Parametric Dynamical Systems: Transient Analysis and Data Driven Modeling

Alexander R. Grimm

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 Güğercin, Chair
Christopher A. Beattie
Matthias Chung
Zlatko Drmač
Mark P. Embree
Eric de Sturler

May 9, 2018
Blacksburg, Virginia

Keywords: Model Reduction, Interpolation, Approximation, Nonlinear Eigenvalue Problems, Least-Squares, Hardy Spaces, Discretization, Rational Approximation.

Parametric Dynamical Systems: Transient Behavior and Data Driven Modeling

Alexander R. Grimm

(ABSTRACT)

Dynamical systems are a commonly used and studied tool for simulation, optimization and design. In many applications such as inverse problem, optimal control, shape optimization and uncertainty quantification, those systems typically depend on a parameter. The need for high fidelity in the modeling stage leads to large-scale parametric dynamical systems. Since these models need to be simulated for a variety of parameter values, the computational burden they incur becomes increasingly difficult. To address these issues, parametric reduced models have encountered increased popularity in recent years.

We are interested in constructing parametric reduced models that represent the full-order system accurately over a range of parameters. First, we define a global joint error measure in the frequency and parameter domain to assess the accuracy of the reduced model. Then, by assuming a rational form for the reduced model with poles both in the frequency and parameter domain, we derive necessary conditions for an optimal parametric reduced model in this joint error measure. Similar to the nonparametric case, Hermite interpolation conditions at the reflected images of the poles characterize the optimal parametric approximant. This result extends the well-known interpolatory $\mathcal{H}_2$ optimality conditions by Meier and Luenberger to the parametric case. We also develop a numerical algorithm to construct locally optimal reduced models. The theory and algorithm are data-driven, in the sense that only function evaluations of the parametric transfer function are required, not access to the internal dynamics of the full model.

While this first framework operates on the continuous function level, assuming repeated transfer function evaluations are available, in some cases merely frequency samples might be given without an option to re-evaluate the transfer function at desired points; in other words, the function samples in parameter and frequency are fixed. In this case, we construct a parametric reduced model that minimizes a discretized least-squares error in the finite set of measurements. Towards this goal, we extend Vector Fitting (VF) to the parametric case, solving a global least-squares problem in both frequency and parameter. The output of this approach might lead to a moderate size reduced model. In this case, we perform a post-processing step to reduce the output of the parametric VF approach using $\mathcal{H}_2$ optimal model

reduction for a special parametrization. The final model inherits the parametric dependence of the intermediate model, but is of smaller order.

A special case of a parameter in a dynamical system is a delay in the model equation, e.g., arising from a feedback loop, reaction time, delayed response and various other physical phenomena. Modeling such a delay comes with several challenges for the mathematical formulation, analysis, and solution. We address the issue of transient behavior for scalar delay equations. Besides the choice of an appropriate measure, we analyze the impact of the coefficients of the delay equation on the finite time growth, which can be arbitrary large purely by the influence of the delay.

This work received support from NSF grant DMS-1217156, NSF grant DMS-1720257 and the Department of Mathematics at Virginia Tech.

Parametric Dynamical Systems: Transient Behavior and Data Driven Modeling

Alexander R. Grimm

(GENERAL AUDIENCE ABSTRACT)

Mathematical models play an increasingly important role in the sciences for experimental design, optimization and control. These high fidelity models are often computationally expensive and may require large resources, especially for repeated evaluation. Parametric model reduction offers a remedy by constructing models that are accurate over a range of parameters, and yet are much cheaper to evaluate. An appropriate choice of quality measure and form of the reduced model enable us to characterize these high quality reduced models. Our first contribution is a characterization of optimal parametric reduced models and an efficient implementation to construct them.

While this first framework assumes we have access to repeated evaluations of the full model, in some cases merely measurement data might be available. In this case, we construct a parametric model that fits the measurements in a least squares sense. The output of this approach might lead to a moderate size reduced model, which we address with a post-processing step that reduces the model size while maintaining important properties.

A special case of a parameter is a delay in the model equation, e.g., arising from a feedback loop, reaction time, delayed response and various other physical phenomena. While asymptotically stable solutions eventually vanish, they might grow large before asymptotic behavior takes over; this leads to the notion of transient behavior, which is our main focus for a simple class of delay equations. Besides the choice of an appropriate measure, we analyze the impact of the structure of the delay equation on the transient growth, which can be arbitrary large purely by the influence of the delay.

To my parents who always supported me.
To my three best friends: Alex, who always stood by my side, Sara, my longest friend, our relationship goes back to childhood memories, and Spencer, whom I've shared many nights of deep discussions with.
And to Tracie, who touched my heart like no other.

For [a] model there is no need to ask the question "Is the model true?".
If "truth" is to be the "whole truth" the answer must be "No".
The only question of interest is "Is the model illuminating and useful?"

— George E. P. Box, 1978

"Well, I must endure the presence of a few caterpillars
if I wish to become acquainted with the butterflies. "

— Antoine de Saint-Exupry, The Little Prince

Acknowledgements

First, I would like to thank my sponsors: NSF and the Department of Mathematics at Virginia Tech for their support.

To my advisor, Serkan Gugercin, a big thank you for your incredible patience and never ending task to keep me focused. You not only helped my research immensely and guided me through what seemed like a jungle of mathematics, but gave me an incredible new and open look on the field of Model Reduction.

Christopher Beattie always had an open door and was open for discussions on what turned out to be tangential topics at times. On several occasions, you shifted my point of view just the right amount to put things into perspective.

Mark Embree helped me especially with the chapter on delay equations. The shortest summary of research with you would be "Keep it simple", for our understanding of the simplest possible (but interesting) case opens the door to understanding everything. A lesson well taken!

I wish to thank all members of my thesis committee for taking the time out of their schedule to provide me with useful feedback, comments, corrections and help on my dissertation.

Special thanks goes to Benjamin Unger for his support in my writing as well as Philip Schulze for his comments. I also like to thank my fellow graduate students for fruitful discussions and comments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Dynamical systems are a common tool in modeling physical phenomena. Particularly linear systems are used in many applications. In this chapter, we introduce the main concepts used throughout this dissertation, followed by a summary of contributions and an outline of the thesis.

1.1 Linear System Theory

A dynamical system is commonly modeled by an internal state $\mathbf{x}(t)$, input/force $u(t)$ and output/observation $y(t)$. Throughout this thesis, we assume systems are discrete in space,

i.e., $\mathbf{x}(t) \in \mathbb{R}^n$, $t \geq 0$. A linear, time invariant dynamical system is represented by

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t), \quad \text{and} \\ y(t) &= \mathbf{c}^\top \mathbf{x}(t), \end{aligned} \tag{1.1.1}$$

where $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$ are the system matrices. We assume $\mathbf{x}(0) = \mathbf{0}$ and $\mathbf{E}$ to be nonsingular. Using the Laplace transformation, the input-output map of (1.1.1) in the frequency domain is given by

$$\mathbf{Y}(s) = \mathcal{H}(s)\mathbf{U}(s), \tag{1.1.2}$$

where the *transfer function* $\mathcal{H}(s)$ of a stable system is defined by

$$\mathcal{H}(s) = \mathbf{c}^\top (s\mathbf{E} - \mathbf{A})^{-1} \mathbf{b}, \qquad s \notin \sigma(\mathbf{A}, \mathbf{E}), \tag{1.1.3}$$

where $\sigma(\mathbf{A}, \mathbf{E})$ denotes the spectrum of the matrix pencil $s\mathbf{E} - \mathbf{A}$, also denoted by $(\mathbf{A}, \mathbf{E})$. Using high fidelity models, for example, arising from high accuracy finite element discretizations, yield a large state-space dimension n, making the time simulation of (1.1.1) computationally challenging. One remedy is model reduction.

1.2 Projection-based Model Reduction

Model reduction has become a common and effective tool in many branches of applied sciences to construct high fidelity models that are computationally tractable, i.e., to find a surrogate model for (1.1.1) with a smaller state-space dimension $r \ll n$; see [3, 19]. More

precisely, we want to find

$$\left.\begin{aligned} \widehat{\mathbf{E}}\dot{\widehat{\mathbf{x}}}(t) &= \widehat{\mathbf{A}}\widehat{\mathbf{x}}(t) + \widehat{\mathbf{b}}u(t), \\ \widehat{y}(t) &= \widehat{\mathbf{c}}^\top \widehat{\mathbf{x}}(t), \end{aligned}\right\} \quad \text{with} \quad \widehat{\mathcal{H}}(s) = \widehat{\mathbf{c}}^\top \left(s\widehat{\mathbf{E}} - \widehat{\mathbf{A}}\right)^{-1} \widehat{\mathbf{b}}, \tag{1.2.1}$$

where $\widehat{\mathbf{x}}(t) \in \mathbb{R}^r$, $\widehat{\mathbf{E}}, \widehat{\mathbf{A}} \in \mathbb{R}^{r\times r}$, $\widehat{\mathbf{b}}, \widehat{\mathbf{c}} \in \mathbb{R}^r$ ($r \ll n$). Observe that (1.2.1) has the same structure as (1.1.1). For a good approximation, we require $\widehat{y} \approx y$ which, in turn, implies $\widehat{\mathcal{H}} \approx \mathcal{H}$.

Projection based model reduction, in particular, assumes the state-space representation of (1.1.1) is available, i.e., we have access to the system matrices $(\mathbf{E}, \mathbf{A}, \mathbf{b}, \mathbf{c})$; see, for example, [3, 106]. With the model reduction basis matrices $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n\times r}$, the reduced system matrices (1.2.1) can be found by

$$\widehat{\mathbf{E}} = \mathbf{W}^\top \mathbf{E}\mathbf{V}, \qquad \widehat{\mathbf{A}} = \mathbf{W}^\top \mathbf{A}\mathbf{V}, \qquad \widehat{\mathbf{b}} = \mathbf{W}^\top \mathbf{b}, \quad \text{and} \quad \widehat{\mathbf{c}} = \mathbf{V}^\top \mathbf{c}. \tag{1.2.2}$$

In the literature, a number of ways to find $\mathbf{V}$ and $\mathbf{W}$ in (1.2.2) have been investigated. If time samples of the state vector $\mathbf{x}(t)$ are available, *proper orthogonal decomposition* (POD) [21, 71, 83, 87] is a widely used tool. POD is based on a snapshot matrix $\mathbf{X} := [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_M)] \in \mathbb{R}^{n\times M}$. A (truncated) singular value decomposition of $\mathbf{X}$ leads to $\mathbf{V}$ and $\mathbf{W} = \mathbf{V}$, chosen as the r leading left singular vectors of $\mathbf{X}$.

From a system theoretic perspective, we can categorize the states $\mathbf{x}(t) \in \mathbb{C}^n$ by their reachability and observability, using the (infinite) reachability and observability *Gramians*. These yield a balancing state-space transformation $\mathbf{T}$, transforming the matrices $\widetilde{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1}$, $\widetilde{\mathbf{E}} = \mathbf{T}\mathbf{E}\mathbf{T}^{-1}$, $\widetilde{\mathbf{b}} = \mathbf{T}\mathbf{b}$ and $\widetilde{\mathbf{c}}^\top = \mathbf{c}^\top\mathbf{T}^{-1}$, so that the state vector is ordered according to

reachability and observability. In *balanced truncation* [94, 95] unreachable / unobservable states are truncated to reduce the state-space dimension while capturing the main information of the system in the input-output map.

Rational interpolation methods construct an approximation $\widehat{\mathcal{H}}(s)$ by interpolating the original model $\mathcal{H}(s)$ at certain points $\sigma_1, \ldots, \sigma_r$. Interpolation conditions to choose those points optimally are implemented in the Iterative Rational Krylov Algorithm (IRKA) [59] as a fixed point iteration. This approach is presented in more detail in Section 2.4.

1.3 Data Driven Model Reduction

The state-space representation (1.1.1) may not always be available, but only transfer function evaluations $\mathcal{H}(s)$ as in (1.1.3) may be accessible. We then speak of *data driven modeling* or *data driven model reduction* or, at times of a *black-box* approach. In this setting, the optimal rational interpolation from [59] can be adapted to find an optimal approximant by using only (repeated) transfer function evaluations (TF-IRKA) instead of system matrices [14].

