive with IRKA[ $\mathbf{L}_i \mathbf{U}_i$ ] in terms of GMRES iterations. Moreover, the computational cost of computing a new incomplete LU for every system is much too expensive while IRKA[ $\mathbf{M}_i$ ] involves a reasonable computational cost even in the large-scale dynamical setting. Nevertheless, IRKA[ $\mathbf{M}_i$ ] still requires  $n$  least squares problems for every update computed. Therefore, we ideally want to avoid computing a new  $\mathbf{M}_i$  for every shift. In IRKA[ $\mathbf{M}_{\frac{r}{2}+1}$ ], we consider only computing one SAI update and observe a substantial increase in GMRES iterations in Table 3.4. As a result, the aim of the next section is to determine a method to dynamically decrease the number of SAI updates computed without causing a substantial increase in GMRES iterations.

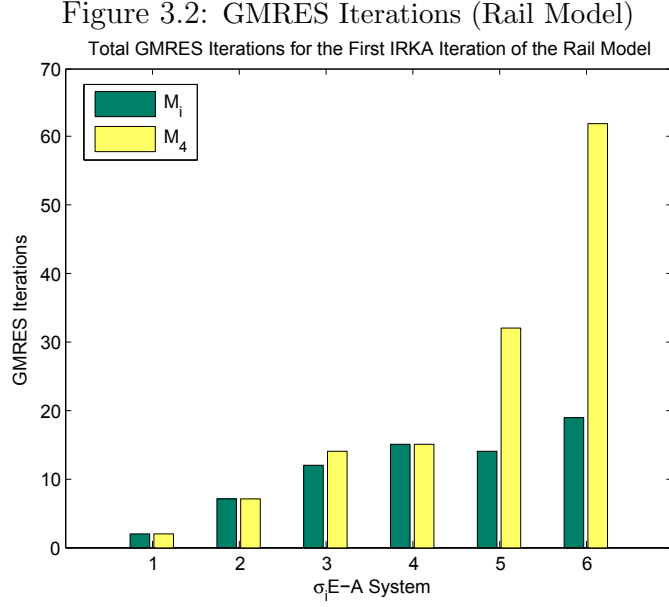
Table 3.4: Total GMRES Iterations for  $\mathbf{V}_r$  in IRKA

Method	Rail	CD	1r
$\mathbf{L}_i\mathbf{U}_i$	84	2000	624
$\mathbf{M}_i$	490	2000	624
$\mathbf{M}_{\frac{r}{2}+1}$	860	56799	25794

### 3.5 Using $\|\Delta\mathbf{E}\mathbf{M}_i\|$ to Update

To further investigate  $\text{IRKA}[\mathbf{M}_i]$  and  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$ , we report in Figures 3.2 - 3.4 the total number of GMRES iterations per linear system of the first IRKA iteration. In the results, we observe a steady increase in the number of GMRES iterations as the shifts move away from the initial shift, namely the shift for which the incomplete LU preconditioner is computed. In Figure 3.3, the CD Model shows a drop in GMRES iterations for the next five shifts after the SAI update  $\mathbf{M}_{\frac{r}{2}+1}$  is computed; although, a similar reduction is not observed for the Rail and 1r models. Nevertheless, the CD Model indicates that for certain models and shift selections the computation of just one additional  $\mathbf{M}_i$  preconditioner can noticeably reduce the number of GMRES iterations. Due to the variation in the effectiveness of the SAI update in reducing the number of GMRES iterations, the aim of this section is to develop a measure that determines when to compute a new SAI preconditioner.

One would perhaps expect that as the shifts  $\sigma_j$  are further from  $\sigma_{\frac{r}{2}+1}$ , the preconditioner quality would decrease. To quantify the relationship between shifts, let  $\Delta = |\sigma_p - \sigma_j|$  where  $\sigma_p$  is the shift for which the SAI preconditioner is computed. In Table 3.5, we report the relative  $\Delta$  values associated with the first iteration of IRKA. For example,  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  for

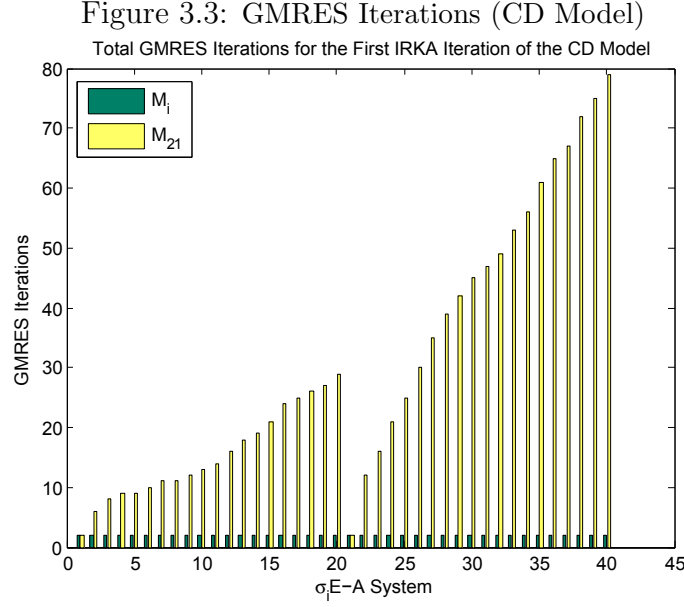


the CD Model required over 28 times more iterations than  $\text{IRKA}[\mathbf{M}_i]$  while  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  for the Rail Model resulted in roughly twice as many iterations as  $\text{IRKA}[\mathbf{M}_i]$ . However,  $\frac{|\sigma_p - \sigma_j|}{|\sigma_p|}$  is of the same order or larger for the Rail Model than  $\frac{|\sigma_p - \sigma_j|}{|\sigma_p|}$  for the CD Model. Therefore, Table 3.5 implies that the difference in shifts does not solely determine the preconditioner's quality. To find an appropriate measure, we first consider two coefficient matrices of the IRKA iteration:  $\mathbf{K}_i = \sigma_i\mathbf{E} - \mathbf{A}$  and  $\mathbf{K}_j = \sigma_j\mathbf{E} - \mathbf{A}$ . Let  $\mathbf{R}_i = \mathbf{K}_i\mathbf{M}_i - \mathbf{K}_0$ . The preconditioner  $\mathbf{M}_i$  is obtained by solving

$$\min \|\mathbf{K}_i\mathbf{M}_i - \mathbf{K}_0\|_F = \min \|\mathbf{R}_i\|_F.$$

If we consider using  $\mathbf{M}_i$  for  $\mathbf{K}_j$ , then

$$\mathbf{K}_j\mathbf{M}_i - \mathbf{K}_0 = \mathbf{K}_i\mathbf{M}_i + (\sigma_j - \sigma_i)\mathbf{E}\mathbf{M}_i - \mathbf{K}_0 = \mathbf{R}_i + \Delta\mathbf{E}\mathbf{M}_i.$$



Thus, we have

$$\|\mathbf{K}_j \mathbf{M}_i - \mathbf{K}_0\|_F \leq \|\mathbf{R}_i\|_F + \|\Delta \mathbf{E} \mathbf{M}_i\|_F,$$

suggesting that if  $\|\mathbf{R}_i\|_F$  and  $\|\Delta \mathbf{E} \mathbf{M}_i\|_F$  are small, then  $\|\mathbf{K}_j \mathbf{M}_i - \mathbf{K}_0\|_F$  will also be small.

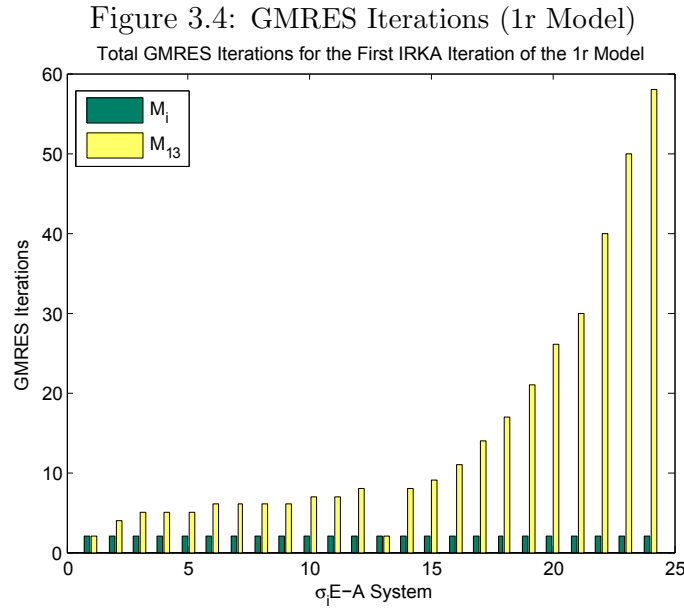
Using this observation, we return to the comparison of  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  and  $\text{IRKA}[\mathbf{M}_i]$ . To quantify the relationship between the methods, we define the following quantities:

$$F_G = \frac{\text{Total GMRES iterations using IRKA}[\mathbf{M}_{\frac{r}{2}+1}]}{\text{Total GMRES iterations using IRKA}[\mathbf{M}_i]}$$

and

$$F_M = \frac{\text{Total least squares problems to compute IRKA}[\mathbf{M}_{\frac{r}{2}+1}]}{\text{Total least squares problems to compute IRKA}[\mathbf{M}_i]}.$$

In Tables 3.6 - 3.8, we report the  $F_G$  factor in the first column. The second column gives  $\|\mathbf{R}_i + \Delta \mathbf{E} \mathbf{M}_i\|_F$ , which is bounded by the sum of the third and fourth columns. From



these tables, we note that for all models the number of GMRES iterations for  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  increases as  $\|\mathbf{R}_i + \Delta\mathbf{E}\mathbf{M}_i\|_F$  and  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F$  increases. Moreover, we notice that  $\|\mathbf{R}_i\|_F$  is small for all models, suggesting  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F$  is more relevant in predicting the behavior of the  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  method. Since we would like to minimize  $\|\mathbf{R}_i + \Delta\mathbf{E}\mathbf{M}_i\|_F$ , this suggests that a new SAI preconditioner update should be computed once  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F$  is large. Of course, one of the pivotal questions is how large should  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F$  be to merit computing a new preconditioner. From Tables 3.6 - 3.8, there does not seem to be an exact connection between the size of  $F_G$  and  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F$ . With the 1r Model, for example,  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F = 1.08 \times 10^2$  and  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  requires over 25 times as many iterations, whereas  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F = 3.06 \times 10^2$  and  $\text{IRKA}[\mathbf{M}_{\frac{r}{2}+1}]$  only requires 6 times as many iterations for the CD Model. To investigate the relationship between  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F$  and the convergence of GMRES, we compute a new SAI preconditioner only when  $\|\Delta\mathbf{E}\mathbf{M}_i\|_F > P_{tol}$ . In Table 3.9, the factors of additional GMRES

Table 3.5:  $\frac{|\sigma_p - \sigma_j|}{|\sigma_p|}$ 

Shift	Rail	CD	1r
$\sigma_{\frac{r}{2}+2}$	$1.28 \times 10^1$	$2.66 \times 10^{-1}$	$4.93 \times 10^{-1}$
$\sigma_{\frac{r}{2}+3}$	$1.90 \times 10^2$	$6.04 \times 10^{-1}$	1.23
$\sigma_{\frac{r}{2}+4}$	-	1.03	2.32
$\sigma_{\frac{r}{2}+5}$	-	1.57	3.96
$\sigma_{\frac{r}{2}+6}$	-	2.26	6.41
$\sigma_{\frac{r}{2}+7}$	-	3.12	$1.01 \times 10^1$
$\sigma_{\frac{r}{2}+8}$	-	4.22	$1.55 \times 10^1$
$\sigma_{\frac{r}{2}+9}$	-	5.61	$2.36 \times 10^1$
$\sigma_{\frac{r}{2}+10}$	-	7.38	$3.57 \times 10^1$
$\sigma_{\frac{r}{2}+11}$	-	9.61	$5.38 \times 10^1$
$\sigma_{\frac{r}{2}+12}$	-	$1.24 \times 10^1$	$8.09 \times 10^1$
$\sigma_{\frac{r}{2}+13}$	-	$1.60 \times 10^1$	-
$\sigma_{\frac{r}{2}+14}$	-	$2.05 \times 10^1$	-
$\sigma_{\frac{r}{2}+15}$	-	$2.63 \times 10^1$	-
$\sigma_{\frac{r}{2}+16}$	-	$3.36 \times 10^1$	-
$\sigma_{\frac{r}{2}+17}$	-	$4.28 \times 10^1$	-
$\sigma_{\frac{r}{2}+18}$	-	$5.44 \times 10^1$	-
$\sigma_{\frac{r}{2}+19}$	-	$6.92 \times 10^1$	-
$\sigma_{\frac{r}{2}+20}$	-	$8.79 \times 10^1$	-

iterations and preconditioner updates are presented. For the Rail Model,  $F_G$  remains roughly at 1 while there is a significant reduction in  $F_M$ . As a result, the Rail Model suggests that we can compute significantly fewer  $\mathbf{M}_i$  preconditioners without increasing the number of GMRES iterations. Meanwhile, the CD and 1r models give a different conclusion. In order for the factor of GMRES iterations to remain roughly at 1, a new preconditioner needs to be computed for almost every system. Otherwise, these models show that GMRES requires up to 40 times more iterations than with IRKA[ $\mathbf{M}_i$ ]. To avoid this increase in GMRES iterations, we note that for all models  $F_G$  is around one when  $P_{tol} \leq 10$ , suggesting that

Table 3.6:  $F_G$  and  $\|\mathbf{R}_i + \Delta\mathbf{EM}_i\|_F$  (Rail Model)

Shift	$F_G$	$\ \mathbf{R}_i + \Delta\mathbf{EM}_i\ _F$	$\ \mathbf{R}_i\ _F$	$\ \Delta\mathbf{EM}_i\ _F$
$\sigma_{\frac{r}{2}+1}$	1	$8.81 \times 10^{-6}$	$8.81 \times 10^{-6}$	0
$\sigma_{\frac{r}{2}+2}$	2.28	$1.28 \times 10^{-3}$	$8.81 \times 10^{-6}$	$1.28 \times 10^{-3}$
$\sigma_{\frac{r}{2}+3}$	3.26	$1.90 \times 10^{-2}$	$8.81 \times 10^{-6}$	$1.90 \times 10^{-2}$

larger  $P_{tol}$  values have the potential to yield a significant increase in the total number of GMRES iterations.

### 3.6 Bellavia et al. Updates

In this section, we consider updating the preconditioner through an approach suggested by Bellavia et al. in [11]. While [11] considers updating preconditioners for the Jacobian in Newton-Krylov methods, the authors mention that the results also apply to sequences of nonsymmetric linear systems. Similar work has been presented by [15], [18], [19], [20], [37], [38], and [63] for different types of systems as well. The idea is to begin with a factorized seed matrix and seed preconditioner; the seed preconditioner is then used to construct preconditioners for subsequent systems. In the interpolatory framework, we apply these updates to the sequences:

$$\mathbf{K}_{seed} = \sigma_0 \mathbf{E} - \mathbf{A}, \quad \mathbf{K}_1 = \sigma_1 \mathbf{E} - \mathbf{A}, \quad \dots, \quad \mathbf{K}_k = \sigma_k \mathbf{E} - \mathbf{A}.$$



Table 3.7:  $F_G$  and  $\|\mathbf{R}_i + \Delta\mathbf{EM}_i\|_F$  (CD Model)

Shift	$F_G$	$\ \mathbf{R}_i + \Delta\mathbf{EM}_i\ _F$	$\ \mathbf{R}_i\ _F$	$\ \Delta\mathbf{EM}_i\ _F$
$\sigma_{\frac{r}{2}+1}$	1	$2.84 \times 10^{-11}$	$2.84 \times 10^{-11}$	0
$\sigma_{\frac{r}{2}+2}$	6	$3.06 \times 10^2$	$2.84 \times 10^{-11}$	$3.06 \times 10^2$
$\sigma_{\frac{r}{2}+3}$	8	$6.93 \times 10^2$	$2.84 \times 10^{-11}$	$6.93 \times 10^2$
$\sigma_{\frac{r}{2}+4}$	10.5	$1.18 \times 10^3$	$2.84 \times 10^{-11}$	$1.18 \times 10^3$
$\sigma_{\frac{r}{2}+5}$	12.5	$1.80 \times 10^3$	$2.84 \times 10^{-11}$	$1.80 \times 10^3$
$\sigma_{\frac{r}{2}+6}$	15	$2.59 \times 10^3$	$2.84 \times 10^{-11}$	$2.59 \times 10^3$
$\sigma_{\frac{r}{2}+7}$	17.5	$3.59 \times 10^3$	$2.84 \times 10^{-11}$	$3.59 \times 10^3$
$\sigma_{\frac{r}{2}+8}$	19.5	$4.85 \times 10^3$	$2.84 \times 10^{-11}$	$4.85 \times 10^3$
$\sigma_{\frac{r}{2}+9}$	21	$6.45 \times 10^3$	$2.84 \times 10^{-11}$	$6.45 \times 10^3$
$\sigma_{\frac{r}{2}+10}$	22.5	$8.47 \times 10^3$	$2.84 \times 10^{-11}$	$8.47 \times 10^3$
$\sigma_{\frac{r}{2}+11}$	23.5	$1.10 \times 10^4$	$2.84 \times 10^{-11}$	$1.10 \times 10^4$
$\sigma_{\frac{r}{2}+12}$	24.5	$1.43 \times 10^4$	$2.84 \times 10^{-11}$	$1.43 \times 10^4$
$\sigma_{\frac{r}{2}+13}$	26.5	$1.84 \times 10^4$	$2.84 \times 10^{-11}$	$1.84 \times 10^4$
$\sigma_{\frac{r}{2}+14}$	28	$2.36 \times 10^4$	$2.84 \times 10^{-11}$	$2.36 \times 10^4$
$\sigma_{\frac{r}{2}+15}$	30.5	$3.02 \times 10^4$	$2.84 \times 10^{-11}$	$3.02 \times 10^4$
$\sigma_{\frac{r}{2}+16}$	32.5	$3.85 \times 10^4$	$2.84 \times 10^{-11}$	$3.85 \times 10^4$
$\sigma_{\frac{r}{2}+17}$	33.5	$4.91 \times 10^4$	$2.84 \times 10^{-11}$	$4.91 \times 10^4$
$\sigma_{\frac{r}{2}+18}$	36	$6.25 \times 10^4$	$2.84 \times 10^{-11}$	$6.25 \times 10^4$
$\sigma_{\frac{r}{2}+19}$	37.5	$7.94 \times 10^4$	$2.84 \times 10^{-11}$	$7.94 \times 10^4$
$\sigma_{\frac{r}{2}+20}$	39.5	$1.01 \times 10^5$	$2.84 \times 10^{-11}$	$1.01 \times 10^5$

Following the notation of [11], we let

$$\mathbf{K}_{seed}^{-1} = \mathbf{W}\mathbf{D}^{-1}\mathbf{Z}^T$$

where  $\mathbf{D}$  is a diagonal matrix and  $\mathbf{W}$  and  $\mathbf{Z}$  are upper triangular matrices with unitary diagonals. We then let

$$\Delta_k = \mathbf{K}_k - \mathbf{K}_{seed} = (\sigma_k - \sigma_0)\mathbf{E}$$

Table 3.8:  $F_G$  and  $\|\mathbf{R}_i + \Delta\mathbf{EM}_i\|_F$  (1r Model)

Shift	$F_G$	$\ \mathbf{R}_i + \Delta\mathbf{EM}_i\ _F$	$\ \mathbf{R}_i\ _F$	$\ \Delta\mathbf{EM}_i\ _F$
$\sigma_{\frac{r}{2}+1}$	1	$4.50 \times 10^{-12}$	$4.50 \times 10^{-12}$	0
$\sigma_{\frac{r}{2}+2}$	4	$9.92 \times 10^{-1}$	$4.50 \times 10^{-12}$	$9.92 \times 10^{-1}$
$\sigma_{\frac{r}{2}+3}$	4.5	2.47	$4.50 \times 10^{-12}$	2.47
$\sigma_{\frac{r}{2}+4}$	5.5	4.68	$4.50 \times 10^{-12}$	4.68
$\sigma_{\frac{r}{2}+5}$	7	7.98	$4.50 \times 10^{-12}$	7.98
$\sigma_{\frac{r}{2}+6}$	8.5	$1.29 \times 10^1$	$4.50 \times 10^{-12}$	$1.29 \times 10^1$
$\sigma_{\frac{r}{2}+7}$	10.5	$2.02 \times 10^1$	$4.50 \times 10^{-12}$	$2.02 \times 10^1$
$\sigma_{\frac{r}{2}+8}$	13	$3.12 \times 10^1$	$4.50 \times 10^{-12}$	$3.12 \times 10^1$
$\sigma_{\frac{r}{2}+9}$	15	$4.76 \times 10^1$	$4.50 \times 10^{-12}$	$4.76 \times 10^1$
$\sigma_{\frac{r}{2}+10}$	20	$7.20 \times 10^1$	$4.50 \times 10^{-12}$	$7.20 \times 10^1$
$\sigma_{\frac{r}{2}+11}$	25	$1.08 \times 10^2$	$4.50 \times 10^{-12}$	$1.08 \times 10^2$
$\sigma_{\frac{r}{2}+12}$	29	$1.63 \times 10^2$	$4.50 \times 10^{-12}$	$1.63 \times 10^2$

and

$$\mathbf{E}_k = \mathbf{Z}^T \Delta_k \mathbf{W}.$$

Then the inverse of the  $\mathbf{K}_k$  matrix is given as follows:

$$\mathbf{K}_k^{-1} = (\mathbf{K}_{seed} + \Delta_k)^{-1} = (\mathbf{Z}^{-T} \mathbf{D} \mathbf{W}^{-1} + \mathbf{Z}^{-T} \mathbf{E}_k \mathbf{W}^{-1})^{-1} = \mathbf{W}(\mathbf{D} + \mathbf{E}_k)^{-1} \mathbf{Z}^T. \quad (3.6.1)$$

Since  $\mathbf{W}$  and  $\mathbf{Z}$  tend to be dense, [11] suggests using sparse approximations for these quantities in order to mitigate the computational costs. These sparse matrices are then used to create a computationally feasible preconditioner. Let  $\tilde{\mathbf{D}}$  be a nonsingular diagonal approximation to  $\mathbf{D}$  while  $\tilde{\mathbf{W}}$  and  $\tilde{\mathbf{Z}}$  are sparse approximations to  $\mathbf{W}$  and  $\mathbf{Z}$ , respectively. Then a preconditioner for  $\mathbf{K}_{seed}$  is

$$\mathbf{P}_{seed} = \tilde{\mathbf{W}} \tilde{\mathbf{D}}^{-1} \tilde{\mathbf{Z}}^T, \quad (3.6.2)$$

Table 3.9: Factor of Additional GMRES Iterations and Preconditioners

$P_{tol}$	Rail $F_G$	Rail $F_M$	CD $F_G$	CD $F_M$	1r $F_G$	1r $F_M$
$10^{-5}$	1.00	0.71	1.00	1.00	1.00	1.00
$10^{-4}$	1.01	0.43	1.00	1.00	1.00	1.00
$10^{-3}$	1.11	0.14	1.00	1.00	1.00	1.00
$10^{-2}$	1.11	0.14	1.00	1.00	1.00	1.00
$10^{-1}$	1.11	0.14	1.00	1.00	1.01	0.99
$10^0$	1.11	0.14	1.00	1.00	1.02	0.99
$10^1$	1.11	0.14	1.00	1.00	1.02	0.99
$10^2$	1.11	0.14	1.03	0.99	14.02	0.44
$10^3$	1.11	0.14	4.88	0.70	26.39	0.22
$10^4$	1.11	0.14	12.00	0.38	40.50	0.08
$10^5$	1.11	0.14	21.47	0.14	40.50	0.08

implying that a preconditioner for  $\mathbf{K}_k$  is

$$\mathbf{P}_k = \widetilde{\mathbf{W}}(\widetilde{\mathbf{D}} + \widetilde{\mathbf{E}}_k)^{-1}\widetilde{\mathbf{Z}}^T, \quad (3.6.3)$$

where  $\widetilde{\mathbf{E}}_k$  is a sparse approximation of  $\widetilde{\mathbf{Z}}^T \widetilde{\Delta}_k \widetilde{\mathbf{W}}$  and  $\widetilde{\Delta}_k$  is a sparse approximation of  $\Delta_k$ .

In [11], the authors present upper bounds for the accuracy of the preconditioner, which we state below without proof.

**Theorem 3.1.** [11] Let  $\mathbf{P}_{seed}$  and  $\mathbf{P}_k$  be given as in (3.6.2) and (3.6.3). Let  $f$  and  $g$  be linear operators dictating the sparsification of the matrix's entries. Define  $o_f(\mathbf{M}) = \mathbf{M} - f(\mathbf{M})$  and  $o_g(\mathbf{M}) = \mathbf{M} - g(\mathbf{M})$  to be linear operators. Define  $\Theta_1 = -o_g(\widetilde{\mathbf{Z}}^T \Delta_k \widetilde{\mathbf{W}})$  and  $\Theta_2 = -g(\widetilde{\mathbf{Z}}^T o_f(\Delta_k) \widetilde{\mathbf{W}})$ . Furthermore, let  $\nu = \|\widetilde{\mathbf{Z}}^{-T}\| \|\widetilde{\mathbf{W}}^{-1}\|$ . Assume that  $\mathbf{P}_{seed}$  satisfies

$$\|\mathbf{K}_{seed} - \mathbf{P}_{seed}^{-1}\| = \varepsilon \|\mathbf{K}_{seed}\|$$

for some  $\varepsilon > 0$ . Then the following upper bound holds:

$$\|\mathbf{K}_k - \mathbf{P}_k^{-1}\| \leq \varepsilon \|\mathbf{K}_{seed}\| + \nu(\|\Theta_1\| + \|\Theta_2\|). \quad (3.6.4)$$

Furthermore, if  $\|\mathbf{K}_k - \mathbf{K}_{seed}\| > \varepsilon \|\mathbf{K}_{seed}\|$ , then

$$\|\mathbf{K}_k - \mathbf{P}_k^{-1}\| \leq \frac{\varepsilon \|\mathbf{K}_{seed}\| + \nu(\|\Theta_1\| + \|\Theta_2\|)}{\|\mathbf{K}_k - \mathbf{K}_{seed}\| - \varepsilon \|\mathbf{K}_{seed}\|} \|\mathbf{K}_k - \mathbf{P}_{seed}^{-1}\|. \quad (3.6.5)$$

Another theoretical result of [11] relies on the assumption that the preconditioner updates are designed for a sequence of Jacobians appearing in the Newton-Krylov methods. Assuming the standard convergence assumptions, namely that the Jacobian is Lipschitz continuous in a ball  $B(x^*, r)$  centered at  $x^*$  with radius  $r$  and the sequence of iterates  $x_k \in B(x^*, \delta)$  for  $k \geq 0$  and  $\delta > 0$ , [11] proves several theorems about the spectrum of the preconditioned Jacobian. Fortunately, these results, namely Theorem 4.3, Corollary 4.4, and Corollary 4.5 of [11], apply to IRKA as well by noting the presence of the Lipschitz constant in the IRKA framework. Consider applying right preconditioning to a sequence of IRKA shifts  $\{\sigma_k\}$  so that  $\sigma_k \rightarrow \sigma^*$  where  $\sigma^*$  is the optimal  $\mathcal{H}_2$  shift. Note that  $\mathbf{K}(\sigma_i) = \sigma_i \mathbf{E} - \mathbf{A}$  is Lipschitz continuous with Lipschitz constant  $\|\mathbf{E}\|$ . Due to this Lipschitz constant, we employ similar reasoning as in [11] to state and prove the following theorem and corollaries.

**Theorem 3.2.** *Let  $\{\sigma_k\}$  be a sequence of shifts generated by IRKA. Let  $\sigma_{seed} = \sigma_0$ . Define  $\mathbf{K}(\sigma_i) = \sigma_i \mathbf{E} - \mathbf{A}$ . Assume that for  $k \geq 0$ ,  $\sigma_k \in B(\sigma^*, \delta)$  where  $\sigma^*$  is the optimal  $\mathcal{H}_2$  shift. Further assume that the preconditioner  $\mathbf{P}_{seed}$  in (3.6.2) satisfies  $\|\mathbf{K}_{seed} - \mathbf{P}_{seed}^{-1}\| \leq \varepsilon$  for some*

positive  $\varepsilon$ . Then the right preconditioned system  $\mathbf{K}_k \mathbf{P}_k$  can be written as

$$\mathbf{K}_k \mathbf{P}_k = \mathbf{I} + \mathbf{R}_k \mathbf{P}_k, \quad \mathbf{R}_k = \mathbf{K}_k - \mathbf{P}_k^{-1}$$

where

$$\|\mathbf{R}_k\| = \varepsilon + 2c\|\mathbf{E}\|\delta$$

and

$$\|\mathbf{R}_k \mathbf{P}_k\| \leq \zeta(\varepsilon + 2c\|\mathbf{E}\|\delta) \|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\| \quad (3.6.6)$$

for  $\zeta = \|\tilde{\mathbf{Z}}\| \|\tilde{\mathbf{W}}\|$  and positive scalar  $c$ .

*Proof:* The proof follows exactly as in [11]. Instead of relying on the Lipschitz continuity of the Jacobian as in [11], the Lipschitz continuity of the matrix  $\mathbf{K}(\sigma_i) = \sigma_i \mathbf{E} - \mathbf{A}$  with Lipschitz constant  $\|\mathbf{E}\|$  yields the result.

**Corollary 3.6.1.** *Let the assumptions of Theorem 3.2 hold. Then there exists  $\hat{\delta}$  and  $\hat{\varepsilon}$  such that for all  $0 < \delta \leq \hat{\delta}$  and  $0 < \varepsilon \leq \hat{\varepsilon}$ , the eigenvalues of  $\mathbf{K}_k \mathbf{P}_k$  are clustered at 1 in the right half complex plane with radius  $\rho = \zeta(\varepsilon + 2c\|\mathbf{E}\|\delta) \|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\|$  for all  $k$ .*

*Proof.* Following the same reasoning as in [11], the result follows from (3.6.6) and letting  $\hat{\delta}$  and  $\hat{\varepsilon}$  be such that for all  $0 < \delta \leq \hat{\delta}$  and  $0 < \varepsilon \leq \hat{\varepsilon}$ , we have

$$\rho = \zeta(\varepsilon + 2c\|\mathbf{E}\|\delta) \|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\| < 1,$$

implying the eigenvalues are clustered at 1 in the right half complex plane with radius  $\rho$ .  $\square$

**Corollary 3.6.2.** *Let the assumptions of Theorem 3.2 hold. Let  $\tilde{\mathbf{E}}_k$  and  $\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k$  be diagonal approximations to  $\mathbf{E}_k$  and  $\mathbf{D} + \mathbf{E}_k$ , respectively. For any diagonal matrix  $\mathbf{Q}$ , denote the  $i^{\text{th}}$  entry of  $\mathbf{Q}$  by  $(\mathbf{Q})_i$ . Then  $\mathbf{K}_k \mathbf{P}_k = \mathbf{I} + \mathbf{R}_k \mathbf{P}_k$  where*

$$\|\mathbf{R}_k \mathbf{P}_k\| \leq \frac{\zeta(\varepsilon + 2c\|\mathbf{E}\|\delta)}{\min_i |(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)_i|}. \quad (3.6.7)$$

*Proof.* As in [11], the proof follows directly from (3.6.6) and noting that

$$\|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\| = \frac{1}{\min_i |(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)_i|}.$$

$\square$

By proving the clustering of the eigenvalues for sufficiently small  $\hat{\delta}$  and  $\hat{\varepsilon}$ , the Bellavia et al. approach is promising. It is important to emphasize, however, the assumptions of the theorem. The quantities  $\hat{\delta}$  and  $\hat{\varepsilon}$  depend on the distance between the shifts and the accuracy of the preconditioner, respectively. Therefore, the shifts may need to be very close together with a highly accurate preconditioner in order for the clustering of eigenvalues to be observed. In the next section, we apply the Bellavia et al. approach to three models to answer the question of if the Bellavia et al. update is numerically effective.

### 3.7 Numerical Results for the Bellavia et al. Update

We applied our proposed preconditioner to the Rail, CD and 1r models as discussed in Section 3.2. To compute the initial factorized form, [11] suggests two techniques: ILU and AINV preconditioners. For the ILU preconditioner, we have

$$\mathbf{K}_{seed} \approx \mathbf{L}\mathbf{D}\mathbf{U}^T,$$

where  $\mathbf{D}$  is a diagonal matrix and the matrices  $\mathbf{L}$  and  $\mathbf{U}$  are unit lower triangular. To obtain the factorized form of (3.6.2), take  $\tilde{\mathbf{Z}} \approx \mathbf{L}^{-T}$  and  $\tilde{\mathbf{W}} \approx \mathbf{U}^{-T}$ . For more details, see [69]. The other method, AINV, employs the biconjugate Gram-Schmidt orthogonalization process to obtain

$$\mathbf{K}_{seed} = \mathbf{M}^{-T}\mathbf{D}\mathbf{N}^{-1}. \quad (3.7.1)$$

Taking  $\tilde{\mathbf{W}} = \mathbf{N}$ ,  $\tilde{\mathbf{D}} = \mathbf{D}$ , and  $\tilde{\mathbf{Z}} = \mathbf{M}$ , (3.7.1) fits into the appropriate factorized framework. For more information about the AINV preconditioner, see [17]. We consider applying both factorizations and obtained similar results for almost all of the models. However, the AINV factorization was prone to occasional breakdowns; therefore, we present our results obtained using the ILU method with zero level of fill-in to compute the seed preconditioner. For the systems studied, the  $\text{ilu}(0)$  factorization yielded similar convergence as the  $\text{ilutp}$  preconditioner.

To sparsify the matrices, [11] suggests several different types of banded approximations. In

our numerical work, we used a diagonal sparsification of the  $\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k$  and kept the factors  $\tilde{\mathbf{W}}$  and  $\tilde{\mathbf{Z}}$  as dense matrices. Then the definition of  $\mathbf{\Delta}_k$  implies that the updated preconditioner is defined by

$$\mathbf{P}_k = \mathbf{W}(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\mathbf{Z}^T, \quad (3.7.2)$$

where  $\tilde{\mathbf{\Delta}}_k = (\sigma_k - \sigma_0)\text{diag}(\mathbf{E})$  and  $\tilde{\mathbf{E}}_k = (\sigma_k - \sigma_0)\text{diag}(\mathbf{Z}^T\text{diag}(\mathbf{E})\mathbf{W})$ . As noted in [11], a breakdown of the preconditioner update occurs if  $\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k$  is singular. Therefore, [11] recommends either shifting the entries or reusing the preconditioner update from the previous iteration. In our numerical results, we implemented the latter approach and updated the preconditioner only if

$$\min_i |(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)_i| > 10^{-4}\|\mathbf{K}_0\|_1. \quad (3.7.3)$$

In Table 3.10, we report the total number of systems throughout the IRKA iteration for which the update was abandoned. For reference, the third column of Table 3.10 reports the number of systems solved throughout the IRKA iteration. It is important to note that all models were plagued with a singular  $\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k$  matrix at some point in the IRKA iteration.

Table 3.10: Instances of Abandoned Update

Model	Update Abandoned	Systems Solved
Rail	5	42
CD	110	1000
1r	198	312



The IRKA algorithm with Bellavia et al. updates is given in IRKA[ $\mathbf{B}_i$ ] (Algorithm 3.7.1). In applying the Bellavia et al. update, we only compute an initial LU preconditioner for the first shift of the first IRKA iteration. The remaining shifts are preconditioned using the update.