Assume further that our access to $\mathcal{H}(s)$ is restricted to a finite set of measurements of $\mathcal{H}(s)$ at sampling points $\xi_1, \ldots \xi_m$ in s. Given such measurements, $\{\xi_i, \mathcal{H}(\xi_i)\}_{i=1}^m \subset \mathbb{C} \times \mathbb{C}$, we aim to construct a model $\widehat{\mathcal{H}}(s)$ that approximates the given measurements in an appropriate sense.

Interpolation and least-squares approximation are two widely used approaches. First, for

interpolation, $\widehat{\mathcal{H}}(s)$ can be constructed by matching the given measurements $\{\xi_i, \mathcal{H}(\xi_i)\}_{i=1}^m$ exactly:

$$\widehat{\mathcal{H}}(\xi_i) = \mathcal{H}(\xi_i), \qquad i = 1, \ldots, m. \tag{1.3.1}$$

This leads to the Loewner framework as introduced in [74, 90] and presented in Section 2.5.1.

Next, for least-squares approximation, we construct $\widehat{\mathcal{H}}(s)$ to solve the least squares problem

$$\sum_{i=1}^{m} \left| \widehat{\mathcal{H}}(\xi_i) - \mathcal{H}(\xi_i) \right|^2 \to \min, \tag{1.3.2}$$

which leads to the Vector Fitting algorithm introduced by [107] and presented in Section 2.5.2.

1.4 Model Reduction of Parametric Dynamical Systems

In applications, dynamical systems often depend on a parameter. Examples include vibration models with varying damping or convection-diffusion models with varying diffusion coefficients. To reflect the parametric dependency, consider

$$\begin{aligned} \mathbf{E}(p)\dot{\mathbf{x}}(t;p) &= \mathbf{A}(p)\mathbf{x}(t;p) + \mathbf{b}(p)u(t), \\ y(t;p) &= \mathbf{c}^\top(p)\mathbf{x}(t;p), \end{aligned} \tag{1.4.1}$$

with transfer function

$$\mathcal{H}(s,p) = \mathbf{c}^\top(p)\left(s\mathbf{E}(p) - \mathbf{A}(p)\right)^{-1}\mathbf{b}(p). \tag{1.4.2}$$

For parametric systems, the problem now has two variables: time/frequency and a parameter, thus increasing the computational burden even more if function evaluations over several parameters are required, for example in optimization [18, 20]

One might ask of a useful reduced model to carry the parametric dependence of (1.4.1), with reduced system matrices:

$$\left.\begin{aligned}\widehat{\mathbf{E}}(p)\dot{\widehat{\mathbf{x}}}(t;p) &= \widehat{\mathbf{A}}(p)\widehat{\mathbf{x}}(t;p) + \widehat{\mathbf{b}}(p)u(t),\\ \widehat{y}(t;p) &= \widehat{\mathbf{c}}^{\top}(p)\widehat{\mathbf{x}}(t;p),\end{aligned}\right\} \quad \text{with} \quad \widehat{\mathcal{H}}(s,p) = \widehat{\mathbf{c}}^{\top}(p)\left(s\widehat{\mathbf{E}}(p) - \widehat{\mathbf{A}}(p)\right)^{-1}\widehat{\mathbf{b}}(p). \tag{1.4.3}$$

This clearly increases the complexity of the task to construct a parametric reduced model that approximates the original function. In addition, we aim for reduced models that are valid not only for a fixed parameter value, but for p ranging over a parameter domain.

We review common approaches to construct a parametric reduced model as in (1.4.3). For more details, we refer to an overview of projection based methods by Benner, Gugercin and Wilcox in [18].

In Chapter 3 and Chapter 4, we extend the non-parametric approaches from Section 1.3 to the parametric case.

1.5 Delay as a Parameter and Transient Analysis

As a particular example of a parametric dynamical system, delay equations are of great interest in both theory and applications. A simple example is the scalar delay equation

$$\dot{x}(t) = ax(t) + bx(t-\tau), \quad t \geq 0 \quad \text{with} \quad x(t) = \phi(t), \qquad t \in [-\tau, 0]. \tag{1.5.1}$$

Here the derivative $\dot{x}(t)$ depends on the state of the system $x(t-\tau)$ as well as $x(t)$.

Applications range from biology, chemistry, finance and traffic pattern analysis to control implementations and regulators. We remark that those applications usually come in the form of delay dynamical systems with inputs and observations, similar to (1.1.1). For our analysis of transient behavior, we focus on the evolution of the state $x(t)$ over time without inputs and outputs.

Asymptotically stable solutions, that is $x(t) \to 0$ for $t \to \infty$ may grow large for finite t before following the asymptotic behavior. Transient analysis aims to characterize the magnitude of $\sup_{t\geq 0} |\mathbf{x}(t)|$ for asymptotically stable solutions. For ODEs, transient growth can only appear in dimensions $n \geq 2$ [114]. For delay differential equations, however, even the scalar case, can exhibit arbitrarily large transient growth [110].

Insight into the stability and asymptotic behavior of a delay system can be gained through spectral analysis, which leads to a nonlinear eigenvalue problem (NLEVP) [64]. The NLEVP is used to analyze the influence of perturbations of the coefficients and the delay parameters on stability. Using the NLEVP, however, gives limited insight into transient growth of a particular solution.

We analyze a rather simple case of a delay equation: the scalar, linear, constant coefficient case. Here, other influences from higher dimensional state-space or nonlinear behavior are deliberately excluded, to concentrate on the influence of the delay.

Even in this simple case, we show in Chapter 5 that solutions to (asymptotically) stable systems can exhibit enormous growth over finite time, for a specific set of coefficients for the terms in the delay equation. Furthermore, we give methods to find such configurations and investigate the corresponding transient behavior.

1.6 Motivating Examples

To illustrate our approaches, we use several typical examples from the NICONET benchmark collection [27, 81]. We introduce them in this section to familiarize the reader with the types of problems we address in this thesis.

1.6.1 Vibrating Cantilever Beam

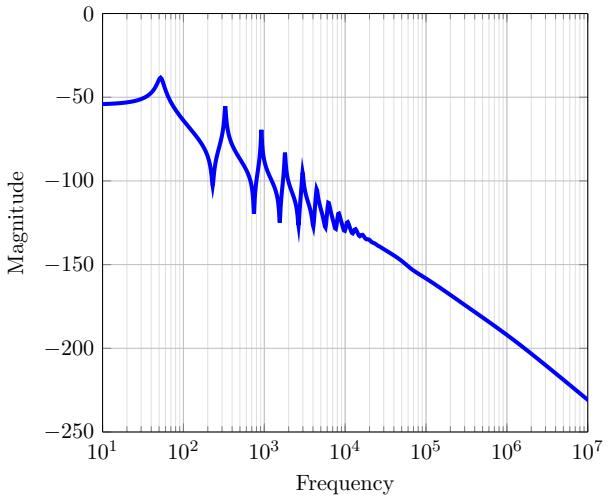
Consider a vibrating cantilever beam [96] with proportional damping, modeled by the second order dynamical system

$$\mathbf{M}\ddot{\mathbf{x}}(t) + (\mathbf{M} + p\mathbf{K})\,\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{b}F(t), \qquad y(t) = \mathbf{c}^\top \mathbf{x}(t), \tag{1.6.1}$$

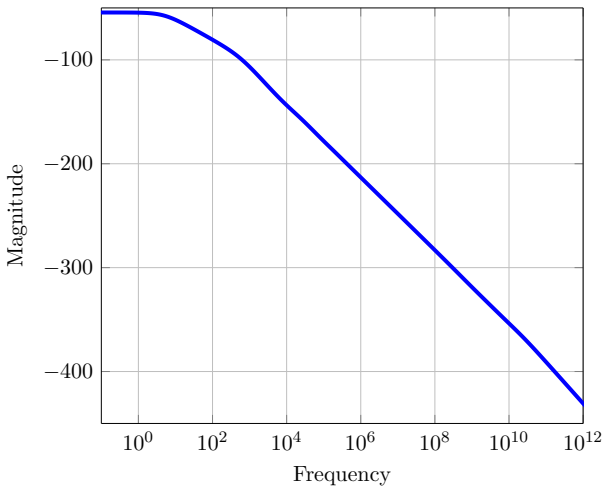
where F represents a forcing function, and the parameter p represents a damping coefficient, normalized, so that $p \in [0, 1]$. We can rewrite (1.6.1) into a system of first order equations

$$\underbrace{\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} \end{bmatrix}}_{=:\mathbf{E}} \begin{bmatrix} \dot{\mathbf{x}}(t) \\ \ddot{\mathbf{x}}(t) \end{bmatrix} = \left(\underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{K} & -\mathbf{M} \end{bmatrix}}_{=:\mathbf{A}_0} + p \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{K} \end{bmatrix}}_{=:\mathbf{A}_1} \right) \begin{bmatrix} \mathbf{x}(t) \\ \dot{\mathbf{x}}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix} F(t)$$
$$y(t) = \begin{bmatrix} \mathbf{c}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}(t) \\ \dot{\mathbf{x}}(t) \end{bmatrix} \tag{1.6.2}$$

We illustrate the influence of p on the system by showing Bode plots in Figure 1.1.



(a) $p = 0.0001$



(b) $p = 0.2$

Figure 1.1: Bode plot for the vibrating Cantilever beam model (1.6.2) for different choices of the damping coefficient p.

The Bode plots in Figure 1.1 illustrate how the dynamics of the model (1.6.1) change with different damping coefficients.

1.6.2 Convection-Diffusion Equation

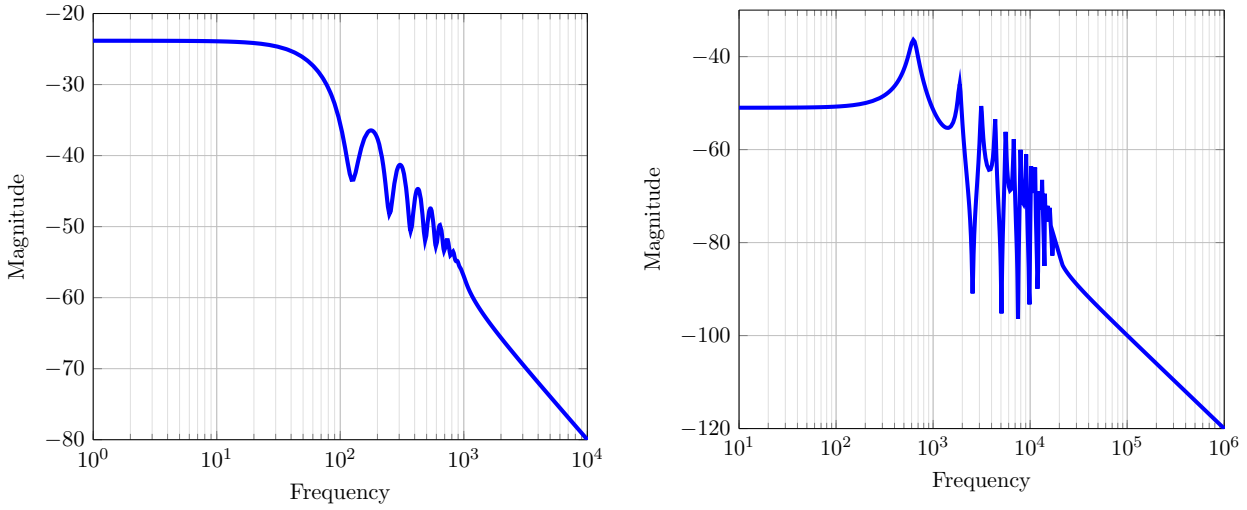
We model convection-diffusion on a rectangle $\Omega = [0,1] \times [0,1]$ by the following partial differential equation [11]. Boundary conditions are chosen as Dirichlet conditions. The resulting PDE is

$$\begin{aligned} \frac{\partial \phi(t;\mathbf{x})}{\partial t} &= \Delta\phi(t,\mathbf{x}) + \mathbf{p}\cdot\nabla\phi(t,\mathbf{x}) + b(\mathbf{x})u(t) && \mathbf{x}\in\Omega, \\ \phi(\mathbf{x}) &= 0, && \mathbf{x}\in\partial\Omega, \end{aligned} \tag{1.6.3}$$

where $b(\mathbf{x})$ represents the characteristic function for a point source in Ω. The parameter $\mathbf{p} = \begin{bmatrix} p_1 & p_2 \end{bmatrix}^\top$ represent the magnitude of convection in x and y direction. Discretizing (1.6.3) with a finite difference scheme leads to the dynamical system

$$\begin{aligned} \dot{\mathbf{x}}(t) &= (\mathbf{A}_0 + p_1\mathbf{A}_1 + p_2\mathbf{A}_2)\,\mathbf{x}(t) + \mathbf{b}u(t), \\ y(t) &= \mathbf{c}^\top\mathbf{x}(t). \end{aligned} \tag{1.6.4}$$

To illustrate (1.6.4), we show a sample transfer function Bode plot in Figure 1.2.



(a) $p_1 = 1$ (b) $p_1 = 20$

Figure 1.2: Bode plot for the convection-diffusion model (1.6.4) for different values of p_1 and fixed $p_2 = 0.1$.

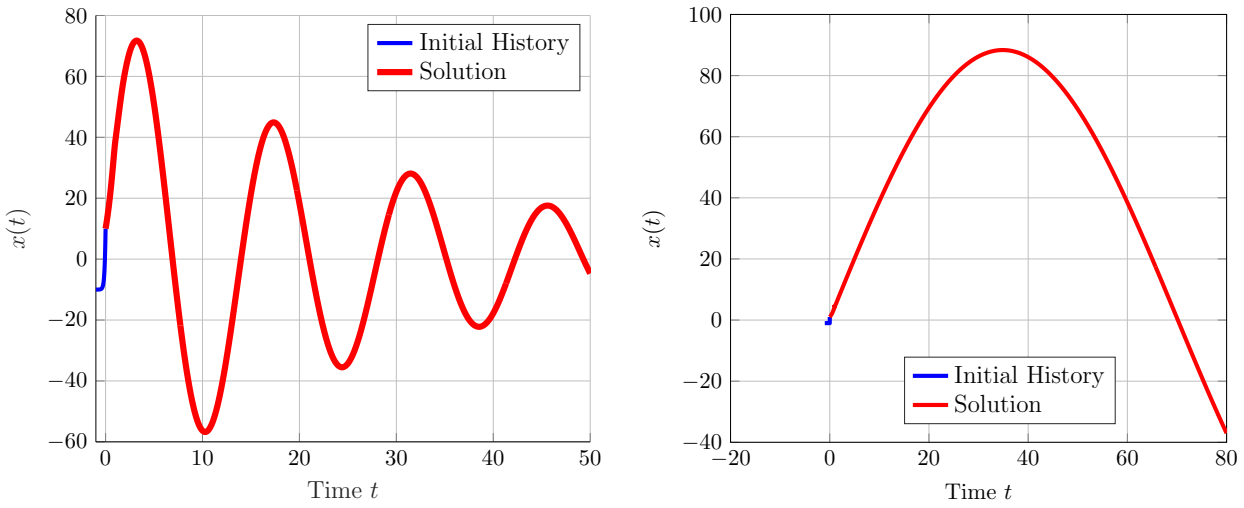
In Figure 1.2, we show the variation of the transfer function $\mathcal{H}(s, p)$ over the parameter range. Since we focus on a scalar parameter, we fix p_2 and vary $p = p_1$.

1.6.3 Delay as a Parameter

The particular case when the parameter represents a delay is illustrated by the scalar delay differential equation

$$\begin{aligned} \dot{x}(t) &= ax(t) + bx(t-\tau), \qquad \tau > 0,\ a, b, \in \mathbb{R}, \\ x(t) &= \phi(t+\tau), \qquad t \in [-\tau, 0]. \end{aligned} \tag{1.6.5}$$

Observe that to solve (1.6.5), we require an initial history function $\phi(t) : [0, \tau] \to \mathbb{R}$. In Figure 1.3, we show a time plot of the initial history function and solution to (1.6.5) for different parameter configurations (a, b) and $\tau = 1$.



(a) $(a, b) = (0.5, -1)$ (b) $(a, b) = (-1, 1)$

Figure 1.3: Solution to (1.6.5) for $\phi(t) = 2e^{20t} - 1$ and different (a, b).

Observe how the solution in Figure 1.3b with $(a, b) = (-1, 1)$ grows larger and over a longer time span compared to the solution in Figure 1.3a with $(a, b) = (0.5, -1)$. Examining the effect of parameter choices (a, b) on the transient behavior is the main subject of our investigation concerning delay equations.

1.7 Summary of Contributions and Organization

This thesis achieves three main goals:

1. Extending rational interpolation methods to the parametric case, with a particular choice of error norms and structure for the approximating system (Chapter 3).

2. Using parametric least squares approximation for measurement based system modeling, given a fixed set of measurements (Chapter 4).

3. Analyzing and understanding transient behavior in delay differential equations (as opposed to non-delay equations) as an example of a particular parametric system (Chapter 5).

The work presented here distinguishes itself in the following ways:

- We extend the interpolatory rational $\mathcal{H}_2$ optimal approximation to parametric setting and construct a parametric model with respect to a joint optimality measure in frequency and parameter domain.

- For a given fixed set of measurement data, we solve a joint discrete least-squares problem in frequency and parameter.

- We emphasize that both cases aim for a global optimality measure in the frequency and parameter simultaneously.

- For the case that the parameter represents a delay, we analyze transient behavior in detail. Therefore, we propose the use of a transformation matrix. This approach is considerably simpler than other methods and enables further insight into transient behavior and corresponding values of the coefficients.

The structure of the document is as follows: Prerequisites and current approaches are reviewed in Chapter 2.

An interpolatory approach to parametric model reduction with joint optimality conditions is investigated in Chapter 3. Here we derive first order optimality conditions in the parametric setting. This allows us to find jointly selected optimal sampling points in both frequency and parameter space with respect to a global optimality measure. We implement our method with a gradient descent algorithm and demonstrate its approximation properties on various examples.

In Chapter 4, we focus on measurement based modeling. For the parametric case, our approach relies on local models at each parameter sample point, which are combined to solve a *global* least-squares problem.

Delay systems are investigated in Chapter 5, as a particular example of parametric systems where the delay is viewed as the parameter. We chose a spectral discretization to evolve the system in time, which leads to new insight into transient behavior of the solution. A machinery is presented to construct coefficients for strong transient growth for asymptotically stable models.

We conclude with finishing remarks and future work in Chapter 6.

Chapter 2

Background

In this chapter, we summarize the necessary prerequisites from linear system theory and model reduction, as well as current approaches for non-parametric model reduction, which form the base for the parametric case. Starting with a brief overview of non-parametric model reduction approaches, we continue with an overview on current parametric model reduction methods for comparison to our approaches in later chapters.

2.1 Notation

Let us introduce a slightly non-standard notation for subsets of $\mathbb{C}$, where $\Re(z)$ denotes the real part and $\Im(z)$ the imaginary part of $z \in \mathbb{C}$:

$$\begin{aligned}
\mathbb{C}_R &:= \{z \in \mathbb{C} : \ \Re(z) > 0\} && \text{the open right half plane,} \\
\mathbb{C}_U &:= \{z \in \mathbb{C} : \ \Im(z) > 0\} && \text{the open upper half plane,} \\
\mathbb{C}_L &:= \{z \in \mathbb{C} : \ \Re(z) < 0\} && \text{the open left half plane,} \\
\mathbb{D} &:= \{z \in \mathbb{C} : \ |z| < 1\} && \text{the (open) unit disc,} \\
\overline{\mathbb{D}} &:= \{z \in \mathbb{C} : \ |z| \leq 1\} && \text{the (closed) unit disc.}
\end{aligned} \tag{2.1.1}$$

Furthermore, transfer functions will be denoted by $\mathcal{H}$ or $\mathcal{G}$, with corresponding reduced models $\widehat{\mathcal{H}}$, $\widehat{\mathcal{G}}$. We emphasize that we operate in the frequency, or Fourier domain in this chapter, and use s for the frequency variable or z if the emphasis is on the complex analysis point of view. A common notation is $\mathbb{C}_-$ and $\mathbb{C}_+$ for $\mathbb{C}_L$ and $\mathbb{C}_R$. Koosis in [80], for example, uses $\mathbb{C}_+ := \{z \in \mathbb{C} : \ \Im(z) > 0\}$. To avoid any confusion, we use the notation in (2.1.1) instead.

2.2 Linear System Theory

The main objects in this dissertation are linear dynamical systems. By linear we mean $\dot{\mathbf{x}}(t)$ depends linearly on the state variable $\mathbf{x}(t)$ and input $u(t)$. Discretizing a physical system in space using finite elements or finite differences, for example, results in a system that we

make precise in the following definition.

Definition 2.2.1 A *linear dynamical system* $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ (sometimes denoted by $\mathbf{\Sigma}(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ [4]) is given by

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t); \\ & \qquad\qquad \mathbf{x}(0) = \mathbf{x}_0, \\ \mathbf{y}(t) &= \mathbf{C}^\top \mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \end{aligned} \tag{2.2.1}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ denotes the state variable, $\mathbf{u}(t) \in \mathbb{R}^m$ the input, and $\mathbf{y}(t) \in \mathbb{R}^\nu$ the output, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, $\mathbf{C} \in \mathbb{R}^{n\times \nu}$ and $\mathbf{D} \in \mathbb{R}^{\nu\times m}$. The system has n internal states, m inputs and ν outputs. Further, we assume that $\mathbf{E}$ is nonsingular. For the case that $m = p = 1$, we speak about a *single-input-single-output system* (SISO). Similarly for $m > 1$, $\nu > 1$ a *multi-input-multi-output system* (MIMO), $m = 1$ and $\nu > 1$ a *multi-input-single-output system* and $m > 1$, $\nu = 1$ a *single-input-multi-output system.*

Unless otherwise mentioned, we assume the initial state $\mathbf{x}_0$ to be zero. This restriction can be lifted [15, 67].

The solution of (2.2.1) can be expressed via a convolution integral with kernel $\mathbf{h}(t)$:

$$\mathcal{S} : \mathbf{u}(t) \mapsto \mathbf{y}(t) = \mathcal{S}(\mathbf{u})(t) := \int_{-\infty}^{\infty} \mathbf{h}(t-\tau)\mathbf{u}(\tau)\,\mathrm{d}\tau, \qquad t \in \mathbb{R}. \tag{2.2.2}$$

Explicitly, for $\mathbf{E}$ invertible, the integral kernel has the form $\mathbf{h}(t) = \mathbf{C}^\top \mathrm{e}^{t\mathbf{E}^{-1}\mathbf{A}}\mathbf{E}^{-1}\mathbf{B} + \mathbf{D}\delta(t)$, where $\delta(t)$ denotes a delta impulse.

Definition 2.2.2 Given a dynamical system $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$, we refer to the function

$$\mathcal{H}(s) := \mathbf{C}^\top (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}, \qquad s \notin \sigma(\mathbf{A}, \mathbf{E}), \tag{2.2.3}$$

as the *transfer function* or *external description* of the dynamical system.

The transfer function relates the Laplace transforms $\mathbf{U}(s)$ of input functions and Laplace transforms $\mathbf{Y}(s)$ of output functions by

$$\mathbf{Y}(s) = \mathcal{H}(s)\mathbf{U}(s). \tag{2.2.4}$$

Moreover, the transfer function $\mathcal{H}(s)$ relates to the integral kernel $\mathbf{h}(t)$ from (2.2.2) by

$$\mathcal{H}(s) = \mathcal{L}[\mathbf{h}(t)](s), \qquad s \notin \sigma(\mathbf{A}, \mathbf{E}). \tag{2.2.5}$$

Choosing the input $\mathbf{u}(t) = \delta(t)$ to be a delta impulse and noting that $\mathcal{L}[\delta(t)](s) = 1$, the transformation to the frequency domain yields

$$\mathbf{Y}(s) = \mathcal{H}(s)\mathbf{U}(s) = \mathcal{H}(s) \cdot 1 = \mathbf{C}^\top (s\mathbf{E} - \mathbf{A})^{-1} \mathbf{B} + \mathbf{D}. \tag{2.2.6}$$

Hence the delta impulse as a particular input yields precisely the transfer function as the frequency domain output.