**Algorithm 3.7.1. IRKA[ $\mathbf{B}_i$ ]: IRKA with Bellavia et al. Updates**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2. Compute  $\mathbf{L}_1\mathbf{U}_1 \approx (\sigma_1\mathbf{E} - \mathbf{A})$ .
3. For  $i = 2, \dots, r$ , compute updates,  $\mathbf{P}_i$  as defined in (3.7.2).
4. Solve  $(\sigma_1\mathbf{E} - \mathbf{A})\mathbf{U}_1^{-1}\mathbf{L}_1^{-1}\mathbf{v}_1 = \mathbf{B}\mathbf{b}_1$  and  $(\sigma_1\mathbf{E} - \mathbf{A})^T\mathbf{L}_1^{-T}\mathbf{U}_1^{-T}\mathbf{w}_1 = \mathbf{C}^T\mathbf{c}_1$ .
5. For  $i = 2, \dots, r$ , solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{P}_i\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and  $(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{P}_i^T\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$  if  $\min_i |(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)_i| > 10^{-4}\|\mathbf{K}_0\|_1$ .  
Otherwise, solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{U}_1^{-1}\mathbf{L}_1^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and  $(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{L}_1^{-T}\mathbf{U}_1^{-T}\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$ .
6.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  and  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .
7. while (not converged)
  - (a)  $\mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C}\mathbf{V}_r$ .
  - (b) Compute  $\mathbf{Y}^T\mathbf{A}_r\mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T\mathbf{E}_r\mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda\mathbf{E}_r - \mathbf{A}_r$ .

(c)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ .

(d) For  $i = 1, \dots, r$ , compute updates  $\mathbf{P}_i$  as defined in (3.7.2).

(e) For  $i = 1, \dots, r$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{P}_i \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and

$$(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{P}_i^T \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i \text{ if } \min_i |(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)_i| > 10^{-4} \|\mathbf{K}_0\|_1.$$

Otherwise, solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_1^{-1} \mathbf{L}_1^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .

(f)  $\mathbf{V}_r = [ \mathbf{v}_1, \dots, \mathbf{v}_r ]$  and  $\mathbf{W}_r = [ \mathbf{w}_1, \dots, \mathbf{w}_r ]$ .

$$8. \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r.$$

Since [11] uses Theorem 3.1 to suggest that the update possesses the potential to be better than freezing the preconditioner, we also consider IRKA $[\mathbf{L}_0 \mathbf{U}_0]$ , where the incomplete LU is computed for the smallest shift of the first IRKA iteration and then used for all remaining linear systems. In Table 3.11, we report the total number of GMRES iterations for  $\mathbf{V}_r$  throughout the IRKA iteration, which are representative of those required for  $\mathbf{W}_r$ . There is no theoretical reason to expect that IRKA $[\mathbf{B}_i]$  will outperform IRKA $[\mathbf{L}_i \mathbf{U}_i]$ ; however, we include this data as a lower bound measure since the preconditioner in IRKA $[\mathbf{L}_i \mathbf{U}_i]$  is computed very accurately and frequently. For the Rail Model, IRKA $[\mathbf{B}_i]$  results in fewer iterations than freezing the preconditioner. Meanwhile, the CD and 1r models are examples for which the Bellavia et al. update involves substantially more GMRES iterations in comparison to IRKA $[\mathbf{L}_0 \mathbf{U}_0]$ . In fact, the upper bounds of Theorem 3.1 and Theorem 3.2 suggest this convergence behavior. As noted in [11], the first bound (3.6.4) suggests that the accuracy of the preconditioner depends on two factors: the accuracy of the original preconditioner,

Table 3.11: Total GMRES Iterations for  $\mathbf{V}_r$ 

Model	IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]	IRKA[ $\mathbf{L}_0\mathbf{U}_0$ ]	IRKA[ $\mathbf{B}_i$ ]
Rail	84	1334	1271
CD	2000	31425	87861
1r	624	23286	25803

$\mathbf{P}_{seed}$ , and the terms  $\|\Theta_1\|$  and  $\|\Theta_2\|$ , which quantify the entries eliminated in the sparsification. Furthermore, the second bound suggests that if  $\nu(\|\Theta_1\| + \|\Theta_2\|)$  is small, then the updated preconditioner,  $\mathbf{P}_k$ , may be more accurate than the original preconditioner,  $\mathbf{P}_{seed}$ , and so  $\mathbf{P}_k$  would be a better preconditioner than simply freezing the preconditioner. Our numerical results as reported in Table 3.12 support the theoretical conclusions of Theorem 3.1. For all of the models, the  $\|\mathbf{K}_k - \mathbf{K}_{seed}\| > \varepsilon\|\mathbf{K}_{seed}\|$  assumption is satisfied, implying that both bound (3.6.4) and (3.6.5) hold. Even though all models begin with accurate seed preconditioners, the norm of  $\|\mathbf{K}_k - \mathbf{P}_k^{-1}\|$  is quite large for all models except the Rail Model. Moreover,  $\nu(\|\Theta_1\| + \|\Theta_2\|)$  is relatively small for the Rail Model in comparison to the CD and 1r models. As a result, these bounds suggest that the update for the Rail Model could be more effective than a frozen preconditioner while the update for the CD and 1r models will probably not be effective. Moreover, it is important to emphasize that a small  $\|\mathbf{K}_k - \mathbf{P}_k^{-1}\|$  quantity does not necessarily guarantee a good preconditioner since  $\|\mathbf{K}_k\mathbf{P}_k - \mathbf{I}\|$  could still be large due to a mismatch of the singular values and vectors of  $\mathbf{P}_k$  and  $\mathbf{K}_k$ . The convergence behavior can also be explained through (3.6.6) of Theorem 3.2, which gives that

Table 3.12: Preconditioner Update Bounds

Model	$\ \mathbf{K}_{seed} - \mathbf{P}_{seed}^{-1}\ $	$\ \mathbf{K}_k - \mathbf{P}_k^{-1}\ $	(3.6.4)	(3.6.5)	$\nu(\ \Theta_1\  + \ \Theta_2\ )$
Rail	$1.58 \times 10^{-5}$	$5.41 \times 10^{-3}$	$3.33 \times 10^{-2}$	$3.34 \times 10^{-2}$	$3.33 \times 10^{-2}$
CD	$6.26 \times 10^{-10}$	$1.20 \times 10^7$	$5.24 \times 10^8$	$5.24 \times 10^8$	$5.24 \times 10^8$
1r	$1.49 \times 10^{-12}$	$8.34 \times 10^5$	$2.69 \times 10^9$	$2.69 \times 10^9$	$2.69 \times 10^9$

$\mathbf{K}_k \mathbf{P}_k = \mathbf{I} + \mathbf{R}_k \mathbf{P}_k$  with

$$\|\mathbf{R}_k \mathbf{P}_k\| \leq \zeta(\varepsilon + 2c\|\mathbf{E}\|\delta)\|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\|. \quad (3.7.4)$$

As discussed previously, this upper bound suggests that if the quantities  $\hat{\delta}$  and  $\hat{\varepsilon}$  are selected sufficiently small, then the eigenvalues of  $\mathbf{K}_k \mathbf{P}_k$  will be clustered at 1 in the right half complex plane. Tables 3.13 - 3.15 report  $\|\mathbf{R}_k \mathbf{P}_k\|$ ,  $\zeta$ ,  $\|\mathbf{R}_k\|$  and  $\|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\|$ . For models where the preconditioner update resulted in substantially more GMRES iterations, we observe that  $\|\mathbf{R}_k \mathbf{P}_k\|$  is large. While Corollary 3.6.1 gives that the eigenvalues of  $\mathbf{K}_k \mathbf{P}_k$  will be clustered at 1 in the right half complex plane, the result depends on selecting  $\hat{\delta}$  and  $\hat{\varepsilon}$  so that  $0 < \delta \leq \hat{\delta}$  and  $0 < \varepsilon \leq \hat{\varepsilon}$  are such that  $\rho = \zeta(\varepsilon + 2c\|\mathbf{E}\|\delta)\|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\| < 1$ . However, as Table 3.14 and Table 3.15 show, the magnitudes of  $\|(\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\|$ ,  $\|\mathbf{R}_k\|$ , and  $\zeta$  are so large that the  $\hat{\delta}$  and  $\hat{\varepsilon}$  values need to be selected extremely small to ensure the conclusion of the eigenvalues being clustered at 1. Since  $\hat{\delta}$  and  $\hat{\varepsilon}$  depend on the distance between shifts and the accuracy of the seed preconditioner, respectively, sufficiently small  $\hat{\delta}$  and  $\hat{\varepsilon}$  values could very well imply that the shifts would need to be very close together with a highly accurate seed preconditioner in order to obtain the ideal spectrum. Therefore, even though Corollary 3.6.1

seems theoretically promising, our numerical results present a bleaker reality.

Table 3.13:  $\|\mathbf{R}_k\mathbf{P}_k\|$  and Associated Terms for the Last IRKA Iteration of the Rail Model

Shift	$\ \mathbf{R}_k\mathbf{P}_k\ $	$\zeta$	$\ \mathbf{R}_k\ $	$\ (\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\ $
$\sigma_1\mathbf{E} - \mathbf{A}$	1.48	3.80	$1.58 \times 10^{-5}$	$2.24 \times 10^5$
$\sigma_2\mathbf{E} - \mathbf{A}$	1.46	3.77	$1.58 \times 10^{-5}$	$2.23 \times 10^5$
$\sigma_3\mathbf{E} - \mathbf{A}$	1.39	3.44	$1.59 \times 10^{-5}$	$2.01 \times 10^5$
$\sigma_4\mathbf{E} - \mathbf{A}$	1.17	3.07	$1.57 \times 10^{-5}$	$1.20 \times 10^5$
$\sigma_5\mathbf{E} - \mathbf{A}$	$5.42 \times 10^{-1}$	1.98	$2.13 \times 10^{-5}$	$9.94 \times 10^4$
$\sigma_6\mathbf{E} - \mathbf{A}$	$8.34 \times 10^{-1}$	1.94	$3.69 \times 10^{-3}$	$3.49 \times 10^4$

### 3.8 Effect of Initial Preconditioner

In this next section, we consider implementing the preconditioner updates, but now with an initial incomplete LU decomposition based on varying levels of fill-in. For a fill-in level of one, the updated preconditioners result in noticeably larger GMRES iterations for all models. As a result, our goal is to select certain points in the IRKA iteration to compute a new incomplete LU in order to obtain better convergence for the updated preconditioned systems as well. We consider two types of requirements for computing a new incomplete LU. The first technique involves a static condition and results in IRKA[**Mid**, **f**] (Algorithm 3.8.1). In the algorithm below, the matrices,  $\mathbf{P}_i$  and  $\tilde{\mathbf{P}}_i$ , denote the entire computed preconditioner, including both the update and the  $\mathbf{L}_i$  and  $\mathbf{U}_i$  factors if appropriate. This algorithm computes an incomplete LU decomposition with level-f fill-in for the matrices  $\mathbf{L}_1\mathbf{U}_1 \approx \sigma_1\mathbf{E} - \mathbf{A}$  and  $\mathbf{L}_{\frac{r}{2}+1}\mathbf{U}_{\frac{r}{2}+1} \approx \sigma_{\frac{r}{2}+1}\mathbf{E} - \mathbf{A}$  at each IRKA iteration. In constructing preconditioner updates

for  $\sigma_i \mathbf{E} - \mathbf{A}$ , the decompositions  $\mathbf{L}_1 \mathbf{U}_1$  and  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1}$  are used for  $i = 2, \dots, \frac{r}{2}$  and  $i = \frac{r}{2} + 2, \dots, r$ , respectively. We also considered computing  $\mathbf{L}_1 \mathbf{U}_1 \approx \sigma_1 \mathbf{E} - \mathbf{A}$  and  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1} \approx \sigma_{\frac{r}{2}+1} \mathbf{E} - \mathbf{A}$  for only the first IRKA iteration and then using the preconditioners throughout all remaining IRKA iterations; however, this did not noticeably impact the convergence of GMRES. As a result, we only present the results for IRKA[Mid, f].

**Algorithm 3.8.1. IRKA[Mid, f]: IRKA with  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1}$**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2. Compute  $\mathbf{L}_1 \mathbf{U}_1 \approx \sigma_1 \mathbf{E} - \mathbf{A}$  and  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1} \approx \sigma_{\frac{r}{2}+1} \mathbf{E} - \mathbf{A}$  with level-f fill-in.
3. For  $i = 1, \frac{r}{2} + 1$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_i^{-1} \mathbf{L}_i^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_i^{-T} \mathbf{U}_i^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
4. For  $i = 2, \dots, \frac{r}{2}$ , use  $\mathbf{L}_1 \mathbf{U}_1$  to compute the updated preconditioners,  $\mathbf{P}_i$  and  $\tilde{\mathbf{P}}_i$ , and solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{P}_i \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \tilde{\mathbf{P}}_i \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
5. For  $i = \frac{r}{2} + 2, \dots, r$ , use  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1}$  to compute the updated preconditioners,  $\mathbf{P}_i$  and  $\tilde{\mathbf{P}}_i$ , and solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{P}_i \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \tilde{\mathbf{P}}_i \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
6.  $\mathbf{V}_r = [ \mathbf{v}_1, \dots, \mathbf{v}_r ]$  and  $\mathbf{W}_r = [ \mathbf{w}_1, \dots, \mathbf{w}_r ]$ .
7. while (not converged)
  - (a)  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ .
  - (b) Compute  $\mathbf{Y}^T \mathbf{A}_r \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_r \mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_r - \mathbf{A}_r$ .

- (c)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ .
- (d) Compute  $\mathbf{L}_1 \mathbf{U}_1 \approx \sigma_1 \mathbf{E} - \mathbf{A}$  and  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1} \approx \sigma_{\frac{r}{2}+1} \mathbf{E} - \mathbf{A}$  with level- $f$  fill-in.
- (e) For  $i = 1, \frac{r}{2} + 1$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_i^{-1} \mathbf{L}_i^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  
 $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_i^{-T} \mathbf{U}_i^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
- (f) For  $i = 2, \dots, \frac{r}{2}$ , use  $\mathbf{L}_1 \mathbf{U}_1$  to compute the updated preconditioners,  $\mathbf{P}_i$  and  $\tilde{\mathbf{P}}_i$ ,  
and solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{P}_i \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \tilde{\mathbf{P}}_i \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
- (g) For  $i = \frac{r}{2} + 2, \dots, r$ , use  $\mathbf{L}_{\frac{r}{2}+1} \mathbf{U}_{\frac{r}{2}+1}$  to compute the updated preconditioners,  $\mathbf{P}_i$   
and  $\tilde{\mathbf{P}}_i$ , and solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{P}_i \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \tilde{\mathbf{P}}_i \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$ .
- (h)  $\mathbf{V}_r = [ \mathbf{v}_1, \dots, \mathbf{v}_r ]$  and  $\mathbf{W}_r = [ \mathbf{w}_1, \dots, \mathbf{w}_r ]$ .

$$8. \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r.$$

The next method, IRKA[ $\mathbf{G}, \mathbf{f}$ ], is detailed in Algorithm 3.8.2, which computes an initial preconditioner for the first shift of the first IRKA iteration. Then the algorithm dynamically computes updates based on the convergence of GMRES. During the first IRKA step, the update is computed if the number of GMRES iterations exceeds twice the number of GMRES iterations required for the first shift. After the first IRKA iteration, a new incomplete LU for  $\sigma_1 \mathbf{E} - \mathbf{A}$  of the current iteration and a preconditioner update are computed only for shifts resulting in more than twice the number of GMRES iterations required for  $\sigma_1 \mathbf{E} - \mathbf{A}$ . In this way, the preconditioners are dynamically updated based on the convergence of GMRES. In the algorithm below,  $\mathbf{P}_i$  and  $\tilde{\mathbf{P}}_i$  denote the entire preconditioner.

**Algorithm 3.8.2. IRKA[G,f]: IRKA with Updates Based on GMRES**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2. Compute  $\mathbf{L}_1\mathbf{U}_1 \approx (\sigma_1\mathbf{E} - \mathbf{A})$ .
3. Let  $\mathbf{L}_i\mathbf{U}_i = \mathbf{L}_1\mathbf{U}_1$  for  $i = 2, \dots, r$ .
4. For  $i = 1, \dots, r$ , solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{U}_i^{-1}\mathbf{L}_i^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and  $(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{L}_i^{-T}\mathbf{U}_i^{-T}\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$  as follows:
  - (a) Let  $q$  and  $l$  be the total number of GMRES iterations required to solve  $(\sigma_1\mathbf{E} - \mathbf{A})\mathbf{U}_1^{-1}\mathbf{L}_1^{-1}\mathbf{v}_1 = \mathbf{B}\mathbf{b}_1$  and  $(\sigma_1\mathbf{E} - \mathbf{A})^T\mathbf{L}_1^{-T}\mathbf{U}_1^{-T}\mathbf{w}_1 = \mathbf{C}^T\mathbf{c}_1$ .
  - (b) Solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{U}_i^{-1}\mathbf{L}_i^{-1}\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and  $(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{L}_i^{-T}\mathbf{U}_i^{-T}\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$  with the maximum number of GMRES iterations equal to  $2q$  or  $2l$ , respectively.
  - (c) If GMRES does not converge within  $2q$  or  $2l$  iterations, compute an update preconditioner,  $\mathbf{P}_i$  or  $\tilde{\mathbf{P}}_i$ , and solve  $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{P}_i\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$  and  $(\sigma_i\mathbf{E} - \mathbf{A})^T\tilde{\mathbf{P}}_i\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$  using GMRES.
5.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  and  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .
6. while (not converged)
  - (a)  $\mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$ ,  $\mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C}\mathbf{V}_r$ .
  - (b) Compute  $\mathbf{Y}^T\mathbf{A}_r\mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T\mathbf{E}_r\mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda\mathbf{E}_r - \mathbf{A}_r$ .



(c)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ .

(d) For  $i = 1, \dots, r$ , solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_i^{-1} \mathbf{L}_i^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$

and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_i^{-T} \mathbf{U}_i^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$  as follows:

i. Let  $q$  and  $l$  be the total number of GMRES iterations required to solve

$$(\sigma_1 \mathbf{E} - \mathbf{A}) \mathbf{U}_1^{-1} \mathbf{L}_1^{-1} \mathbf{v}_1 = \mathbf{B} \mathbf{b}_1 \text{ and } (\sigma_1 \mathbf{E} - \mathbf{A})^T \mathbf{L}_1^{-T} \mathbf{U}_1^{-T} \mathbf{w}_1 = \mathbf{C}^T \mathbf{c}_1.$$

ii. Solve  $(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{U}_i^{-1} \mathbf{L}_i^{-1} \mathbf{v}_i = \mathbf{B} \mathbf{b}_i$  and  $(\sigma_i \mathbf{E} - \mathbf{A})^T \mathbf{L}_i^{-T} \mathbf{U}_i^{-T} \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i$  with the maximum number of GMRES iterations equal to  $2q$  and  $2l$ , respectively.

iii. If GMRES does not converge within  $2q$  or  $2l$  iterations, compute

$\mathbf{L}_i \mathbf{U}_i = \sigma_i \mathbf{E} - \mathbf{A}$ , an updated preconditioner,  $\mathbf{P}_i$  or  $\tilde{\mathbf{P}}_i$ , and solve

$$(\sigma_i \mathbf{E} - \mathbf{A}) \mathbf{P}_i \mathbf{v}_i = \mathbf{B} \mathbf{b}_i \text{ and } (\sigma_i \mathbf{E} - \mathbf{A})^T \tilde{\mathbf{P}}_i \mathbf{w}_i = \mathbf{C}^T \mathbf{c}_i \text{ using GMRES.}$$

(e)  $\mathbf{V}_r = [ \mathbf{v}_1, \dots, \mathbf{v}_r ]$  and  $\mathbf{W}_r = [ \mathbf{w}_1, \dots, \mathbf{w}_r ]$ .

$$7. \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r.$$

We compare these methods with the previously proposed algorithms. Since we obtained similar results for  $\mathbf{W}_r$ , only the GMRES iterations required for  $\mathbf{V}_r$  are recorded. Table 3.16, Table 3.18, and Table 3.20 report the total number of GMRES iterations for  $\mathbf{V}_r$ . In the tables, IRKA[k,f] refers to applying IRKA[ $\mathbf{L}_i \mathbf{U}_i$ ], IRKA[ $\mathbf{M}_i$ ] and IRKA[ $\mathbf{B}_i$ ]. In addition, Table 3.17, Table 3.19 and Table 3.21 illustrate the total number of preconditioners computed during the iteration. In examining this data, there are two relationships we would like to explore, namely the performance of IRKA[**Mid**, **f**] and IRKA[**G**, **f**] with the updates and the effect of the preconditioner's accuracy.

### 3.8.1 Effect of Preconditioner Accuracy

Intuitively, we expect a direct relationship between the accuracy of the initial preconditioner and the overall GMRES convergence. For the Rail Model, such a relationship is observed, namely for all methods,  $\text{IRKA}[\cdot, 2]$  results in a similar or smaller number of GMRES iterations than  $\text{IRKA}[\cdot, 1]$ . For the CD and 1r models, the accuracy of the preconditioner is more peculiar. If the SAI update is used for the CD and 1r models, then the number of GMRES iterations for  $\text{IRKA}[\cdot, 2]$  is less than or equal to the corresponding number for  $\text{IRKA}[\cdot, 1]$  as anticipated. With the Bellavia et al. update, however, GMRES actually experiences worse convergence for  $\text{IRKA}[\cdot, 2]$  than  $\text{IRKA}[\cdot, 1]$ . Therefore, this suggests that the additional computations required for obtaining a more accurate preconditioner do not always correspond to improved GMRES convergence.

### 3.8.2 Effect of $\text{IRKA}[\text{Mid}, \mathbf{f}]$ and $\text{IRKA}[\mathbf{G}, \mathbf{f}]$ on SAI Updates

The results indicate that the SAI method performs best with  $\text{IRKA}[\text{Mid}, \mathbf{f}]$ . In fact,  $\text{IRKA}[\text{Mid}, 2]$  with SAI updates resulted in fewer GMRES iterations than  $\text{IRKA}[\mathbf{k}, 2]$  for all models. This is quite impressive since  $\text{IRKA}[\text{Mid}, \mathbf{f}]$  requires substantially fewer incomplete LU decompositions than  $\text{IRKA}[\mathbf{k}, 2]$ ; for example,  $\text{IRKA}[\mathbf{k}, 2]$  for the 1r Model, involves over fifteen times as many incomplete LU decompositions as  $\text{IRKA}[\text{Mid}, \mathbf{f}]$ . Therefore, this data suggests that  $\text{IRKA}[\text{Mid}, \mathbf{f}]$  offers computational savings in terms of both GMRES iterations and incomplete LU decompositions computed. In many ways, the superior performance of

IRKA[**Mid**, **f**] for the Rail Model is expected as IRKA[**Mid**, **f**] computes more preconditioners. In examining Table 3.19 and Table 3.21, however, the superiority of IRKA[**Mid**, **f**] is not easily explained since IRKA[**G**, **f**] computes over twelve and six times more preconditioners than IRKA[**Mid**, **f**], respectively. A possible explanation is the additional  $2l$  and  $2q$  GMRES iterations involved in IRKA[**G**, **f**] before deciding to compute a new LU decomposition outweighs the savings associated with the new preconditioner for these models. Since all models display a noticeable increase with IRKA[**G**, **f**], the conclusion is that the SAI update is best implemented with IRKA[**Mid**, **f**].

### 3.8.3 Effect of IRKA[**Mid**, **f**] and IRKA[**G**, **f**] on Bellavia et al. Updates

From our findings, the effect of IRKA[**Mid**, **f**] and IRKA[**G**, **f**] seems to depend on the initial quality of the Bellavia et al. update. For the Rail Model, convergence of the Bellavia et al. update is similar to other methods, and we observe that the update benefits most from the additional preconditioners computed by IRKA[**Mid**, **f**]. For the CD and 1r models, however, the bounds presented in Theorem 3.1 and Theorem 3.2 predict that the Bellavia et al. update is a poor preconditioner, which is supported by the numerical results of Table 3.18 and Table 3.20. By computing a new seed preconditioner when these poor preconditioners are encountered in the IRKA iteration, IRKA[**G**, **f**] results in fewer GMRES iterations than IRKA[**Mid**, **f**], which continues to use the same initial seed preconditioner until the  $\frac{r}{2} + 1$

system. While  $\text{IRKA}[\mathbf{G}, \mathbf{f}]$  always results in more GMRES iterations than  $\text{IRKA}[k, \mathbf{f}]$ , it is important to emphasize that the former method computes only about sixty percent of the incomplete LU decompositions that  $\text{IRKA}[k, \mathbf{f}]$  computes. Especially if the cost of computing an incomplete LU is large, this savings may outweigh the additional GMRES iterations. Therefore, the models considered indicate that  $\text{IRKA}[\mathbf{G}, \mathbf{f}]$  with Bellavia et al. updates is an appropriate method, yielding a reasonable number of GMRES iterations and incomplete LU decompositions when compared to other methods.

Especially considering the large number of GMRES iterations required for the SAI and Bellavia et al. update with  $\text{IRKA}[k, \mathbf{f}]$ , these observations illustrate that the updates become more competitive when implemented in conjunction with either  $\text{IRKA}[\mathbf{Mid}, \mathbf{f}]$  or  $\text{IRKA}[\mathbf{G}, \mathbf{f}]$ .

Table 3.14:  $\|\mathbf{R}_k \mathbf{P}_k\|$  and Associated Terms for the Last IRKA Iteration of the CD Model

Shift	$\ \mathbf{R}_k \mathbf{P}_k\ $	$\zeta$	$\ \mathbf{R}_k\ $	$\ (\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\ $
$\sigma_1 \mathbf{E} - \mathbf{A}$	8.55	$9.95 \times 10^3$	$2.26 \times 10^1$	$9.76 \times 10^{-1}$
$\sigma_2 \mathbf{E} - \mathbf{A}$	8.20	$9.94 \times 10^3$	$2.26 \times 10^1$	$7.75 \times 10^{-1}$
$\sigma_3 \mathbf{E} - \mathbf{A}$	$1.49 \times 10^1$	$9.93 \times 10^3$	$4.36 \times 10^1$	$6.14 \times 10^{-1}$
$\sigma_4 \mathbf{E} - \mathbf{A}$	$1.50 \times 10^1$	$9.91 \times 10^3$	$4.77 \times 10^1$	$4.87 \times 10^{-1}$
$\sigma_5 \mathbf{E} - \mathbf{A}$	$9.06 \times 10^1$	$9.88 \times 10^3$	$9.42 \times 10^5$	$1.30 \times 10^{-1}$
$\sigma_6 \mathbf{E} - \mathbf{A}$	$9.52 \times 10^1$	$9.85 \times 10^3$	$1.27 \times 10^6$	$3.37 \times 10^{-2}$
$\sigma_7 \mathbf{E} - \mathbf{A}$	$9.39 \times 10^1$	$9.81 \times 10^3$	$1.26 \times 10^6$	$2.34 \times 10^{-2}$
$\sigma_8 \mathbf{E} - \mathbf{A}$	$9.25 \times 10^1$	$9.76 \times 10^3$	$1.44 \times 10^6$	$1.71 \times 10^{-2}$
$\sigma_9 \mathbf{E} - \mathbf{A}$	$9.08 \times 10^1$	$9.70 \times 10^3$	$1.43 \times 10^6$	$1.55 \times 10^{-2}$
$\sigma_{10} \mathbf{E} - \mathbf{A}$	$8.85 \times 10^1$	$9.63 \times 10^3$	$1.50 \times 10^6$	$1.39 \times 10^{-2}$
$\sigma_{11} \mathbf{E} - \mathbf{A}$	$8.61 \times 10^1$	$9.53 \times 10^3$	$1.48 \times 10^6$	$1.35 \times 10^{-2}$
$\sigma_{12} \mathbf{E} - \mathbf{A}$	$9.30 \times 10^1$	$9.41 \times 10^3$	$3.70 \times 10^6$	$5.23 \times 10^{-3}$
$\sigma_{13} \mathbf{E} - \mathbf{A}$	$9.13 \times 10^1$	$9.26 \times 10^3$	$3.64 \times 10^6$	$6.79 \times 10^{-3}$
$\sigma_{14} \mathbf{E} - \mathbf{A}$	$1.13 \times 10^1$	$9.08 \times 10^3$	$2.45 \times 10^2$	$4.64 \times 10^{-2}$
$\sigma_{15} \mathbf{E} - \mathbf{A}$	$8.88 \times 10^1$	$8.85 \times 10^3$	$5.43 \times 10^6$	$9.23 \times 10^{-3}$
$\sigma_{16} \mathbf{E} - \mathbf{A}$	$8.70 \times 10^1$	$8.58 \times 10^3$	$5.27 \times 10^6$	$5.98 \times 10^{-3}$
$\sigma_{17} \mathbf{E} - \mathbf{A}$	$1.03 \times 10^2$	$8.25 \times 10^3$	$8.22 \times 10^6$	$9.96 \times 10^{-3}$
$\sigma_{18} \mathbf{E} - \mathbf{A}$	$9.91 \times 10^1$	$7.86 \times 10^3$	$7.75 \times 10^6$	$8.58 \times 10^{-3}$
$\sigma_{19} \mathbf{E} - \mathbf{A}$	$8.55 \times 10^1$	$7.41 \times 10^3$	$8.66 \times 10^6$	$6.55 \times 10^{-3}$
$\sigma_{20} \mathbf{E} - \mathbf{A}$	$8.31 \times 10^1$	$6.89 \times 10^3$	$8.08 \times 10^6$	$4.10 \times 10^{-3}$
$\sigma_{21} \mathbf{E} - \mathbf{A}$	$7.92 \times 10^1$	$6.30 \times 10^3$	$8.11 \times 10^6$	$2.58 \times 10^{-3}$
$\sigma_{22} \mathbf{E} - \mathbf{A}$	$7.62 \times 10^1$	$5.66 \times 10^3$	$7.35 \times 10^6$	$2.10 \times 10^{-3}$
$\sigma_{23} \mathbf{E} - \mathbf{A}$	$7.24 \times 10^1$	$4.98 \times 10^3$	$6.79 \times 10^6$	$2.58 \times 10^{-3}$
$\sigma_{24} \mathbf{E} - \mathbf{A}$	$6.88 \times 10^1$	$4.29 \times 10^3$	$5.95 \times 10^6$	$2.46 \times 10^{-3}$
$\sigma_{25} \mathbf{E} - \mathbf{A}$	$2.19 \times 10^1$	$3.60 \times 10^3$	$2.75 \times 10^7$	$1.58 \times 10^{-3}$
$\sigma_{26} \mathbf{E} - \mathbf{A}$	$2.17 \times 10^1$	$2.94 \times 10^3$	$2.25 \times 10^7$	$7.06 \times 10^{-4}$
$\sigma_{27} \mathbf{E} - \mathbf{A}$	$2.00 \times 10^1$	$2.33 \times 10^3$	$1.94 \times 10^7$	$4.06 \times 10^{-4}$
$\sigma_{28} \mathbf{E} - \mathbf{A}$	$1.97 \times 10^1$	$1.80 \times 10^3$	$1.50 \times 10^7$	$8.63 \times 10^{-4}$
$\sigma_{29} \mathbf{E} - \mathbf{A}$	$1.58 \times 10^1$	$1.35 \times 10^3$	$1.42 \times 10^7$	$5.76 \times 10^{-4}$
$\sigma_{30} \mathbf{E} - \mathbf{A}$	$1.55 \times 10^1$	$9.93 \times 10^2$	$1.04 \times 10^7$	$3.47 \times 10^{-4}$
$\sigma_{31} \mathbf{E} - \mathbf{A}$	8.20	$7.11 \times 10^2$	$1.49 \times 10^7$	$3.63 \times 10^{-4}$
$\sigma_{32} \mathbf{E} - \mathbf{A}$	8.10	$4.98 \times 10^2$	$1.04 \times 10^7$	$1.88 \times 10^{-4}$
$\sigma_{33} \mathbf{E} - \mathbf{A}$	7.70	$3.42 \times 10^2$	$7.48 \times 10^6$	$2.34 \times 10^{-4}$
$\sigma_{34} \mathbf{E} - \mathbf{A}$	7.52	$2.32 \times 10^2$	$5.09 \times 10^6$	$2.05 \times 10^{-4}$
$\sigma_{35} \mathbf{E} - \mathbf{A}$	6.65	$1.55 \times 10^2$	$3.82 \times 10^6$	$1.70 \times 10^{-4}$
$\sigma_{36} \mathbf{E} - \mathbf{A}$	6.38	$1.02 \times 10^2$	$2.56 \times 10^6$	$1.46 \times 10^{-4}$
$\sigma_{37} \mathbf{E} - \mathbf{A}$	5.10	$6.74 \times 10^1$	$3.40 \times 10^6$	$1.06 \times 10^{-4}$
$\sigma_{38} \mathbf{E} - \mathbf{A}$	4.17	$4.42 \times 10^1$	$2.22 \times 10^6$	$8.07 \times 10^{-5}$
$\sigma_{39} \mathbf{E} - \mathbf{A}$	3.69	$2.90 \times 10^1$	$2.40 \times 10^6$	$5.59 \times 10^{-5}$
$\sigma_{40} \mathbf{E} - \mathbf{A}$	4.39	$1.92 \times 10^1$	$1.55 \times 10^6$	$4.64 \times 10^{-5}$

Table 3.15:  $\|\mathbf{R}_k \mathbf{P}_k\|$  and Associated Terms for the Last IRKA Iteration of the 1r Model

Shift	$\ \mathbf{R}_k \mathbf{P}_k\ $	$\zeta$	$\ \mathbf{R}_k\ $	$\ (\tilde{\mathbf{D}} + \tilde{\mathbf{E}}_k)^{-1}\ $
$\sigma_1 \mathbf{E} - \mathbf{A}$	1.99	$3.76 \times 10^9$	$7.75 \times 10^{-1}$	$1.00 \times 10^3$
$\sigma_2 \mathbf{E} - \mathbf{A}$	1.99	$1.69 \times 10^9$	$7.75 \times 10^{-1}$	$6.70 \times 10^2$
$\sigma_3 \mathbf{E} - \mathbf{A}$	5.12	$7.58 \times 10^8$	1.99	$4.49 \times 10^2$
$\sigma_4 \mathbf{E} - \mathbf{A}$	5.12	$3.40 \times 10^8$	1.99	$3.01 \times 10^2$
$\sigma_5 \mathbf{E} - \mathbf{A}$	5.91	$1.53 \times 10^8$	2.30	$2.02 \times 10^2$
$\sigma_6 \mathbf{E} - \mathbf{A}$	5.91	$6.86 \times 10^7$	2.30	$1.35 \times 10^2$
$\sigma_7 \mathbf{E} - \mathbf{A}$	6.40	$3.08 \times 10^7$	2.49	$9.05 \times 10^1$
$\sigma_8 \mathbf{E} - \mathbf{A}$	6.40	$1.38 \times 10^7$	2.49	$6.06 \times 10^1$
$\sigma_9 \mathbf{E} - \mathbf{A}$	6.60	$6.21 \times 10^6$	2.57	$4.06 \times 10^1$
$\sigma_{10} \mathbf{E} - \mathbf{A}$	6.59	$2.79 \times 10^6$	2.57	$2.72 \times 10^1$
$\sigma_{11} \mathbf{E} - \mathbf{A}$	$1.00 \times 10^1$	$1.25 \times 10^6$	3.91	$1.82 \times 10^1$
$\sigma_{12} \mathbf{E} - \mathbf{A}$	9.94	$5.65 \times 10^5$	3.91	$1.22 \times 10^1$
$\sigma_{13} \mathbf{E} - \mathbf{A}$	$1.41 \times 10^1$	$2.56 \times 10^5$	5.63	8.19
$\sigma_{14} \mathbf{E} - \mathbf{A}$	$1.36 \times 10^1$	$1.17 \times 10^5$	5.63	5.48
$\sigma_{15} \mathbf{E} - \mathbf{A}$	$1.80 \times 10^1$	$5.43 \times 10^4$	7.93	3.67
$\sigma_{16} \mathbf{E} - \mathbf{A}$	$1.59 \times 10^1$	$2.61 \times 10^4$	7.94	2.46
$\sigma_{17} \mathbf{E} - \mathbf{A}$	$8.11 \times 10^2$	$1.32 \times 10^4$	$1.98 \times 10^5$	$7.18 \times 10^{-1}$
$\sigma_{18} \mathbf{E} - \mathbf{A}$	$8.05 \times 10^2$	$7.05 \times 10^3$	$9.35 \times 10^4$	$3.57 \times 10^{-1}$
$\sigma_{19} \mathbf{E} - \mathbf{A}$	$2.15 \times 10^2$	$4.00 \times 10^3$	$1.74 \times 10^5$	$2.40 \times 10^{-1}$
$\sigma_{20} \mathbf{E} - \mathbf{A}$	$2.15 \times 10^2$	$2.39 \times 10^3$	$9.20 \times 10^4$	$2.47 \times 10^{-1}$
$\sigma_{21} \mathbf{E} - \mathbf{A}$	$1.97 \times 10^2$	$1.48 \times 10^3$	$5.72 \times 10^4$	$9.05 \times 10^{-2}$
$\sigma_{22} \mathbf{E} - \mathbf{A}$	$1.95 \times 10^2$	$9.37 \times 10^2$	$3.51 \times 10^4$	$8.75 \times 10^{-2}$
$\sigma_{23} \mathbf{E} - \mathbf{A}$	$1.55 \times 10^2$	$6.05 \times 10^2$	$2.84 \times 10^4$	$7.20 \times 10^{-2}$
$\sigma_{24} \mathbf{E} - \mathbf{A}$	$1.52 \times 10^2$	$3.96 \times 10^2$	$1.88 \times 10^4$	$5.38 \times 10^{-2}$

Table 3.16: Rail Model: Total GMRES Iterations for  $\mathbf{V}_r$ 

Method	IRKA[ $\mathbf{L}_i \mathbf{U}_i$ ]	SAI	Bellavia
IRKA[k,1]	1250	1553	1600
IRKA[Mid, 1]	-	1317	1398
IRKA[G, 1]	-	1606	1705
IRKA[k,2]	404	533	1336
IRKA[Mid, 2]	-	365	545
IRKA[G, 2]	-	1157	1776

Table 3.17: Rail Model: Incomplete LU Decompositions Computed for  $\mathbf{V}_r$ 

Method	IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]	SAI	Bellavia
IRKA[k,f]	42	1	1
IRKA[ <b>Mid</b> , f]	-	14	14
IRKA[ <b>G</b> , <b>1</b> ]	-	1	3
IRKA[ <b>G</b> , <b>2</b> ]	-	1	3

Table 3.18: CD Model: Total GMRES Iterations for  $\mathbf{V}_r$ 

Method	IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]	SAI	Bellavia
IRKA[k,1]	2000	20760	85358
IRKA[ <b>Mid</b> , <b>1</b> ]	-	2000	76636
IRKA[ <b>G</b> , <b>1</b> ]	-	6940	5872
IRKA[k,2]	2000	20760	85357
IRKA[ <b>Mid</b> , <b>2</b> ]	-	1920	92528
IRKA[ <b>G</b> , <b>2</b> ]	-	6665	8448

Table 3.19: CD Model: Incomplete LU Decompositions Computed for  $\mathbf{V}_r$ 

Method	IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]	SAI	Bellavia
IRKA[k,f]	1000	1	1
IRKA[ <b>Mid</b> , f]	-	80	80
IRKA[ <b>G</b> , <b>1</b> ]	-	983	644
IRKA[ <b>G</b> , <b>2</b> ]	-	983	644

Table 3.20: 1r Model: Total GMRES Iterations for  $\mathbf{V}_r$ 

Method	IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]	SAI	Bellavia
IRKA[k,1]	624	5688	40934
IRKA[ <b>Mid</b> , <b>1</b> ]	-	624	24912
IRKA[ <b>G</b> , <b>1</b> ]	-	2135	1878
IRKA[k,2]	624	5688	40934
IRKA[ <b>Mid</b> , <b>2</b> ]	-	576	35281
IRKA[ <b>G</b> , <b>2</b> ]	-	1981	2520

Table 3.21: 1r Model: Incomplete LU Decompositions Computed for  $\mathbf{V}_r$ 

Method	IRKA[ $\mathbf{L}_i\mathbf{U}_i$ ]	SAI	Bellavia
IRKA[k,f]	312	1	1
IRKA[Mid, f]	-	48	48
IRKA[G, 1]	-	298	186
IRKA[G, 2]	-	298	186



# Chapter 4

## Interpolatory Methods for DAEs

In this chapter, we consider systems modeled by differential algebraic equations, where the transfer function,  $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ , is characterized by a singular  $\mathbf{E}$  matrix. Systems with a singular  $\mathbf{E}$  matrix arise naturally in modeling various physical processes, such as electrical circuits, linearized and semidiscretized Navier-Stokes equations, multibody systems, and semidiscretized partial differential equations. The aim of this chapter is to consider applying interpolatory methods to DAE systems. To do so, Section 4.1 provides a brief summary of important properties associated with DAEs and presents a counterexample to illustrate that the simple application of Theorem 1.1 may result in unbounded model reduction errors. One of the main contributions of this chapter is Theorem 4.1, which defines a reduced-order model so that the model reduction errors remain bounded and interpolation is achieved. From Theorem 4.1, an extension of the IRKA framework to DAEs is presented. While the resulting algorithm proves effective for the reduction of DAEs, it relies on com-

putationally expensive projectors. As a result, the remainder of the chapter is devoted to employing the characteristics associated with index-1 and Hessenberg index-2 DAEs to circumvent the explicit computation of these projectors.

## 4.1 Interpolatory Model Reduction of DAEs

In this section, we provide a summary of the important theoretical properties of DAEs. A DAE system is of the following form:

$$\mathbf{H}(s) : \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \end{cases} \quad (4.1.1)$$

where  $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{B} \in \mathbb{R}^{n \times m}$ ,  $\mathbf{C} \in \mathbb{R}^{p \times n}$ , and the  $\mathbf{E}$  matrix is singular. Closely related to the characterization of the DAE is its differentiation index. For first-order equations, the index refers to the total number of derivatives required in order to obtain an explicit ODE [22]. An alternative definition of the index is given through the Weierstrass canonical form. Assuming that  $\lambda\mathbf{E} - \mathbf{A}$  is a regular pencil, namely there exists  $\lambda$  such that  $\det(\lambda\mathbf{E} - \mathbf{A}) \neq 0$ , the Weierstrass canonical form provides the existence of matrices  $\mathbf{S}$  and  $\mathbf{T}$  such that

$$\mathbf{E} = \mathbf{S} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \mathbf{T}^{-1}, \quad \text{and} \quad \mathbf{A} = \mathbf{S} \begin{bmatrix} \mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_\infty} \end{bmatrix} \mathbf{T}^{-1}$$

where  $\mathbf{J}$  is a Jordan block corresponding to the finite eigenvalues of  $\lambda\mathbf{E} - \mathbf{A}$  and  $\mathbf{N}$  is a nilpotent submatrix corresponding to the infinite eigenvalues. The index of nilpotency provides an alternative expression for the index of the DAE, namely the index of the DAE is  $\nu$  if and only if  $\mathbf{N}^{\nu-1} \neq 0$  and  $\mathbf{N}^\nu = 0$ . If  $\mathbf{E}$  is nonsingular, then the index is 0; therefore, DAEs are characterized by an index greater than or equal to 1. See [24] for more details. More importantly, the quantities  $n_f$  and  $n_\infty$  give the dimension of the deflating subspaces of  $\lambda\mathbf{E} - \mathbf{A}$  corresponding to the finite and infinite eigenvalues. The spectral projectors onto the right and left deflating subspaces of  $\lambda\mathbf{E} - \mathbf{A}$  corresponding to the finite eigenvalues are defined as

$$\mathbf{\Pi}_r = \mathbf{T} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{T}^{-1}, \quad \text{and} \quad \mathbf{\Pi}_l = \mathbf{S} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{S}^{-1}.$$

The complementary projectors,  $\mathbf{\Pi}_{r,\infty}$  and  $\mathbf{\Pi}_{l,\infty}$ , are the spectral projectors onto the right and left deflating subspaces of  $\lambda\mathbf{E} - \mathbf{A}$  corresponding to the infinite eigenvalues and are defined as

$$\mathbf{\Pi}_{r,\infty} = \mathbf{I} - \mathbf{T} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{T}^{-1}, \quad \text{and} \quad \mathbf{\Pi}_{l,\infty} = \mathbf{I} - \mathbf{S} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{S}^{-1}.$$

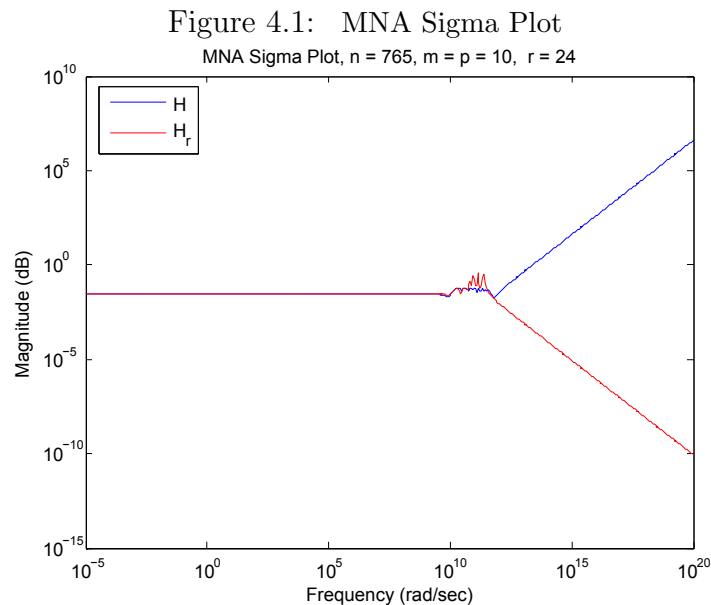
Details regarding the DAE index and Weierstrass canonical form are in [71].

Due to the singularity of  $\mathbf{E}$ , an additive decomposition of the transfer function  $\mathbf{H}(s)$  exists

as

$$\mathbf{H}(s) = \mathbf{R}(s) + \mathbf{P}(s)$$

where  $\mathbf{R}(s)$  is the strictly proper part and  $\mathbf{P}(s)$  is the polynomial part. Since interpolation results, such as Theorem 1.1, do not require the matrix  $\mathbf{E}$  to be nonsingular, interpolation methods can be applied to DAEs as long as  $\sigma\mathbf{E} - \mathbf{A}$  is nonsingular. For example, we applied Theorem 1.1 with  $M = N = 1$  and  $\sigma_i = \mu_i$  for  $i = 1, \dots, 24$  to a model of dimension  $n = 765$  with  $p = 10$  and  $m = 10$  that describes an electrical circuit and was obtained through modified nodal analysis. As the sigma plot shown in Figure 4.1 illustrates,  $\mathbf{H}_r(s)$  fails to capture the behavior of  $\mathbf{H}(s)$  for higher frequencies, leading to unbounded model reduction errors,  $\|\mathbf{H}(s) - \mathbf{H}_r(s)\|_{\mathcal{H}_2}$  and  $\|\mathbf{H}(s) - \mathbf{H}_r(s)\|_{\mathcal{H}_\infty}$ . In general, the key issue in



applying Theorem 1.1 to the DAE is that  $\mathbf{W}_r$  and  $\mathbf{V}_r$  tend to be of rank  $r$ ; therefore, the

reduced quantity  $\mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$  will be nonsingular provided  $r < \text{rank}(\mathbf{E})$ . More importantly, the nonsingularity of  $\mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$  implies that the reduced-order model is actually an ODE even though the full-order model is a DAE. Since the polynomial part of an ODE is zero, the reduced model does not account for the polynomial part of  $\mathbf{H}(s)$ . One way to introduce the presence of a polynomial part in the reduced model is simply to take  $\mathbf{P}_r(s) = \mathbf{P}(s)$  and define the reduced-order model as

$$\mathbf{H}_r(s) = \mathbf{R}_r(s) + \mathbf{P}(s).$$

In doing so, the polynomial parts of the full and reduced models cancel in the error term, resulting in the model reduction error only being expressed in terms of the strictly proper parts, namely  $\mathbf{H}(s) - \mathbf{H}_r(s) = \mathbf{R}(s) - \mathbf{R}_r(s)$ . As a result, the condition that  $\mathbf{P}_r(s) = \mathbf{P}(s)$  reduces the interpolation of the DAE to only requiring interpolation of the strictly proper part. The next theorem utilizes this observation to present a new interpolation result for DAEs. This theorem is extremely noteworthy as it provides the theoretical basis for how to achieve both interpolation and a bounded model reduction error.

**Theorem 4.1.** *Given the full-order model  $\mathbf{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$ , define  $\mathbf{\Pi}_l$  and  $\mathbf{\Pi}_r$  to be the spectral projectors onto the left and right deflating subspaces of the pencil  $\lambda\mathbf{E} - \mathbf{A}$  corresponding to the finite eigenvalues. Let the columns of  $\mathbf{W}_\infty$  and  $\mathbf{V}_\infty$  span the left and right deflating subspaces of the pencil  $\lambda\mathbf{E} - \mathbf{A}$  corresponding to the infinite eigenvalues. Let*

$\sigma$  and  $\mu$  be interpolation points and  $\mathbf{b} \in \mathbb{R}^m$  and  $\mathbf{c} \in \mathbb{R}^l$ . If

$$((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E})^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{b} \in \text{Ran}(\mathbf{V}_f) \quad \text{for } j = 1, \dots, N$$

$$((\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{E}^T)^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{C}^T\mathbf{c} \in \text{Ran}(\mathbf{W}_f) \quad \text{for } j = 1, \dots, M,$$

then with the choice of  $\mathbf{W} = [\mathbf{W}_f \quad \mathbf{W}_\infty]$  and  $\mathbf{V} = [\mathbf{V}_f \quad \mathbf{V}_\infty]$ , we have  $\mathbf{P}(s) = \mathbf{P}_r(s)$  and

$$1) \mathbf{H}^{(l)}(\sigma)\mathbf{b} = \mathbf{H}_r^{(l)}(\sigma)\mathbf{b} \quad \text{for } l = 0, \dots, N.$$

$$2) \mathbf{c}^T\mathbf{H}^{(l)}(\sigma) = \mathbf{c}^T\mathbf{H}_r^{(l)}(\sigma) \quad \text{for } l = 0, \dots, M.$$

$$3) \mathbf{c}^T\mathbf{H}^{(l)}(\sigma)\mathbf{b} = \mathbf{c}^T\mathbf{H}_r^{(l)}(\sigma)\mathbf{b} \quad \text{for } l = 0, \dots, M + N + 1.$$

*Proof.* Define  $\mathbf{B}_p = \mathbf{\Pi}_l\mathbf{B}$  and  $\mathbf{C}_p = \mathbf{C}\mathbf{\Pi}_r$ . Let

$$((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E})^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p\mathbf{b} \in \text{Ran}(\mathbf{V}_p) \quad \text{for } j = 1, \dots, N \quad (4.1.2)$$

$$((\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{E}^T)^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{C}_p^T\mathbf{c} \in \text{Ran}(\mathbf{W}_p) \quad \text{for } j = 1, \dots, M. \quad (4.1.3)$$

Then properties of spectral projectors imply that  $\text{span}\{\mathbf{V}_f \quad \mathbf{V}_\infty\} = \text{span}\{\mathbf{V}_p \quad \mathbf{V}_\infty\}$  and  $\text{span}\{\mathbf{W}_f \quad \mathbf{W}_\infty\} = \text{span}\{\mathbf{W}_p \quad \mathbf{W}_\infty\}$ . Hence, we will prove that (4.1.2) and (4.1.3) lead to the theorem's conclusion. Let the transfer functions  $\mathbf{H}(s)$  and  $\mathbf{H}_r(s)$  be additively decomposed as

$$\mathbf{H}(s) = \mathbf{R}(s) + \mathbf{P}(s) \quad \text{and} \quad \mathbf{H}_r(s) = \mathbf{R}_r(s) + \mathbf{P}_r(s)$$

where  $\mathbf{R}(s)$  and  $\mathbf{R}_r(s)$  denote the strictly proper parts and  $\mathbf{P}(s)$  and  $\mathbf{P}_r(s)$  are the polynomial

parts of  $\mathbf{H}(s)$  and  $\mathbf{H}_r(s)$ , respectively. Let the pencil  $\lambda\mathbf{E} - \mathbf{A}$  be transformed into the Weierstrass canonical form

$$\mathbf{E} = \mathbf{S} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix} \mathbf{T}^{-1}, \quad \text{and} \quad \mathbf{A} = \mathbf{S} \begin{bmatrix} \mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_\infty} \end{bmatrix} \mathbf{T}^{-1}, \quad (4.1.4)$$

where  $\mathbf{S}$  and  $\mathbf{T}$  are nonsingular and  $\mathbf{N}$  is nilpotent. Then the projectors  $\mathbf{\Pi}_l$  and  $\mathbf{\Pi}_r$  can be represented as

$$\mathbf{\Pi}_l = \mathbf{S} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{S}^{-1}, \quad \mathbf{\Pi}_r = \mathbf{T} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{T}^{-1}. \quad (4.1.5)$$

With this form, the transfer function  $\mathbf{H}(s)$  is

$$\mathbf{H}(s) = \mathbf{CT} \begin{bmatrix} s\mathbf{I}_{n_f} - \mathbf{J} & \mathbf{0} \\ \mathbf{0} & s\mathbf{N} - \mathbf{I}_{n_\infty} \end{bmatrix} \mathbf{S}^{-1}\mathbf{B} + \mathbf{D}. \quad (4.1.6)$$

Let  $\mathbf{T} = [\mathbf{T}_1 \quad \mathbf{T}_2]$  and  $\mathbf{S}^{-1} = [\mathbf{S}_1 \quad \mathbf{S}_2]^T$  be partitioned conformally to  $\mathbf{E}$  and  $\mathbf{A}$ . Then (4.1.6) gives

$$\mathbf{H}(s) = \mathbf{CT}_1(s\mathbf{I}_{n_f} - \mathbf{J})^{-1}\mathbf{S}_1^T\mathbf{B} + \mathbf{CT}_2(s\mathbf{N} - \mathbf{I}_{n_\infty})^{-1}\mathbf{S}_2^T\mathbf{B} + \mathbf{D}. \quad (4.1.7)$$

Without loss of generality, assume that  $\mathbf{W}_\infty^T = [\mathbf{0}, \mathbf{I}_{n_\infty}]\mathbf{S}^{-1}$  and  $\mathbf{V}_\infty = \mathbf{T}[\mathbf{0}, \mathbf{I}_{n_\infty}]^T$ . By

(4.1.7), the polynomial part of  $\mathbf{H}(s)$  then has the form

$$\mathbf{P}(s) = \mathbf{C}\mathbf{V}_\infty(s\mathbf{W}_\infty^T\mathbf{E}\mathbf{V}_\infty - \mathbf{W}_\infty^T\mathbf{A}\mathbf{V}_\infty)^{-1}\mathbf{W}_\infty^T\mathbf{B} + \mathbf{D}.$$

To see that  $\mathbf{P}(s) = \mathbf{P}_r(s)$ , note that the matrices  $\mathbf{W} = [\mathbf{W}_p \quad \mathbf{W}_\infty]$  and  $\mathbf{V} = [\mathbf{V}_p \quad \mathbf{V}_\infty]$  yield a reduced transfer function  $\mathbf{H}_r(s) = \mathbf{C}_r(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r$ , where

$$\begin{aligned} \mathbf{E}_r = \mathbf{W}^T\mathbf{E}\mathbf{V} &= \begin{bmatrix} \mathbf{W}_p^T\mathbf{E}\mathbf{V}_p & \mathbf{W}_p^T\mathbf{E}\mathbf{V}_\infty \\ \mathbf{W}_\infty^T\mathbf{E}\mathbf{V}_p & \mathbf{W}_\infty^T\mathbf{E}\mathbf{V}_\infty \end{bmatrix}, & \mathbf{B}_r = \mathbf{W}^T\mathbf{B} &= \begin{bmatrix} \mathbf{W}_p^T\mathbf{B} \\ \mathbf{W}_\infty^T\mathbf{B} \end{bmatrix}, \\ \mathbf{A}_r = \mathbf{W}^T\mathbf{A}\mathbf{V} &= \begin{bmatrix} \mathbf{W}_p^T\mathbf{A}\mathbf{V}_p & \mathbf{W}_p^T\mathbf{A}\mathbf{V}_\infty \\ \mathbf{W}_\infty^T\mathbf{A}\mathbf{V}_p & \mathbf{W}_\infty^T\mathbf{A}\mathbf{V}_\infty \end{bmatrix}, & \mathbf{C}_r = \mathbf{C}\mathbf{V} &= [\mathbf{C}\mathbf{V}_p, \mathbf{C}\mathbf{V}_\infty]. \end{aligned}$$

Due to the properties of spectral projectors, we have

$$\mathbf{E}\mathbf{\Pi}_r = \mathbf{\Pi}_l\mathbf{E} \quad \text{and} \quad \mathbf{A}\mathbf{\Pi}_r = \mathbf{\Pi}_l\mathbf{A}. \quad (4.1.8)$$

Coupled with the assumptions that

$$\begin{aligned} ((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E})^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p\mathbf{b} &\in \text{Ran}(\mathbf{V}_p) \quad \text{for } j = 1, \dots, N \\ ((\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{E}^T)^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{C}_p^T\mathbf{c} &\in \text{Ran}(\mathbf{W}_p) \quad \text{for } j = 1, \dots, M, \end{aligned}$$



(4.1.8) gives that  $\text{Ran}(\mathbf{V}_p) \in \text{Ran}(\mathbf{\Pi}_r)$  and  $\text{Ran}(\mathbf{W}_p) \in \text{Ran}(\mathbf{\Pi}_l^T)$ . This observation gives

$$\mathbf{V}_p = \mathbf{\Pi}_r \mathbf{V}_p \quad \mathbf{W}_p = \mathbf{\Pi}_l^T \mathbf{W}_p \quad \mathbf{V}_\infty = (\mathbf{I} - \mathbf{\Pi}_r) \mathbf{V}_\infty \quad \mathbf{W}_\infty = (\mathbf{I} - \mathbf{\Pi}_l^T) \mathbf{W}_\infty. \quad (4.1.9)$$

Using (4.1.8) and (4.1.9), we have

$$\mathbf{W}_p^T \mathbf{E} \mathbf{V}_\infty = \mathbf{0}, \quad \mathbf{W}_\infty^T \mathbf{E} \mathbf{V}_p = \mathbf{0}, \quad \mathbf{W}_p^T \mathbf{A} \mathbf{V}_\infty = \mathbf{0}, \quad \mathbf{W}_\infty^T \mathbf{A} \mathbf{V}_p = \mathbf{0}.$$

Thus,  $\mathbf{H}_r(s)$  is expressed as

$$\begin{aligned} \mathbf{H}_r(s) &= \mathbf{C} \mathbf{V}_p (s \mathbf{W}_p^T \mathbf{E} \mathbf{V}_p - \mathbf{W}_p^T \mathbf{A} \mathbf{V}_p)^{-1} \mathbf{W}_p^T \mathbf{B} \\ &+ \mathbf{C} \mathbf{V}_\infty (s \mathbf{W}_\infty^T \mathbf{E} \mathbf{V}_\infty - \mathbf{W}_\infty^T \mathbf{A} \mathbf{V}_\infty)^{-1} \mathbf{W}_\infty^T \mathbf{B} + \mathbf{D}. \end{aligned}$$

This implies that

$$\mathbf{R}_r(s) = \mathbf{C} \mathbf{V}_p (s \mathbf{W}_p^T \mathbf{E} \mathbf{V}_p - \mathbf{W}_p^T \mathbf{A} \mathbf{V}_p)^{-1} \mathbf{W}_p^T \mathbf{B} \quad (4.1.10)$$

and that the polynomial parts of the full and reduced models match since

$$\mathbf{P}_r(s) = \mathbf{C} \mathbf{V}_\infty (s \mathbf{W}_\infty^T \mathbf{E} \mathbf{V}_\infty - \mathbf{W}_\infty^T \mathbf{A} \mathbf{V}_\infty)^{-1} \mathbf{W}_\infty^T \mathbf{B} + \mathbf{D} = \mathbf{P}(s).$$

Since  $\mathbf{P}(s) = \mathbf{P}_r(s)$ , the proof of the interpolation result reduces to proving  $\mathbf{R}(\sigma) = \mathbf{R}_r(\sigma)$ .

To prove this, we first note that (4.1.4) and (4.1.5) imply that

$$\mathbf{C}_p(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p = \mathbf{CT} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} s\mathbf{I}_{n_f} - \mathbf{J} & \mathbf{0} \\ \mathbf{0} & s\mathbf{N} - \mathbf{I}_{n_\infty} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}_{n_f} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{S}^{-1}\mathbf{B}.$$

Coupled with (4.1.7), this gives

$$\mathbf{C}_p(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p = \mathbf{CT}_1(\sigma\mathbf{I}_{n_f} - \mathbf{J})^{-1}\mathbf{S}_1^T\mathbf{B} = \mathbf{R}(\sigma). \quad (4.1.11)$$

Further note that  $\mathbf{V}_p = \mathbf{\Pi}_r\mathbf{V}_p$  and  $\mathbf{W}_p = \mathbf{\Pi}_l^T\mathbf{W}_p$  imply

$$\mathbf{C}_p\mathbf{V}_p = \mathbf{C}\mathbf{\Pi}_r\mathbf{V}_p = \mathbf{C}\mathbf{V}_p \quad \text{and} \quad \mathbf{W}_p^T\mathbf{B}_p = \mathbf{W}_p^T\mathbf{\Pi}_l\mathbf{B} = \mathbf{W}_p^T\mathbf{B}. \quad (4.1.12)$$

Due to the assumptions

$$((\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{E})^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p\mathbf{b} \in \text{Ran}(\mathbf{V}_p) \quad \text{for } j = 1, \dots, N$$

$$((\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{E}^T)^{j-1}(\sigma\mathbf{E} - \mathbf{A})^{-T}\mathbf{C}_p^T\mathbf{c} \in \text{Ran}(\mathbf{W}_p) \quad \text{for } j = 1, \dots, M$$

holding, Theorem 1.1 gives

$$\begin{aligned} \mathbf{C}_p\mathbf{V}_p(\sigma\mathbf{W}_p^T\mathbf{E}\mathbf{V}_p - \mathbf{W}_p^T\mathbf{A}\mathbf{V}_p)^{-1}\mathbf{W}_p^T\mathbf{B}_p\mathbf{b} &= \mathbf{C}_p(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p\mathbf{b} \\ \mathbf{c}^T\mathbf{C}_p\mathbf{V}_p(\sigma\mathbf{W}_p^T\mathbf{E}\mathbf{V}_p - \mathbf{W}_p^T\mathbf{A}\mathbf{V}_p)^{-1}\mathbf{W}_p^T\mathbf{B}_p &= \mathbf{c}^T\mathbf{C}_p(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_p. \end{aligned}$$

Using (4.1.11) and (4.1.12), the above equations become

$$\begin{aligned}\mathbf{CV}_p(\sigma\mathbf{W}_p^T\mathbf{E}\mathbf{V}_p - \mathbf{W}_p^T\mathbf{A}\mathbf{V}_p)^{-1}\mathbf{W}_p^T\mathbf{B}\mathbf{b} &= \mathbf{R}(\sigma)\mathbf{b} \\ \mathbf{c}^T\mathbf{CV}_p(\sigma\mathbf{W}_p^T\mathbf{E}\mathbf{V}_p - \mathbf{W}_p^T\mathbf{A}\mathbf{V}_p)^{-1}\mathbf{W}_p^T\mathbf{B} &= \mathbf{c}^T\mathbf{R}(\sigma).\end{aligned}$$

Coupled with (4.1.10), this gives

$$\mathbf{R}_r(\sigma)\mathbf{b} = \mathbf{R}(\sigma)\mathbf{b} \quad \text{and} \quad \mathbf{c}^T\mathbf{R}_r(\sigma) = \mathbf{c}^T\mathbf{R}(\sigma). \quad (4.1.13)$$

Since both (a) and (b) of Theorem 1.1 hold, we have  $\mathbf{c}^T\mathbf{R}'_r(\sigma)\mathbf{b} = \mathbf{c}^T\mathbf{R}'(\sigma)\mathbf{b}$ . The remainder of the proof follows by induction.  $\square$

*Remark:* The one-sided interpolation result from Theorem 1.1 does not hold. If  $\mathbf{W}_p = \mathbf{V}_p$ , for example, then

$$\begin{aligned}\mathbf{H}_r(s) &= \mathbf{CV}_p(s\mathbf{V}_p^T\mathbf{E}\mathbf{V}_p - \mathbf{V}_p^T\mathbf{A}\mathbf{V}_p)^{-1}\mathbf{V}_p^T\mathbf{B} + \mathbf{CV}_\infty(s\mathbf{W}_\infty^T\mathbf{E}\mathbf{V}_\infty - \mathbf{W}_\infty^T\mathbf{A}\mathbf{V}_\infty)^{-1}\mathbf{W}_\infty^T\mathbf{B} + \mathbf{D} + \\ &\quad \mathbf{CV}_p(s\mathbf{V}_p^T\mathbf{E}\mathbf{V}_p - \mathbf{V}_p^T\mathbf{A}\mathbf{V}_p)^{-1}(s\mathbf{V}_p^T\mathbf{E}\mathbf{V}_\infty - \mathbf{V}_p^T\mathbf{A}\mathbf{V}_\infty)(s\mathbf{W}_\infty^T\mathbf{E}\mathbf{V}_\infty - \mathbf{W}_\infty^T\mathbf{A}\mathbf{V}_\infty)^{-1}\mathbf{W}_\infty^T\mathbf{B}.\end{aligned}$$

Due to the inclusion of  $\mathbf{V}_\infty$  and  $\mathbf{W}_\infty$ , we still have

$$\mathbf{P}(s) = \mathbf{CV}_\infty(s\mathbf{W}_\infty^T\mathbf{E}\mathbf{V}_\infty - \mathbf{W}_\infty^T\mathbf{A}\mathbf{V}_\infty)^{-1}\mathbf{W}_\infty^T\mathbf{B} + \mathbf{D},$$

and a similar argument as in the two-sided case yields

$$\mathbf{c}^T \mathbf{C} \mathbf{V}_p (\sigma \mathbf{V}_p^T \mathbf{E} \mathbf{V}_p - \mathbf{V}_p^T \mathbf{A} \mathbf{V}_p)^{-1} \mathbf{V}_p^T \mathbf{B} \mathbf{b} = \mathbf{c}^T \mathbf{R}(\sigma) \mathbf{b}.$$

These observations give that

$$\mathbf{c}^T \mathbf{H}(\sigma) \mathbf{b} - \mathbf{c}^T \mathbf{H}_r(\sigma) \mathbf{b} =$$

$$\mathbf{c}^T \mathbf{C} \mathbf{V}_p (s \mathbf{V}_p^T \mathbf{E} \mathbf{V}_p - \mathbf{W}_\infty^T \mathbf{A} \mathbf{V}_\infty)^{-1} (s \mathbf{V}_p^T \mathbf{E} \mathbf{V}_\infty - \mathbf{V}_p^T \mathbf{A} \mathbf{V}_\infty) (s \mathbf{W}_\infty^T \mathbf{E} \mathbf{V}_\infty - \mathbf{W}_\infty^T \mathbf{A} \mathbf{V}_\infty)^{-1} \mathbf{W}_\infty^T \mathbf{B} \mathbf{b}.$$

For the two-sided case, the  $(s \mathbf{V}_p^T \mathbf{E} \mathbf{V}_\infty - \mathbf{V}_p^T \mathbf{A} \mathbf{V}_\infty)$  term becomes  $s \mathbf{W}_p^T \mathbf{E} \mathbf{V}_\infty - \mathbf{W}_p^T \mathbf{A} \mathbf{V}_\infty = \mathbf{0}$ , and so  $\mathbf{c}^T \mathbf{H}(\sigma) \mathbf{b} - \mathbf{c}^T \mathbf{H}_r(\sigma) \mathbf{b} = 0$ . In the one-sided case, however, this error term does not vanish, implying that  $\mathbf{R}_r(s) \neq \mathbf{P}(s)$ .

Using Theorem 4.1, an algorithm for optimal  $\mathcal{H}_2$  reduction of DAEs in the interpolatory framework is presented in Algorithm 4.1.1. It is important to emphasize that one of the main issues is the computation of the deflating subspaces. For large-scale dynamical systems, computing  $\mathbf{\Pi}_{r,\infty}$  and  $\mathbf{\Pi}_{l,\infty}$  may not be computationally desirable or even feasible. See [35], [36], [55], [56], and [77]. Fortunately, the remainder of this chapter discusses the utilization of properties associated with index-1 and Hessenberg index-2 DAEs to circumvent the explicit computation of the projectors.

**Algorithm 4.1.1. Interpolatory  $\mathcal{H}_2$  Optimal Model Reduction Method for DAE Descriptor Systems**

1. Make an initial selection of the interpolation points  $\{\sigma_i\}_{i=1}^r$ , and tangent directions  $\{\mathbf{b}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i\}_{i=1}^r$ .

2. Compute  $\mathbf{B}_p = \mathbf{\Pi}_l \mathbf{B}$  and  $\mathbf{C}_p = \mathbf{C} \mathbf{\Pi}_r$ .

3.  $\mathbf{V}_f = [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B}_p \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B}_p \mathbf{b}_r]$ ,

$$\mathbf{W}_f = \left[ (\mathbf{c}_1^T \mathbf{C}_p (\sigma_1 \mathbf{E} - \mathbf{A})^{-1})^T, \dots, (\mathbf{c}_r^T \mathbf{C}_p (\sigma_r \mathbf{E} - \mathbf{A})^{-1})^T \right].$$

4. while (not converged)

(a)  $\mathbf{A}_r^{sp} = \mathbf{W}_f^T \mathbf{A} \mathbf{V}_f$ ,  $\mathbf{E}_r^{sp} = \mathbf{W}_f^T \mathbf{E} \mathbf{V}_f$ ,  $\mathbf{B}_r^{sp} = \mathbf{W}_f^T \mathbf{B}$ , and  $\mathbf{C}_r^{sp} = \mathbf{C} \mathbf{V}_f$ .

(b) Compute  $\mathbf{A}_r^{sp} \mathbf{x}_i = \tilde{\lambda}_i \mathbf{E}_r^{sp} \mathbf{x}_i$  and  $\mathbf{y}_i^T \mathbf{A}_r^{sp} = \tilde{\lambda}_i \mathbf{y}_i^T \mathbf{E}_r^{sp}$  with  $\mathbf{y}_i^T \mathbf{E}_r \mathbf{x}_j = \delta_{ij}$

where  $\mathbf{y}_i$  and  $\mathbf{x}_i$  are left and right eigenvectors associated with  $\tilde{\lambda}_i$ .

(c)  $\sigma_i \leftarrow -\tilde{\lambda}_i$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{y}_i^T \mathbf{B}_r^{sp}$  and  $\mathbf{c}_i \leftarrow \mathbf{C}_r^{sp} \mathbf{x}_i$ , for  $i = 1, \dots, r$ .