If $\mathcal{H}(s)$ has finitely many poles, we can express $\mathcal{H}(s)$ in pole-residue form as

$$\mathcal{H}(s) = \sum_{i=1}^{n} \frac{\boldsymbol{\phi}_i}{s - \lambda_i}, \qquad \boldsymbol{\phi}_i \in \mathbb{C}^{\nu \times m}, \tag{2.2.7}$$

where the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $(\mathbf{A}, \mathbf{E})$ are simple. In fact, if $\mathbf{E} = \mathbf{I}$ and $\mathbf{A}$ is diagonal, then $\phi_i = \mathbf{c}_i \mathbf{b}_i^\top$ are rank 1 matrices.

For ease of exposition, we focus on SISO systems, where $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times 1}$, $\mathbf{D} \in \mathbb{R}$ and use lower case $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{d}$ in (2.2.1) and (2.2.3) instead.

2.2.1 Time-Discrete Dynamical Systems

Instead of advancing the state of a system continuously in time, we can consider a discrete time movement, for example from t_k to t_{k+1}. Such a discrete dynamical system $(\mathbf{E}_\Delta, \mathbf{A}_\Delta, \mathbf{b}_\Delta, \mathbf{c}_\Delta, \mathbf{d}_\Delta)$ may arise naturally from a physical phenomenon or come from a time-discretization of a continuous system. We denote the system matrices of such a time-discrete system with a subscript Δ, so the system is

$$\begin{aligned} \mathbf{E}_\Delta \mathbf{x}_{k+1} &= \mathbf{A}_\Delta \mathbf{x}_k + \mathbf{b}_\Delta \mathbf{u}_k; \\ \mathbf{y}_k &= \mathbf{c}_\Delta^\top \mathbf{x}_k + \mathbf{d}_\Delta \mathbf{u}_k, \end{aligned} \qquad k = 1, 2, \ldots. \tag{2.2.8}$$

Stability for such time-discrete systems corresponds to the eigenvalues of the pencil $(\mathbf{A}_\Delta, \mathbf{E}_\Delta)$ being inside the open unit disc, i.e., $\sigma(\mathbf{A}_\Delta, \mathbf{E}_\Delta) \subset \mathbb{D}$.

In the following subsection, we introduce the function spaces for transfer functions $\mathcal{H}(s)$.

2.2.2 Function Spaces and Norms

Hardy spaces have been introduced by F. Riesz [103] and named after G.H. Hardy in reference to [66]. Our focus on SISO dynamical systems leads to the scalar valued *Hardy space* $\mathcal{H}_2$ with the corresponding inner product and (induced) norm.

Definition 2.2.3 For the open right half plane, we define the space by

$$\mathcal{H}_2(\mathbb{C}_R) = \left\{ \mathcal{H} : \mathbb{C} \to \mathbb{C} : \ \mathcal{H} \text{ analytic in } \mathbb{C}_R, \ \sup_{x>0} \int_{-\infty}^{\infty} |\mathcal{H}(x + \mathrm{i}y)|^2 \, \mathrm{d}y < \infty \right\}, \tag{2.2.9}$$

with the inner product and induced norm as

$$\langle \mathcal{H},\ \mathcal{G} \rangle_{\mathcal{H}_2(\mathbb{C}_R)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \overline{\mathcal{H}(\mathrm{i}\omega)} \mathcal{G}(\mathrm{i}\omega)\, \mathrm{d}\omega, \qquad \|\mathcal{H}\|^2_{\mathcal{H}_2(\mathbb{C}_R)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathcal{H}(\mathrm{i}\omega)|^2\, \mathrm{d}\omega. \quad (2.2.10)$$

Recall that a dynamical system is said to be asymptotically stable if all the eigenvalues of the matrix pencil $(\mathbf{A}, \mathbf{E})$ lie in the open left half plane $\mathbb{C}_L$. Hence the transfer function $\mathcal{H}(s)$ of a stable system only has poles in $\mathbb{C}_L$ and is analytic in $\mathbb{C}_R$, which is one motivation for Definition 2.2.3.

Definition 2.2.4 Consider the Hardy space on the open unit disc:

$$\mathcal{H}_2(\mathbb{D}) = \left\{ \mathcal{H} : \mathbb{C} \to \mathbb{C} :\ \mathcal{H} \text{ analytic in } \mathbb{D},\ \sup_{0<r<1} \int_0^{2\pi} |\mathcal{H}(r\mathrm{e}^{\mathrm{i}\theta})|^2\, \mathrm{d}\theta < \infty \right\}, \quad (2.2.11)$$

with the inner product and induced norm

$$\langle \mathcal{H},\ \mathcal{G} \rangle_{\mathcal{H}_2(\mathbb{D})} = \frac{1}{2\pi} \int_0^{2\pi} \overline{\mathcal{H}(e^{i\theta})} \mathcal{G}(\mathrm{e}^{\mathrm{i}\theta})\, \mathrm{d}\theta, \qquad \|\mathcal{H}\|^2_{\mathcal{H}_2(\mathbb{D})} = \frac{1}{2\pi} \int_0^{2\pi} |\mathcal{H}(\mathrm{e}^{\mathrm{i}\theta})|^2\, \mathrm{d}\theta. \quad (2.2.12)$$

To be clear, we note that $\mathcal{H}_2(\mathbb{D})$ contains functions that are analytic on $\mathbb{D}$. Recall that stability for discrete time systems requires $\sigma(\mathbf{A}_\Delta, \mathbf{E}_\Delta) \subset \mathbb{D}$, leading to the complementary Hardy space $\mathcal{H}_2(\mathbb{C}\backslash\overline{\mathbb{D}}) = L_2(\partial\mathbb{D})\backslash\mathcal{H}_2(\mathbb{D})$.

Let us further introduce the notation of *rational* Hardy spaces [47, 79]:

$$\begin{aligned} \mathcal{RH}_2(\mathbb{C}_R) &:= \{ \mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R) :\ \mathcal{H} \text{ is rational } \}, \quad \text{and} \\ \mathcal{RH}_2(\mathbb{D}) &:= \{ \mathcal{H} \in \mathcal{H}_2(\mathbb{D}) :\ \mathcal{H} \text{ is rational } \}. \end{aligned} \quad (2.2.13)$$

Example 2.2.5 To highlight the difference between $\mathcal{RH}_2(\mathbb{C}_R)$ and $\mathcal{H}_2(\mathbb{C}_R)$, consider $f(z) = \mathrm{e}^{-|z|}$. It is clear that $f(z) \in \mathcal{H}_2(\mathbb{C}_R)$ since the function has no poles and certainly a finite integral on $\mathrm{i}\mathbb{R}$. But $f \in \mathcal{H}_2(\mathbb{C}_R)\backslash\mathcal{RH}_2(\mathbb{C}_R)$.

We simplify notation and write $\mathcal{H}_2$ for $\mathcal{H}_2(\mathbb{C}_R)$ and $\mathcal{H}_2(\mathbb{D})$ when the underlying space if clear from the context.

Why are we interested in the Hardy spaces above in the context of model reduction? The following lemma ([3]) establishes a connection between time domain error and approximation error in the norm (2.2.9).

Lemma 2.2.6 *Let $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R)$ and $u \in L_2(\mathbb{R}^+)$. Then we can bound the norm of the output $y(t)$ in (2.2.1) as*

$$\|y\|_{L_\infty} \leq \|\mathcal{H}\|_{\mathcal{H}_2} \cdot \|u\|_{L_2}. \tag{2.2.14}$$

Therefore the L_∞ time domain energy of the output can be bounded by the $\mathcal{H}_2$ norm of the transfer function and the L_2 energy of the input $u(t)$.

Proof Let $U(s)$ be the Laplace transformation of the input $u(t)$, $Y(s)$ that of the output $y(t)$. Using Hölder's inequality, we compute

$$\begin{aligned}
\|y\|_{L\infty} &= \max_{t>0} |y(t)| \\
&= \max_{t>0} \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\mathrm{i}\omega) \mathrm{e}^{\mathrm{i}\omega t} \, \mathrm{d}\omega \right| \\
&\leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |Y(\mathrm{i}\omega)| \, \mathrm{d}\omega \\
&\leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathcal{H}(\mathrm{i}\omega)| \, |U(\mathrm{i}\omega)| \, \mathrm{d}\omega \\
&\leq \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathcal{H}(\mathrm{i}\omega)|^2 \, \mathrm{d}\omega \right)^{1/2} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |U(\mathrm{i}\omega)|^2 \, \mathrm{d}\omega \right)^{1/2} \\
&\leq \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |\mathcal{H}(\mathrm{i}\omega)|^2 \, \mathrm{d}\omega \right)^{1/2} \left(\int_{0}^{\infty} |u(t)|^2 \, \mathrm{d}t \right)^{1/2} \\
&= \|\mathcal{H}\|_{\mathcal{H}_2} \cdot \|u\|_{L_2}.
\end{aligned} \tag{2.2.15}$$

■

Corollary 2.2.7 *Let $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R)$ with output $\mathbf{y}$, $\widehat{\mathcal{H}} \in \mathcal{H}_2(\mathbb{C}_R)$ with output $\widehat{\mathbf{y}}$ and $u \in L_2(\mathbb{R}^+)$. Then Lemma 2.2.6 implies*

$$\|\mathbf{y} - \widehat{\mathbf{y}}\|_{L_\infty} \leq \|\mathcal{H} - \widehat{\mathcal{H}}\|_{\mathcal{H}_2} \|u\|_{L_2}. \tag{2.2.16}$$

For a special case of $\mathcal{H}(s)$, the $\mathcal{H}_2$ norm has a simple expression that we present in the following lemma.

Lemma 2.2.8 *Given a transfer function $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R)$ in pole-residue form $\mathcal{H}(s) = \sum_{k=1}^{n} \frac{\phi_k}{s - \lambda_k}$ and $\mathcal{G} \in \mathcal{H}_2(\mathbb{C}_R)$, the $\mathcal{H}_2$ inner product and norm can be expressed as*

$$\langle \mathcal{H},\ \mathcal{G} \rangle_{\mathcal{H}_2} = \sum_{i=1}^{n} \overline{\phi_i} \mathcal{G}\left(-\overline{\lambda_i}\right), \quad \textit{and} \quad \|\mathcal{H}\|_{\mathcal{H}_2} = \sqrt{\sum_{i=1}^{n} \overline{\phi_i} \mathcal{H}\left(-\overline{\lambda_i}\right)}. \tag{2.2.17}$$

Proof We compute the $\mathcal{H}_2$ inner product by direct computation. First consider a simple pole $\mathcal{H}_0(s) = \dfrac{1}{s - \lambda_0}$. Using the residue theorem on the contour Γ_R, a semi-circle of radius R on the imaginary axis large enough to enclose the pole λ_0, we see that

$$\begin{aligned}
\langle \mathcal{H}_0,\ \mathcal{G} \rangle_{\mathcal{H}_2} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\mathrm{i}\omega - \lambda_0} \mathcal{G}(\mathrm{i}\omega)\, \mathrm{d}\omega \\
&= \frac{-1}{2\pi \mathrm{i}} \int_{\Gamma_R} \frac{1}{-z - \overline{\lambda_0}} \mathcal{G}(z)\, \mathrm{d}z \\
&= \operatorname{Res}\left[\frac{-1}{-z - \overline{\lambda_0}} \mathcal{G}(z), \quad z = -\overline{\lambda_0} \right] \\
&= \lim_{z \to -\overline{\lambda_0}} \frac{-(z + \overline{\lambda_0})}{-z - \overline{\lambda_0}} \mathcal{G}(z) = \mathcal{G}(-\overline{\lambda_0}).
\end{aligned} \tag{2.2.18}$$

By linearity of the integral, we find that

$$\langle \mathcal{H},\ \mathcal{G} \rangle_{\mathcal{H}_2} = \sum_{k=1}^{n} \overline{\phi_k} \left\langle \frac{1}{s - \lambda_k},\ \mathcal{G} \right\rangle_{\mathcal{H}_2} = \sum_{k=1}^{n} \overline{\phi_k} \mathcal{G}(-\overline{\lambda_k}), \tag{2.2.19}$$

which proves the claim for the inner product. By definition of the induced norm, it then follows that

$$\|\mathcal{H}\|_{\mathcal{H}_2} = \sqrt{\langle \mathcal{H},\ \mathcal{H} \rangle_{\mathcal{H}_2}} = \sqrt{\sum_{k=1}^{n} \overline{\phi_k} \mathcal{H}(-\overline{\lambda_k})}. \tag{2.2.20}$$

■

With our focus on $\mathcal{H}_2$ model reduction, we assume $\mathbf{d} = 0$, since for $\|\mathcal{H} - \widehat{\mathcal{H}}\|_{\mathcal{H}_2} < \infty$, it needs to hold that $\widehat{\mathbf{d}} = \mathbf{d}$.

2.3 Model Reduction

In this section, we review the main concepts of model reduction for linear, time-invariant dynamical systems. Many approaches are available to approximate various types of models. We focus on models applicable to linear dynamical systems with particular focus on approaches that can be extended to the parametric case.

2.3.1 The Gramians

Reachability and observability Gramians [3] characterize the states of the system $\mathbf{x}(t) \in \mathbb{R}^n$ by how *reachable* and *observable* they are. For reachability, the metric is the input energy required to steer the system $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ from 0 to a target state $\mathbf{x}_1 \in \mathbb{R}^n$. High required energy represents a difficult to reach state. The observability Gramian uses the adjoint concept and gauges the energy observed from driving a given state to 0. To simplify notation, we assume $\mathbf{E} = \mathbf{I}$.

Definition 2.3.1 Given a dynamical system $(\mathbf{A}, \mathbf{b}, \mathbf{c})$, we define the *reachability Gramian*

$$\mathcal{P} := \int_0^\infty e^{\mathbf{A}\tau}\mathbf{b}\mathbf{b}^\top \mathrm{e}^{\mathbf{A}^\top\tau} d\tau, \tag{2.3.1}$$

and the *observability Gramian*

$$\mathcal{Q} := \int_0^\infty e^{\mathbf{A}^\top\tau}\mathbf{c}\mathbf{c}^\top \mathrm{e}^{\mathbf{A}\tau} d\tau. \tag{2.3.2}$$

The Gramians satisfy the following Lyapunov equations.

Lemma 2.3.2 *The Gramians $\mathcal{P}$ (2.3.1) and $\mathcal{Q}$ (2.3.2) satisfy*

$$\begin{aligned} \mathbf{A}\mathcal{P} + \mathcal{P}\mathbf{A}^\top + \mathbf{b}\mathbf{b}^\top = 0, \quad \textit{and} \\ \mathbf{A}^\top\mathcal{Q} + \mathcal{Q}\mathbf{A} + \mathbf{c}\mathbf{c}^\top = 0. \end{aligned} \tag{2.3.3}$$

Note that the Lyapunov equaions (2.3.3) have a unique solution if and only if the eigenvalues of $\mathbf{A}$ and $\mathbf{A}^\top$ satisfy $\lambda_i(\mathbf{A}) + \lambda_j(\mathbf{A}^\top) \neq 0$ for $i, j = 1, \ldots, n$.

For the discrete time case, the following definitions parallel Definition 2.3.1 in the discrete time case.

Definition 2.3.3 Let $\Delta(\mathbf{A}_d, \mathbf{b}_d, \mathbf{c}_d)$ be a discrete dynamical system. Define the (discrete) *observability Gramian*

$$\mathcal{Q}_\Delta := \sum_{k=0}^{\infty} (\mathbf{A}_d^\top)^k \mathbf{c}_d \mathbf{c}_d^\top \mathbf{A}_d^k, \tag{2.3.4}$$

and the (discrete) *reachability Gramian*

$$\mathcal{P}_\Delta := \sum_{k=0}^{\infty} \mathbf{A}_d^k \mathbf{b}_d \mathbf{b}_d^\top (\mathbf{A}_d^\top)^k. \tag{2.3.5}$$

Instead of Lyapunov equations, for the discrete time case, the Gramians satisfy so called *Stein* equations as shown in the following lemma.

Lemma 2.3.4 *The discrete gramians $\mathcal{P}_\Delta$ and $\mathcal{Q}_\Delta$ from Definition 2.3.3 satisfy*

$$\begin{aligned} \mathbf{A}_d \mathcal{P}_\Delta \mathbf{A}_d^\top + \mathbf{b}_d \mathbf{b}_d^\top &= \mathcal{P}_\Delta, \quad \textit{and} \\ \mathbf{A}_d^\top \mathcal{Q}_\Delta \mathbf{A}_d + \mathbf{c}_d \mathbf{c}_d^\top &= \mathcal{Q}_\Delta. \end{aligned} \tag{2.3.6}$$

The Stein equations in (2.3.6) have a unique solution if and only if the eigenvalues of $\mathbf{A}_d$ and $\mathbf{A}_d^\top$ satisfy $\lambda_i(\mathbf{A}) \cdot \lambda_j(\mathbf{A}^\top) \neq 1$ for $i, j = 1, \ldots, n$.

We extend the notion of Gramians from (2.3.3) to the parametric case in Section 3.2.3.

2.3.2 Projection Based Model Reduction

In case a state-space representation as in (2.2.1) is available, a common approach to model reduction is projection.

The underlying assumption is that the relevant dynamics of the state $\mathbf{x}(t)$ of the system $(\mathbf{E}, \mathbf{A}, \mathbf{b}, \mathbf{c})$ evolve in a low dimensional subspace $\mathcal{V}$, where $\dim \mathcal{V} = r \ll n$, spanned by the columns of the matrix $\mathbf{V}$, so that $\mathbf{x}(t) \approx \mathbf{V}\widehat{\mathbf{x}}(t)$. We consider a *Petrov-Galerkin projection* onto an r-dimensional subspace $\mathcal{V}$. Let $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times r}$, where $\mathbf{V}$ represents the target subspace for $\widehat{\mathbf{x}}(t)$ and $\mathbf{W}$ the *trial subspace*, along which we project.

The projected system matrices have the form

$$\widehat{\mathbf{E}} = \mathbf{W}^\top \mathbf{E} \mathbf{V}, \quad \widehat{\mathbf{A}} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \widehat{\mathbf{b}} = \mathbf{W}^\top \mathbf{b}, \quad \text{and} \quad \widehat{\mathbf{c}} = \mathbf{V}^\top \mathbf{c}. \tag{2.3.7}$$

Many approaches have been investigated to constructing $\mathbf{V}$ and $\mathbf{W}$. One particularly useful type of projection based model reduction is *balanced truncation* [94, 95] Since balanced truncation is not the main topic of this dissertation, we refer the reader to [3, 58] and the references therein. In this thesis, we mainly focus on interpolation.

2.4 Interpolatory Model Reduction

In this section, we construct a reduced model $\widehat{\mathcal{H}}(\mathrm{s})$ by matching the original model $\mathcal{H}(s)$ at some selected points. One can construct the projection matrices $\mathbf{V}$ and $\mathbf{W}$ in such a way that $\widehat{\mathbf{H}}(s)$ interpolates the system at certain points. Let $\sigma_1, \ldots, \sigma_r$ be chosen interpolation points. Let $\mathbf{V}$ and $\mathbf{W}$ be such that

$$\begin{aligned} \left[(\sigma_1\mathbf{E}-\mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r\mathbf{E}-\mathbf{A})^{-1}\mathbf{b}\right] &\subset \mathrm{span}(\mathbf{V}), \qquad \text{and} \\ \left[(\sigma_1\mathbf{E}-\mathbf{A})^{-\top}\mathbf{c}, \ldots, (\sigma_r\mathbf{E}-\mathbf{A})^{-\top}\mathbf{c}\right] &\subset \mathrm{span}(\mathbf{W}). \end{aligned} \tag{2.4.1}$$

Then $\widehat{\mathcal{H}}(\sigma_i) = \mathcal{H}(\sigma_i)$ and $\widehat{\mathcal{H}}'(\sigma_i) = \mathcal{H}'(\sigma_i)$ for $i = 1 \ldots, r$. This leads to the concept of interpolatory model reduction, where the reduced system is constructed by enforcing an exact match with the original model at certain selected points.

To make this discussion largely self-contained, we review some results in (non-parametric) model reduction that are used in the parametric case in a similar way. Consider the non-parametric dynamical system with a single input and output

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t), \qquad \mathbf{y}(t) = \mathbf{c}^\top\mathbf{x}(t), \tag{2.4.2}$$

with the initial condition $\mathbf{x}(0) = 0$ and matrices $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{n\times n}$ and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^{n\times 1}$, $\mathbf{E}$ invertible, focusing on the SISO case. The transfer function mapping Laplace transforms $\mathbf{U}(s)$ of inputs to Laplace transforms $\mathbf{Y}(s)$ of outputs is $\mathbf{H}(s) = \mathbf{c}^\top (s\mathbf{E} - \mathbf{A})^{-1} \mathbf{b}$. Our goal is to arrive at a reduced model of order $r \ll n$ with the same structure:

$$\widehat{\mathbf{E}}\dot{\widehat{\mathbf{x}}}(t) = \widehat{\mathbf{A}}\widehat{\mathbf{x}}(t) + \widehat{\mathbf{b}}u(t), \qquad \widehat{\mathbf{y}}(t) = \widehat{\mathbf{c}}^\top \widehat{\mathbf{x}}(t), \tag{2.4.3}$$

and associated transfer function $\widehat{\mathcal{H}}(s) = \widehat{\mathbf{c}}^\top \left(s\widehat{\mathbf{E}} - \widehat{\mathbf{A}}\right)^{-1} \widehat{\mathbf{b}}$ and $\widehat{\mathbf{A}}, \widehat{\mathbf{E}} \in \mathbb{R}^{r\times r}$ and $\widehat{\mathbf{b}}, \widehat{\mathbf{c}} \in \mathbb{R}^r$ and we assume $\widehat{\mathbf{E}}$ to be invertible. The reduced model should approximate the output well, given the reduced order r, for any input $u(t)$; more precisely $\|y(t) - \widehat{y}(t)\|_{L_\infty} \to \min$. Note here that the reduced model does not depend on the input function $u(t)$ but should be a good approximation for all inputs $u(t)$.

Recall from Lemma 2.2.6 that the time-domain L_∞ norm of the output, the L_2 energy of the input and frequency domain quantities are related through

$$\|y - \widehat{y}\|_{L_\infty} \leq \|\mathcal{H} - \widehat{\mathcal{H}}\|_{\mathcal{H}_2} \|u\|_{L_2}. \tag{2.4.4}$$

In essence, a small difference in the $\mathcal{H}_2$ norm between transfer functions guarantees a low maximum mismatch of the outputs rements over time. For more details, see [3].

Constructing a good reduced model corresponds to minimizing $\|\mathcal{H} - \widehat{\mathcal{H}}\|_{\mathcal{H}_2}$. Optimality conditions for this case have been derived by Meier and Luenberger in 1967.

Theorem 2.4.1 ([91], SISO Optimality Conditions) *Let $\widehat{\mathcal{H}}(s) = \sum_{i=1}^r \frac{\phi_i}{s - \lambda_i} \in \mathcal{H}_2$ be a transfer function in pole-residue form. Then $\widehat{\mathcal{H}}(s)$ minimizes $\|\mathcal{H} - \widehat{\mathcal{H}}\|_{\mathcal{H}_2}$ if*

$$\mathcal{H}(-\lambda_i) = \widehat{\mathcal{H}}(-\lambda_i), \qquad \textit{and} \qquad \mathcal{H}'(-\lambda_i) = \widehat{\mathcal{H}}'(-\lambda_i), \qquad i = 1, \ldots, r. \tag{2.4.5}$$

The Iterative Rational Krylov Algorithm (IRKA, [59] - Algorithm 2.4.1) performs an iterative refinement of an initial pole selection $\{\lambda_i\}_{i=1}^r$, rather than sampling the transfer function at greedy points. Upon convergence, the resulting model $\widehat{\mathcal{H}}(s)$ is guaranteed to be locally optimal in the $\mathcal{H}_2$ sense because local maxima and saddle points are repellant [46]. We emphasize that IRKA *automatically* selects interpolation points in the frequency domain.

Note that the reduced model interpolates the original model at the *mirror images* of the reduced poles (which are not known a priori).