(d)  $\mathbf{V}_f = [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B}_p \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B}_p \mathbf{b}_r]$ ,

$$\mathbf{W}_f = \left[ (\mathbf{c}_1^T \mathbf{C}_p (\sigma_1 \mathbf{E} - \mathbf{A})^{-1})^T, \dots, (\mathbf{c}_r^T \mathbf{C}_p (\sigma_r \mathbf{E} - \mathbf{A})^{-1})^T \right].$$

5. Compute  $\mathbf{W}_\infty$  and  $\mathbf{V}_\infty$  such that  $\text{Im}(\mathbf{W}_\infty) = \text{Im}(\mathbf{I} - \mathbf{\Pi}_l)^T$  and

$$\text{Im}(\mathbf{V}_\infty) = \text{Im}(\mathbf{I} - \mathbf{\Pi}_r).$$

6. Set  $\mathbf{V} = [\mathbf{V}_f \ \mathbf{V}_\infty]$  and  $\mathbf{W} = [\mathbf{W}_f \ \mathbf{W}_\infty]$ .

7.  $\mathbf{E}_r = \mathbf{W}^T \mathbf{E} \mathbf{V}$ ,  $\mathbf{A}_r = \mathbf{W}^T \mathbf{A} \mathbf{V}$ ,  $\mathbf{B}_r = \mathbf{W}^T \mathbf{B}$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}$ ,  $\mathbf{D}_r = \mathbf{D}$ .

## 4.2 Index-1 DAEs

We consider the following index-1 differential algebraic equation:

$$\mathbf{G}(s) : \begin{cases} \mathbf{E}_{11}\dot{\mathbf{x}}_1(t) + \mathbf{E}_{12}\dot{\mathbf{x}}_2(t) = \mathbf{A}_{11}\mathbf{x}_1(t) + \mathbf{A}_{12}\mathbf{x}_2(t) + \mathbf{B}_1\mathbf{u}(t) \\ \mathbf{0} = \mathbf{A}_{21}\mathbf{x}_1(t) + \mathbf{A}_{22}\mathbf{x}_2(t) + \mathbf{B}_2\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}_1\mathbf{x}_1(t) + \mathbf{C}_2\mathbf{x}_2(t) \end{cases} \quad (4.2.1)$$

where the state is  $\mathbf{x}(t) = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \in \mathbb{R}^n$  with  $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ ,  $\mathbf{x}_2 \in \mathbb{R}^{n_2}$  and  $n_1 + n_2 = n$ , the input is  $\mathbf{u}(t) \in \mathbb{R}^m$ , the output is  $\mathbf{y}(t) \in \mathbb{R}^p$ , and  $\mathbf{E}_{11}, \mathbf{A}_{11} \in \mathbb{R}^{n_1 \times n_1}$ ,  $\mathbf{E}_{12}, \mathbf{A}_{12} \in \mathbb{R}^{n_1 \times n_2}$ ,  $\mathbf{A}_{21} \in \mathbb{R}^{n_2 \times n_1}$ ,  $\mathbf{A}_{22} \in \mathbb{R}^{n_2 \times n_2}$ ,  $\mathbf{B}_1 \in \mathbb{R}^{n_1 \times m}$ ,  $\mathbf{B}_2 \in \mathbb{R}^{n_2 \times m}$ ,  $\mathbf{C}_1 \in \mathbb{R}^{p \times n_1}$ ,  $\mathbf{C}_2 \in \mathbb{R}^{p \times n_2}$ . We assume that  $\mathbf{A}_{22}$  and  $\mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$  are nonsingular.

For these systems, Theorem 1.1 can be applied to construct a reduced model  $\mathbf{G}_r(s)$ . As the previous example illustrates,  $\mathbf{G}_r(s)$  will oftentimes be strictly proper, implying that the polynomial parts of  $\mathbf{G}(s)$  and  $\mathbf{G}_r(s)$  will not match. However, if we assume that  $\mathbf{G}(s)$  is an index-1 DAE, then the polynomial part is a constant matrix as the next lemma states and proves.

**Lemma 4.2.** *Let  $\mathbf{G}(s)$  be an index-1 DAE such that  $\mathbf{A}_{22}$  and  $\mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$  are invertible.*

*Then the polynomial part of  $\mathbf{G}(s)$  is a constant matrix, namely*

$$\lim_{s \rightarrow \infty} \mathbf{G}(s) = \mathbf{C}_1\mathbf{M}_1\mathbf{B}_2 + \mathbf{C}_2\mathbf{M}_2\mathbf{B}_2$$

where

$$\mathbf{M}_1 = (\mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{E}_{12}\mathbf{A}_{22}^{-1} \quad (4.2.2)$$

and

$$\mathbf{M}_2 = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{E}_{11} - \mathbf{E}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{E}_{12}\mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}. \quad (4.2.3)$$

*Proof.* Consider

$$\begin{bmatrix} s\mathbf{E}_{11} - \mathbf{A}_{11} & s\mathbf{E}_{12} - \mathbf{A}_{12} \\ -\mathbf{A}_{21} & -\mathbf{A}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}$$

or equivalently

$$\begin{bmatrix} s\mathbf{E}_{11} - \mathbf{A}_{11} & s\mathbf{E}_{12} - \mathbf{A}_{12} \\ -\mathbf{A}_{21} & -\mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}.$$

Then we have

$$(s\mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{z}_1 + (s\mathbf{E}_{12} - \mathbf{A}_{12})\mathbf{z}_2 = \mathbf{B}_1 \quad (4.2.4)$$

$$-\mathbf{A}_{21}\mathbf{z}_1 - \mathbf{A}_{22}\mathbf{z}_2 = \mathbf{B}_2. \quad (4.2.5)$$

Solving (4.2.5) for  $\mathbf{z}_2$  gives  $\mathbf{z}_2 = -\mathbf{A}_{22}^{-1}(\mathbf{B}_2 + \mathbf{A}_{21}\mathbf{z}_1)$ , and so

$$\mathbf{z}_1 = [(s\mathbf{E}_{11} - \mathbf{A}_{11}) - (s\mathbf{E}_{12} - \mathbf{A}_{12})\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1} [\mathbf{B}_1 + (s\mathbf{E}_{12} - \mathbf{A}_{12})\mathbf{A}_{22}^{-1}\mathbf{B}_2],$$

implying that

$$\lim_{s \rightarrow \infty} \mathbf{z}_1 = (\mathbf{E}_{11} - \mathbf{E}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21})^{-1} \mathbf{E}_{12} \mathbf{A}_{22}^{-1} \mathbf{B}_2.$$

Coupled with (4.2.5), this gives

$$\lim_{s \rightarrow \infty} \mathbf{z}_2 = \left[ -\mathbf{A}_{22}^{-1} \mathbf{A}_{21} (\mathbf{E}_{11} - \mathbf{E}_{12} \mathbf{A}_{22}^{-1} \mathbf{A}_{21})^{-1} \mathbf{E}_{12} \mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1} \right] \mathbf{B}_2.$$

Finally, note that  $\lim_{s \rightarrow \infty} \mathbf{G}(s) = \lim_{s \rightarrow \infty} \mathbf{C}_1 \mathbf{z}_1 + \mathbf{C}_2 \mathbf{z}_2$ . □

**Lemma 4.3.** [3] Given  $\mathbf{G}(s)$  as in (4.2.1),  $r$  distinct points  $\{\sigma_i\}_{i=1}^r$ , left tangential directions  $\{\mathbf{c}_i\}_{i=1}^r$ , and right tangential directions  $\{\mathbf{b}_i\}_{i=1}^r$ , let  $\mathbf{V}_r \in \mathbb{C}^{n \times r}$  and  $\mathbf{W}_r \in \mathbb{C}^{n \times r}$  be as defined in (1.4.7) and (1.4.8), respectively. Define  $\mathcal{B}$  and  $\mathcal{C}$  to be the matrices composed of the tangential directions as

$$\mathcal{B} = [\mathbf{b}_1, \dots, \mathbf{b}_r], \quad \text{and} \quad \mathcal{C}^T = [\mathbf{c}_1, \dots, \mathbf{c}_r]^T, \quad (4.2.6)$$

and define the reduced-order model quantities as

$$\begin{aligned} \mathbf{E}_r &= \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, & \mathbf{A}_r &= \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r + \mathcal{C}^T \mathbf{D}_r \mathcal{B}, \\ \mathbf{B}_r &= \mathbf{W}_r^T \mathbf{B} - \mathcal{C}^T \mathbf{D}_r, & \mathbf{C}_r &= \mathbf{C} \mathbf{V}_r - \mathbf{D}_r \mathcal{B}, \\ \mathbf{D}_r &= \mathbf{C}_1 \mathbf{M}_1 \mathbf{B}_2 + \mathbf{C}_2 \mathbf{M}_2 \mathbf{B}_2 \end{aligned} \quad (4.2.7)$$

where  $\mathbf{E}_r$  is nonsingular. Then the polynomial parts of  $\mathbf{G}_r(s) = \mathbf{C}_r (s \mathbf{E}_r - \mathbf{A}_r)^{-1} \mathbf{B}_r + \mathbf{D}_r$



and  $\mathbf{G}(s)$  match, and  $\mathbf{G}_r$  satisfies

$$\mathbf{G}(\sigma_i)\mathbf{b}_i = \mathbf{G}_r(\sigma_i)\mathbf{b}_i, \quad \mathbf{c}_i^T \mathbf{G}(\sigma_i) = \mathbf{c}_i^T \mathbf{G}_r(\sigma_i), \quad \mathbf{c}_i^T \mathbf{G}'(\sigma_i)\mathbf{b}_i = \mathbf{c}_i^T \mathbf{G}_r'(\sigma_i)\mathbf{b}_i \quad (4.2.8)$$

for  $i = 1, \dots, r$ , provided  $\sigma_i \mathbf{E} - \mathbf{A}$  and  $\sigma_i \mathbf{E}_r - \mathbf{A}_r$  are nonsingular.

*Proof.* By Lemma 4.2, choosing  $\mathbf{D}_r = \mathbf{C}_1 \mathbf{M}_1 \mathbf{B}_2 + \mathbf{C}_2 \mathbf{M}_2 \mathbf{B}_2$  ensures that the polynomial parts of  $\mathbf{G}(s)$  and  $\mathbf{G}_r(s)$  match since

$$\lim_{s \rightarrow \infty} \mathbf{G}(s) = \lim_{s \rightarrow \infty} \mathbf{G}_r(s) = \mathbf{D}_r,$$

and the interpolation as stated in (4.2.8) holds due to Theorem 3 of [3].  $\square$

Applying Lemma 4.3 results in Algorithm 4.2.1, which delineates the method to achieve Hermite interpolation of index-1 DAEs.

#### Algorithm 4.2.1. Interpolatory Index-1 Model Reduction Method

1. Make an initial selection of the interpolation points  $\{\sigma_i\}_{i=1}^r$ , and tangent directions  $\{\mathbf{b}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i\}_{i=1}^r$ .

2.  $\mathbf{V}_r = [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_r]$ ,

$$\mathbf{W}_r = \left[ (\mathbf{c}_1^T \mathbf{C} (\sigma_1 \mathbf{E} - \mathbf{A})^{-1})^T, \dots, (\mathbf{c}_r^T \mathbf{C} (\sigma_r \mathbf{E} - \mathbf{A})^{-1})^T \right].$$

3. Define  $\mathbf{D}_r = \mathbf{C}_1 \mathbf{M}_1 \mathbf{B}_2 + \mathbf{C}_2 \mathbf{M}_2 \mathbf{B}_2$  where  $\mathbf{M}_1$  and  $\mathbf{M}_2$  are defined in (4.2.2) and (4.2.3), respectively.

4. Define  $\mathcal{B} = [ \mathbf{b}_1, \dots, \mathbf{b}_r ]$ , and  $\mathcal{C}^T = [ \mathbf{c}_1, \dots, \mathbf{c}_r ]^T$ .

5.  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r$ ,  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r + \mathcal{C}^T \mathbf{D}_r \mathcal{B}$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B} - \mathcal{C}^T \mathbf{D}_r$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r - \mathbf{D}_r \mathcal{B}$ .

Of course, Lemma 4.3 also provides the theoretical basis for an IRKA iteration to reduce index-1 DAEs. One option is to simply apply the IRKA iteration to the DAE and then define the reduced-order model as

$$\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r + \mathcal{C}^T \mathbf{D}_r \mathcal{B}, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B} - \mathcal{C}^T \mathbf{D}_r, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r - \mathbf{D}_r \mathcal{B}$$

once the IRKA iteration has converged. While the addition of the  $\mathbf{D}_r$  term upon convergence results in a bounded model reduction error, this addition also shifts the spectrum, and thereby the resulting reduced model will not satisfy the optimal  $\mathcal{H}_2$  necessary conditions. In order to obtain the optimal  $\mathcal{H}_2$  model, the  $\mathbf{D}_r$  term must be included as part of the IRKA iteration as shown in Algorithm 4.2.2. Then upon convergence of Algorithm 4.2.2 the reduced-order model will satisfy the optimal  $\mathcal{H}_2$  conditions and have a bounded model reduction error.

**Algorithm 4.2.2. IRKA for Index-1 DAEs (IRKA-D):**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2.  $\mathbf{V}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_r ]$ .

$$3. \mathbf{W}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_r ].$$

4. Define  $\mathbf{D}_r = \mathbf{C}_1 \mathbf{M}_1 \mathbf{B}_2 + \mathbf{C}_2 \mathbf{M}_2 \mathbf{B}_2$  where  $\mathbf{M}_1$  and  $\mathbf{M}_2$  are defined in (4.2.2) and (4.2.3), respectively.

$$5. \text{ Define } \mathbf{B} = [ \mathbf{b}_1, \dots, \mathbf{b}_r ], \text{ and } \mathcal{C}^T = [ \mathbf{c}_1, \dots, \mathbf{c}_r ]^T.$$

6. while (not converged)

$$(a) \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r + \mathcal{C}^T \mathbf{D}_r \mathbf{B}, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B} - \mathcal{C}^T \mathbf{D}_r, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r - \mathbf{D}_r \mathbf{B}.$$

(b) Compute  $\mathbf{Y}^T \mathbf{A}_r \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_r \mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_r - \mathbf{A}_r$ .

$$(c) \sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r) \text{ for } i = 1, \dots, r, \mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r, \text{ and } \mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i.$$

$$(d) \mathbf{V}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_r ].$$

$$(e) \mathbf{W}_r = [ (\sigma_1 \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_1, \dots, (\sigma_r \mathbf{E} - \mathbf{A})^{-T} \mathbf{C}^T \mathbf{c}_r ].$$

$$7. \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E} \mathbf{V}_r, \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A} \mathbf{V}_r + \mathcal{C}^T \mathbf{D}_r \mathbf{B}, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B} - \mathcal{C}^T \mathbf{D}_r, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r - \mathbf{D}_r \mathbf{B}.$$

**Theorem 4.4.** Suppose  $\mathbf{G}(s)$  and  $\mathbf{G}_r(s)$  are real stable dynamical systems. Let

$$\mathbf{G}_r(s) = \sum_{i=1}^r \frac{1}{s - \hat{\lambda}_i} \mathbf{c}_i \mathbf{b}_i^T + \mathbf{D}_r$$

where  $\{\hat{\lambda}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i \mathbf{b}_i^T\}_{i=1}^r$  are the simple poles and residues of  $\mathbf{G}_r(s)$ , respectively. Then the reduced-order model,  $\mathbf{G}_r(s)$ , obtained with IRKA-D (Algorithm 4.2.2) satisfies the first-order necessary conditions of the  $\mathcal{H}_2$  optimality problem.

*Proof.* The proof follows immediately from the shift selection prescribed in IRKA-D coupled with Lemma 4.3. □

## 4.3 Numerical Results for Index-1 DAEs

In this section, we consider index-1 DAE models of electrical circuits resulting from modified nodal analysis. We are interested in comparing Algorithm 4.1.1 and Algorithm 4.2.2. To emphasize the distinction between the algorithms, we will refer to Algorithm 4.1.1 and Algorithm 4.2.2 as IRKA-P and IRKA-D, respectively. From a computation perspective, IRKA-P requires substantially more work than IRKA-D. The aim of this section is to study if IRKA-D provides models of high fidelity, indicating that computational savings are possible without significant loss of accuracy in the reduced-order model.

### 4.3.1 TL1 Model

The first model, the TL1 Model, describes a transmission line with  $n = 600$  and was reduced to  $r = 16$ . IRKA-D converged in four iterations. In Figure 4.2, the bode plots indicate a nice match between the full and reduced models. The overall model reduction errors as displayed in Table 4.1 show that the method yields acceptable errors in both the transfer function and the strictly proper part. As a result, the TL1 Model supports the argument that the computation of the spectral projectors is unnecessary since IRKA-D provides a model of high fidelity.

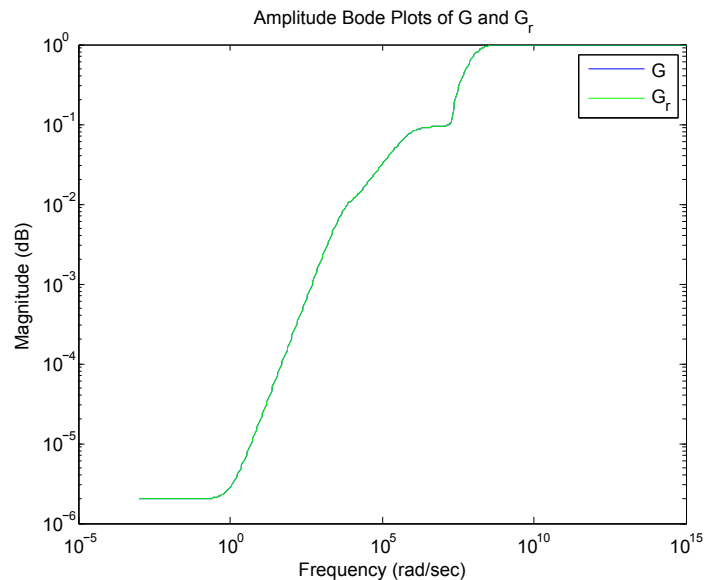
Figure 4.2: TL1 Model: Amplitude Bode Plots of  $\mathbf{G}(s)$  and  $\mathbf{G}_r(s)$ 

Table 4.1: Model Reduction Errors for the TL1 Model

Method	$\frac{\ \mathbf{G}-\mathbf{G}_r\ _{\mathcal{H}_\infty}}{\ \mathbf{G}\ _{\mathcal{H}_\infty}}$	$\frac{\ \mathbf{G}^{sp}-\mathbf{G}_r^{sp}\ _{\mathcal{H}_\infty}}{\ \mathbf{G}^{sp}\ _{\mathcal{H}_\infty}}$
IRKA-D	$9.36 \times 10^{-5}$	$1.46 \times 10^{-4}$

### 4.3.2 TL2 Model

The second model, the TL2 Model, describes an RLC circuit using modified nodal analysis. This results in a SISO model of dimension  $n = 400$ , which was reduced to  $r = 20$ . IRKA-D required only three iterations to yield the bode plots for the full and reduced models as given in Figure 4.3. These bode plots indicate a good match between the full and reduced models. In addition, Table 4.2 indicates that the  $\frac{\|\mathbf{G}-\mathbf{G}_r\|_{\mathcal{H}_\infty}}{\|\mathbf{G}\|_{\mathcal{H}_\infty}}$  and  $\frac{\|\mathbf{G}-\mathbf{G}_r^{sp}\|_{\mathcal{H}_\infty}}{\|\mathbf{G}^{sp}\|_{\mathcal{H}_\infty}}$  errors are satisfactory. Therefore, these models illustrate that explicit computation of the projectors can be circumvented without any significant loss of accuracy in the reduction of the index-1 model.

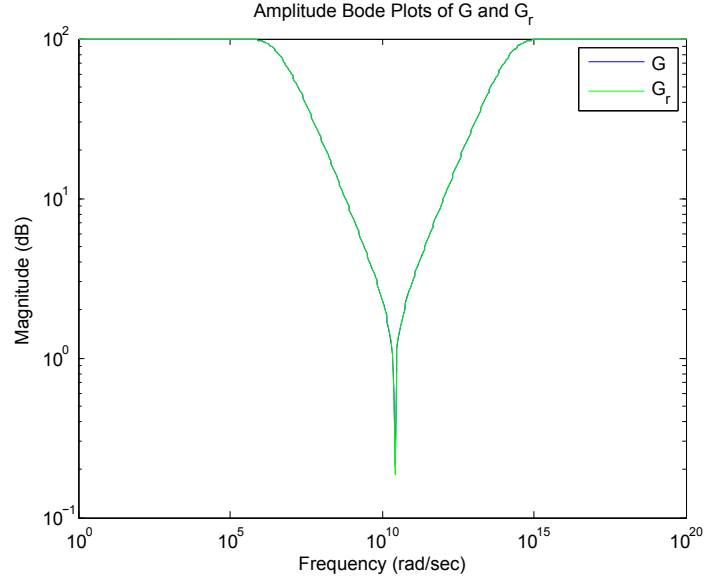
Figure 4.3: TL2 Model: Amplitude Bode Plots of  $\mathbf{G}(s)$  and  $\mathbf{G}_r(s)$ 

Table 4.2: Model Reduction Errors for the TL2 Model

Method	$\frac{\ \mathbf{G}-\mathbf{G}_r\ _{\mathcal{H}_\infty}}{\ \mathbf{G}\ _{\mathcal{H}_\infty}}$	$\frac{\ \mathbf{G}^{sp}-\mathbf{G}_r^{sp}\ _{\mathcal{H}_\infty}}{\ \mathbf{G}^{sp}\ _{\mathcal{H}_\infty}}$
IRKA-D	$4.87 \times 10^{-4}$	$4.57 \times 10^{-4}$

## 4.4 Hessenberg Index-2 DAEs

In this next section, we consider Hessenberg index-2 differential algebraic equations. These equations take the form:

$$\mathbf{H}(s) : \begin{cases} \mathbf{E}_{11}\dot{\mathbf{x}}_1(t) = \mathbf{A}_{11}\mathbf{x}_1(t) + \mathbf{A}_{12}\mathbf{x}_2(t) + \mathbf{B}_1\mathbf{u}(t) \\ \mathbf{0} = \mathbf{A}_{12}^T\mathbf{x}_1(t) + \mathbf{B}_2\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}_1\mathbf{x}_1(t) + \mathbf{C}_2\mathbf{x}_2(t) + \mathbf{D}\mathbf{u}(t) \end{cases} \quad (4.4.1)$$

where the state is  $\mathbf{x}(t) = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \in \mathbb{R}^n$  with  $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ ,  $\mathbf{x}_2 \in \mathbb{R}^{n_2}$  and  $n_1 + n_2 = n$ , the input is  $\mathbf{u}(t) \in \mathbb{R}^m$ , the output is  $\mathbf{y}(t) \in \mathbb{R}^p$ , and  $\mathbf{E}_{11}, \mathbf{A}_{11} \in \mathbb{R}^{n_1 \times n_1}$ ,  $\mathbf{A}_{12} \in \mathbb{R}^{n_1 \times n_2}$ ,  $\mathbf{B}_1 \in \mathbb{R}^{n_1 \times m}$ ,  $\mathbf{B}_2 \in \mathbb{R}^{n_2 \times m}$ ,  $\mathbf{C}_1 \in \mathbb{R}^{p \times n_1}$ ,  $\mathbf{C}_2 \in \mathbb{R}^{p \times n_2}$ , and  $\mathbf{D} \in \mathbb{R}^{p \times m}$ . For Hessenberg index-2 equations, the matrix  $\mathbf{E}_{11}$  is a symmetric positive definite matrix and  $\mathbf{A}_{12}$  is of full rank.

In [49], the authors considered applying balanced truncation to system (4.4.1). The aim of this next section is to reduce (4.4.1) using interpolatory model reduction. To begin, consider the system (4.4.1) with  $\mathbf{B}_2 = \mathbf{0}$  as the case of  $\mathbf{B}_2 \neq \mathbf{0}$  is similar. Following the exposition of [49], we assume  $\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12}$  is nonsingular and define the oblique projector as

$$\mathbf{\Pi} = \mathbf{I} - \mathbf{A}_{12} (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{A}_{12}^T \mathbf{E}_{11}^{-1}.$$

Then system (4.4.1) can be written in terms of  $\mathbf{\Pi}$  as follows:

$$\begin{cases} \mathbf{\Pi} \mathbf{E}_{11} \mathbf{\Pi}^T \dot{\mathbf{x}}_1(t) &= \mathbf{\Pi} \mathbf{A}_{11} \mathbf{\Pi}^T \mathbf{x}_1(t) + \mathbf{\Pi} \mathbf{B}_1 \mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C} \mathbf{\Pi}^T \mathbf{x}_1(t) + \tilde{\mathbf{D}} \mathbf{u}(t) \\ \mathbf{\Pi}^T \mathbf{x}(0) &= \mathbf{\Pi}^T \mathbf{x}_0 \end{cases} \quad (4.4.2)$$

where  $\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2 (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{11}$  and  $\tilde{\mathbf{D}} = \mathbf{D} - \mathbf{C}_2 (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{B}_1$ .

By decomposing  $\mathbf{\Pi}$  as

$$\mathbf{\Pi} = \mathbf{\Theta}_l \mathbf{\Theta}_r^T \quad \text{with} \quad \mathbf{\Theta}_l, \mathbf{\Theta}_r \in \mathbb{R}^{n_1 \times (n_1 - n_2)} \quad \text{such that} \quad \mathbf{\Theta}_l^T \mathbf{\Theta}_r = \mathbf{I}, \quad (4.4.3)$$

and defining  $\tilde{\mathbf{x}}_1 = \Theta_l^T \mathbf{x}_1 \in \mathbb{R}^{n_1 - n_2}$ , system (4.4.2) becomes

$$\begin{cases} \Theta_r^T \mathbf{E}_{11} \Theta_r \dot{\tilde{\mathbf{x}}}_1(t) &= \Theta_r^T \mathbf{A}_{11} \Theta_r \tilde{\mathbf{x}}_1(t) + \Theta_r^T \mathbf{B}_1 \mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C} \Theta_r \tilde{\mathbf{x}}_1(t) + \tilde{\mathbf{D}} \mathbf{u}(t) \\ \tilde{\mathbf{x}}_1(0) &= \Theta_l^T \mathbf{x}_0. \end{cases} \quad (4.4.4)$$

As noted in [49], the reduction of the DAE system in (4.4.1) is equivalent to the reduction of systems (4.4.2) or (4.4.4). However, the beauty of this equivalence lies in the observation that systems (4.4.2) and (4.4.4) are ODEs, namely the algebraic component of the system has been moved to the  $\tilde{\mathbf{D}}$  term. Therefore, standard model reduction procedures for ODEs can be applied to systems (4.4.2) and (4.4.4), and the reduced model obtained will be equivalent to directly reducing the DAE system in (4.4.1). It is important to emphasize that even though (4.4.2) and (4.4.4) are equivalent to (4.4.1), the ultimate goal of this chapter is to develop an interpolatory model reduction method that does not require the explicit computation of either  $\mathbf{\Pi}$  or  $\Theta_r$ . To do so, the motivation for our method comes from [49], which defines

$$\tilde{\mathbf{E}} = \mathbf{\Pi} \mathbf{E}_{11} \mathbf{\Pi}^T, \quad \tilde{\mathbf{A}} = \mathbf{\Pi} \mathbf{A}_{11} \mathbf{\Pi}^T, \quad \tilde{\mathbf{B}} = \mathbf{\Pi} \mathbf{B}_1, \quad \tilde{\mathbf{C}} = \mathbf{C} \mathbf{\Pi}^T \quad (4.4.5)$$

and then proves several key properties of the matrix  $\tilde{\mathbf{E}} + \mu \tilde{\mathbf{A}}$ . In the remainder of this section, we are interested in applying interpolatory methods where the matrix of interest will be shown to be  $\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}}$ . To implement interpolatory methods, we rely on the theoretical results of [49]; however, we present these results in terms of  $\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}}$  instead of  $\tilde{\mathbf{E}} + \mu \tilde{\mathbf{A}}$ .



**Lemma 4.5.** *Let  $\Theta_r$  be the matrix defined in (4.4.4) and let  $\sigma \in \mathbb{C}^-$  be such that  $\sigma\Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r$  is nonsingular. The matrix defined as*

$$(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I := \Theta_r (\sigma \Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \quad (4.4.6)$$

*satisfies*

$$(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I (\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}}) = \Pi^T \quad \text{and} \quad (\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}}) (\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I = \Pi.$$

*Similarly, the matrix defined as*

$$(\sigma \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I := \Theta_r (\sigma \Theta_r^T \mathbf{E}_{11}^T \Theta_r - \Theta_r^T \mathbf{A}_{11}^T \Theta_r)^{-1} \Theta_r^T \quad (4.4.7)$$

*satisfies*

$$(\sigma \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I (\sigma \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T) = \Pi^T \quad \text{and} \quad (\sigma \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T) (\sigma \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I = \Pi.$$

*Proof.* Following a similar argument to that in [49], the proof of the first equality follows directly from the definitions (4.4.3) and (4.4.6).

$$\begin{aligned}
(\sigma\tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I(\sigma\tilde{\mathbf{E}} - \tilde{\mathbf{A}}) &= \Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}\Theta_r - \Theta_r^T\mathbf{A}_{11}\Theta_r)^{-1}\Theta_r^T\Pi(\sigma\mathbf{E}_{11} - \mathbf{A}_{11})\Pi^T \\
&= \Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}\Theta_r - \Theta_r^T\mathbf{A}_{11}\Theta_r)^{-1}\Theta_r^T(\sigma\mathbf{E}_{11} - \mathbf{A}_{11})\Theta_r\Theta_l^T \\
&= \Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}\Theta_r - \Theta_r^T\mathbf{A}_{11}\Theta_r)^{-1}(\sigma\Theta_r^T\mathbf{E}_{11}\Theta_r - \Theta_r^T\mathbf{A}_{11}\Theta_r)\Theta_l^T \\
&= \Theta_r\Theta_l^T \\
&= \Pi^T.
\end{aligned}$$

The remaining equalities follow similarly. □

At first glance, the definition of the inverse restricted to the subspace  $\Pi$  may seem irrelevant to the reduction of the DAE system (4.4.1). However, recall that reducing (4.4.1) is equivalent to reducing system (4.4.2). To reduce (4.4.2), Theorem 1.1 requires the inverses of  $(\sigma\tilde{\mathbf{E}} - \tilde{\mathbf{A}})$  and  $(\sigma\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)$ . However, these inverses do not exist. As a result, definitions (4.4.6) and (4.4.7) become pivotal in order to achieve the interpolation of system (4.4.2) and thereby interpolation of system (4.4.1) as shown in the next theorem.

**Theorem 4.6.** *Let  $\mathbf{V}_r, \mathbf{W}_r$  be full rank. Let  $s = \sigma, \mu \in \mathbb{C}$  be such that the matrices  $s\boldsymbol{\Theta}_r^T \mathbf{E}_{11} \boldsymbol{\Theta}_r - \boldsymbol{\Theta}_r^T \mathbf{A}_{11} \boldsymbol{\Theta}_r$  and  $s\mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r - \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r$  are invertible.*

*Define the transfer functions*

$$\begin{aligned}\tilde{\mathbf{H}}(s) &= \mathbf{C} \boldsymbol{\Theta}_r (s\boldsymbol{\Theta}_r^T \mathbf{E}_{11} \boldsymbol{\Theta}_r - \boldsymbol{\Theta}_r^T \mathbf{A}_{11} \boldsymbol{\Theta}_r)^{-1} \boldsymbol{\Theta}_r^T \mathbf{B}_1 + \tilde{\mathbf{D}} \\ \tilde{\mathbf{H}}_r(s) &= \mathbf{C} \mathbf{V}_r (s\mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r - \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r)^{-1} \mathbf{W}_r^T \mathbf{B}_1 + \tilde{\mathbf{D}}.\end{aligned}$$

*Let  $\mathbf{b} \in \mathbb{C}^m$  and  $\mathbf{c} \in \mathbb{C}^l$  be fixed nontrivial vectors.*

- i) If  $(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^l \tilde{\mathbf{B}} \mathbf{b} \in \text{Ran}(\mathbf{V}_r)$ , then  $\tilde{\mathbf{H}}(\sigma) \mathbf{b} = \tilde{\mathbf{H}}_r(\sigma) \mathbf{b}$ .*
- ii) If  $(\mu \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^l \tilde{\mathbf{C}}^T \mathbf{c} \in \text{Ran}(\mathbf{W}_r)$ , then  $\mathbf{c}^T \tilde{\mathbf{H}}(\mu) = \mathbf{c}^T \tilde{\mathbf{H}}_r(\mu)$ .*
- iii) If (i) and (ii) hold and  $\sigma = \mu$ , then  $\mathbf{c}^T \tilde{\mathbf{H}}'(\sigma) \mathbf{b} = \mathbf{c}^T \tilde{\mathbf{H}}'_r(\sigma) \mathbf{b}$ .*

*Remark:* Before presenting the proof, it is important to emphasize that this interpolation result is not the usual interpolation result as given in Theorem 1.1. In this theorem, the reducing matrices  $\mathbf{V}_r$  and  $\mathbf{W}_r$  are in terms of system (4.4.2), namely  $\tilde{\mathbf{A}}, \tilde{\mathbf{E}}, \tilde{\mathbf{B}}$  and  $\tilde{\mathbf{C}}$  are used to construct  $\mathbf{V}_r$  and  $\mathbf{W}_r$ . Such a construction only implies that system (4.4.2) is interpolated. However, Theorem 4.6 states that this also yields interpolation of the original DAE (4.4.1) which is described in terms of  $\mathbf{E}_{11}, \mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{C}_1, \mathbf{C}_2$ , and  $\tilde{\mathbf{D}}$ . Also, it is important to emphasize that unlike Theorem 4.1, where only the two-sided interpolation result held for the DAE system, Theorem 4.6 states that the one-sided result holds for the case of Hessenberg index-2 DAEs.