Algorithm 2.4.1 Iterative Rational Krylov Algorithm - SISO - State Space

1. Make initial selection of sampling points $\{\sigma_1, \ldots, \sigma_r\} \subset \mathbb{C}_R$, closed under conjugation

2. Construct projection matrices

$$\begin{aligned} \mathbf{V} &= [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1}\mathbf{b}] \\ \mathbf{W} &= [(\sigma_1 \mathbf{E} - \mathbf{A})^{-\top}\mathbf{c}, \ldots, (\sigma_r \mathbf{E} - \mathbf{A})^{-\top}\mathbf{c}] \end{aligned} \tag{2.4.6}$$

3. Until convergence

   (a) Project system matrices $\widehat{\mathbf{A}} = \mathbf{W}^\top \mathbf{A}\mathbf{V}$, $\widehat{\mathbf{E}} = \mathbf{W}^\top \mathbf{E}\mathbf{V}$, $\widehat{\mathbf{b}} = \mathbf{W}^\top \mathbf{b}$ and $\widehat{\mathbf{c}} = \mathbf{V}^\top \mathbf{c}$

   (b) Compute eigenvalues $\lambda_1, \ldots, \lambda_r$ of the pencil $(\widehat{\mathbf{A}}, \widehat{\mathbf{E}})$

   (c) Assign $\sigma_i \leftarrow -\lambda_i$ for $i = 1, \ldots, r$

   (d) Update projection matrices as in (2.4.1)

$$\begin{aligned} \mathbf{V} &= [(\sigma_1 \mathbf{E} - \mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r \mathbf{E} - \mathbf{A})^{-1}\mathbf{b}] \\ \mathbf{W} &= [(\sigma_1 \mathbf{E} - \mathbf{A})^{-\top}\mathbf{c}, \ldots, (\sigma_r \mathbf{E} - \mathbf{A})^{-\top}\mathbf{c}] \end{aligned} \tag{2.4.7}$$

4. Set $\widehat{\mathbf{A}} = \mathbf{W}^\top \mathbf{A}\mathbf{V}$, $\widehat{\mathbf{E}} = \mathbf{W}^\top \mathbf{E}\mathbf{V}$, $\widehat{\mathbf{b}} = \mathbf{W}^\top \mathbf{b}$ and $\widehat{\mathbf{c}} = \mathbf{V}^\top \mathbf{c}$

For completeness of this exposition, IRKA can also be performed for MIMO systems. In this case, in addition to sampling points σ_i, left and right tangent directions for interpolation are selected.

Algorithm 2.4.2 Iterative Rational Krylov Algorithm - MIMO - State Space

1. Make initial selection of sampling points $\{\sigma_1, \ldots, \sigma_r\} \subset \mathbb{C}_R$, closed under conjugation, right tangent directions $\mathbf{r}_1, \ldots, \mathbf{r}_r$ and left tangent directions $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_r$.

2. Construct projection matrices

$$\begin{aligned} \mathbf{V} &= [(\sigma_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_1, \ldots, (\sigma_r\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_r] \\ \mathbf{W} &= [(\sigma_1\mathbf{E} - \mathbf{A})^{-\top}\mathbf{C}\boldsymbol{\ell}_1, \ldots, (\sigma_r\mathbf{E} - \mathbf{A})^{-\top}\mathbf{C}\boldsymbol{\ell}_r] \end{aligned} \tag{2.4.8}$$

3. Until convergence

   (a) Project system matrices $\widehat{\mathbf{A}} = \mathbf{W}^\top\mathbf{A}\mathbf{V}$, $\widehat{\mathbf{E}} = \mathbf{W}^\top\mathbf{E}\mathbf{V}$, $\widehat{\mathbf{B}} = \mathbf{W}^\top\mathbf{B}$ and $\widehat{\mathbf{C}} = \mathbf{V}^\top\mathbf{C}$

   (b) Compute pole-residue expansion of $\widehat{\mathcal{H}}(s)$:

$$\widehat{\mathcal{H}}(s) = \sum_{i=1}^{r} \frac{\widehat{\boldsymbol{\ell}}_i \widehat{\mathbf{r}}_i^{\top}}{s - \lambda_i} \tag{2.4.9}$$

   (c) Assign $\sigma_i \leftarrow -\lambda_i$, $\mathbf{r}_i \leftarrow \widehat{\mathbf{r}}_i$ and $\boldsymbol{\ell}_i \leftarrow \widehat{\boldsymbol{\ell}}_i$ for $i = 1, \ldots, r$

   (d) Update projection matrices as in (2.4.1)

$$\begin{aligned} \mathbf{V} &= [(\sigma_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_1, \ldots, (\sigma_r\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r}_r] \\ \mathbf{W} &= [(\sigma_1\mathbf{E} - \mathbf{A})^{-\top}\mathbf{C}\boldsymbol{\ell}_1, \ldots, (\sigma_r\mathbf{E} - \mathbf{A})^{-\top}\mathbf{C}\boldsymbol{\ell}_r] \end{aligned} \tag{2.4.10}$$

4. Set $\widehat{\mathbf{A}} = \mathbf{W}^\top\mathbf{A}\mathbf{V}$, $\widehat{\mathbf{E}} = \mathbf{W}^\top\mathbf{E}\mathbf{V}$, $\widehat{\mathbf{B}} = \mathbf{W}^\top\mathbf{B}$ and $\widehat{\mathbf{C}} = \mathbf{V}^\top\mathbf{C}$

The improvement of the model subspace given by the poles $\{\lambda_i\}_{i=1}^r$ during IRKA is not monotone; the error may increase at one iteration and decrease at the next - even if the

algorithm converges. An improvement of this behavior is made by a trust region method [13] that ensures a strict improvement of the error function with each pole update step until convergence.

2.5 Data Driven Methods

In this section, we introduce some of the main methods to construct a system of desired order given a set of measurements instead of an internal or external description of a given model. There is little assumption of where the measurements are taken from, whether it be numerical simulations or real, physical measurements. The methods presented here can be categorized in two ways: interpolatory methods and least-squares fit. First, we present the Loewner framework [90] as a popular method to constructs a system that matches the set of measurements exactly. The second method, called Vector Fitting [63], presented here construct a least-squares approximant to the data set.

2.5.1 The Loewner Framework

Let $\{\xi_i, \mathcal{H}(\xi_i)\}_{i=1}^m \subset \mathbb{C} \times \mathbb{C}$ be a set of frequency-measurement pairs. Those need not come from a *known* transfer function $\mathcal{H}(s)$ but can be physical measurements. The following lemma is from [74].

Lemma 2.5.1 *Given measurements $\{\xi_i, M_i\}_{i=1}^m$ and a partition of those in (almost) equal*

sets

$$\begin{aligned}\{\xi_1,\ldots,\xi_m\} &= \{\mu_1,\ldots,\mu_{\underline{m}}\}\cup\{\nu_1,\ldots,\nu_{\overline{m}}\}\\ \{\mathcal{H}(\xi_1),\ldots,\mathcal{H}(\xi_m)\} &= \{\mathcal{H}(\mu_1),\ldots,\mathcal{H}(\mu_{\underline{m}})\}\cup\{\mathcal{H}(\nu_1),\ldots,\mathcal{H}(\nu_{\overline{m}})\},\end{aligned} \tag{2.5.1}$$

with $\mu_i \neq \nu_j$, $i = 1,\ldots,\underline{m}$, $j = 1,\ldots,\overline{m}$, $\underline{m}+\overline{m} = m$. *Measurement points and values are partitioned in the same way. Define the* Loewner matrix *(divided difference matrix)* $\mathbb{L}$ *as*

$$\mathbb{L} := \begin{bmatrix} \dfrac{\mathcal{H}(\mu_1)-\mathcal{H}(\nu_1)}{\mu_1-\nu_1} & \cdots & \dfrac{\mathcal{H}(\mu_1)-\mathcal{H}(\nu_{\overline{m}})}{\mu_1-\nu_{\overline{m}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\mathcal{H}(\mu_{\underline{m}})-\mathcal{H}(\nu_1)}{\mu_{\underline{m}}-\nu_1} & \cdots & \dfrac{\mathcal{H}(\mu_{\underline{m}})-\mathcal{H}(\nu_{\overline{m}})}{\mu_{\underline{m}}-\nu_{\overline{m}}} \end{bmatrix}. \tag{2.5.2}$$

Let $\mathbf{c}$ *be in the null space of* $\mathbb{L}$, *i.e.,* $\mathbb{L}\mathbf{c} = 0$ *and* $b_i = \mathcal{H}(\xi_i)$. *Then the barycentric rational function*

$$r(s) = \frac{n(s)}{d(s)} = \frac{\sum_{i=1}^{\overline{m}} \dfrac{\mathbf{c}_i b_i}{s-\nu_i}}{\sum_{i=1}^{\overline{m}} \dfrac{\mathbf{c}_i}{s-\nu_i}}, \tag{2.5.3}$$

interpolates the measurements, meaning $r(\xi_i) = \mathcal{H}(\xi_i)$, $i = 1,\ldots,m$.

So far, we split up the measurements points into two distinct sets (and the measurements accordingly). If derivative information on $\mathcal{H}(s)$ is available, we can introduce the notion of a *shifted Loewner matrix* [90].

Lemma 2.5.2 *Let* $\{\xi_i, \mathcal{H}(\xi_i), \mathcal{H}'(\xi_i)\}_{i=1}^{m} \subset \mathbb{C}\times\mathbb{C}\times\mathbb{C}$ *be measurement data. Define the Loewner and shifted Loewner matrices as*

$$[\mathbb{L}]_{i,j} := \begin{cases} \dfrac{\mathcal{H}(\xi_i)-\mathcal{H}(\xi_j)}{\xi_i-\xi_j}, & i\neq j; \\ \mathcal{H}'(\xi_i), & i = j; \end{cases} \quad \textit{and} \quad [\mathbb{L}_s]_{i,j} := \begin{cases} \dfrac{\xi_i\mathcal{H}(\xi_i)-\mathcal{H}(\xi_j)\xi_j}{\xi_i-\xi_j}, & i\neq j; \\ \xi_i\mathcal{H}'(\xi_i), & i = j. \end{cases} \tag{2.5.4}$$

Further let

$$\widehat{\mathbf{b}} = \widehat{\mathbf{c}} = [\mathcal{H}(\xi_1), \ldots, \mathcal{H}(\xi_m)]. \tag{2.5.5}$$

Then the transfer function $\mathcal{G}$ given by

$$\mathcal{G}(s) := \widehat{\mathbf{c}}^\top (\mathbb{L}_s - s\mathbb{L})^{-1} \widehat{\mathbf{b}} \tag{2.5.6}$$

interpolates the measurement data in both a Lagrange sense ($\mathcal{G}(\xi_i) = \mathcal{H}(\xi_i)$) and Hermite sense ($\mathcal{G}'(\xi_i) = \mathcal{H}'(\xi_i)$).

We can view Lemma 2.5.1 as a special case of Lemma 2.5.2 with repeated measurements $[\xi_1, \xi_1, \xi_2, \xi_2, \ldots, \xi_m, \xi_m]^\top \in \mathbb{C}^{2m}$.

Using the Loewner framework, we can construct a state space model given any set of sampling points of a full order transfer function $\mathcal{H}(s)$. With that, Algorithm 2.4.1 can be extended to not require an internal formulation of the dynamical system but merely repeated function and derivative evaluations of any given $\mathcal{H}(s)$ [14].

2.5.2 Least Squares Approximation

In preparation for our extension to the parametric case, we review the non-parametric approximation methods of the Snathanan-Koerner iteration and Vector Fitting.

Problem 2.5.3 *We consider the following problem.*

For a given data set $\{\xi_i,\ \mathcal{H}(\xi_i)\}_{i=1}^{m_s} \subset \mathbb{C}\times\mathbb{C}$*, find a stable rational transfer function* $\widehat{\mathcal{H}}(s)$ *of degree* r *that minimizes the (weighted) discrete least squares problem*

$$\sum_{i=1}^{m_s} w_i \left|\widehat{\mathcal{H}}(\xi_i) - \mathcal{H}(\xi_i)\right|^2 \to \min, \tag{2.5.7}$$

for some weights $w_i > 0$*,* $i = 1, \ldots, m_s$*.*

Note that the data may not come from a *known* transfer function $\mathcal{H}(s)$, but can be physical measurements. The problem (2.5.7) above can be considered for MIMO systems, replacing $|\cdot|$ by $\|\cdot\|_F$ in (2.5.7). We focus our attention on the SISO case, .

2.5.3 The SK-Iteration

One way to solve Problem 2.5.3 is the Sanathanan-Koerner iteration (SK Iteration) [105]. A proper rational function $\widehat{\mathcal{H}}(s)$ is fitted to the measurements in a least squares sense with the representation

$$\widehat{\mathcal{H}}(s) = \frac{n(s)}{d(s)} = \frac{\displaystyle\sum_{k=0}^{r-1} \alpha_k s^k}{\displaystyle\sum_{k=0}^{r} \beta_k s^k}, \qquad \alpha_k, \beta_k \in \mathbb{C},\ k = 1, \ldots, r. \tag{2.5.8}$$

Hereby we call $n(s) = \displaystyle\sum_{k=0}^{r-1} \alpha_k s^k$ the numerator and $d(s) = \displaystyle\sum_{k=0}^{r} \beta_k s^k$ the denominator. If $d(\xi_i) \neq 0$, $i = 1, \ldots, m$, we rearrange (2.5.7) to arrive at the equivalent minimization problem

$$\sum_{i=1}^{m_s} \frac{w_i}{|d(\xi_i)|^2} \left|n(\xi_i) - d(\xi_i)\mathcal{H}(\xi_i)\right|^2 \to \min. \tag{2.5.9}$$

Observe that (2.5.9) is a *nonlinear* least squares problem. Let $\mathbf{x} = [\alpha_0, \ldots, \alpha_{r-1}, \beta_0, \ldots, \beta_r]^\top = [\boldsymbol{\alpha}, \boldsymbol{\beta}]^\top \in \mathbb{C}^{2r+1}$ be the vector of numerator and denominator coefficients in (2.5.8) and

$\mathbf{b} = [\mathcal{H}(\xi_1), \ldots, \mathcal{H}(\xi_m)]^\top \in \mathbb{C}^{m\times 1}$, the measurement vector. The least squares problem (2.5.9) is solved for $\mathbf{x}$ via a sqeuence of linear least squares problems

$$\|\Delta^{(k)} \left(\mathcal{A}\mathbf{x}^{(k+1)} - \mathbf{b}\right)\|_2^2 \to \min, \qquad k = 0, 1, 2, \ldots, \tag{2.5.10}$$

with the matrix

$$\mathcal{A} = \begin{bmatrix} 1 & \xi_1 & \xi_1^2 & \cdots & \xi_1^{r-1} & -\mathcal{H}(\xi_1)\xi_1 & -\mathcal{H}(\xi_1)\xi_1^2 & \cdots & -\mathcal{H}(\xi_1)\xi_1^r \\ 1 & \xi_2 & \xi_2^2 & \cdots & \xi_2^{r-1} & -\mathcal{H}(\xi_2)\xi_2 & -\mathcal{H}(\xi_2)\xi_2^2 & \cdots & -\mathcal{H}(\xi_2)\xi_2^r \\ \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_m & \xi_m^2 & \cdots & \xi_m^{r-1} & -\mathcal{H}(\xi_m)\xi_m & -\mathcal{H}(\xi_m)\xi_m^2 & \cdots & -\mathcal{H}(\xi_m)\xi_m^r \end{bmatrix}. \tag{2.5.11}$$

Here the weights $\Delta^{(k)}$ in (2.5.10) are iteratively refined based on the numerator $d^{(k-1)}(s)$ from the previous iteration:

$$\Delta^{(k)} = \operatorname{diag}\left\{ \frac{w_i}{|d^{(k-1)}(\xi_i)|^2} \right\}_{i=1}^{m}, \qquad k = 1, 2, \ldots. \tag{2.5.12}$$

We note that the superscripts $d^{(k)}$ represent iteration indices, *not* derivatives here. The procedure is summarized in Algorithm 2.5.1.

Algorithm 2.5.1 Sanathanan-Koerner (SK) Iteration - Monomial Basis

INPUT: Weights w_i, $i = 0, \ldots, m$, measurement vector $\mathbf{b} = [\mathcal{H}(\xi_1), \ldots, \mathcal{H}(\xi_m)]^\top$.

OUTPUT: Rational function $\widehat{\mathcal{H}}(s)$.

1. Set $\Delta^{(0)} = \mathrm{diag}[1, \ldots, 1] \in \mathbb{C}^{m \times m}$.

2. Assemble $\mathcal{A}$ from (2.5.11) as

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_0 & \mathcal{A}_1 \end{bmatrix}, \qquad [\mathcal{A}_0]_{i,j} = \xi_j^{i-1}, \qquad [\mathcal{A}_1]_{i,j} = -\mathcal{H}(\xi_j)\xi_j^{i-1} \tag{2.5.13}$$

3. For $k = 0, 1, 2, \ldots$, until convergence

   (a) Solve the least squares problem

$$\| \Delta^{(k)} \left(\mathcal{A}\mathbf{x}^{(k+1)} - \mathbf{b} \right) \|^2 \to \min \tag{2.5.14}$$

   (b) Assemble $\Delta^{(k+1)} = \mathrm{diag} \left\{ \dfrac{w_i}{d^{(k)}(\xi_i)} \right\}_{i=1}^{m}$

4. Assemble the approximant

$$\widehat{\mathcal{H}}(s) = \frac{\sum_{k=0}^{r-1} \alpha_k s^k}{\sum_{k=0}^{r-1} \beta_k s^k} \tag{2.5.15}$$

The SK iteration remains valid for other representations of the approximating function $\widehat{\mathcal{H}}(s)$. It is well known [22, 118] that the *barycentric* representation of rational functions has numerical advantages over the representation in (2.5.8), such as increased numerical stability

of function evaluation [70]. Thus instead of (2.5.8), it is standard to use the representation

$$\widehat{\mathcal{H}}(s) = \frac{n(s)}{d(s)} = \frac{\sum_{k=1}^{r} \dfrac{\psi_k}{s-\lambda_k}}{1 + \sum_{k=1}^{r} \dfrac{\varphi_k}{s-\lambda_k}}. \tag{2.5.16}$$

Here $\lambda_k \in \mathbb{C}$ are the poles of $n(s)$ and $d(s)$ (*not* of $\widehat{\mathcal{H}}(s)$), ψ_k and φ_k are referred to as *residues*, $k = 1, \ldots, r$ of $n(s)$ and $d(s)$, respectively. Even though in (2.5.8), both $n(s)$ and $d(s)$ are polynomials, in (2.5.16) they represent rational functions; we keep the notation $n(s)$ and $d(s)$ for the numerator and denominator functions of $\widehat{\mathcal{H}}(s)$. With $\widehat{\mathcal{H}}(s)$ as in (2.5.16), the least squares problem (2.5.10) has the left hand side matrix

$$\mathcal{A} = \begin{bmatrix} \dfrac{1}{\xi_1-\lambda_1} & \dfrac{1}{\xi_1-\lambda_2} & \cdots & \dfrac{1}{\xi_1-\lambda_r} & \dfrac{-\mathcal{H}(\xi_1)}{\xi_1-\lambda_1} & \dfrac{-\mathcal{H}(\xi_1)}{\xi_1-\lambda_2} & \cdots & \dfrac{-\mathcal{H}(\xi_1)}{\xi_1-\lambda_r} \\ \dfrac{1}{\xi_2-\lambda_1} & \dfrac{1}{\xi_2-\lambda_2} & \cdots & \dfrac{1}{\xi_2-\lambda_r} & \dfrac{-\mathcal{H}(\xi_2)}{\xi_2-\lambda_1} & \dfrac{-\mathcal{H}(\xi_2)}{\xi_2-\lambda_2} & \cdots & \dfrac{-\mathcal{H}(\xi_2)}{\xi_2-\lambda_r} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ \dfrac{1}{\xi_m-\lambda_1} & \dfrac{1}{\xi_m-\lambda_2} & \cdots & \dfrac{1}{\xi_m-\lambda_r} & \dfrac{-\mathcal{H}(\xi_m)}{\xi_m-\lambda_1} & \dfrac{-\mathcal{H}(\xi_m)}{\xi_m-\lambda_2} & \cdots & \dfrac{-\mathcal{H}(\xi_m)}{\xi_m-\lambda_r} \end{bmatrix}. \tag{2.5.17}$$

Since the poles λ_i $(i = 1, \ldots, r)$ are fixed a priori and remain unchanged throughout the SK-iteration, the unknowns are the residues for $d(s)$ and $n(s)$, namely $\mathbf{x} := [\boldsymbol{\psi}^\top,\ \boldsymbol{\varphi}^\top]^\top = [\phi, \ldots, \phi_r, \varphi_1, \ldots, \varphi_r]^\top \in \mathbb{C}^{2r}$. The implementation in Algorithm 2.5.1 extends by replacing $\mathcal{A}$ by the modified version in (2.5.16) for the barycentric form of $\widehat{\mathcal{H}}(s)$.

2.5.4 Vector Fitting

In the SK-iteration, the poles λ_i $(i = 1, \ldots, r)$ of the approximating function $\widehat{\mathcal{H}}(s)$ remain fixed. A natural next step is to choose the poles λ_i from (2.5.16) adaptively. Such a pole selection necessarily improves the conditioning of the linear least squares problem (2.5.14) [62].

The Vector Fitting algorithm [63] also transforms the non linear least squares problem (2.5.9) into a sequence of linear least squares problems with an additional updating of the poles λ_i in (2.5.16), using the zeros of the denominator $d(s) = 1 + \sum_{k=1}^{r} \frac{\varphi_k}{s - \lambda_k}$. Note that the zeros of $d(s)$ can be computed through an eigenvalue problem, detailed in the following lemma [51].

Lemma 2.5.4 *Let $d : \mathbb{C} \to \mathbb{C}$ be a rational function with simple zeros of the form*

$$d(s) = 1 + \sum_{i=1}^{n} \frac{\varphi_i}{s - \lambda_i}, \quad \textit{with} \quad \boldsymbol{\lambda} := [\lambda_1, \ldots, \lambda_n]^\top, \qquad \boldsymbol{\varphi} := [\varphi_1, \ldots, \varphi_n]^\top, \tag{2.5.18}$$

where $\varphi_i \neq 0$, $i = 1, \ldots, n$. Further define

$$\mathbf{A} := \operatorname{diag}(\boldsymbol{\lambda}) - \boldsymbol{\varphi}\mathbf{e}^\top = \begin{bmatrix} \lambda_1 - \boldsymbol{\varphi}_1 & -\boldsymbol{\varphi}_1 & \cdots & -\boldsymbol{\varphi}_1 & -\boldsymbol{\varphi}_1 \\ -\boldsymbol{\varphi}_2 & \lambda_2 - \boldsymbol{\varphi}_2 & \ldots & -\boldsymbol{\varphi}_2 & -\boldsymbol{\varphi}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\boldsymbol{\varphi}_{n-1} & -\boldsymbol{\varphi}_{n-1} & \cdots & \lambda_{n-1} - \boldsymbol{\varphi}_{n-1} & -\boldsymbol{\varphi}_{n-1} \\ -\boldsymbol{\varphi}_n & -\boldsymbol{\varphi}_n & \cdots & -\boldsymbol{\varphi}_n & \lambda_n - \boldsymbol{\varphi}_n \end{bmatrix}. \tag{2.5.19}$$

Then the (simple) eigenvalues of $\mathbf{A}$ are the (simple) zeros of $d(s)$.

With this preparation, we are now able to present Vector Fitting in Algorithm 2.5.2.

Algorithm 2.5.2 Vector Fitting (VF) Iteration

INPUT: Weights w_i, $i = 0, \ldots, m$, measurement vector $\mathbf{b} = [\mathcal{H}(\xi_1), \ldots, \mathcal{H}(\xi_m)]^\top$.

OUTPUT: Rational function $\widehat{\mathcal{H}}(s)$.