*Proof.* To prove (i) – (iii), we first define

$$\widehat{\mathbf{H}}_r(s) = \mathbf{C}\Theta_r\widetilde{\mathbf{V}}_r(s\widetilde{\mathbf{W}}_r^T\Theta_r^T\mathbf{E}_{11}\Theta_r\widetilde{\mathbf{V}}_r - \widetilde{\mathbf{W}}_r^T\Theta_r^T\mathbf{A}_{11}\Theta_r\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T\Theta_r^T\mathbf{B}_1 + \widetilde{\mathbf{D}}$$

and set  $\widetilde{\mathbf{V}}_r = \Theta_l^T\mathbf{V}$  and  $\widetilde{\mathbf{W}}_r = \Theta_l^T\mathbf{W}_r$ . Since  $\Theta_l^T\Theta_r = \mathbf{I}$ ,

$$\mathbf{V}_r = \Theta_r\widetilde{\mathbf{V}}_r \quad \text{and} \quad \mathbf{W}_r = \Theta_r\widetilde{\mathbf{W}}_r. \quad (4.4.8)$$

This implies that  $\widehat{\mathbf{H}}_r(s) = \widetilde{\mathbf{H}}_r(s)$ . To prove (i), we note that (4.4.3) implies that

$$\Theta_r^T\widetilde{\mathbf{B}} = \Theta_r^T\mathbf{\Pi}\mathbf{B}_1 = \Theta_r^T\Theta_l\Theta_r^T\mathbf{B}_1 = \Theta_r^T\mathbf{B}_1. \quad (4.4.9)$$

Now let  $(\sigma\widetilde{\mathbf{E}} - \widetilde{\mathbf{A}})^l\widetilde{\mathbf{B}}\mathbf{b} \in \text{Ran}(\mathbf{V}_r)$ . Then there exists  $\mathbf{q} \in \mathbb{R}^r$  such that

$$(\sigma\widetilde{\mathbf{E}} - \widetilde{\mathbf{A}})^l\widetilde{\mathbf{B}}\mathbf{b} = \mathbf{V}_r\mathbf{q}.$$

By the definition of (4.4.6), we have

$$\Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}\Theta_r - \Theta_r^T\mathbf{A}_{11}\Theta_r)^{-1}\Theta_r^T\widetilde{\mathbf{B}}\mathbf{b} = \mathbf{V}_r\mathbf{q}.$$

Using (4.4.8) and (4.4.9), the above equation is written in terms of  $\Theta_r$  as:

$$\Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}\Theta_r - \Theta_r^T\mathbf{A}_{11}\Theta_r)^{-1}\Theta_r^T\Theta_l\Theta_r^T\mathbf{B}_1\mathbf{b} = \Theta_r\widetilde{\mathbf{V}}_r\mathbf{q}.$$

Left multiplying by  $\Theta_l^T$  gives

$$\Theta_l^T [\Theta_r (\sigma \Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \Theta_l \Theta_r^T \mathbf{B}_1 \mathbf{b} = \Theta_r \tilde{\mathbf{V}}_r \mathbf{q}],$$

implying

$$(\sigma \Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \mathbf{B}_1 \mathbf{b} = \tilde{\mathbf{V}}_r \mathbf{q}$$

by (4.4.3). Hence,  $(\sigma \Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \mathbf{B}_1 \mathbf{b} \in \text{Ran}(\tilde{\mathbf{V}}_r)$ . Combined with Theorem 1.1, we have that  $\tilde{\mathbf{H}}(\sigma) \mathbf{b} = \hat{\mathbf{H}}_r(\sigma) \mathbf{b}$ . Finally, noting that  $\hat{\mathbf{H}}_r(s) = \tilde{\mathbf{H}}_r(s)$  gives the result.

To prove (ii), we first note that (4.4.3) implies that

$$\Theta_r^T \tilde{\mathbf{C}}^T = \Theta_r^T \mathbf{\Pi} \mathbf{C}^T = \Theta_r^T \Theta_l \Theta_r^T \mathbf{C}^T = \Theta_r^T \mathbf{C}^T. \quad (4.4.10)$$

Let  $(\mu \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c} \in \text{Ran}(\mathbf{W}_r)$ , then there exists  $\mathbf{q} \in \mathbb{R}^r$  such that

$$(\mu \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c} = \mathbf{W}_r \mathbf{q}.$$

By the definition of (4.4.7),

$$\Theta_r (\sigma \Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \tilde{\mathbf{C}}^T \mathbf{c} = \mathbf{W}_r \mathbf{q}.$$

Using (4.4.8) and (4.4.10), we have

$$\Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}^T\Theta_r - \Theta_r^T\mathbf{A}_{11}^T\Theta_r)^{-1}\Theta_r^T\Theta_l\Theta_r^T\mathbf{C}^T\mathbf{c} = \Theta_r\widetilde{\mathbf{W}}_r\mathbf{q}.$$

Left multiplying by  $\Theta_l^T$  gives

$$\Theta_l^T[\Theta_r(\sigma\Theta_r^T\mathbf{E}_{11}^T\Theta_r - \Theta_r^T\mathbf{A}_{11}^T\Theta_r)^{-1}\Theta_r^T\Theta_l\Theta_r^T\mathbf{C}^T\mathbf{c} = \Theta_r\widetilde{\mathbf{W}}_r\mathbf{q}],$$

implying

$$(\sigma\Theta_r^T\mathbf{E}_{11}^T\Theta_r - \Theta_r^T\mathbf{A}_{11}^T\Theta_r)^{-1}\Theta_r^T\mathbf{C}^T\mathbf{c} = \widetilde{\mathbf{W}}_r\mathbf{q}$$

by (4.4.3). Hence,  $(\sigma\Theta_r^T\mathbf{E}_{11}^T\Theta_r - \Theta_r^T\mathbf{A}_{11}^T\Theta_r)^{-1}\Theta_r^T\mathbf{C}^T\mathbf{c} \in \text{Ran}(\widetilde{\mathbf{W}}_r)$ . Combined with Theorem

1.1, we have that  $\mathbf{c}^T\widetilde{\mathbf{H}}(\sigma) = \mathbf{c}^T\widehat{\mathbf{H}}_r(\sigma)$ . Noting that  $\widehat{\mathbf{H}}_r(s) = \widetilde{\mathbf{H}}_r(s)$  yields the conclusion.

To prove part (iii), we note that parts (i) and (ii) holding imply that (a) and (b) of Theorem 1.1 hold, namely,

$$\widetilde{\mathbf{H}}(\sigma)\mathbf{b} = \widehat{\mathbf{H}}_r(\sigma)\mathbf{b} \quad \text{and} \quad \mathbf{c}^T\widetilde{\mathbf{H}}(\sigma) = \mathbf{c}^T\widehat{\mathbf{H}}_r(\sigma).$$

Hence, part (c) of Theorem 1.1 implies

$$\mathbf{c}^T\widetilde{\mathbf{H}}'(\sigma)\mathbf{b} = \mathbf{c}^T\widehat{\mathbf{H}}_r'(\sigma)\mathbf{b}.$$

Since  $\widehat{\mathbf{H}}_r(s) = \widetilde{\mathbf{H}}_r(s)$ , we have

$$\mathbf{c}^T \widetilde{\mathbf{H}}'(\sigma) \mathbf{b} = \mathbf{c}^T \widetilde{\mathbf{H}}'_r(\sigma) \mathbf{b}.$$

□

Alternatively, we can consider interpolation through the Sylvester equation framework. For a first-order descriptor system,  $\mathbf{H}(s)$ , with  $\mathcal{K}(s) = s\mathbf{E} - \mathbf{A}$ ,  $\mathcal{B}(s) = \mathbf{B}$  and  $\mathcal{C}(s) = \mathbf{C}$ , Theorem 1.1 with  $l = 0$  is represented in terms of the Sylvester equation as described in the next result.

**Lemma 4.7.** *Let  $\sigma_i, \mu_i \in \mathbb{C}$  such that  $s\mathbf{E} - \mathbf{A}$  and  $s\mathbf{E}_r - \mathbf{A}_r$  are invertible for  $s = \sigma_i, \mu_i, i = 1, \dots, r$ . Define  $\boldsymbol{\Sigma} = \text{diag}(\sigma_i)$ ,  $\boldsymbol{\Upsilon} = \text{diag}(\mu_i)$ ,  $\mathcal{B} = [\mathbf{b}_1, \dots, \mathbf{b}_r]$ ,  $\mathcal{C} = [\mathbf{c}_1, \dots, \mathbf{c}_r]$ , where  $\mathbf{b}_i \in \mathbb{C}^m$  and  $\mathbf{c}_i \in \mathbb{C}^l$  are fixed nontrivial vectors, then*

a) *If  $\mathbf{V}_r$  solves the Sylvester equation*

$$\mathbf{A}\mathbf{V}_r - \mathbf{E}\mathbf{V}_r\boldsymbol{\Sigma} + \mathbf{B}\mathcal{B} = \mathbf{0}, \quad (4.4.11)$$

*then  $(\sigma_i\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{b}_i = (\sigma_i\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r\mathbf{b}_i$ .*

b) *If  $\mathbf{W}_r$  solves the Sylvester equation*

$$\mathbf{A}^T\mathbf{W}_r - \mathbf{E}^T\mathbf{W}_r\boldsymbol{\Upsilon} + \mathbf{C}^T\mathcal{C} = \mathbf{0}, \quad (4.4.12)$$

*then  $\mathbf{c}_i^T\mathbf{C}(\mu_i\mathbf{E} - \mathbf{A})^{-1} = \mathbf{c}_i^T\mathbf{C}_r(\mu_i\mathbf{E}_r - \mathbf{A}_r)^{-1}$ .*

c) If both (a) and (b) hold with  $\sigma_i = \mu_i$  then  $\mathbf{c}_i^T \mathbf{H}'(\sigma_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(\sigma_i) \mathbf{b}_i$ .

*Proof.* The result follows readily from Theorem 1.1 with  $M, N = 1$  and noting that the  $i^{\text{th}}$  column of (4.4.11) is

$$\mathbf{A} \mathbf{v}_i - \sigma_i \mathbf{E} \mathbf{v}_i + \mathbf{B} \mathbf{b}_i = \mathbf{0},$$

implying  $\mathbf{v}_i = (\sigma_i \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{b}_i$ , and the  $i^{\text{th}}$  column of (4.4.12) is

$$\mathbf{A}^T \mathbf{w}_i - \mu_i \mathbf{E}^T \mathbf{w}_i + \mathbf{C}^T \mathbf{c}_i = \mathbf{0},$$

implying  $\mathbf{w}_i = (\mu_i \mathbf{E}^T - \mathbf{A}^T)^{-1} \mathbf{C}^T \mathbf{c}_i$ . □

Using Lemma 4.7, Theorem 4.6 can be reinterpreted in terms of the Sylvester equations as shown in the next lemma.

**Lemma 4.8.** *Let  $\mathbf{V}_r, \mathbf{W}_r$  be full rank. For  $i = 1, \dots, r$ , let  $s = \sigma_i, \mu_i \in \mathbb{C}$  be such that*

*$s \mathbf{\Theta}_r \mathbf{E}_{11} \mathbf{\Theta}_r - \mathbf{\Theta}_r^T \mathbf{A}_{11} \mathbf{\Theta}_r$  and  $s \mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r - \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r$  are invertible.*

a) *Let  $\mathbf{V}_r = \mathbf{\Theta}_r \tilde{\mathbf{V}}_r$  with  $\mathbf{v}_i = (\sigma_i \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^T \tilde{\mathbf{B}} \mathbf{b}_i$ . Then  $\mathbf{V}_r$  satisfies*

$$\mathbf{\Theta}_r^T \mathbf{A}_{11} \mathbf{\Theta}_r \tilde{\mathbf{V}}_r - \mathbf{\Theta}_r^T \mathbf{E}_{11} \mathbf{\Theta}_r \tilde{\mathbf{V}}_r \Sigma + \mathbf{\Theta}_r^T \mathbf{B}_1 \mathbf{B} = \mathbf{0}, \quad (4.4.13)$$

*implying that  $\tilde{\mathbf{H}}(\sigma_i) \mathbf{b}_i = \tilde{\mathbf{H}}_r(\sigma_i) \mathbf{b}_i$ .*



b) Let  $\mathbf{W}_r = \Theta_r \widetilde{\mathbf{W}}_r$  with  $\mathbf{w}_i = (\mu_i \widetilde{\mathbf{E}}^T - \widetilde{\mathbf{A}}^T)^I \widetilde{\mathbf{C}}^T \mathbf{c}_i$ . Then  $\mathbf{W}_r$  satisfies

$$\Theta_r^T \mathbf{A}_{11}^T \Theta_r \widetilde{\mathbf{W}}_r - \Theta_r^T \mathbf{E}_{11}^T \Theta_r \widetilde{\mathbf{W}}_r \Upsilon + \Theta_r^T \mathbf{C}^T \mathcal{C} = \mathbf{0}, \quad (4.4.14)$$

implying that  $\mathbf{c}_i^T \widetilde{\mathbf{H}}(\mu_i) = \mathbf{c}_i^T \widetilde{\mathbf{H}}_r(\mu_i)$ .

c) If both (a) and (b) hold with  $\sigma_i = \mu_i$  then  $\mathbf{c}_i^T \widetilde{\mathbf{H}}'(\sigma_i) \mathbf{b}_i = \mathbf{c}_i^T \widetilde{\mathbf{H}}_r'(\sigma_i) \mathbf{b}_i$ .

*Proof.* Define the transfer functions

$$\begin{aligned} \widetilde{\mathbf{H}}(s) &= \mathbf{C} \Theta_r (s \Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \mathbf{B}_1 + \widetilde{\mathbf{D}} \\ \widetilde{\mathbf{H}}_r(s) &= \mathbf{C} \mathbf{V}_r (s \mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r - \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r)^{-1} \mathbf{W}_r^T \mathbf{B}_1 + \widetilde{\mathbf{D}} \\ \widehat{\mathbf{H}}_r(s) &= \mathbf{C} \Theta_r \widetilde{\mathbf{V}}_r (s \widetilde{\mathbf{W}}_r^T \Theta_r^T \mathbf{E}_{11} \Theta_r \widetilde{\mathbf{V}}_r - \widetilde{\mathbf{W}}_r^T \Theta_r^T \mathbf{A}_{11} \Theta_r \widetilde{\mathbf{V}}_r)^{-1} \widetilde{\mathbf{W}}_r^T \Theta_r^T \mathbf{B}_1 + \widetilde{\mathbf{D}}. \end{aligned}$$

First, note by (4.4.3),

$$\Theta_r^T \Pi = \Theta_r^T \Theta_l \Theta_r^T = \Theta_r^T. \quad (4.4.15)$$

Let  $\mathbf{v}_i = (\sigma_i \widetilde{\mathbf{E}} - \widetilde{\mathbf{A}})^J \widetilde{\mathbf{B}} \mathbf{b}_i$ . Left multiplying by  $(\sigma_i \widetilde{\mathbf{E}} - \widetilde{\mathbf{A}})$  gives

$$(\sigma_i \widetilde{\mathbf{E}} - \widetilde{\mathbf{A}}) \mathbf{v}_i = \Pi \widetilde{\mathbf{B}} \mathbf{b}_i = \widetilde{\mathbf{B}} \mathbf{b}_i$$

by Lemma 4.5 and the projector property of  $\Pi$ , namely  $\Pi^2 = \Pi$ . Rearranging, gives

$\tilde{\mathbf{A}}\mathbf{v}_i - \sigma_i\tilde{\mathbf{E}}\mathbf{v}_i + \tilde{\mathbf{B}}\mathbf{b}_i = \mathbf{0}$ , or equivalently the matrix equation

$$\tilde{\mathbf{A}}\mathbf{V}_r - \tilde{\mathbf{E}}_r\mathbf{V}_r\Sigma + \tilde{\mathbf{B}}\mathcal{B} = \mathbf{0}.$$

Multiplying by  $\Theta_r^T$  and using (4.4.5) along with the definition of  $\mathbf{V}_r$ , we have

$$\Theta_r^T\Pi\mathbf{A}_{11}\Pi^T\Theta_r\tilde{\mathbf{V}}_r - \Theta_r^T\Pi\mathbf{E}_{11}\Pi^T\Theta_r\tilde{\mathbf{V}}_r\Sigma + \Theta_r^T\Pi\mathbf{B}_1\mathcal{B} = \mathbf{0}.$$

Applying (4.4.15) yields

$$\Theta_r^T\mathbf{A}_{11}\Theta_r\tilde{\mathbf{V}}_r - \Theta_r^T\mathbf{E}_{11}\Theta_r\tilde{\mathbf{V}}_r\Sigma + \Theta_r^T\mathbf{B}_1\mathcal{B} = \mathbf{0}.$$

Therefore,  $\mathbf{V}_r$  satisfies the Sylvester equation required for interpolation of  $\tilde{\mathbf{H}}(s)\mathbf{b}_i$ , implying

$\tilde{\mathbf{H}}(\sigma_i)\mathbf{b}_i = \hat{\mathbf{H}}_r(\sigma_i)\mathbf{b}_i$  by Lemma 4.7. Since  $\Theta_i^T\Theta_r = \mathbf{I}$  and  $\mathbf{V}_r = \Theta_r\tilde{\mathbf{V}}_r$ , we have

$\hat{\mathbf{H}}_r(s)\mathbf{b}_i = \tilde{\mathbf{H}}_r(s)\mathbf{b}_i$ . Therefore,  $\tilde{\mathbf{H}}(\sigma_i)\mathbf{b}_i = \tilde{\mathbf{H}}_r(\sigma_i)\mathbf{b}_i$ . For the proof of part (b), a similar

argument is used. Let  $\mathbf{w}_i = (\mu_i\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^T\tilde{\mathbf{C}}^T\mathbf{c}_i$ . Left multiplying by  $(\mu_i\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)$  gives

$$(\mu_i\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)\mathbf{w}_i = \Pi\tilde{\mathbf{C}}^T\mathbf{c}_i = \tilde{\mathbf{C}}^T\mathbf{c}_i$$

by Lemma 4.5 and the projector property of  $\Pi$ , namely  $\Pi^2 = \Pi$ . Rearranging, gives

$\tilde{\mathbf{A}}^T\mathbf{w}_i - \mu_i\tilde{\mathbf{E}}^T\mathbf{w}_i + \tilde{\mathbf{C}}^T\mathbf{c}_i = \mathbf{0}$ , or equivalently the matrix equation

$$\tilde{\mathbf{A}}^T\mathbf{W}_r - \tilde{\mathbf{E}}_r^T\mathbf{W}_r\Upsilon + \tilde{\mathbf{C}}^T\mathcal{C} = \mathbf{0}.$$

Multiplying by  $\Theta_r^T$  and using (4.4.5) and the definition of  $\mathbf{W}_r$ , we have

$$\Theta_r^T \Pi \mathbf{A}_{11}^T \Pi^T \Theta_r \widetilde{\mathbf{W}}_r - \Theta_r^T \Pi \mathbf{E}_{11}^T \Pi^T \Theta_r \widetilde{\mathbf{W}}_r \Upsilon + \Theta_r^T \Pi \mathbf{C}^T \mathcal{C} = \mathbf{0}.$$

Applying (4.4.15) gives

$$\Theta_r^T \mathbf{A}_{11}^T \Theta_r \widetilde{\mathbf{W}}_r - \Theta_r^T \mathbf{E}_{11}^T \Theta_r \widetilde{\mathbf{W}}_r \Upsilon + \Theta_r^T \mathbf{C}^T \mathcal{C} = \mathbf{0}.$$

Therefore,  $\mathbf{W}_r$  satisfies the Sylvester equation required for interpolation of  $\widetilde{\mathbf{H}}(s)$ , implying  $\mathbf{c}_i^T \widetilde{\mathbf{H}}(\sigma_i) = \mathbf{c}_i^T \widehat{\mathbf{H}}_r(\sigma_i)$  by Lemma 4.7. Since  $\Theta_l^T \Theta_r = \mathbf{I}$  and  $\mathbf{W}_r = \Theta_r \widetilde{\mathbf{W}}_r$ , we have  $\mathbf{c}_i^T \widehat{\mathbf{H}}_r(s) = \mathbf{c}_i^T \widetilde{\mathbf{H}}_r(s)$ . Therefore,  $\mathbf{c}_i^T \widetilde{\mathbf{H}}(\sigma_i) = \mathbf{c}_i^T \widetilde{\mathbf{H}}_r(\sigma_i)$ . The proof of part (c) follows immediately from Lemma 4.7 once (a) and (b) are established.  $\square$

#### 4.4.1 Related Computational Issues to the Reduction of Hessenberg Index-2 DAEs

In [49], the authors present lemmas related to the computation of the matrix  $(\widetilde{\mathbf{E}} + \mu \widetilde{\mathbf{A}})^I$ . For interpolatory model reduction, Theorem 4.6 shows that the quantities of interest are

$$(\sigma \widetilde{\mathbf{E}} - \widetilde{\mathbf{A}})^I \widetilde{\mathbf{B}} \mathbf{b} \quad \text{and} \quad (\mu \widetilde{\mathbf{E}}^T - \widetilde{\mathbf{A}}^T)^I \widetilde{\mathbf{C}}^T \mathbf{c}. \quad (4.4.16)$$

In the following lemmas, we will use the notation and reasoning of [49] to present the results of [49] as they relate to the computation of (4.4.16).

**Lemma 4.9.** [49] *The matrix  $\mathbf{Z}$  satisfies  $\mathbf{Z} = \mathbf{\Pi}^T \mathbf{Z}$  and  $\mathbf{\Pi}(\sigma \mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{\Pi}^T \mathbf{Z} = \mathbf{\Pi} \mathbf{F}$  if and only if*

$$\begin{bmatrix} \sigma \mathbf{E}_{11} - \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Z} \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{0} \end{bmatrix}. \quad (4.4.17)$$

*Proof.* First note the projector properties of  $\mathbf{\Pi}$  imply that

$$\mathbf{A}_{12}^T \mathbf{z} = \mathbf{0} \quad \text{if and only if} \quad \mathbf{\Pi}^T \mathbf{z} = \mathbf{z}. \quad (4.4.18)$$

If  $\mathbf{Z} = \mathbf{\Pi}^T \mathbf{Z}$  satisfies  $\mathbf{\Pi}(\sigma \mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{\Pi}^T \mathbf{Z} = \mathbf{\Pi} \mathbf{F}$ , then  $\mathbf{\Pi}[(\sigma \mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{Z} - \mathbf{F}] = \mathbf{0}$ , implying the columns of  $(\sigma \mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{Z} - \mathbf{F}$  are in  $\text{Null}(\mathbf{\Pi}) = \text{Ran}(\mathbf{A}_{12})$ . This provides the existence of  $\mathbf{\Gamma}$  such that

$$(\sigma \mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{Z} - \mathbf{F} = -\mathbf{A}_{12}\mathbf{\Gamma}.$$

Rearranging, the above equation gives the first block of equations:

$$(\sigma \mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{Z} + \mathbf{A}_{12}\mathbf{\Gamma} = \mathbf{F}.$$

The second block directly follows from  $\mathbf{Z} = \mathbf{\Pi}^T \mathbf{Z}$  and projector properties of  $\mathbf{\Pi}$  as stated in (4.4.18). If (4.4.17) is satisfied, then  $\mathbf{A}_{12}^T \mathbf{Z} = \mathbf{0}$ , implying  $\mathbf{Z} = \mathbf{\Pi}^T \mathbf{Z}$  by (4.4.18). Since

$$\mathbf{\Pi}\mathbf{A}_{12} = \mathbf{0},$$

$$\mathbf{\Pi}\mathbf{F} = \mathbf{\Pi}((\sigma\mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{\Pi}^T\mathbf{Z} + \mathbf{A}_{12}\mathbf{\Gamma}) = \mathbf{\Pi}(\sigma\mathbf{E}_{11} - \mathbf{A}_{11})\mathbf{\Pi}^T\mathbf{Z}.$$

□

**Lemma 4.10.** [49] *The matrix  $\mathbf{Q}$  satisfies  $\mathbf{Q} = \mathbf{\Pi}^T\mathbf{Q}$  and  $\mathbf{\Pi}(\sigma\mathbf{E}_{11}^T - \mathbf{A}_{11}^T)\mathbf{\Pi}^T\mathbf{Q} = \mathbf{\Pi}\mathbf{G}$  if and only if*

$$\begin{bmatrix} \sigma\mathbf{E}_{11}^T - \mathbf{A}_{11}^T & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q} \\ \mathbf{\Lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{G} \\ \mathbf{0} \end{bmatrix}. \quad (4.4.19)$$

*Proof.* The proof follows a similar argument as employed in the proof of Lemma 4.9. □

**Lemma 4.11.** [49] *Let  $\sigma \in \mathbb{C}^-$ . Then the vector*

$$\mathbf{z} = (\sigma\tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I\tilde{\mathbf{B}}\mathbf{b} \quad (4.4.20)$$

*solves*

$$\begin{bmatrix} \sigma\mathbf{E}_{11} - \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1\mathbf{b} \\ \mathbf{0} \end{bmatrix}. \quad (4.4.21)$$

*Proof.* By Lemma 4.9, the matrix  $\mathbf{Z}$  obtained by solving (4.4.21) must satisfy  $\mathbf{z} = \mathbf{\Pi}^T \mathbf{z}$  and

$$(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}}) \mathbf{z} = \mathbf{\Pi} \mathbf{B}_1 \mathbf{b} = \tilde{\mathbf{B}} \mathbf{b}.$$

Left multiplying by  $(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I$  and using Lemma 4.5 gives

$$\mathbf{\Pi}^T \mathbf{z} = (\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I \tilde{\mathbf{B}} \mathbf{b}.$$

Finally, noting  $\mathbf{z} = \mathbf{\Pi}^T \mathbf{z}$  yields (4.4.20). □

**Lemma 4.12.** [49] *Let  $\mu \in \mathbb{C}^-$ . Then the vector*

$$\mathbf{q} = (\mu \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c} \tag{4.4.22}$$

*solves*

$$\begin{bmatrix} \mu \mathbf{E}_{11}^T - \mathbf{A}_{11}^T & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{q} \\ \mathbf{\Lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^T \mathbf{c} \\ \mathbf{0} \end{bmatrix}. \tag{4.4.23}$$

*Proof.* By Lemma 4.10, the vector  $\mathbf{q}$  obtained by solving (4.4.23) must satisfy  $\mathbf{q} = \mathbf{\Pi}^T \mathbf{q}$  and

$$(\mu \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T) \mathbf{q} = \mathbf{\Pi} \mathbf{C}^T \mathbf{c} = \tilde{\mathbf{C}}^T \mathbf{c}.$$

Left multiplying by  $(\mu\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I$  and using Lemma 4.5 gives

$$\mathbf{\Pi}^T \mathbf{q} = (\mu\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c}.$$

Finally, noting  $\mathbf{q} = \mathbf{\Pi}^T \mathbf{q}$  yields (4.4.22).  $\square$

From a computational perspective of implementing Theorem 4.6, these lemmas are extremely important. To achieve interpolation, Theorem 4.6 relies on computing the quantities  $(\sigma\tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I \tilde{\mathbf{B}}\mathbf{b}$  and  $(\sigma\tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c}$ , both of which involve the computation of  $\mathbf{\Pi}$  and  $\mathbf{\Theta}_r$ . However, Lemma 4.11 and Lemma 4.12 illustrate that the computation of  $\mathbf{\Pi}$  and  $\mathbf{\Theta}_r$  is unnecessary and only the following linear systems need to be solved to achieve interpolation of the DAE:

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11} - \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \mathbf{b}_i \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sigma_i \mathbf{E}_{11}^T - \mathbf{A}_{11}^T & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^T \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}.$$

These observations yield the algorithm for interpolatory model reduction of Hessenberg index-2 DAEs as described by Algorithm 4.4.1.

**Algorithm 4.4.1. Interpolatory Model Reduction Method of Hessenberg Index-2 DAEs**

1. Make an initial selection of the interpolation points  $\{\sigma_i\}_{i=1}^r$ , and tangent directions  $\{\mathbf{b}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i\}_{i=1}^r$ .

2. For  $i = 1, \dots, r$  solve

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11} - \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \boldsymbol{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \mathbf{b}_i \\ \mathbf{0} \end{bmatrix}$$

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11}^T - \mathbf{A}_{11}^T & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \boldsymbol{\Lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^T \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}.$$

3.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$ ,  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .

4.  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r$ ,  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}_1$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ ,  $\mathbf{D}_r = \tilde{\mathbf{D}}$ .

#### 4.4.2 IRKA for Hessenberg Index-2 DAEs

In implementing IRKA for Hessenberg index-2 DAEs, the direct computation of  $\boldsymbol{\Pi}$  and  $\boldsymbol{\Theta}_r$  is also undesirable. Fortunately,  $\mathbf{V}_r = \boldsymbol{\Pi}^T \mathbf{V}_r$  and  $\mathbf{W}_r = \boldsymbol{\Pi}^T \mathbf{W}_r$  give that

$$\mathbf{E}_r = \mathbf{W}_r^T \boldsymbol{\Pi} \mathbf{E}_{11} \boldsymbol{\Pi}^T \mathbf{V}_r = \mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r, \quad \text{and} \quad \mathbf{A}_r = \mathbf{W}_r^T \boldsymbol{\Pi} \mathbf{A}_{11} \boldsymbol{\Pi}^T \mathbf{V}_r = \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r.$$

Therefore, the computation of the generalized eigenvalue problems in IRKA can be implemented without explicitly computing the projector  $\boldsymbol{\Pi}$ . These observations result in Algorithm 4.4.2.



**Algorithm 4.4.2. IRKA for Hessenberg Index-2 DAEs**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .

2. For  $i = 1, \dots, r$  solve

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11} - \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \mathbf{b}_i \\ \mathbf{0} \end{bmatrix}$$

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11}^T - \mathbf{A}_{11}^T & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \mathbf{\Lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^T \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}.$$

3.  $\mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$ ,  $\mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r]$ .

4.  $\mathbf{E}_r = \mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r$ ,  $\mathbf{A}_r = \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}_1$ ,  $\mathbf{C}_r = \mathbf{C} \mathbf{V}_r$ ,  $\mathbf{D}_r = \tilde{\mathbf{D}}$ .

5. while (not converged)

(a) Compute  $\mathbf{Y}^T \mathbf{A}_r \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_r \mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_r - \mathbf{A}_r$ .

(b)  $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \leftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ .

(c) For  $i = 1, \dots, r$  solve

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11} - \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \mathbf{b}_i \\ \mathbf{0} \end{bmatrix}$$

$$\begin{bmatrix} \sigma_i \mathbf{E}_{11}^T - \mathbf{A}_{11}^T & \mathbf{A}_{12} \\ \mathbf{A}_{12}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{w}_i \\ \boldsymbol{\Lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^T \mathbf{c}_i \\ \mathbf{0} \end{bmatrix}.$$

$$(d) \mathbf{V}_r = [\mathbf{v}_1, \dots, \mathbf{v}_r], \quad \mathbf{W}_r = [\mathbf{w}_1, \dots, \mathbf{w}_r].$$

$$(e) \mathbf{E}_r = \mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r, \mathbf{A}_r = \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}_1, \mathbf{C}_r = \mathbf{C} \mathbf{V}_r, \mathbf{D}_r = \tilde{\mathbf{D}}.$$

**Theorem 4.13.** *Suppose  $\mathbf{H}(s)$  and  $\mathbf{H}_r$  are real stable dynamical systems. Let*

$$\mathbf{H}_r(s) = \sum_{i=1}^r \frac{1}{s - \hat{\lambda}_i} \mathbf{c}_i \mathbf{b}_i^T$$

where  $\mathbf{E}_r, \mathbf{A}_r, \mathbf{C}_r$  and  $\mathbf{B}_r$  are obtained from Algorithm 4.4.2 and  $\{\hat{\lambda}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i \mathbf{b}_i^T\}_{i=1}^r$  are the simple poles and residues of  $\mathbf{H}_r(s)$ , respectively. Then upon convergence of Algorithm 4.4.2,  $\mathbf{H}_r(s)$  satisfies the first-order  $\mathcal{H}_2$  optimality conditions, namely for  $i = 1, 2, \dots, r$

- 1)  $\mathbf{H}(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{H}_r(-\hat{\lambda}_i) \mathbf{b}_i$
- 2)  $\mathbf{c}_i^T \mathbf{H}(-\hat{\lambda}_i) = \mathbf{c}_i^T \mathbf{H}_r(-\hat{\lambda}_i)$
- 3)  $\mathbf{c}_i^T \mathbf{H}'(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(-\hat{\lambda}_i) \mathbf{b}_i.$

*Proof.* Steps (5c) and (5d) of Algorithm 4.4.2 coupled with Lemma 4.11 and Lemma 4.12 give that  $(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I \tilde{\mathbf{B}} \mathbf{b} \in \text{Ran}(\mathbf{V}_r)$  and  $(\sigma \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c} \in \text{Ran}(\mathbf{W}_r)$ . Since the new shift and direction iterate is given by  $\sigma_i \longleftarrow -\hat{\lambda}_i(\mathbf{A}_r, \mathbf{E}_r)$  for  $i = 1, \dots, r$ ,  $\mathbf{b}_i^T \longleftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathbf{B}_r$ , and  $\mathbf{c}_i^T \longleftarrow \mathbf{C}_r \mathbf{X} \mathbf{e}_i$ , in Algorithm 4.4.2, Theorem 4.6 implies that for  $i = 1, \dots, r$ , we have

- 1)  $\mathbf{H}(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{H}_r(-\hat{\lambda}_i) \mathbf{b}_i$
- 2)  $\mathbf{c}_i^T \mathbf{H}(-\hat{\lambda}_i) = \mathbf{c}_i^T \mathbf{H}_r(-\hat{\lambda}_i)$

$$3) \mathbf{c}_i^T \mathbf{H}'(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(-\hat{\lambda}_i) \mathbf{b}_i. \quad \square$$

### 4.4.3 $\mathbf{B}_2 \neq \mathbf{0}$ Case

As shown in [49], the  $\mathbf{B}_2 \neq \mathbf{0}$  case is similar to if  $\mathbf{B}_2 = \mathbf{0}$ . For the nontrivial  $\mathbf{B}_2$  case, the authors of [49] decompose the initial condition as follows:

$$\mathbf{x}(t) = \mathbf{x}_0(t) + \mathbf{x}_g(t)$$

where  $\mathbf{x}_g(t) = -\mathbf{E}_{11}^{-1} \mathbf{A}_{12} (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{B}_2 \mathbf{u}(t)$  with  $\mathbf{x}_0(t)$  satisfying  $\mathbf{A}_{12}^T \mathbf{x}_0(t) = \mathbf{0}$ . This leads to

$$\left\{ \begin{array}{l} \mathbf{\Pi} \mathbf{E}_{11} \mathbf{\Pi}^T \dot{\mathbf{x}}_0(t) = \mathbf{\Pi} \mathbf{A}_{11} \mathbf{\Pi}^T \mathbf{x}_0(t) + \mathbf{\Pi} \mathbf{B} \mathbf{u}(t) \\ \mathbf{\Pi}^T \mathbf{x}_0(0) = \mathbf{\Pi}^T (\mathbf{x}_0 - \mathbf{x}_g(0)) \\ \mathbf{y}(t) = \mathbf{C} \mathbf{\Pi}^T \mathbf{x}_0(t) + \tilde{\mathbf{D}} \mathbf{u}(t) - \mathbf{C}_2 (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{B}_2 \dot{\mathbf{u}}(t) \end{array} \right. \quad (4.4.24)$$

where

$$\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2 (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{11} \quad (4.4.25)$$

$$\mathbf{B} = \mathbf{B}_1 - \mathbf{A}_{11} \mathbf{E}_{11}^{-1} \mathbf{A}_{12} (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{B}_2 \quad (4.4.26)$$

$$\tilde{\mathbf{D}} = \mathbf{D} - \mathbf{C}_2 (\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{B}_1. \quad (4.4.27)$$

Therefore, the  $\mathbf{B}_2 \neq 0$  case extends to the interpolation framework as well. To see this, first define

$$\tilde{\mathbf{E}} = \Pi \mathbf{E}_{11} \Pi^T \quad \tilde{\mathbf{A}} = \Pi \mathbf{A}_{11} \Pi^T \quad \tilde{\mathbf{B}} = \Pi \mathbf{B} \quad \tilde{\mathbf{C}} = \mathbf{C} \Pi^T.$$

In this case, Theorem 4.14 reformulates Theorem 4.6 for the  $\mathbf{B}_2 \neq \mathbf{0}$  case, and the proof of Theorem 4.14 follows similarly as the proof for Theorem 4.6.

**Theorem 4.14.** *Let  $\mathbf{V}_r, \mathbf{W}_r$  be full rank. Let  $s = \sigma, \mu \in \mathbb{C}$  be such that the matrices  $s\Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r$  and  $s\mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r - \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r$  are invertible.*

*Define the transfer functions*

$$\begin{aligned} \tilde{\mathbf{H}}(s) &= \mathbf{C} \Theta_r (s\Theta_r^T \mathbf{E}_{11} \Theta_r - \Theta_r^T \mathbf{A}_{11} \Theta_r)^{-1} \Theta_r^T \mathbf{B} + \hat{\mathbf{D}} \\ \tilde{\mathbf{H}}_r(s) &= \mathbf{C} \mathbf{V}_r (s\mathbf{W}_r^T \mathbf{E}_{11} \mathbf{V}_r - \mathbf{W}_r^T \mathbf{A}_{11} \mathbf{V}_r)^{-1} \mathbf{W}_r^T \mathbf{B} + \hat{\mathbf{D}} \end{aligned}$$

where  $\hat{\mathbf{D}} = \tilde{\mathbf{D}} - s\mathbf{C}_2(\mathbf{A}_{12}^T \mathbf{E}_{11}^{-1} \mathbf{A}_{12})^{-1} \mathbf{B}_2$  and  $\tilde{\mathbf{D}}$  is defined in (4.4.27). Let  $\mathbf{b} \in \mathbb{C}^m$  and  $\mathbf{c} \in \mathbb{C}^l$  be fixed nontrivial vectors.

- i) If  $(\sigma \tilde{\mathbf{E}} - \tilde{\mathbf{A}})^I \tilde{\mathbf{B}} \mathbf{b} \in \text{Ran}(\mathbf{V}_r)$ , then  $\tilde{\mathbf{H}}(\sigma) \mathbf{b} = \tilde{\mathbf{H}}_r(\sigma) \mathbf{b}$ .
- ii) If  $(\mu \tilde{\mathbf{E}}^T - \tilde{\mathbf{A}}^T)^I \tilde{\mathbf{C}}^T \mathbf{c} \in \text{Ran}(\mathbf{W}_r)$ , then  $\mathbf{c}^T \tilde{\mathbf{H}}(\mu) = \mathbf{c}^T \tilde{\mathbf{H}}_r(\mu)$ .
- iii) If (i) and (ii) hold and  $\sigma = \mu$ , then  $\mathbf{c}^T \tilde{\mathbf{H}}'(\sigma) \mathbf{b} = \mathbf{c}^T \tilde{\mathbf{H}}'_r(\sigma) \mathbf{b}$ .

## 4.5 Numerical Results for Hessenberg Index-2 DAEs

In this section, we consider model reduction of the Oseen equations, which describe the flow of a viscous and incompressible fluid. The goal of this section is to apply interpolatory methods to these models and compare the results to those obtained with balanced truncation as discussed in [49]. For more details about these models, see [49].

### 4.5.1 Problem 1

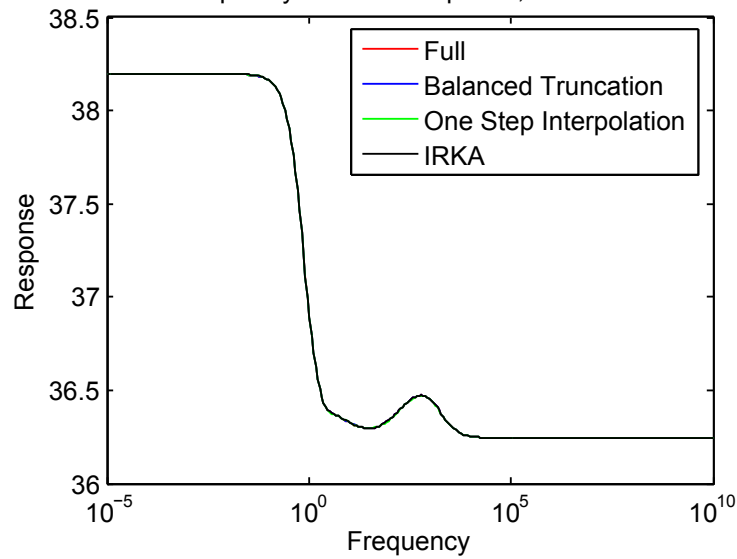
The Problem 1 Model is of the form:

$$\mathbf{H}(s) : \begin{cases} \mathbf{E}_{11}\dot{\mathbf{x}}_1(t) &= \mathbf{A}_{11}\mathbf{x}_1(t) + \mathbf{A}_{12}\mathbf{x}_2(t) + \mathbf{B}_1\mathbf{u}(t) \\ \mathbf{0} &= \mathbf{A}_{12}^T\mathbf{x}_1(t) + \mathbf{B}_2\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}_1\mathbf{x}_1(t) \end{cases} \quad (4.5.1)$$

where  $\mathbf{E}_{11}, \mathbf{A}_{11} \in \mathbb{R}^{5520 \times 5520}$ ,  $\mathbf{A}_{12} \in \mathbb{R}^{5520 \times 761}$ ,  $\mathbf{B}_1 \in \mathbb{R}^{5520 \times 6}$ ,  $\mathbf{B}_2 \in \mathbb{R}^{761 \times 6}$ ,  $\mathbf{C}_1 \in \mathbb{R}^{5520}$ . By definition of the reduced model, the polynomial parts of the full and reduced models match; therefore, Table 4.3 only presents the error in the strictly proper parts associated with reducing to order  $r = 14$ . The error for one step of interpolation was noticeably larger than the error for balanced truncation and IRKA. In addition to the sigma plots in Figure 4.4, we also used controls  $\mathbf{u}_i(t) = \sin(it)$  for  $i = 1, \dots, 6$  to obtain the time domain plots as illustrated in Figure 4.5, and both of these figures indicate a close match between the full and reduced models.

Table 4.3: Oseen Equations: Problem 1

Method	$\frac{\ \mathbf{H}^{sp} - \mathbf{H}_r^{sp}\ _{\mathcal{H}_\infty}}{\ \mathbf{H}^{sp}\ _{\mathcal{H}_\infty}}$
Balanced Truncation	$2.70 \times 10^{-5}$
One Step Interpolation	$3.09 \times 10^{-2}$
IRKA	$9.20 \times 10^{-5}$

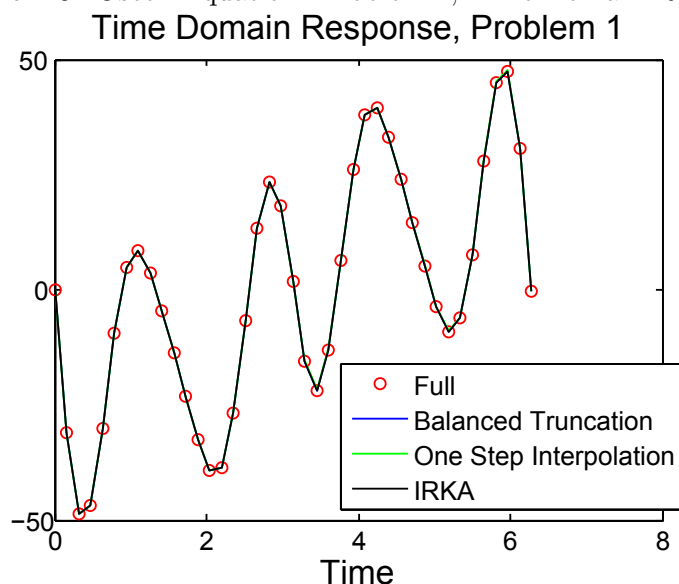
Figure 4.4: Oseen Equation: Problem 1, Frequency Response  
Frequency Domain Response, Problem 1

### 4.5.2 Problem 2

The DAE for Problem 2 is of the form

$$\mathbf{H}(s) : \begin{cases} \mathbf{E}_{11}\dot{\mathbf{x}}_1(t) = \mathbf{A}_{11}\mathbf{x}_1(t) + \mathbf{A}_{12}\mathbf{x}_2(t) + \mathbf{B}_1\mathbf{u}(t) \\ \mathbf{0} = \mathbf{A}_{12}^T\mathbf{x}_1(t) + \mathbf{B}_2\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}_1\mathbf{x}_1(t) + \mathbf{C}_2\mathbf{x}_2(t) \end{cases} \quad (4.5.2)$$

Figure 4.5: Oseen Equation: Problem 1, Time Domain Response



where  $\mathbf{E}_{11}, \mathbf{A}_{11} \in \mathbb{R}^{5520 \times 5520}$ ,  $\mathbf{A}_{12} \in \mathbb{R}^{5520 \times 761}$ ,  $\mathbf{B}_1 \in \mathbb{R}^{5520 \times 6}$ ,  $\mathbf{B}_2 \in \mathbb{R}^{761 \times 6}$ ,  $\mathbf{C}_1 \in \mathbb{R}^{2 \times 5520}$ ,  $\mathbf{C}_2 \in \mathbb{R}^{2 \times 761}$ .

Therefore, unlike the MISO model described by Problem 1, Problem 2 is a MIMO model and was reduced to order  $r = 14$ . In Table 4.4, the model reduction errors indicate the superiority of balanced truncation and IRKA. The importance of the IRKA iteration is also displayed in the time domain response plots resulting from controls  $\mathbf{u}_i(t) = \sin(it)$  for  $i = 1, \dots, 6$ . For the first output, Figure 4.6 depicts a complete mismatch between the full and reduced models obtained by one step of interpolation; this difference is emphasized in Figure 4.7 where only one step of interpolation and balanced truncation are shown. It is important to note that this discrepancy is removed by the IRKA iteration, and a nice match between the full and reduced models is observed in Figure 4.8. While the time domain response associated with output one exemplifies the importance of IRKA, Figure 4.9 depicts

all model reduction methods considered, namely balanced truncation, one step interpolation and IRKA, resulting in similar time domain responses for output two. Also, the sigma plot is shown in Figure 4.10, and all methods are observed to give similar results as the full-order model. Therefore, these results indicate that interpolatory methods are competitive with those proposed in [49].

Table 4.4: Oseen Equations: Problem 2

Method	$\frac{\ \mathbf{H}^{sp} - \mathbf{H}_r^{sp}\ _{\mathcal{H}_\infty}}{\ \mathbf{H}^{sp}\ _{\mathcal{H}_\infty}}$
Balanced Truncation	$2.64 \times 10^{-5}$
One Step Interpolation	$\infty$
IRKA	$5.15 \times 10^{-5}$

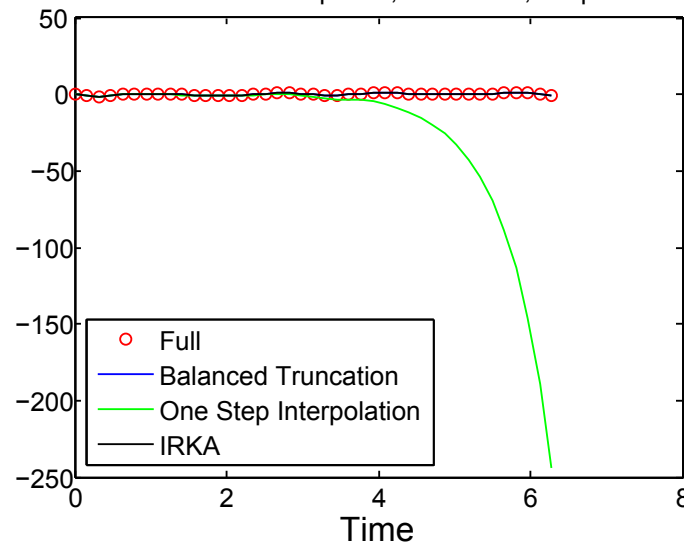
Figure 4.6: Oseen Equation: Problem 2, Time Domain Response, First Output  
Time Domain Response, Problem 2, Output 1



Figure 4.7: Oseen Equation: Problem 2, Time Domain Response, First Output: Balanced Truncation and One Step Interpolation

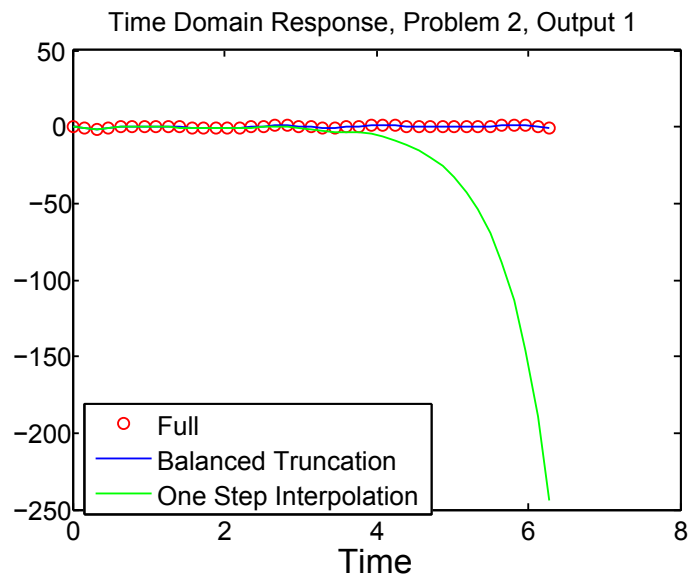


Figure 4.8: Oseen Equation: Problem 2, Time Domain Response, First Output: Balanced Truncation and IRKA

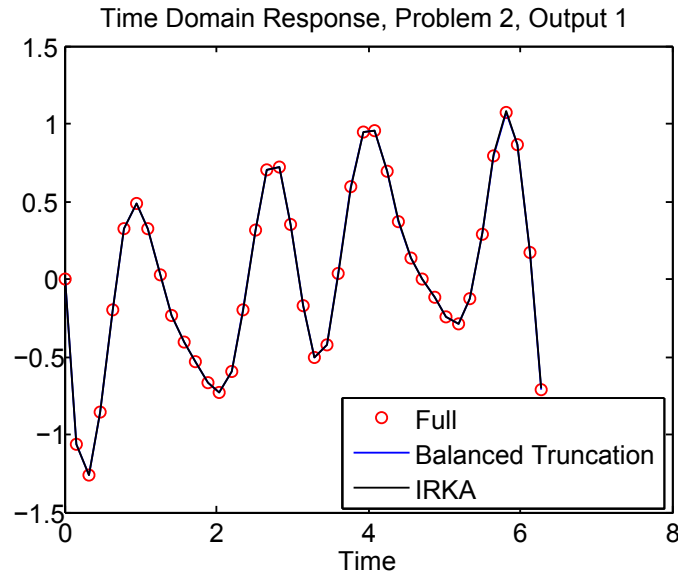


Figure 4.9: Oseen Equation: Problem 2, Time Domain Response, Second Output

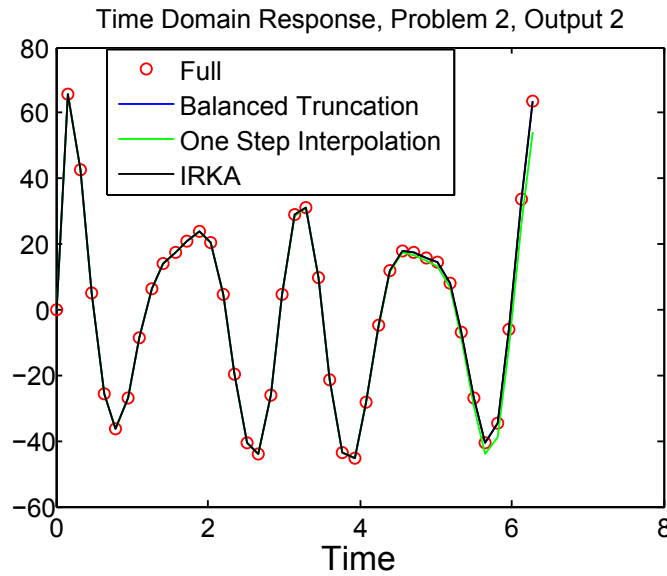
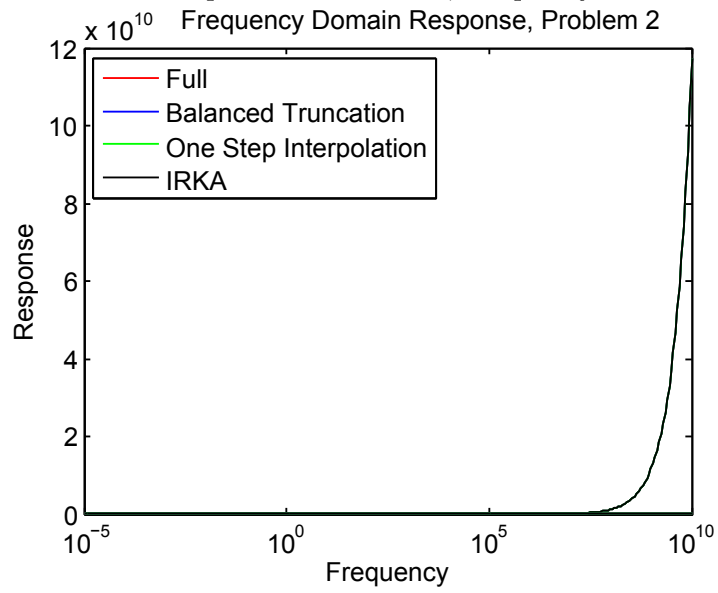


Figure 4.10: Oseen Equation: Problem 2, Frequency Domain Response



## Chapter 5

# Model Reduction of Second-order Systems

The aim of this chapter is to consider several frameworks for reducing second-order systems. In Section 5.2, an overview of existing methods for the reduction of second-order systems is provided. Then in Section 5.3, we use the first-order IRKA iteration to develop an IRKA framework for second-order systems. In Section 5.4, additional implementation issues of the algorithms proposed in Section 5.3 are studied. Finally, Section 5.4 provides a numerical study of four models to compare the methods discussed in this chapter.

## 5.1 Second-Order Systems

Second-order systems are represented in state-space as

$$\mathbf{H}(s) : \begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + \mathbf{G}\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}_1\mathbf{x}(t) + \mathbf{C}_2\dot{\mathbf{x}}(t), \end{cases} \quad (5.1.1)$$

where  $\mathbf{M}, \mathbf{G}, \mathbf{K} \in \mathbb{R}^{n \times n}$  and  $\mathbf{B} \in \mathbb{R}^{n \times m}$ ,  $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{p \times n}$ ,  $\mathbf{x}(t) \in \mathbb{R}^n$  is the *state*,  $\mathbf{u}(t) \in \mathbb{R}^m$  is the *input*, and  $\mathbf{y}(t) \in \mathbb{R}^p$  is the *output*.

The transfer function  $\mathbf{H}(s)$  is then given by

$$\mathbf{H}(s) = (\mathbf{C}_1 + s\mathbf{C}_2)(s^2\mathbf{M} + s\mathbf{G} + \mathbf{K})^{-1}\mathbf{B}.$$

The second-order system is called asymptotically stable provided the matrix polynomial  $\mathbf{P}(\lambda) = \lambda^2\mathbf{M} + \lambda\mathbf{G} + \mathbf{K}$  is stable.

Many physical phenomena are modeled with second-order systems; for example, structural vibrations can be modeled in this way where  $\mathbf{M}$ ,  $\mathbf{G}$ , and  $\mathbf{K}$  describe the structure's mass, damping and stiffness distributions, respectively. Also, electrical circuits are frequently described as second-order systems. Due to the nature of the physical phenomena modeled, oftentimes  $\mathbf{M}$ ,  $\mathbf{G}$ , and  $\mathbf{K}$  are symmetric positive definite. See [8], [3], [79], [67] for more examples.

For these second-order systems, the aim is to obtain a reduced-order model of order  $r \ll n$

of the same form as the full-order model, namely:

$$\mathbf{H}_r(s) : \begin{cases} \mathbf{M}_r \ddot{\mathbf{x}}_r(t) + \mathbf{G}_r \dot{\mathbf{x}}_r(t) + \mathbf{K}_r \mathbf{x}_r(t) = \mathbf{B}_r \mathbf{u}(t) \\ \mathbf{y}_r(t) = \mathbf{C}_{1r} \mathbf{x}_r(t) + \mathbf{C}_{2r} \dot{\mathbf{x}}_r(t), \end{cases} \quad (5.1.2)$$

where  $\mathbf{W}_r, \mathbf{V}_r \in \mathbb{R}^{n \times r}$  are used to compute the reduced quantities

$$\mathbf{M}_r = \mathbf{W}_r^T \mathbf{M} \mathbf{V}_r, \quad \mathbf{G}_r = \mathbf{W}_r^T \mathbf{G} \mathbf{V}_r, \quad \mathbf{K}_r = \mathbf{W}_r^T \mathbf{K} \mathbf{V}_r, \quad \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \quad \mathbf{C}_{1r} = \mathbf{C}_1 \mathbf{V}_r, \quad \mathbf{C}_{2r} = \mathbf{C}_2 \mathbf{V}_r.$$

Taking  $\mathcal{K}(s) = s^2 \mathbf{M} + s \mathbf{G} + \mathbf{K}$ ,  $\mathcal{B}(s) = \mathbf{B}$ , and  $\mathcal{C}(s) = \mathbf{C}_1 + s \mathbf{C}_2$ , system (5.1.1) fits into the generalized coprime realization as defined by (1.3.3). Therefore, Theorem 1.1 can be applied in the following way: given a set of interpolation points  $\{\sigma_i\}_{i=1}^r, \{\mu_i\}_{i=1}^r \subset \mathbb{C}$  and sets of right-tangential directions,  $\{\mathbf{b}_i\}_{i=1}^r \subset \mathbb{C}^m$ , and left-tangential directions,  $\{\mathbf{c}_i\}_{i=1}^r \subset \mathbb{C}^p$ , define

$$\mathbf{V}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_r ], \quad (5.1.3)$$

$$\mathbf{W}_r^T = \begin{bmatrix} \mathbf{c}_1^T (\mathbf{C}_1 + \mu_1 \mathbf{C}_2) (\mu_1^2 \mathbf{M} + \mu_1 \mathbf{G} + \mathbf{K})^{-1} \\ \vdots \\ \mathbf{c}_r^T (\mathbf{C}_1 + \mu_r \mathbf{C}_2) (\mu_r^2 \mathbf{M} + \mu_r \mathbf{G} + \mathbf{K})^{-1} \end{bmatrix}. \quad (5.1.4)$$

Then the reduced model,  $\mathbf{H}_r(s)$ , defined in (5.1.2) tangentially interpolates  $\mathbf{H}(s)$ , namely

$$\mathbf{H}(\sigma_i) \mathbf{b}_i = \mathbf{H}_r(\sigma_i) \mathbf{b}_i \quad \text{and} \quad \mathbf{c}_i^T \mathbf{H}(\sigma_i) = \mathbf{c}_i^T \mathbf{H}_r(\sigma_i).$$

Furthermore, if  $\mu_i = \sigma_i$ , then  $\mathbf{H}_r(s)$  bitangentially interpolates  $\mathbf{H}(s)$  at  $s = \sigma_i$ , namely

$$\mathbf{c}_i^T \mathbf{H}'(\sigma_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(\sigma_i) \mathbf{b}_i.$$

For models where  $\mathbf{M}$ ,  $\mathbf{G}$ , and  $\mathbf{K}$  are symmetric positive definite, the reduced-order model defined in terms of  $\mathbf{V}_r$  and  $\mathbf{W}_r$  will not necessarily be such that  $\mathbf{M}_r$ ,  $\mathbf{G}_r$  and  $\mathbf{K}_r$  are symmetric positive definite. For these cases, taking  $\mathbf{W}_r = \mathbf{V}_r$  allows for the symmetry and positive definiteness to be preserved. Moreover, the reduced model is guaranteed to be stable in this case.

Another way to analyze second-order systems is to convert the second-order system into a first-order system by letting

$$\mathbf{q}(t) = \begin{bmatrix} \mathbf{x}(t) \\ \dot{\mathbf{x}}(t) \end{bmatrix} \in \mathbb{R}^{2n}.$$

Then the first-order state-space realization is given by

$$\begin{cases} \mathcal{E}_{2n} \dot{\mathbf{q}}(t) = \mathcal{A}_{2n} \mathbf{q}(t) + \mathcal{B}_{2n} \mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}_{2n} \mathbf{q}(t), \end{cases} \quad (5.1.5)$$

where

$$\mathcal{E}_{2n} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix}, \quad \mathcal{A}_{2n} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{K} & -\mathbf{G} \end{bmatrix}, \quad \mathcal{B}_{2n} = \begin{bmatrix} \mathbf{0} \\ \mathbf{B} \end{bmatrix}, \quad \mathbf{C}_{2n} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix}.$$

Denote the associated transfer function by  $\mathcal{H}_{2n}(s) = \mathbf{C}_{2n}(s\mathbf{E}_{2n} - \mathbf{A}_{2n})^{-1}\mathbf{B}_{2n}$ . Of course, since  $\mathcal{H}_{2n}(s)$  is simply the first-order representation of  $\mathbf{H}(s)$ , we have  $\mathcal{H}_{2n}(s) = \mathbf{H}(s)$ .

Applying Theorem 1.1 to  $\mathcal{H}_{2n}(s)$  requires constructing matrices

$$\mathbf{V}_{r,1} = [\mathbf{v}_1, \dots, \mathbf{v}_r] = [(\sigma_1\mathbf{E}_{2n} - \mathbf{A}_{2n})^{-1}\mathbf{B}_{2n}\mathbf{b}_1, \dots, (\sigma_r\mathbf{E}_{2n} - \mathbf{A}_{2n})^{-1}\mathbf{B}_{2n}\mathbf{b}_r],$$

$$\mathbf{W}_{r,1}^T = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_r^T \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1^T \mathbf{C}_{2n} (\sigma_1\mathbf{E}_{2n} - \mathbf{A}_{2n})^{-1} \\ \vdots \\ \mathbf{c}_r^T \mathbf{C}_{2n} (\sigma_r\mathbf{E}_{2n} - \mathbf{A}_{2n})^{-1} \end{bmatrix}.$$

Here we use the subscript to emphasize that  $\mathbf{V}_{r,1}$  and  $\mathbf{W}_{r,1}$  are the reducing matrices for the first-order representation. Of course, one of the main issues is that the second-order model is now being reduced to a first-order model. To remedy this issue, [27] suggests splitting  $\mathbf{V}_{r,1}$  and  $\mathbf{W}_{r,1}$  as

$$\mathbf{V}_{r,1} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{W}_{r,1} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix},$$

and then defining the reducing matrices as

$$\widetilde{\mathbf{V}}_r = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{W}}_r = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix},$$

where  $\widetilde{\mathbf{W}}_r^T \widetilde{\mathbf{V}}_r = \mathbf{I}_{2n}$  and  $\mathbf{W}_1^T \mathbf{V}_2$  is assumed to be invertible. The reduced-order model defined



by

$$\begin{cases} \widetilde{\mathbf{W}}_r^T \boldsymbol{\varepsilon}_{2n} \widetilde{\mathbf{V}}_r \dot{\mathbf{q}}(t) = \widetilde{\mathbf{W}}_r^T \mathcal{A}_{2n} \widetilde{\mathbf{V}}_r \mathbf{q}(t) + \widetilde{\mathbf{W}}_r^T \mathcal{B}_{2n} \mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{c}_{2n} \widetilde{\mathbf{V}}_r \mathbf{q}(t), \end{cases} \quad (5.1.6)$$

where

$$\begin{aligned} \widetilde{\mathbf{W}}_r^T \boldsymbol{\varepsilon}_{2n} \widetilde{\mathbf{V}}_r &= \begin{bmatrix} \mathbf{W}_1^T \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2^T \mathbf{M} \mathbf{V}_2 \end{bmatrix} & \widetilde{\mathbf{W}}_r^T \mathcal{A}_{2n} \widetilde{\mathbf{V}}_r &= \begin{bmatrix} \mathbf{0} & \mathbf{W}_1^T \mathbf{V}_2 \\ -\mathbf{W}_2^T \mathbf{K} \mathbf{V}_1 & -\mathbf{W}_2^T \mathbf{G} \mathbf{V}_2 \end{bmatrix} \\ \widetilde{\mathbf{W}}_r^T \mathcal{B}_{2n} &= \begin{bmatrix} \mathbf{0} \\ \mathbf{W}_2^T \mathbf{B} \end{bmatrix} & \mathbf{c}_{2n} \widetilde{\mathbf{V}}_r &= \begin{bmatrix} \mathbf{C}_1 \mathbf{V}_1 & \mathbf{C}_2 \mathbf{V}_2 \end{bmatrix} \end{aligned}$$

corresponds to a second-order model as shown in [27]. For more details about second-order models, see [8], [68], [31],[29], [27], [61], and [4].

## 5.2 Balanced Truncation Methods for Second-order Systems

First, we consider converting the second-order model into its corresponding first-order model,  $\mathcal{H}_{2n}(s) = \mathbf{c}_{2n}(s\boldsymbol{\varepsilon}_{2n} - \mathcal{A}_{2n})^{-1}\mathcal{B}_{2n}$  as in (5.1.5). Once the original model is represented as a first-order model of order  $2n$ , any generic model reduction method, such as IRKA or balanced truncation, may be implemented. To apply IRKA in the first-order framework,

take  $\mathbf{E} = \mathbf{E}_{2n}$ ,  $\mathbf{A} = \mathbf{A}_{2n}$ ,  $\mathbf{B} = \mathbf{B}_{2n}$ , and  $\mathbf{C} = \mathbf{C}_{2n}$  in Algorithm 1.7.1. Implementing balanced truncation in the first-order framework requires the reachability and controllability gramians,  $\mathbf{P}$  and  $\mathbf{Q}$ , respectively, which are the solutions to the computationally intensive Lyapunov equations:

$$\mathbf{E}_{2n}\mathbf{P}\mathbf{A}_{2n}^T + \mathbf{A}_{2n}\mathbf{P}\mathbf{E}_{2n}^T = -\mathbf{B}_{2n}\mathbf{B}_{2n}^T, \quad \mathbf{E}_{2n}^T\mathbf{Q}\mathbf{A}_{2n} + \mathbf{A}_{2n}^T\mathbf{Q}\mathbf{E}_{2n} = -\mathbf{C}_{2n}^T\mathbf{C}_{2n}. \quad (5.2.1)$$

For a reachable, observable and stable first-order system, a balancing transformation is computed in order to give

$$\mathbf{T}\mathbf{P}\mathbf{T}^T = \mathbf{T}^{-T}\mathbf{Q}\mathbf{T}^{-1} = \text{diag}(\nu_1, \dots, \nu_r)$$

where  $\nu_i$  is the  $i^{\text{th}}$  Hankel singular value defined as

$$\nu_i = \sqrt{\lambda_i(\mathbf{P}\mathbf{Q})}.$$

To obtain the reduced model, the states associated with the smallest singular values, namely those that correspond to the states that are both difficult to reach and difficult to observe, are eliminated. See [2] for implementation details.

While balanced truncation and IRKA may provide small model reduction errors, one of the key concerns is that theoretical or computational issues may preclude the first-order reduced model from being expressed as a second-order model as discussed in [64]. As a

result, the second-order structure, and thereby the physical interpretation of the model, is lost even though the model reduction error may be small. Since the preservation of this second-order structure is pivotal for many models, research has been devoted to the development of balanced truncation methods for second-order systems. To do so, balanced truncation methods for the second-order framework as discussed in [68] begin by converting the second-order model to its corresponding first-order state-space realization as in (5.1.5) and then obtain the controllability and observability gramians,  $\mathcal{P}$  and  $\mathcal{Q}$ , respectively, by solving (5.2.1). The gramians are then partitioned into four  $n \times n$  blocks as follows:

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_p & \mathcal{P}_{12} \\ \mathcal{P}_{12}^T & \mathcal{P}_v \end{bmatrix}, \quad \mathcal{Q} = \begin{bmatrix} \mathcal{Q}_p & \mathcal{Q}_{12} \\ \mathcal{Q}_{12}^T & \mathcal{Q}_v \end{bmatrix}.$$

The quantities  $\mathcal{P}_p$  and  $\mathcal{P}_v$  are called the position and velocity controllability gramians of the second-order system, and  $\mathcal{Q}_p$  and  $\mathcal{Q}_v$  are called the position and velocity observability gramians of the second-order system, respectively. Using these gramians, the following singular values are defined, provided the second-order system is asymptotically stable:

- a) The *position singular values* of (5.1.1), denoted by  $\eta_j^p$ , are defined as the square roots of the eigenvalues of the matrix  $\mathcal{P}_p \mathcal{Q}_p$ , namely  $\eta_j^p = \sqrt{\lambda(\mathcal{P}_p \mathcal{Q}_p)}$ .
- b) The *velocity singular values* of (5.1.1), denoted by  $\eta_j^v$ , are defined as the square roots of the eigenvalues of the matrix  $\mathcal{P}_v \mathbf{M}^T \mathcal{Q}_v \mathbf{M}$ , namely  $\eta_j^v = \sqrt{\lambda(\mathcal{P}_v \mathbf{M}^T \mathcal{Q}_v \mathbf{M})}$ .
- c) The *position-velocity singular values* of (5.1.1), denoted by  $\eta_j^{pv}$ , are defined as the square roots of the eigenvalues of the matrix  $\mathcal{P}_p \mathbf{M}^T \mathcal{Q}_v \mathbf{M}$ , namely  $\eta_j^{pv} = \sqrt{\lambda(\mathcal{P}_p \mathbf{M}^T \mathcal{Q}_v \mathbf{M})}$ .

d) The *velocity-position singular values* of (5.1.1), denoted by  $\eta_j^{vp}$ , are defined as the square roots of the eigenvalues of the matrix  $\mathcal{P}_v \mathcal{Q}_p$ , namely  $\eta_j^{vp} = \sqrt{\lambda(\mathcal{P}_v \mathcal{Q}_p)}$ .

Using (a) - (d), the following balanced realizations are defined:

- 1) System (5.1.1) is *position balanced* if  $\mathcal{P}_p = \mathcal{Q}_p$ .
- 2) System (5.1.1) is *velocity balanced* if  $\mathcal{P}_v = \mathcal{Q}_v$ .
- 3) System (5.1.1) is *position-velocity balanced* if  $\mathcal{P}_p = \mathcal{Q}_v$ .
- 4) System (5.1.1) is *velocity-position balanced* if  $\mathcal{P}_v = \mathcal{Q}_p$ .

In order to compute the reduced-order model, the full-order system is converted into one of the forms as given in (1) - (4) and then the appropriate position and velocity components corresponding to the smallest singular values are eliminated. Therefore, these definitions result in several methods, such as second-order balanced truncation with position balancing (SOBTp), second-order balanced truncation with position-velocity balancing (SOBTpv) and free velocity second-order balanced truncation (SOBTfv) as discussed in [68] and [64].

For the numerical simulations, we considered obtaining the position-velocity balanced form through the SOBTpv method as proposed in [68]. To implement SOBTpv, a Cholesky factorization

$$\mathcal{P}_p = \mathbf{R}_p \mathbf{R}_p^T \quad \mathcal{Q}_v = \mathbf{L}_v \mathbf{L}_v^T$$

and a singular value decomposition

$$\mathbf{R}_p^T \mathbf{M}^T \mathbf{L}_v = \mathbf{U}_{pv} \boldsymbol{\Sigma}_{pv} \mathbf{V}_{pv}^T$$

are first computed. Then as [68] proves, system (5.1.1) is balanced when the balancing transformation matrices are defined as

$$\mathbf{T}_r = \mathbf{R}_p \mathbf{U}_{pv} \boldsymbol{\Sigma}_{pv}^{-\frac{1}{2}} \quad \text{and} \quad \mathbf{T}_l = \boldsymbol{\Sigma}_{pv}^{-\frac{1}{2}} \mathbf{V}_{pv}^T \mathbf{L}_v^T.$$

From this balancing transformation, an algorithm for balanced truncation of second-order systems, SOBTpv, is presented in [68].

**Algorithm 5.2.1.** [68] **Second-order Balanced Truncation Model Reduction with Position-Velocity Balancing (SOBTpv)**

1. Solve the Lyapunov equations (5.2.1) of dimension  $2n$  to obtain  $\mathcal{P}$  and  $\mathcal{Q}$ .
2. Partition  $\mathcal{P}$  and  $\mathcal{Q}$  as in (5.2).
3. Compute the Cholesky factorizations:  $\mathcal{P}_p = \mathbf{R}_p \mathbf{R}_p^T$     $\mathcal{Q}_v = \mathbf{L}_v \mathbf{L}_v^T$ .
4. Compute the singular value decomposition:  $\mathbf{R}_p^T \mathbf{M}^T \mathbf{L}_v = \mathbf{U}_{pv} \boldsymbol{\Sigma}_{pv} \mathbf{V}_{pv}^T$  with

$$\mathbf{U}_{pv} = \begin{bmatrix} \mathbf{U}_{pv,1} & \mathbf{U}_{pv,2} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{pv} = \begin{bmatrix} \boldsymbol{\Sigma}_{pv,1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{pv,2} \end{bmatrix}, \quad \mathbf{V}_{pv} = \begin{bmatrix} \mathbf{V}_{pv,1} & \mathbf{V}_{pv,2} \end{bmatrix}$$

and  $\Sigma_{pv,1} = \text{diag}(\eta_1^{pv}, \dots, \eta_r^{pv})$  and  $\Sigma_{pv,2} = \text{diag}(\eta_{r+1}^{pv}, \dots, \eta_n^{pv})$ .

5. Define  $\mathbf{W}_r = \mathbf{L}_v \mathbf{V}_{pv,1} \Sigma_{pv,1}^{-\frac{1}{2}}$  and  $\mathbf{V}_r = \mathbf{R}_p \mathbf{U}_{pv,1} \Sigma_{pv,1}^{-\frac{1}{2}}$ .

6.  $\mathbf{M}_r = \mathbf{I}_r$ ,  $\mathbf{G}_r = \mathbf{W}_r^T \mathbf{G} \mathbf{V}_r$ ,  $\mathbf{K}_r = \mathbf{W}_r^T \mathbf{K} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r(s) = \mathbf{C}(s) \mathbf{V}_r$ .

It is important to emphasize that if the matrices  $\mathbf{M}$ ,  $\mathbf{G}$ , and  $\mathbf{K}$  are symmetric positive definite, then SOBTpv and SOBTfv preserve this property. In general, however, stability can not be guaranteed with second-order balanced truncation methods. Moreover, in all cases, a priori upper bounds for second-order balanced truncation have yet to be presented. This is an important deviation from the first-order case, where balanced truncation is lauded for its preservation of stability and a priori upper bound despite its reliance on the costly solution of the Lyapunov equations.

### 5.3 An IRKA framework for Second-order Systems

In this section, we consider extending the IRKA framework to second-order systems. The  $\mathcal{H}_2$  approximation problem for second-order systems is defined as

$$\mathbf{H}_r(s) = \min_{\substack{\deg(\hat{\mathbf{H}}_r)=r \\ \hat{\mathbf{H}}_r \text{ stable}}} \|\mathbf{H}(s) - \hat{\mathbf{H}}_r(s)\|_{\mathcal{H}_2}. \quad (5.3.1)$$

$$\hat{\mathbf{H}}_r = (\mathbf{C}_{1r} + s \mathbf{C}_{2r}) (s^2 \mathbf{M}_r + s \mathbf{G}_r + \mathbf{K}_r)^{-1} \mathbf{B}_r$$

Recall that IRKA converges to a reduced model that satisfies the first-order necessary conditions. However, these conditions as stated in Theorem 1.4 only require a pole-residue

formulation, namely the reduced-order model is defined as

$$\mathbf{H}_r(s) = \sum_{i=1}^r \frac{1}{s - \hat{\lambda}_i} \mathbf{c}_i \mathbf{b}_i^T$$

where  $\{\hat{\lambda}_i\}_{i=1}^r$  and  $\{\mathbf{c}_i \mathbf{b}_i^T\}_{i=1}^r$  are the simple poles and residues of  $\mathbf{H}_r(s)$ , respectively. By expressing  $\mathbf{H}_r(s)$  only in terms of its poles and residues, the model's structure is ignored, implying the reduced model obtained by IRKA will not satisfy (5.3.1), which restricts the reduced-order model to be only of second-order structure. However, if IRKA is applied to the first-order representation of the second-order model,

$$\mathbf{H}(s) = \mathfrak{H}_{2n}(s) = \mathbf{C}_{2n}(s\mathbf{E}_{2n} - \mathbf{A}_{2n})^{-1}\mathbf{B}_{2n},$$

as defined in (5.1.5), then  $\mathfrak{H}_r(s)$  will satisfy the optimal  $\mathcal{H}_2$  problem:

$$\mathfrak{H}_r(s) = \min_{\substack{\deg(\tilde{\mathfrak{H}}_r)=r \\ \tilde{\mathfrak{H}}_{2n}:stable}} \|\mathfrak{H}_{2n}(s) - \tilde{\mathfrak{H}}_r(s)\|_{\mathcal{H}_2}.$$

By reducing the second-order model to a first-order reduced model, the issues as previously discussed arise, namely conversion of the first-order model to a second-order model may be computationally intensive or impossible. As a result, this chapter proposes an IRKA-based algorithm for second-order models that preserves the second-order structure. Applying the IRKA iteration to second-order systems with  $\mathfrak{K}(s) = s^2\mathbf{M} + s\mathbf{G} + \mathbf{K}$ ,  $\mathfrak{B}(s) = \mathbf{B}$ , and  $\mathfrak{C}(s) = \mathbf{C}_1 + s\mathbf{C}_2$  requires using an initial shift selection of  $\sigma_i$  for  $i = 1, \dots, r$  and initial

tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$  to solve

$$\mathbf{V}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_r ] \in \mathbb{R}^{n \times r}$$

$$\mathbf{W}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-T} \mathbf{C}(\sigma_1)^T \mathbf{c}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-T} \mathbf{C}(\sigma_r)^T \mathbf{c}_r ] \in \mathbb{R}^{n \times r}$$

where  $\mathbf{C}(s) = \mathbf{C}_1 + s\mathbf{C}_2$ . Then the intermediate reduced-order model is defined as

$$\mathbf{H}_{2r} = \mathbf{C}_r(s)(s^2 \mathbf{M}_r + s\mathbf{G}_r + \mathbf{K}_r)^{-1} \mathbf{B}_r$$

where  $\mathbf{M}_r = \mathbf{W}_r^T \mathbf{M} \mathbf{V}_r$ ,  $\mathbf{G}_r = \mathbf{W}_r^T \mathbf{G} \mathbf{V}_r$ ,  $\mathbf{K}_r = \mathbf{W}_r^T \mathbf{K} \mathbf{V}_r$ ,  $\mathbf{C}_r(s) = \mathbf{C}_1 \mathbf{V}_r + s\mathbf{C}_2 \mathbf{V}_r$ , and  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ . The reduced model's transfer function is denoted by  $2r$  to emphasize the presence of  $2r$  poles. This follows due to the quadratic polynomial eigenvalue problem associated with the second-order model. The mirror images of these  $2r$  poles then become the shifts for the next iteration of IRKA. Now with  $2r$  shifts, the next step requires computing the  $n \times 2r$  matrices:

$$\mathbf{V}_{2r} = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_{2r}^2 \mathbf{M} + \sigma_{2r} \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_{2r} ] \in \mathbb{R}^{n \times 2r}$$

$$\mathbf{W}_{2r} = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-T} \mathbf{C}(\sigma_1)^T \mathbf{c}_1, \dots, (\sigma_{2r}^2 \mathbf{M} + \sigma_{2r} \mathbf{G} + \mathbf{K})^{-T} \mathbf{C}(\sigma_{2r})^T \mathbf{c}_{2r} ] \in \mathbb{R}^{n \times 2r}.$$

Using  $\mathbf{V}_{2r}$  and  $\mathbf{W}_{2r}$ , the next reduced model obtained is

$$\mathbf{H}_{4r} = \mathbf{C}_{2r}(s)(s^2 \mathbf{M}_{2r} + s\mathbf{G}_{2r} + \mathbf{K}_{2r})^{-1} \mathbf{B}_{2r},$$



where  $\mathbf{M}_{2r} = \mathbf{W}_{2r}^T \mathbf{M} \mathbf{V}_{2r}$ ,  $\mathbf{G}_{2r} = \mathbf{W}_{2r}^T \mathbf{G} \mathbf{V}_{2r}$ ,  $\mathbf{K}_{2r} = \mathbf{W}_{2r}^T \mathbf{K} \mathbf{V}_{2r}$ ,  $\mathbf{C}_{2r}(s) = \mathbf{C}_1 \mathbf{V}_{2r} + s \mathbf{C}_2 \mathbf{V}_{2r}$ , and  $\mathbf{B}_{2r} = \mathbf{W}_{2r}^T \mathbf{B}$ . The key issue is that the presence of  $2r$  rather than  $r$  shifts implies that the reduced-order model will grow by a factor of two at each IRKA iteration. Letting  $k$  denote the number of IRKA iterations, this shift increase implies that the final reduced-order model will be of order  $2^{k-1}r$  when IRKA is initialized with  $r$  shifts and directions. To prevent this growth, the new shift iterate must include only  $r$  shifts even though  $2r$  poles are available.

### 5.3.1 SOR-IRKA

The first IRKA-based algorithm is presented in Algorithm 5.3.1. To initialize,  $r$  shifts and directions are chosen. Then the reducing matrices  $\mathbf{V}_r$  and  $\mathbf{W}_r$  are computed as in (5.1.3). The reduced-order model,  $\mathbf{H}_r(s)$  obtained in Step 4(a), is an  $r^{th}$  order model with  $2r$  poles. To obtain only  $r$  shifts, we first convert  $\mathbf{H}_r(s)$  to its corresponding first-order representation,  $\mathcal{H}_{2r}(s) = \mathbf{C}_{2r}(s\mathcal{E}_{2r} - \mathcal{A}_{2r})^{-1}\mathcal{B}_{2r}$ , where the subscript  $2r$  is used to emphasize the model's dimension. At this point, any of the generic model reduction techniques can be applied to reduce  $\mathcal{H}_{2r}(s)$  to an order  $r$  model, denoted by  $\tilde{\mathbf{H}}_r(s)$ . In our numerical results, we reduced  $\mathcal{H}_{2r}(s)$  using either IRKA (Method  $\mathcal{H}_{1,r}$ ) or balanced truncation (Method  $\mathcal{H}_{2,r}$ ). It is important to keep in mind that the usual drawback of balanced truncation, namely the expensive Lyapunov equations, is no longer a concern as the full-order model in this case is only of dimension  $2r$  with  $r \ll n$ . See [2] for implementation details of balanced truncation. Upon obtaining the  $r^{th}$  order model,  $\tilde{\mathbf{H}}_r(s)$ , the shifts for the next iteration are the mirror

images of  $\tilde{\mathbf{H}}_r(s)$ . Since the shifts are not the mirror images of the poles of the second-order reduced model, the  $\mathcal{H}_2$  optimality first-order necessary conditions will not be satisfied upon convergence. Despite this lack of optimality, this method may be justified as follows: model reduction methods obtain an  $r^{\text{th}}$  order model that captures the behavior of an order  $2r$  model, implying that the shift information of the  $r^{\text{th}}$  order model may best describe the key characteristics of the  $2r$  shifts. This method as detailed in Algorithm 5.3.1 is referred to as SOR-IRKA, where the letter suffix R in the title emphasizes that the intermediate model,  $\mathfrak{H}_{2r}(s)$ , is reduced to obtain the  $r$  shifts.

**Algorithm 5.3.1. Second-order IRKA for MIMO Tangential Interpolation**

**(SOR-IRKA)**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .
2.  $\mathbf{V}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_r ]$ .
3.  $\mathbf{W}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-T} \mathbf{C} (\sigma_1)^T \mathbf{c}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-T} \mathbf{C} (\sigma_r)^T \mathbf{c}_r ]$ .
4. while (not converged)
  - (a)  $\mathbf{M}_r = \mathbf{W}_r^T \mathbf{M} \mathbf{V}_r$ ,  $\mathbf{G}_r = \mathbf{W}_r^T \mathbf{G} \mathbf{V}_r$ ,  $\mathbf{K}_r = \mathbf{W}_r^T \mathbf{K} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r(s) = \mathbf{C}(s) \mathbf{V}_r$ .
  - (b) Convert the second-order reduced model found in Step (4a) to its associated first-order framework,  $\mathfrak{H}_{2r}(s) = \mathbf{C}_{2r}(s \mathbf{E}_{2r} - \mathbf{A}_{2r})^{-1} \mathbf{B}_{2r}$ , as in (5.1.5).

(c) *Intermediate Step: Reduce the order  $2r$  system,  $\mathcal{H}_{2r}(s)$ , to an order  $r$  system,*

$$\tilde{\mathbf{H}}_r(s) = \tilde{\mathbf{C}}_r(s\tilde{\mathbf{E}}_r - \tilde{\mathbf{A}}_r)^{-1}\tilde{\mathbf{B}}_r.$$

(d) *Compute  $\mathbf{Y}^T\tilde{\mathbf{A}}_r\mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T\tilde{\mathbf{E}}_r\mathbf{X} = \mathbf{I}_r$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda\tilde{\mathbf{E}}_r - \tilde{\mathbf{A}}_r$ .*

(e)  $\sigma_i \leftarrow -\lambda_i(\tilde{\mathbf{A}}_r, \tilde{\mathbf{E}}_r)$ ,  $\mathbf{b}_i^T \leftarrow \mathbf{e}_i^T\mathbf{Y}^T\tilde{\mathbf{B}}_r$ , and  $\mathbf{c}_i^T \leftarrow \tilde{\mathbf{C}}_r\mathbf{X}\mathbf{e}_i$  for  $i = 1, \dots, r$ .

(f)  $\mathbf{V}_r = [ (\sigma_1^2\mathbf{M} + \sigma_1\mathbf{G} + \mathbf{K})^{-1}\mathbf{B}\mathbf{b}_1, \dots, (\sigma_r^2\mathbf{M} + \sigma_r\mathbf{G} + \mathbf{K})^{-1}\mathbf{B}\mathbf{b}_r ]$ .

(g)  $\mathbf{W}_r = [ (\sigma_1^2\mathbf{M} + \sigma_1\mathbf{G} + \mathbf{K})^{-T}\mathbf{C}(\sigma_1)^T\mathbf{c}_1, \dots, (\sigma_r^2\mathbf{M} + \sigma_r\mathbf{G} + \mathbf{K})^{-T}\mathbf{C}(\sigma_r)^T\mathbf{c}_r ]$ .

$$5. \mathbf{M}_r = \mathbf{W}_r^T\mathbf{M}\mathbf{V}_r, \mathbf{G}_r = \mathbf{W}_r^T\mathbf{G}\mathbf{V}_r, \mathbf{K}_r = \mathbf{W}_r^T\mathbf{K}\mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}, \mathbf{C}_r(s) = \mathbf{C}(s)\mathbf{V}_r.$$

As discussed, we consider two approaches for Step (4c):

- **Method  $\mathcal{H}_{1,r}$ :** Reduce  $\mathcal{H}_{2r}(s)$  to order  $r$  using IRKA (Algorithm 1.7.1).
- **Method  $\mathcal{H}_{2,r}$ :** Reduce  $\mathcal{H}_{2r}(s)$  to order  $r$  using first-order balanced truncation.

### 5.3.2 SO-IRKA

Another IRKA-based algorithm is proposed in Algorithm 5.3.2. As with SOR-IRKA, the algorithm begins by computing the matrices  $\mathbf{V}_r$  and  $\mathbf{W}_r$  required to achieve Hermite tangential interpolation for a given initial shift and direction selection. The second-order reduced model calculated using  $\mathbf{V}_r$  and  $\mathbf{W}_r$  is then converted to its first-order representation, denoted by  $\mathcal{H}_{2r}(s)$ . Instead of reducing  $\mathcal{H}_{2r}(s)$  as in SOR-IRKA, SO-IRKA computes the

poles of  $\mathcal{H}_{2r}(s)$  and then assigns  $r$  of the mirror images of these poles to be the shifts used in constructing the next  $\mathbf{V}_r$  and  $\mathbf{W}_r$  matrices.

**Algorithm 5.3.2. Second-order IRKA for MIMO Tangential Interpolation (SO-IRKA)**

1. Make an initial shift selection  $\sigma_i$  for  $i = 1, \dots, r$  and initial tangent directions  $\mathbf{b}_1, \dots, \mathbf{b}_r$  and  $\mathbf{c}_1, \dots, \mathbf{c}_r$ .

2.  $\mathbf{V}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_r ]$ .

3.  $\mathbf{W}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-T} \mathbf{C} (\sigma_1)^T \mathbf{c}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-T} \mathbf{C} (\sigma_r)^T \mathbf{c}_r ]$ .

4. while (not converged)

(a)  $\mathbf{M}_r = \mathbf{W}_r^T \mathbf{M} \mathbf{V}_r$ ,  $\mathbf{G}_r = \mathbf{W}_r^T \mathbf{G} \mathbf{V}_r$ ,  $\mathbf{K}_r = \mathbf{W}_r^T \mathbf{K} \mathbf{V}_r$ ,  $\mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}$ ,  $\mathbf{C}_r(s) = \mathbf{C}(s) \mathbf{V}_r$ .

(b) Convert the second-order reduced model found in Step (4a) to its associated first-order framework  $\mathcal{H}_{2r}(s) = \mathbf{C}_{2r}(s \mathbf{E}_{2r} - \mathbf{A}_{2r})^{-1} \mathbf{B}_{2r}$  as in (5.1.5).

(c) Compute  $\mathbf{Y}^T \mathbf{A}_{2r} \mathbf{X} = \text{diag}(\lambda_i)$  and  $\mathbf{Y}^T \mathbf{E}_{2r} \mathbf{X} = \mathbf{I}_{2r}$  where  $\mathbf{Y}^T$  and  $\mathbf{X}$  are left and right eigenvectors of  $\lambda \mathbf{E}_{2r} - \mathbf{A}_{2r}$ .

(d) Shift Selection Step: Using  $\lambda_1, \dots, \lambda_{2r}$ , assign  $\sigma_1, \dots, \sigma_r$  and  $\mu_1, \dots, \mu_r$  (see below).

(e)  $\mathbf{V}_r = [ (\sigma_1^2 \mathbf{M} + \sigma_1 \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_1, \dots, (\sigma_r^2 \mathbf{M} + \sigma_r \mathbf{G} + \mathbf{K})^{-1} \mathbf{B} \mathbf{b}_r ]$ .

(f)  $\mathbf{W} = [ (\mu_1^2 \mathbf{M} + \mu_1 \mathbf{G} + \mathbf{K})^{-T} \mathbf{C} (\mu_1)^T \mathbf{c}_1, \dots, (\mu_r^2 \mathbf{M} + \mu_r \mathbf{G} + \mathbf{K})^{-T} \mathbf{C} (\mu_r)^T \mathbf{c}_r ]$ .

$$5. \mathbf{M}_r = \mathbf{W}_r^T \mathbf{M} \mathbf{V}_r, \mathbf{G}_r = \mathbf{W}_r^T \mathbf{G} \mathbf{V}_r, \mathbf{K}_r = \mathbf{W}_r^T \mathbf{K} \mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T \mathbf{B}, \mathbf{C}_r(s) = \mathbf{C}(s) \mathbf{V}_r.$$

One way to make the selection in Step (4d) is by choosing the same shifts for both  $\mathbf{V}_r$  and  $\mathbf{W}_r$ , which is presented in Method  $\mathcal{H}_{3,r}$ . In this method, the mirror images of the poles closest to the imaginary axis are chosen; however, any of the  $2r$  shifts may be chosen.

#### Method $\mathcal{H}_{3,r}$ : Shift Selection Step (4d) of SO-IRKA (Hermite)

- $\sigma_i \longleftarrow -\lambda_i(\mathcal{A}_{2r}, \mathcal{E}_{2r}), \mathbf{b}_i^T \longleftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathcal{B}_{2r}$  for  $i = 1, \dots, r$ .
- $\mu_i \longleftarrow -\lambda_i(\mathcal{A}_{2r}, \mathcal{E}_{2r}), \mathbf{c}_i^T \longleftarrow \mathcal{C}_{2r} \mathbf{X} \mathbf{e}_i$  for  $i = 1, \dots, r$ .

Since Method  $\mathcal{H}_{3,r}$  ignores the poles furthest away from the imaginary axis, we also consider using all of the  $2r$  shifts by defining the new shift selection as in Method  $\mathcal{H}_{4,r}$ . While Method  $\mathcal{H}_{4,r}$  does not bitangentially interpolate  $\mathbf{H}(s)$  as in Method  $\mathcal{H}_{3,r}$ , this method is motivated by [48], where only tangential interpolation was achieved in order to maintain the structure of the port-Hamiltonian system.

#### Method $\mathcal{H}_{4,r}$ : Shift Selection Step (4d) of SO-IRKA (Lagrange)

- $\sigma_i \longleftarrow -\lambda_i(\mathcal{A}_{2r}, \mathcal{E}_{2r}), \mathbf{b}_i^T \longleftarrow \mathbf{e}_i^T \mathbf{Y}^T \mathcal{B}_{2r}$  for  $i = 1, \dots, r$ .
- $\mu_i \longleftarrow -\lambda_{i+r}(\mathcal{A}_{2r}, \mathcal{E}_{2r}), \mathbf{c}_i^T \longleftarrow \mathcal{C}_{2r} \mathbf{X} \mathbf{e}_{i+r}$  for  $i = 1, \dots, r$ .

It is important to emphasize the distinction between the SOR-IRKA and SO-IRKA algorithms. Since SOR-IRKA uses the poles from the reduced model obtained through the first-order representation, the shift iterates are no longer the mirror images of the poles

of the second-order model. Therefore, SOR-IRKA will not satisfy any of the optimal  $\mathcal{H}_2$  first-order necessary conditions. However, SO-IRKA satisfies a subset of the optimal  $\mathcal{H}_2$  necessary conditions upon convergence.

**Theorem 5.1.** *Suppose  $\mathbf{H}(s)$  and  $\mathbf{H}_r(s)$  are real stable dynamical systems. Let*

$$\mathbf{H}_r(s) = \sum_{i=1}^{2r} \frac{1}{s - \hat{\lambda}_i} \mathbf{c}_i \mathbf{b}_i^T$$

where  $\{\hat{\lambda}_i\}_{i=1}^{2r}$  and  $\{\mathbf{c}_i \mathbf{b}_i^T\}_{i=1}^{2r}$  are the simple poles and residues of  $\mathbf{H}_r(s)$ , respectively, then

a) *SO-IRKA with Method  $\mathcal{H}_{3,r}$  satisfies  $3r$  of the first-order necessary conditions of the  $\mathcal{H}_2$  optimality problem, namely*

1.  $\mathbf{H}(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{H}_r(-\hat{\lambda}_i) \mathbf{b}_i$  for  $i = 1, \dots, r$
2.  $\mathbf{c}_j^T \mathbf{H}(-\hat{\lambda}_j) = \mathbf{c}_j^T \mathbf{H}_r(-\hat{\lambda}_j)$  for  $j = 1, \dots, r$
3.  $\mathbf{c}_i^T \mathbf{H}'(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(-\hat{\lambda}_i) \mathbf{b}_i$  for  $i = 1, \dots, r$ .

b) *SO-IRKA with Method  $\mathcal{H}_{4,r}$  satisfies  $2r$  of the first-order necessary conditions of the  $\mathcal{H}_2$  optimality problem, namely*

1.  $\mathbf{H}(-\hat{\lambda}_i) \mathbf{b}_i = \mathbf{H}_r(-\hat{\lambda}_i) \mathbf{b}_i$  for  $i = 1, \dots, r$
2.  $\mathbf{c}_j^T \mathbf{H}(-\hat{\lambda}_j) = \mathbf{c}_j^T \mathbf{H}_r(-\hat{\lambda}_j)$  for  $j = r + 1, \dots, 2r$ .

*Proof.* By Theorem 1.4, the first-order necessary conditions are given as

1.  $\mathbf{H}(-\hat{\lambda}_i)\mathbf{b}_i = \mathbf{H}_r(-\hat{\lambda}_i)\mathbf{b}_i$  for  $i = 1, \dots, 2r$
2.  $\mathbf{c}_i^T \mathbf{H}(-\hat{\lambda}_i) = \mathbf{c}_i^T \mathbf{H}_r(-\hat{\lambda}_i)$  for  $i = 1, \dots, 2r$
3.  $\mathbf{c}_i^T \mathbf{H}'(-\hat{\lambda}_i)\mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(-\hat{\lambda}_i)\mathbf{b}_i$  for  $i = 1, \dots, 2r$ .

The shift selection prescribed by Method  $\mathcal{H}_{3,r}$  gives that  $\mathbf{H}_r(s)$  tangentially and bitangentially interpolates  $\mathbf{H}(s)$ , namely

1.  $\mathbf{H}(\sigma_i)\mathbf{b}_i = \mathbf{H}_r(\sigma_i)\mathbf{b}_i$  for  $i = 1, \dots, r$
2.  $\mathbf{c}_j^T \mathbf{H}(\sigma_j) = \mathbf{c}_j^T \mathbf{H}_r(\sigma_j)$  for  $j = 1, \dots, r$
3.  $\mathbf{c}_i^T \mathbf{H}'(\sigma_i)\mathbf{b}_i = \mathbf{c}_i^T \mathbf{H}'_r(\sigma_i)\mathbf{b}_i$  for  $i = 1, \dots, r$ .

The shifts used in Method  $\mathcal{H}_{4,r}$  imply that  $\mathbf{H}_r(s)$  tangentially interpolates  $\mathbf{H}(s)$ , namely

1.  $\mathbf{H}(\sigma_i)\mathbf{b}_i = \mathbf{H}_r(\sigma_i)\mathbf{b}_i$  for  $i = 1, \dots, r$
2.  $\mathbf{c}_j^T \mathbf{H}(\sigma_j) = \mathbf{c}_j^T \mathbf{H}_r(\sigma_j)$  for  $j = r + 1, \dots, 2r$ .

The result then follows immediately since the shifts are chosen to be the reflected poles of the second-order model in Method  $\mathcal{H}_{3,r}$  and Method  $\mathcal{H}_{4,r}$ . □

A similar approach to the Shift Selection Step was considered in [21], where the authors applied a global Arnoldi method to the reduction of second-order MIMO systems. While [21] chooses a subset of the eigenvalues based on their proximity to the imaginary axis,

the approach of [21] does not include tangential interpolation, implying that only a limited number of poles can be reflected and that none of the  $\mathcal{H}_2$  optimality conditions are satisfied.

*Remark:* For models where  $\mathbf{M}$ ,  $\mathbf{G}$ , and  $\mathbf{K}$  are symmetric positive definite matrices, setting

$$\mathbf{W}_r = \mathbf{V}_r$$

in both SOR-IRKA (Algorithm 5.3.1) and SO-IRKA (Algorithm 5.3.2) implies that the symmetry and positive definiteness is preserved as well as stability.

## 5.4 Numerical Results for the Effect of the Shift Reduction Step

To thoroughly investigate the *Intermediate Step* and *Shift Selection Step* of SOR-IRKA and SO-IRKA, we reduced a one-dimensional beam model, a building model, and a truss segment model using the methods proposed for SOR-IRKA and SO-IRKA.

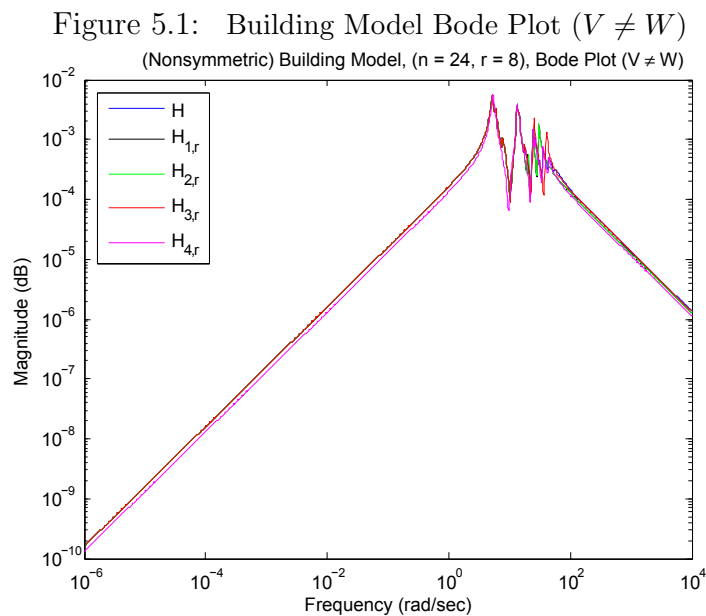


### 5.4.1 Building Model

The Building Model describes the Los Angeles University Hospital's eight floors as a SISO model of the form:

$$\mathbf{H}(s) : \begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + \mathbf{G}\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{B}u(t) \\ y(t) = \mathbf{C}\mathbf{x}(t), \end{cases} \quad (5.4.1)$$

where  $\mathbf{M}, \mathbf{G}, \mathbf{K} \in \mathbb{R}^{24 \times 24}$ ,  $\mathbf{B} \in \mathbb{R}^{24}$ , and  $\mathbf{C} \in \mathbb{R}^{24}$ . The reduced model obtained is a second-order model of dimension  $r = 8$ . While this is a very small model, it serves as an example of a second-order model where  $\mathbf{M}, \mathbf{G}$ , and  $\mathbf{K}$  are not symmetric positive definite. Therefore,  $\mathbf{W}_r$  is not set to be equal to  $\mathbf{V}_r$ , and two-sided reduction is implemented. For more details about the Building Model, see [2] and [3].



In Figure 5.1, the bode plots resulting from Methods  $\mathcal{H}_{1,r} - \mathcal{H}_{4,r}$  illustrate a very close match

Table 5.1: Building Model Errors, reducing from  $2r$  to  $r$ ,  $\mathbf{V}_r \neq \mathbf{W}_r$ 

Method	$\frac{\ H-H_r\ _{\mathcal{H}_\infty}}{\ H\ _{\mathcal{H}_\infty}}$	$\frac{\ H-H_r\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$
$\mathcal{H}_{1,r}$	$2.08 \times 10^{-1}$	$2.57 \times 10^{-1}$
$\mathcal{H}_{2,r}$	$3.17 \times 10^{-1}$	$3.65 \times 10^{-1}$
$\mathcal{H}_{3,r}$	$3.06 \times 10^{-1}$	$3.80 \times 10^{-1}$
$\mathcal{H}_{4,r}$	$3.57 \times 10^{-1}$	$2.76 \times 10^{-1}$

between the full and reduced models for all methods. Furthermore, all of the  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  and  $\frac{\|H-H_r\|_{\mathcal{H}_2}}{\|H\|_{\mathcal{H}_2}}$  errors are of the same order as shown in Table 5.1. Therefore, the Building Model suggests that SOR-IRKA and SO-IRKA yield similar results regardless of how the *Intermediate Step* and *Shift Selection Step* are implemented.

## 5.4.2 Beam Model

The Beam Model is a second-order system with proportional damping of the form:

$$\mathbf{H}(s) : \begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + (\alpha\mathbf{M} + \beta\mathbf{K})\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{B}u(t) \\ y(t) = \mathbf{C}\mathbf{x}(t), \end{cases} \quad (5.4.2)$$

where  $\mathbf{M}, \mathbf{G}, \mathbf{K} \in \mathbb{R}^{500 \times 500}$ ,  $\mathbf{B} \in \mathbb{R}^{500}$ , and  $\mathbf{C} \in \mathbb{R}^{500}$ . The scalars  $\alpha$  and  $\beta$  are proportional damping coefficients set as  $\alpha = 1/100$  and  $\beta = 1/100$ . See [9] for more details about this model.

Since the matrices  $\mathbf{M}$  and  $\mathbf{K}$  are symmetric positive definite, we set  $\mathbf{W}_r = \mathbf{V}_r$ , implying that only one sided reduction occurs in SOR-IRKA and SO-IRKA. In Figure 5.2, the bode plots for the full and reduced models are given. Although the oscillations associated with

the middle frequencies are captured nicely by all methods, a significant mismatch for higher frequencies occurs for Method  $\mathcal{H}_{2,r}$ ,  $\mathcal{H}_{3,r}$  and  $\mathcal{H}_{4,r}$ , indicating that SOR-IRKA with IRKA for the shift reduction step is superior. When examining Table 5.2, however, we observe that the overall model reduction errors,  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  and  $\frac{\|H-H_r\|_{\mathcal{H}_2}}{\|H\|_{\mathcal{H}_2}}$ , are of the same order regardless of the implementation.

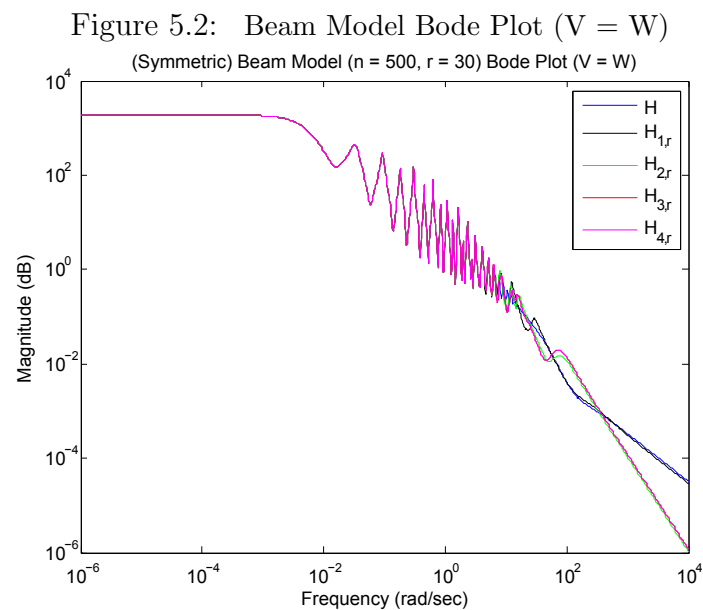


Table 5.2: Beam Model Errors, reducing from  $2r$  to  $r$ ,  $\mathbf{V}_r = \mathbf{W}_r$

Method	$\frac{\ H-H_r\ _{\mathcal{H}_\infty}}{\ H\ _{\mathcal{H}_\infty}}$	$\frac{\ H-H_r\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$
$\mathcal{H}_{1,r}$	$3.47 \times 10^{-4}$	$7.00 \times 10^{-3}$
$\mathcal{H}_{2,r}$	$3.08 \times 10^{-4}$	$5.46 \times 10^{-3}$
$\mathcal{H}_{3,r}$	$2.92 \times 10^{-4}$	$5.83 \times 10^{-3}$
$\mathcal{H}_{4,r}$	$2.92 \times 10^{-4}$	$5.83 \times 10^{-3}$

### 5.4.3 12a Model

The 12a Model describes the second left-side truss segment of the 1r Module from the International Space Station. This is a SISO model of the form:

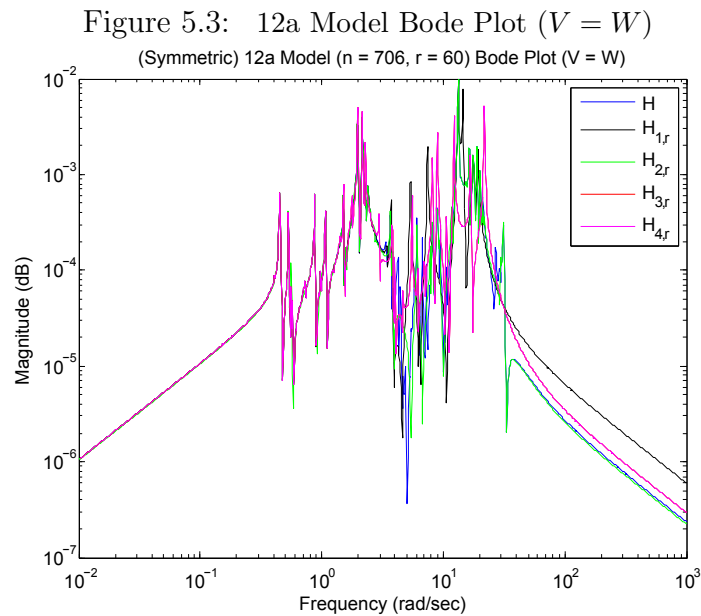
$$\mathbf{H}(s) : \begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + \mathbf{G}\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{B}u(t) \\ y(t) = \mathbf{C}_1\mathbf{x}(t) + \mathbf{C}_2\dot{\mathbf{x}}(t), \end{cases} \quad (5.4.3)$$

where  $\mathbf{M}, \mathbf{G}, \mathbf{K} \in \mathbb{R}^{706 \times 706}$ ,  $\mathbf{B} \in \mathbb{R}^{706}$ , and  $\mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}^{706}$ . The matrices  $\mathbf{M}, \mathbf{G}$  and  $\mathbf{K}$  are symmetric positive definite, so we only consider one-sided implementation of SOR-IRKA and SO-IRKA with  $\mathbf{W}_r = \mathbf{V}_r$ . For more details about this model, see [2] and [3].

In Figure 5.3, we observe that several of the methods fail to fully capture the complexity of the original model. Especially for larger frequencies, Method  $\mathcal{H}_{2,r}$  is the only method that matches the full-order model. This is also observed in Table 5.3, where the  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  and  $\frac{\|H-H_r\|_{\mathcal{H}_2}}{\|H\|_{\mathcal{H}_2}}$  errors associated with  $\mathcal{H}_{2,r}$  are one to two orders smaller than all other methods. Therefore, this model suggests that the intermediate step may noticeably impact the SOR-IRKA iteration.

Table 5.3: 12a Model Errors, reducing from  $2r$  to  $r$ ,  $\mathbf{V}_r = \mathbf{W}_r$

Method	$\frac{\ H-H_r\ _{\mathcal{H}_\infty}}{\ H\ _{\mathcal{H}_\infty}}$	$\frac{\ H-H_r\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$
$\mathcal{H}_{1,r}$	$9.75 \times 10^{-1}$	1.25
$\mathcal{H}_{2,r}$	$4.36 \times 10^{-2}$	$5.76 \times 10^{-2}$
$\mathcal{H}_{3,r}$	$9.92 \times 10^{-1}$	1.32
$\mathcal{H}_{4,r}$	$9.92 \times 10^{-1}$	1.32



## 5.5 Comparison with Balanced Truncation Methods

In this section, we compare the IRKA-based second-order model reduction techniques to the other methods discussed in Section 5.2 for several models. Since we are interested in comparing both first and second-order models,  $r$  refers to the dimension of the first-order reduced model and the dimension of the second-order model's corresponding first-order realization. In the reported data, *Balanced Truncation* denotes converting the given second-order system into a first-order system and then applying the balanced truncation method. *SOR-IRKA* refers to using balanced truncation for the intermediate step of SOR-IRKA.

### 5.5.1 Building Model

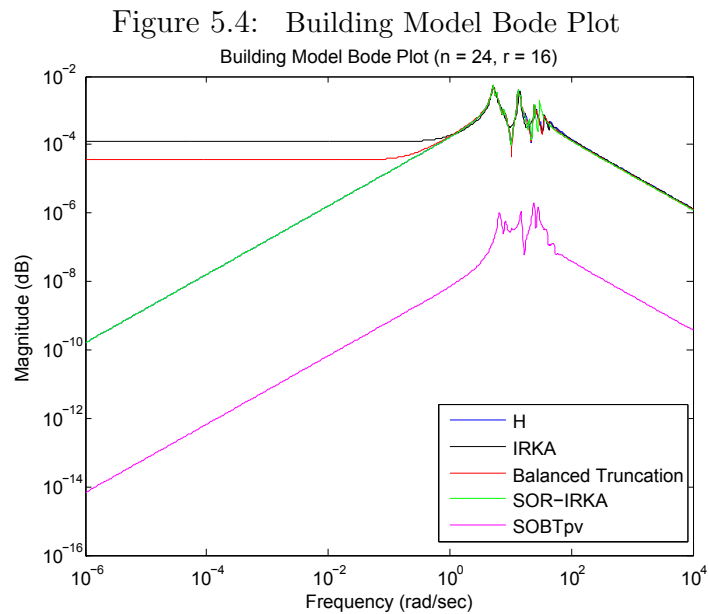
The Building Model, as introduced in Section 5.4, is an example of a nonsymmetric model. The full-order model of dimension  $n = 24$  was reduced to dimension  $r = 16$ . In Table 5.4, the  $\frac{\|H-H_r\|_{\mathcal{H}_2}}{\|H\|_{\mathcal{H}_2}}$  and  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$  errors are displayed, and the bode plots are given in Figure 5.4. Perhaps most noticeable from this data is the huge discrepancy between the full and reduced models associated with the SOBTpv method. First-order balanced truncation results in the smallest  $\mathcal{H}_\infty$  error, but the  $\mathcal{H}_2$  errors for balanced truncation and SOR-IRKA are of the same order. Also, for low frequencies, first-order balanced truncation fails to capture the behavior of the full-order model, leaving SOR-IRKA to be the only method resulting in a reduced model that reflects the behavior of the full-order model.

Table 5.4: Building Model Errors

Method	$\frac{\ H-H_r\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$	$\frac{\ H-H_r\ _{\mathcal{H}_\infty}}{\ H\ _{\mathcal{H}_\infty}}$
Balanced Truncation	$1.01 \times 10^{-1}$	$7.31 \times 10^{-2}$
SOBTpv	$1.92 \times 10^3$	$2.31 \times 10^3$
SOR-IRKA	$3.65 \times 10^{-1}$	$3.17 \times 10^{-1}$

### 5.5.2 Beam Model

Recall from Section 5.4 that the Beam Model is a model of dimension  $n = 500$  where  $\mathbf{M}$ ,  $\mathbf{G}$ , and  $\mathbf{K}$  are symmetric positive definite matrices. To preserve these system properties, we took  $\mathbf{W}_r = \mathbf{V}_r$  and reduced to an order  $r = 60$  model. As Table 5.5 demonstrates, the errors associated with SOBTpv are much larger than those obtained with any of the other



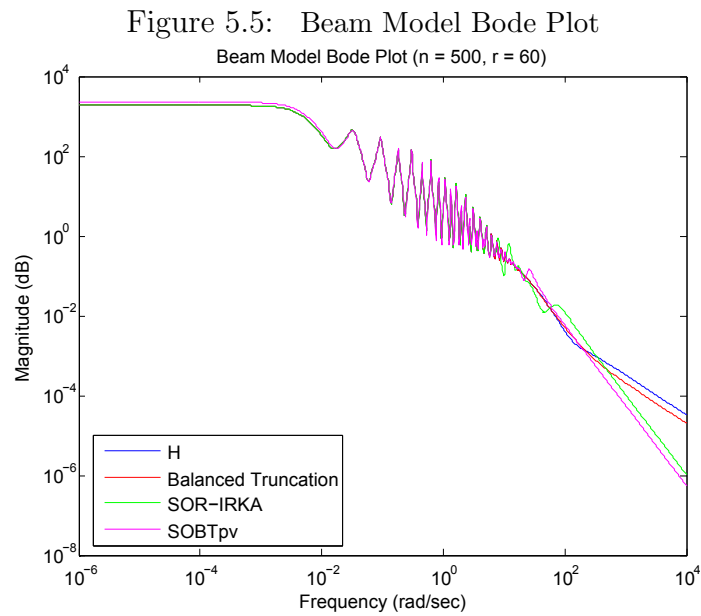
methods while the first-order balanced truncation method results in the smallest errors. Since SOR-IRKA preserves the second-order structure and produces reasonably small errors, we conclude that SOR-IRKA is an appropriate method for the Beam Model.

Table 5.5: Beam Model Errors

Method	$\frac{\ H-H_r\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$	$\frac{\ H-H_r\ _{\mathcal{H}_\infty}}{\ H\ _{\mathcal{H}_\infty}}$
Balanced Truncation	$1.28 \times 10^{-4}$	$1.16 \times 10^{-6}$
SOBTpv	$1.79 \times 10^{-1}$	$1.96 \times 10^{-1}$
SOR-IRKA	$5.46 \times 10^{-3}$	$3.08 \times 10^{-4}$

### 5.5.3 12a Model

We reduced the 12a Model to dimension  $r = 120$ . In Table 5.6, the model reduction errors are presented, and Figure 5.6 provides the comparison between the full and reduced model's bode plots. As illustrated by this data, SOR-IRKA is competitive with both the first-order

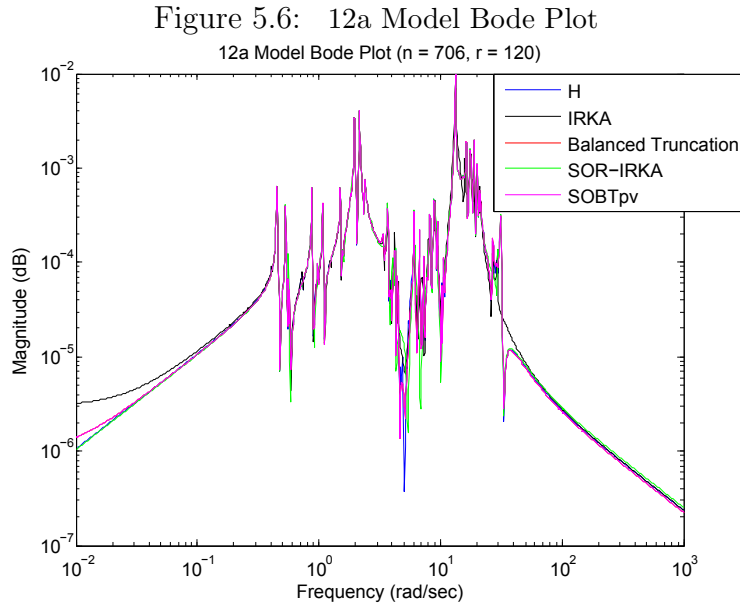


and second-order balanced truncation frameworks. It is important to note, however, that the balanced truncation errors came at the cost of solving two Lyapunov equations of dimension  $2n$ , which may be an intractable problem especially for large-scale systems. Furthermore, the reduced-order models obtained by first-order methods fail to reflect the structure of the full-order model, potentially dimensioning the practicality of the reduced-order model in a real-world setting. Especially for large-scale problems, where solving the Lyapunov equations may be infeasible, these results illustrate the advantages associated with SOR-IRKA.

Table 5.6: 12a Model Errors

Method	$\frac{\ H-H_r\ _{\mathcal{H}_2}}{\ H\ _{\mathcal{H}_2}}$	$\frac{\ H-H_r\ _{\mathcal{H}_\infty}}{\ H\ _{\mathcal{H}_\infty}}$
Balanced Truncation	$1.02 \times 10^{-2}$	$2.87 \times 10^{-3}$
SOBTpv	$1.47 \times 10^{-2}$	$7.01 \times 10^{-3}$
SOR-IRKA	$4.36 \times 10^{-2}$	$5.76 \times 10^{-2}$





#### 5.5.4 Butterfly Gyro Model

As an example of a large-scale dynamical system, we consider applying SOR-IRKA to the Butterfly Gyro Model. This model is provided by Dag Billger through the Oberwolfach Model Reduction Benchmark Collection and models a gyro chip. The model has the following form

$$\mathbf{H}(s) : \begin{cases} \mathbf{M}\ddot{\mathbf{x}}(t) + (\alpha\mathbf{M} + \beta\mathbf{K})\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{B}u(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t), \end{cases} \quad (5.5.1)$$

where  $\mathbf{M}, \mathbf{G}, \mathbf{K} \in \mathbb{R}^{17361 \times 17361}$ ,  $\mathbf{B} \in \mathbb{R}^{17361}$ , and  $\mathbf{C} \in \mathbb{R}^{12 \times 17361}$  and  $\mathbf{M}, \mathbf{G}, \mathbf{K}$  are symmetric positive definite. For our numerical simulations, we took  $\mathbf{G} = \alpha\mathbf{M} + \beta\mathbf{K}$  with  $\alpha = 0$  and  $\beta = 1 \times 10^{-6}$  as suggested by [54]. More details are available in [54]. *It is important to emphasize that the associated Lyapunov equation for this model is of order 34,722, implying that balanced truncation and SOBTpv are not feasible methods.* As a result, we computed

the second-order reduced model of dimension  $r = 30$  by applying SOR-IRKA using  $\mathcal{H}_{1,r}$ . To preserve the symmetry and positive definiteness of the original model, we only employed one-sided reduction with  $\mathbf{W}_r = \mathbf{V}_r$ . The results of SOR-IRKA applied to this model are quite impressive. Not only is the method computationally feasible, but rapid convergence is observed. In Table 5.7, the first, second and final shift iterates of IRKA show that by the second IRKA iteration the shifts have converged. Moreover, the reduced model obtained is of high fidelity with the relative  $\mathcal{H}_\infty$  error,  $\frac{\|H-H_r\|_{\mathcal{H}_\infty}}{\|H\|_{\mathcal{H}_\infty}}$ , equal to  $5.78 \times 10^{-9}$ . Furthermore, an almost perfect match between the full and reduced-order models is observed in Figure 5.7. Especially for this large-scale example, our results illustrate that the SOR-IRKA method provides a second-order model reduction technique that is computationally feasible and reliable.

Figure 5.7: Butterfly Gyro Model Bode Plot

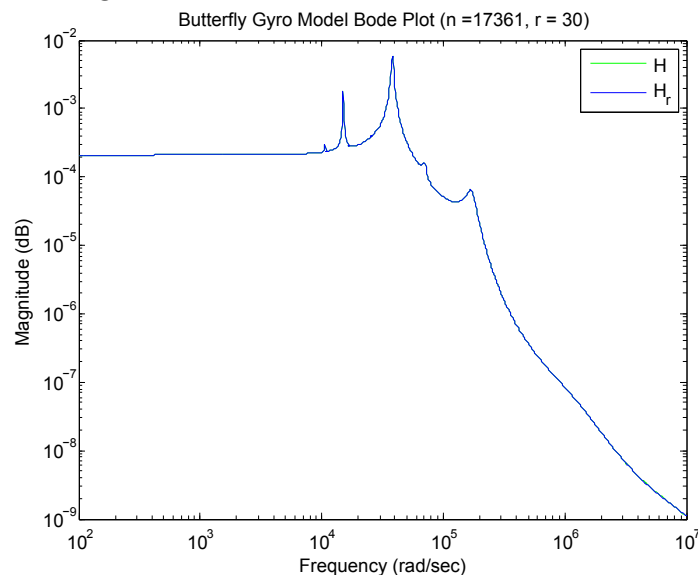


Table 5.7: Shift Iteration for the Butterfly Gyro Model

First IRKA Step	Second IRKA Step	Final IRKA Step
$1.00 \times 10^3$	$5.65 \times 10^1$	$5.65 \times 10^1$
$1.00 \times 10^3$	$5.65 \times 10^1$	$5.65 \times 10^1$
$1.64 \times 10^3$	$1.14 \times 10^2$	$1.14 \times 10^2$
$1.64 \times 10^3$	$1.14 \times 10^2$	$1.14 \times 10^2$
$2.68 \times 10^3$	$1.24 \times 10^2$	$1.24 \times 10^2$
$2.68 \times 10^3$	$1.24 \times 10^2$	$1.24 \times 10^2$
$4.39 \times 10^3$	$3.84 \times 10^2$	$3.84 \times 10^2$
$4.39 \times 10^3$	$3.84 \times 10^2$	$3.84 \times 10^2$
$7.20 \times 10^3$	$7.31 \times 10^2$	$7.31 \times 10^2$
$7.20 \times 10^3$	$7.31 \times 10^2$	$7.31 \times 10^2$
$1.18 \times 10^4$	$9.72 \times 10^2$	$9.72 \times 10^2$
$1.18 \times 10^4$	$9.72 \times 10^2$	$9.72 \times 10^2$
$1.93 \times 10^4$	$2.09 \times 10^3$	$2.09 \times 10^3$
$1.93 \times 10^4$	$2.09 \times 10^3$	$2.09 \times 10^3$
$3.16 \times 10^4$	$2.11 \times 10^3$	$2.11 \times 10^3$
$3.16 \times 10^4$	$2.11 \times 10^3$	$2.11 \times 10^3$
$5.18 \times 10^4$	$2.51 \times 10^3$	$2.51 \times 10^3$
$5.18 \times 10^4$	$2.51 \times 10^3$	$2.51 \times 10^3$
$8.48 \times 10^4$	$4.04 \times 10^3$	$4.04 \times 10^3$
$8.48 \times 10^4$	$4.04 \times 10^3$	$4.04 \times 10^3$
$1.39 \times 10^5$	$4.20 \times 10^3$	$4.20 \times 10^3$
$1.39 \times 10^5$	$4.20 \times 10^3$	$4.20 \times 10^3$
$2.28 \times 10^5$	$7.02 \times 10^3$	$7.02 \times 10^3$
$2.28 \times 10^5$	$7.02 \times 10^3$	$7.02 \times 10^3$
$3.73 \times 10^5$	$9.27 \times 10^3$	$9.27 \times 10^3$
$3.73 \times 10^5$	$9.27 \times 10^3$	$9.27 \times 10^3$
$6.11 \times 10^5$	$9.39 \times 10^3$	$9.39 \times 10^3$
$6.11 \times 10^5$	$9.39 \times 10^3$	$9.39 \times 10^3$
$1.00 \times 10^6$	$1.48 \times 10^4$	$1.48 \times 10^4$
$1.00 \times 10^6$	$1.48 \times 10^4$	$1.48 \times 10^4$

# Chapter 6

## Conclusion

In this dissertation, we studied and developed methods for effective interpolatory model reduction. One of the key bottlenecks of interpolatory methods is the sequence of linear systems required. Especially in the large-scale setting, inexact solves become necessary, and the convergence of these iterative methods often benefits greatly from preconditioning. In this dissertation, we studied the effect of the preconditioner in the model reduction setting and showed that several previously proven results for unpreconditioned inexact solves also hold in a similar manner for the preconditioned case. We also established an important distinction between the types of preconditioning techniques, namely a backward error result assuming only the Petrov-Galerkin framework exists in the case of split preconditioning but not for right and left preconditioning.

Due to the importance of preconditioning, we developed preconditioning techniques specific

to the IRKA iteration. Two updating methods, namely sparse approximate inverses (SAI) and a modification of the update proposed by Bellavia et al. in [11] were studied. The SAI update proved effective for some models, and even resulted in the same number of GMRES iterations as when an incomplete LU was computed. For the Bellavia et al. update, we extended several results to the preconditioned IRKA case. Despite these promising theoretical results, our data indicated that oftentimes the resulting update was close to singular, and therefore should not be used. To remedy these issues, we considered computing additional incomplete LU factorizations along with the update, which resulted in the updating methods being more competitive with computing a new preconditioner at each step.

Perhaps the most important contribution of this dissertation is our study of several theoretical and numerical issues associated with interpolatory model reduction of DAEs. While the main interpolation theorem holds for a DAE system, the model reduction error may potentially be unbounded. As a result, we present a theorem that delineates the conditions required for both interpolation and a bounded model reduction error for DAEs. Using this theorem, we use the IRKA framework to present an algorithm for the reduction of DAEs. Since this algorithm requires the explicit computation of the spectral projectors, we consider exploiting the properties of certain DAEs to avoid this computation. For index-1 and Hessenberg index-2 DAEs, we present theoretical and numerical results illustrating effective reduction of the DAE system without the projectors.

Finally, another main contribution of this dissertation pertains to the study of second-order systems. Many methods for reducing second-order systems rely on the first-order represen-

tation of the second-order system, and thereby computations involving an order  $2n$  model at some point in the reduction process. Moreover, methods that reduce directly the first-order representation of the second-order model may yield a model that can not be converted back to the second-order framework. As a result, the physical meaning of the problem is absent in the reduced-order model even though the model reduction error may be small. To avoid these issues, this dissertation proposes IRKA-based algorithms for second-order systems. Although the resulting model will not satisfy the  $\mathcal{H}_2$  optimal conditions for second-order systems, our numerical results indicate that the methods produce accurate reduced second-order models.

# Bibliography

- [1] K. Ahuja, E. de Sturler, R. Chang, and S. Gugercin, *Recycling BiCG for model reduction*, submitted to SIAM J. Sci. Comput., July 2010.
- [2] A.C. Antoulas, *Lectures on the Approximation of Linear Dynamical Systems*, Advances in Design and Control, SIAM, Philadelphia, 2004.
- [3] A.C. Antoulas, C.A. Beattie and S. Gugercin, *Interpolatory model reduction of large-scale dynamical systems. Efficient Modeling and Control of Large-Scale Systems*, J. Mohammadpour and K. Grigoriadis editors, Springer-Verlag, 2010.
- [4] Z. Bai, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*, Applied Numerical Mathematics, 43(1-2), pp. 9-44, 2002.
- [5] R. Barrett, M. Berry, T. F. Chan, J. Demmel, et al. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, 1994.
- [6] C. Beattie, *Structured perturbations in rational Krylov methods for model reduction*, Workshop on Structured Perturbations and Distance Problems in Matrix Computations,

Bedlewo, Poland, March 2007.

- [7] C.A. Beattie and S. Gugercin, *Inexact solves in Krylov-based model reduction*, in Proceedings of the 45th IEEE Conference on Decision and Control, 2006.
- [8] C.A. Beattie and S. Gugercin, *Interpolatory projection methods for structure-preserving model reduction*, Systems and Control Letters, 58(3), pp. 225-232, 2009.
- [9] C.A. Beattie and S. Gugercin, *Krylov-based model reduction of second-order systems with proportional damping*, Proceedings of the 44th IEEE Conference on Decision and Control, pp. 2278-2283, 2005.
- [10] S. Bellavia, V. De Simone, D. Di Serafino, and B. Morini, *Efficient preconditioner updates for shifted linear systems*, J. SIAM Sci. Comput. 33(4), pp. 1785-1809.
- [11] S. Bellavia, D. Bertaccini and B. Morini, *Nonsymmetric preconditioner updates in Newton-Krylov methods for nonlinear systems*, J. SIAM Sci. Comput. 33(5), pp. 2595-2619, 2011.
- [12] P. Benner, *Solving large-scale control problems*, IEEE Control Systems Magazine, 24(1), pp. 44-59, 2004.
- [13] P. Benner and J. Saak, *Efficient numerical solution of the LQR-problem for the heat equation*, Proc. Appl. Math. Mech., 4, pp. 648-649, 2004.
- [14] M. Benson, J. Krettmann, and M. Wright, *Parallel algorithms for the solution of certain large sparse linear systems*, Internat. J. Comput. Math., 16, pp. 245-260, 1984.



- [15] M. Benzi, and D. Bertaccini, *Approximate inverse preconditioning for shifted linear systems*, BIT, 43, pp. 231-244, 2003.
- [16] M. Benzi and M. Tuma, *A Sparse approximate inverse preconditioner for nonsymmetric linear systems*, J. SIAM Sci. Comput., 19(3), pp. 968-994, 1998.
- [17] M. Benzi and M. Tuma, *A comparative study of sparse approximate inverse preconditioners*, Applied Numerical Mathematics, 30, pp. 305-340, 1999.
- [18] D. Bertaccini, *Efficient preconditioning for sequences of parametric complex symmetric linear systems*, Electron. Trans. Numer. Anal, 18, pp. 49-64, 2004.
- [19] D. Bertaccini and F. Sgallari, *Updating preconditioners for nonlinear deblurring and denoising image restoration*, Applied Numerical Mathematics, 60, pp. 994-1006, 2010.
- [20] P. Birken, J. Duntjer Tebbens, A. Meister, and M. Tuma, *Preconditioner updates applied to CFD model problems*, App. Num. Math., 58, pp. 1628-1641, 2008.
- [21] T. Bonin, H. Fassbender, A. Soppa, M. Zaeh, *A global Arnoldi method for the model reduction of second-order structural dynamical systems*, submitted.
- [22] K.E. Brenan, S.L.V. Campbell, L.R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, SIAM Classics in Applied Mathematics, SIAM, 1996.
- [23] A.E. Bryson and A. Carrier, *Second-order algorithm for optimal model order reduction*, J. Guidance Contr. Dynam., pp. 887-892, 1990.

- [24] R. Byers, T. Geerts, and V. Mehrmann, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35(2), pp. 462-479, 1997.
- [25] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, D. Wilczek, *H<sub>2</sub>-optimal model reduction for large scale discrete dynamical MIMO systems*, J. Comput. and App. Math., 2009.
- [26] C. Calgario, J.P. Chehab, Y. Saad, *Incremental incomplete LU factorizations with applications to time-dependent PDEs*, 2008.
- [27] V. Chahlaoui, K.A. Gallivan, A. Vandendorpe, P. Van Dooren, *Model reduction of second order system*, In: P. Benner, V.Mehrmann, D. Sorensen (eds.) Dimension Reduction of Large-Scale Systems, Lecture Notes in Computational Science and Engineering, 45, pp. 149-172. Springer-Verlag, Berlin/Heidelberg, Germany, 2005.
- [28] E. Chow and Y. Saad, *Approximate inverse preconditioners via Sparse-Sparse Iterations*, SIAM J. Sci. Comput, 19, pp. 995-1023, 1998.
- [29] J.V. Clark, N. Zhou, D. Bindel, L. Schenato, W. Wu, J. Demmel, K.S.J. Pister, *3D MEMS simulation using modified nodal analysis*, In: Proceedings of Microscale Systems: Mechanics and Measurements Symposium, pp. 6875, 2000.
- [30] J.D.F. Cosgrove, J.C. Diaz, and A. Griewank, *Approximate inverse preconditionings for sparse linear systems*, Internat. J. Comput. Math, 44, pp. 91-110, 1992.
- [31] R.R. Craig Jr., *Structral dynamics: an introduction to computer methods*, John Wiley and Sons, 1981.

- [32] E. de Sturler and H. A. van der Vorst, *Reducing the effect of global communication in GMRES( $m$ ) and CG on parallel distributed memory computers*, Applied Numerical Mathematics (IMACS), 18, pp. 441-459, 1995.
- [33] C. De Villemagne and R. Skelton, *Model reduction using a projection formulation*, Internat. J. Control, 40, pp. 2141-2169, 1987.
- [34] S. Demko, W.F. Moss, and P.W. Smith, *Decay rates for inverse of band matrices*, Math. Comput., 43, pp. 491-499, 1984.
- [35] J.W. Demmel and B. Kagstrom, *Stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88(89), pp. 137-186, 1987.
- [36] J.W. Demmel and B. Kagstrom, *The generalized Schur decomposition of an arbitrary pencil  $A - \lambda B$  Parts I and II*, ACM Trans. Math. Software, 19, pp. 160-201, 1993.
- [37] J. Duntjer Tebbens and M. Tuma, *Efficient preconditioning of sequences of nonsymmetric linear systems*, SIAM J. Sci. Comput., 29, pp. 1918-1941, 2007.
- [38] J. Duntjer Tebbens and M. Tuma, *Preconditioner updates for solving sequences of linear systems in matrix-free environment*, Numer. Linear Algebra Appl., 17, pp. 997-1019, 2010.
- [39] P. Feldman and R.W. Freund, *Efficient linear circuit analysis by Padé approximation via a Lanczos method*, IEEE Trans. Computer-Aided Design, 14, pp. 639-649, 1995.

- [40] N.I.M. Gould and J.A. Scott, *On approximate-inverse preconditioners*, Technical Report RAL-95-026, Rutherford Appleton Laboratory, Chilton, UK, 1998.
- [41] Anne Greenbaum, *Iterative methods for solving linear systems*, SIAM, Philadelphia, 1997.
- [42] M. Grote and T. Huckle, *Parallel preconditioning with sparse approximate inverses*, SIAM J. Sci. Comput., 18, pp. 838-853, 1997.
- [43] M. Grote and H. Simon, *Parallel preconditioning and approximate inverses on the Connection Machine*, in Proc. Sixth SIAM Conference on Parallel Processing for Scientific Computing, R.F. Sincover, D.E. Keyes, M.R. Leuze, L.R. Petzold, and D.A. Reed, eds., SIAM, Philadelphia, PA, pp. 519-523, 1993.
- [44] E.J. Grimme, *Krylov projection methods for model reduction*, Ph.D. thesis, University of Illinois, Urbana-Champaign, Urbana, IL, 1997.
- [45] S. Gugercin, *Projection methods for model reduction of large-scale dynamical systems*, Ph.D. thesis, Rice University, Houston, TX, 2002.
- [46] S. Gugercin and A.C. Antoulas, *An  $\mathcal{H}_2$  error expression for the Lanczos procedure*, in Proceedings of the 42nd IEEE Conference on Decision and Control, pp. 1869-1872, 2003.
- [47] S. Gugercin, A.C. Antoulas and C.A. Beattie,  *$\mathcal{H}_2$  model reduction for large-scale linear dynamical systems*, SIAM J. Matrix Anal. and App., 30(2), pp. 609-638, 2008.

- [48] S. Gugercin, R. V. Polyuga, C.A. Beattie and A. van der Schaft, *Structure-preserving tangential-interpolation based model reduction of port-Hamiltonian systems*, accepted to appear in *Automatica*, 2011. Available as arXiv:1101.3485v2.
- [49] M. Heinkenschloss, D.C. Sorensen, and K. Sun, *Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations*, *SIAM J. Sci. Comput.*, 30(2), pp. 1038-1063, 2008.
- [50] N. Higham, *Perturbation theory and backward error for  $AX - XB = C^*$* , *BIT*, 22, pp. 124-136, 1993.
- [51] C. W. Ho, A. Ruehli and P. Brennan, *The modified nodal approach to network analysis*, *IEEE*, 22(6), pp. 504-509, 1975.
- [52] R. Holland, A. Wathen, and G. Shaw, *Sparse approximate inverses and target matrices*, *SIAM J. Sci. Comput.* 26(3), pp. 1000-1011.
- [53] D.C. Hyland and D.S. Bernstein, *The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton and Moore*, *IEE. Trans. Automa. Contr.*, 30(12), pp. 1201-1211, 1985.
- [54] J. Lienemann, D. Billger, E. B. Rudnyi, A. Greiner, and J. G. Korvink, *MEMS compact modeling meets model order reduction: examples of the application of Arnoldi methods to microsystem devices*, the Technical Proceedings of the 2004 Nanotechnology Conference and Trade Show, Nanotech 2004, March 7-11, 2004, Boston, Massachusetts, USA.

- [55] B. Kagstrom, *RGSVD—an algorithm for computing the Kronecker structure and reducing subspaces of singular  $A - \lambda B$  pencils*, SIAM J.Sci. Statist. Comput. 7, pp. 185-211, 1986.
- [56] G. Kalogeropoulos, M. Mitrouli, A. Pentelous, and D. Triantafyllou, *The Weierstrass Canonical Form of a regular matrix pencil: numerical issues and computational techniques*, Springer-Verlag Berlin Heidelberg, pp. 322-329, 2009.
- [57] W. A. Kohler and L. W. Johnson, *Elementary Differential Equations with Boundary Value Problems*, Addison-Wesley; 1st edition, 2004.
- [58] L. Y. Kolotilina, A.A. Nikishin, and A. Y. Yeremin, *Factorized sparse approximate inverse (FSAI) preconditionings for solving 3D FE systems on massively parallel computers II: Iterative construction of FSAI preconditioners*, in Proc. IMACS International Symposium on Iterative Methods in Linear Algebra, R. Beauwens and P. de Groen, eds, North-Holland, Amsterdam, pp. 311-312, 1992.
- [59] L. Y. Kolotilina and A. Y. Yeremin, *Factorized sparse approximate inverse preconditioning I: Theory*, SIAM J. Matrix Anal. Appl., 14, pp. 45-58, 1993.
- [60] L. Y. Kolotilina and A. Y. Yeremin, *Factorized sparse approximate inverse preconditioning II: Solution of 3D FE systems on massively parallel computers*, Internat. J. High Speed Comput., 7, pp. 191-215, 1997.
- [61] J. Korvink, E. Rudnyi, *Oberwolfach benchmark collection*, In: P. Benner, V. Mehrmann, D.C. Sorensen (eds.) Dimension Reduction of Large-Scale Systems, Lec-

- ture Notes in Computational Science and Engineering, 45, pp. 311-315, Springer-Verlag, Berlin/Heidelberg, Germany, 2005.
- [62] L. Meier and D.G. Luenberger, *Approximation of linear constant systems*, IEE. Trans. Automat. Contr., 12, pp. 585-588, 1967.
- [63] G. Meurant, *On the incomplete Cholesky decomposition of a class of perturbed matrices*, SIAM J. Sci. Comput., 23, pp. 419-429, 2001.
- [64] D.G. Meyer, and S. Srinivasan, *Balancing and model reduction for second-order form linear systems*, IEEE Trans. Automat. Control, 41, pp. 1632-1644, 1996.
- [65] M. Parks, E. de Sturler, G. Mackey, D. D. Johnson, and S. Maiti, *Recycling Krylov Subspaces for Sequences of Linear Systems*, SIAM J. Sci. Comput., 28(5), pp. 1651-1674, 2006.
- [66] P. M. Pinsky and N. N. Abboud. *Finite element solution of the transient exterior structural acoustics problem based on the use of radially asymptotic boundary conditions* Computer Methods in Applied Mechanics and Engineering, 85, pp. 311-348, 1991.
- [67] A. Preumont, *Vibration control of active structures: an introduction*, Springer, 2002.
- [68] T. Reis and T. Stykel, *Balanced truncation model reduction of second-order systems*, Technical Report, 376, 2007.
- [69] Yousef Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, 2004.

- [70] J.T. Spanos, M.H. Millman, and D.L. Mingori, *A new algorithm for  $L_2$  optimal model reduction*, *Automatics*, pp. 897-909, 1992.
- [71] G.W. Stewart, J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.
- [72] T. Stykel, *Low-rank iterative methods for projected generalized Lyapunov equations*, *Electronic Transactions on Numerical Analysis*, 30, pp. 187-202, 2008.
- [73] T. Stykel, T. Reis, *Passivity-preserving balanced truncation model reduction of circuit equations*, *Scientific Computing in Electrical Engineering, SCEE 2008*, J. Roos and L.R.J. Costa, eds., *Mathematics in Industry*, 14, Springer-Verlag, Berlin, Heidelberg, pp. 483-490, 2010.
- [74] Lloyd N. Trefethen and David Bau, III, *Numerical Linear Algebra*, SIAM, 1997.
- [75] A. van der Sluis, *Condition numbers and equilibration of matrices*, *Numerische Mathematik*, 14(1), 1423, 1969.
- [76] P. Van Dooren, K. Gallivan, P. Absil,  *$H_2$ -optimal model reduction of MIMO systems*, *Applied Mathematics Letters*, 21(12), pp. 1267-1273, 2008.
- [77] S. Van Huffel, J. Vandewalle, *An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values*, *J. Comput. and App. Math.*, 19, pp. 313-330, 1987.



- [78] S. Wang and E. de Sturler, *Multilevel Sparse Approximate Inverse Preconditioners for Adaptive Mesh Refinement*, Linear Algebra and its Applications, 2009.
- [79] W. Weaver, P. Johnston, *Structural dynamics by finite elements*, Prentice Hall, Upper Saddle River, 1987.
- [80] D.A. Wilson, *Optimum solution of model reduction problem*, in Proc. Inst. Elec. Eng., pp. 1161-1165, 1970.
- [81] W-Y. Yan and J. Lam, *An approximate approach to  $\mathcal{H}_2$  optimal model reduction*, IEEE Trans. Automat. Contr., AC-44, pp. 1341-1358, 1999.
- [82] A. Yousouff and R.E. Skelton, *Covariance equivalent realizations with applications to model reduction of large-scale systems*, in Control and Dynamic Systems, 22, C.T. Leonides, ed., Academic Press, New York, pp. 273-348, 1985.
- [83] A. Yousouff, D.A. Wagie, and R.E. Skelton, *Linear system approximation via covariance equivalent realizations*, J. Math. Anal. Appl., 196, pp. 91-115, 1985.