1. Make initial selection poles λ_i, for $i = 0, \ldots, r$ and set $\Delta = \mathrm{diag}\{w_1, \ldots, w_m\}$.

2. For $k = 1, 2, \ldots$ until convergence

   (a) Assemble $\mathcal{A}^{(k)}$ by (2.5.17)

   (b) Solve the least squares problem for $\mathbf{x} = \begin{bmatrix} \boldsymbol{\psi}^\top & \boldsymbol{\varphi}^\top \end{bmatrix}^\top$

   $$\| \Delta \left(\mathcal{A}^{(k)} \mathbf{x}^{(k+1)} - \mathbf{b} \right) \| \to \min \tag{2.5.20}$$

   (c) Find zeros z_i of $d^{(k)}(s)$ by Lemma 2.5.4, assign poles $\lambda_i^{(k+1)} \leftarrow z_i$, $i = 1, \ldots, r$

3. Assemble the approximant as

$$\widehat{\mathcal{H}}(s) = \frac{\sum_{k=1}^{r} \dfrac{\psi_k}{s - \lambda_k}}{1 + \sum_{k=1}^{r} \dfrac{\varphi_k}{s - \lambda_k}} \tag{2.5.21}$$

More details of the Vector Fitting framework can be found in [23, 30, 34, 35, 61]. It is worth noting that the convergence of Vector Fitting in general remains an open problem [55, 62, 68, 85, 107, 108]. In practice, we observe convergence in few iterations, even with poorly selected initial poles, but that is not always guaranteed. A choice of better conditioned basis functions, such as orthogonal or orthonormal rational functions [2, 36, 77] improves the numerical condition of (2.5.20) and the convergence rate of Algorithm 2.5.2.

2.5.5 Connection to Continuous $\mathcal{H}_2$ Spaces

In the previous section, the frequency sampling points ξ_i were given. When we are free to choose sampling points, e.g., by designing the physical experiment accordingly, there is a remarkable connection to continuous $\mathcal{H}_2$ norm approximation [40]. If the ξ_i are chosen according to a quadrature scheme with corresponding weights $w_i > 0$, the $\mathcal{H}_2$ norm can be approximated by

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} |\mathcal{H}(\mathrm{i}\omega)|^2 \mathrm{d}\omega \approx \sum_{k=1}^{m} w_k |\mathcal{H}(\xi_k)|^2. \tag{2.5.22}$$

Then Vector Fitting can be considered as minimizing a discretized $\mathcal{H}_2$ error measure

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} |\widehat{\mathcal{H}}(\mathrm{i}\omega) - \mathcal{H}(\mathrm{i}\omega)|^2 \mathrm{d}\omega \approx \sum_{k=1}^{m} w_k |\widehat{\mathcal{H}}(\xi_k) - \mathcal{H}(\xi_k)|^2, \tag{2.5.23}$$

where we select ξ_i according to, for example, Boyd / Clenshaw-Curtis points [25] with the corresponding weights w_i, $i = 1, \ldots, m$. For more details, see [40].

2.6 Parametric Model Reduction

Function evaluations $\mathcal{H}(s, p)$ often depend on parameters of the physical model which we represent with the linear, parametric dynamical system

$$\left.\begin{aligned} \mathbf{E}(p)\dot{\mathbf{x}}(t;p) &= \mathbf{A}(p)\mathbf{x}(t;p) + \mathbf{b}(p)u(t); \\ y(t;p) &= \mathbf{c}^\top(p)\mathbf{x}(t;p), \end{aligned}\right\} \quad \text{with} \quad \mathcal{H}(s,p) = \mathbf{c}^\top(p)\left(s\mathbf{E}(p) - \mathbf{A}(p)\right)^{-1}\mathbf{b}(p). \tag{2.6.1}$$

We want to reflect the parametric dependence in our reduced model of the same structure

$$\begin{aligned} \widehat{\mathbf{E}}(p)\dot{\widehat{\mathbf{x}}}(t;p) &= \widehat{\mathbf{A}}(p)\widehat{\mathbf{x}}(t;p) + \widehat{\mathbf{b}}(p)u(t); \\ \widehat{y}(t;p) &= \widehat{\mathbf{c}}^{\top}(p)\widehat{\mathbf{x}}(t;p), \end{aligned} \tag{2.6.2}$$

with transfer function

$$\widehat{\mathcal{H}}(s,p) = \widehat{\mathbf{c}}^{\top}(p)\left(s\widehat{\mathbf{E}}(p) - \widehat{\mathbf{A}}(p)\right)^{-1}\widehat{\mathbf{b}}(p). \tag{2.6.3}$$

In this section we review common approaches for finding low order approximations as in (2.6.2) in the literature. Some of the methods presented here extend concepts from non-parametric model reduction from Section 2.3. A complete survey of parametric model reduction methods is beyond the scope of this document, so we concentrate on relevant approaches in our context. A more extensive review can be found in [18] and the references therein.

To give an overview, we propose to group existing model reduction approaches into the following categories:

- Projection-based approaches using the system matrices $\mathbf{A}(p), \mathbf{E}(p), \mathbf{b}(p)$ and $\mathbf{c}(p)$ of the original system, possibly with a structure in the parametric dependence.
- Data driven approaches that
  1. Use the transfer function $(s,p) \mapsto \mathcal{H}(s,p)$.
  2. Use measurement data, i.e., a set of samples of $\mathcal{H}(s,p)$.

The latter two approaches differ in the number of $\mathcal{H}(s,p)$ evaluations we allow: measurement driven approaches find a parametric model given a finite and fixed set of data, while transfer function approaches allow access to the map $(s,p) \mapsto \mathcal{H}(s,p)$ that can be evaluated as much as necessary. The following subsections discuss each category in more detail.

2.6.1 Projection Based Methods

We assume a state-space description of $\mathcal{H}(s,p)$ is available, i.e., we have access to the system matrices $\mathbf{A}(p)$, $\mathbf{E}(p)$, $\mathbf{b}(p)$ and $\mathbf{c}(p)$. A common assumption is the following affine dependence of the system matrices in p:

$$\mathbf{E}(p) = \sum_{j=1}^{M} \eta_j(p)\mathbf{E}_j, \quad \mathbf{A}(p) = \sum_{j=1}^{M} \alpha_j(p)\mathbf{A}_j, \quad \mathbf{b}(p) = \sum_{j=1}^{M} \beta_j(p)\mathbf{b}_j, \quad \mathbf{c}(p) = \sum_{j=1}^{M} \gamma_j(p)\mathbf{c}_j, \tag{2.6.4}$$

with coefficient functions $\eta_j(p)$, $\alpha_j(p)$, $\beta_j(p)$ and $\gamma_j(p)$, $j = 1, \ldots, M$.

In case that the original model does not have known affine structure, one can construct an approximant of the form in (2.6.4) using, for example, the Empirical Interpolation Method (EIM) [7, 88] and its discrete counterpart, the Discrete Empirical Interpolation Method (DEIM) [28, 39]. Both methods iteratively select interpolation points based on a greedy search algorithm, resulting in a model that has affine dependence in p.

Similar to the non-parametric case, we assume the relevant dynamics of the system evolve in a lower dimensional subspace $\mathcal{V} \subset \mathbb{R}^n$, spanned by the columns of $\mathbf{V} \in \mathbb{R}^{n \times r}$. Following the Petrov-Galerkin projection, a *test subspace* $\mathbf{W}$ is chosen, along which the state space

matrices are projected. There are many approaches to construct $\mathbf{V}$ and $\mathbf{W}$ for parametric systems, see [18, 69, 100]

Using rational interpolation guarantees the model subspace to match the original transfer function at selected frequency and parameter values. For a given set of sampling points $\{p_1, \ldots, p_\ell\}$ in the parameter p and $\{\sigma_1, \ldots, \sigma_m\}$ in the frequency s, *local bases* can be constructed by

$$\begin{aligned}
\mathbf{V}_k &= \left[(\sigma_1 \mathbf{E}(p_k) - \mathbf{A}(p_k))^{-1} \mathbf{b}(p_k), \ldots, (\sigma_m \mathbf{E}(p_k) - \mathbf{A}(p_k))^{-1} \mathbf{b}(p_k) \right], \qquad k = 1, \ldots, \ell, \\
\mathbf{W}_k &= \left[(\sigma_1 \mathbf{E}(p_k) - \mathbf{A}(p_k))^{-\top} \mathbf{c}(p_k), \ldots, (\sigma_m \mathbf{E}(p_k) - \mathbf{A}(p_k))^{-\top} \mathbf{c}(p_k) \right], \qquad k = 1, \ldots, \ell.
\end{aligned} \tag{2.6.5}$$

Then the *global basis* $\mathbf{V}$ and $\mathbf{W}$ are assembled as

$$\mathbf{V} = [\mathbf{V}_1, \ldots, \mathbf{V}_\ell], \qquad \mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_\ell]. \tag{2.6.6}$$

The system matrices (2.6.4) are projected by

$$\begin{aligned}
\widehat{\mathbf{E}}(p) &= \sum_{j=1}^{M} \eta_j(p) \mathbf{W}^\top \mathbf{E}_j \mathbf{V}, & \widehat{\mathbf{A}}(p) &= \sum_{j=1}^{M} \alpha_j(p) \mathbf{W}^\top \mathbf{A}_j \mathbf{V}, \\
\widehat{\mathbf{b}}(p) &= \sum_{j=1}^{M} \beta_j(p) \mathbf{W}^\top \mathbf{b}_j, & \widehat{\mathbf{c}}(p) &= \sum_{j=1}^{M} \gamma_j(p) \mathbf{c}_j \mathbf{V}.
\end{aligned} \tag{2.6.7}$$

Note that (2.6.7) retains the original affine parametric structure of $\mathcal{H}(s, p)$.

The global basis approach (2.6.6) yields a reduced system $\widehat{\mathcal{H}}(s, p)$ that is a Hermite interpolant to the original transfer function at all frequency and parameter sampling points. In

particular, for all σ_i and p_j, it holds:

$$\begin{aligned} \widehat{\mathcal{H}}(\sigma_i, p_j) &= \mathcal{H}(\sigma_i, p_j), \\ \widehat{\mathcal{H}}'(\sigma_i, p_j) &= \mathcal{H}'(\sigma_i, p_j), \quad \text{and} \\ \frac{\partial}{\partial p}\widehat{\mathcal{H}}(\sigma_i, p_j) &= \frac{\partial}{\partial p}\mathcal{H}(\sigma_i, p_j). \end{aligned} \tag{2.6.8}$$

Such interpolation methods [11, 17, 24] can be extended to interpolate higher order derivatives as well.

2.6.2 Parametric Loewner Framework

The Loewner framework from Section 2.5.1 can be extended to the parametric case. We present the SISO version for sampling data $\{\xi_i, \mu_j, \mathcal{H}(\xi_i, \mu_j)\}_{i=1,j=1}^{i=n_s,j=n_p} \subset \mathbb{C} \times \mathbb{C} \times \mathbb{C}$. Consider the following partition into sets:

$$\begin{aligned} \{\xi_1, \ldots, \xi_{n_s}\} &= \{\lambda_1, \ldots, \lambda_{\overline{n_s}}\} \cup \left\{\mu_1, \ldots, \mu_{\underline{n_s}}\right\}, \quad \text{and} \\ \left\{\mu_1, \ldots, \mu_{n_p}\right\} &= \left\{\pi_1, \ldots, \pi_{\overline{n_p}}\right\} \cup \left\{\nu_1, \ldots, \nu_{\underline{n_p}}\right\}. \end{aligned} \tag{2.6.9}$$

We assume the partition (2.6.9) to be distinct and such that $\overline{n_s} + \underline{n_s} = n_s$ and $\overline{n_p} + \underline{n_p} = n_p$. The measurements $\mathcal{H}(\xi_i, \mu_j)$ are partitioned accordingly, denoted in the matrix $\mathbf{\Phi}$:

$$\mathbf{\Phi} = \left[\begin{array}{ccc|ccc} w_{1,1} & \cdots & w_{1,\overline{n_p}} & \phi_{1,\overline{n_p}+1} & \cdots & \phi_{1,n_p} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ w_{\overline{n_s},1} & \cdots & w_{\overline{n_s},\overline{n_p}} & \phi_{\overline{n_s},\overline{n_p}+1} & \cdots & \phi_{\overline{n_s},n_p} \\ \hline \phi_{\overline{n_s}+1,1} & \cdots & \phi_{\overline{n_s}+1,\overline{n_p}} & v_{1,1} & \cdots & v_{1,\underline{n_p}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \phi_{n_s,1} & \cdots & \phi_{n_s,\overline{n_p}} & v_{\underline{n_s},1} & \cdots & v_{\underline{n_s},\underline{n_p}} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{\Phi}_{1,1} & \mathbf{\Phi}_{2,1} \\ \hline \mathbf{\Phi}_{1,2} & \mathbf{\Phi}_{2,2} \end{array}\right]. \tag{2.6.10}$$

With this notation, we are able to present the two-variable Loewner interpolation result in the following lemma.

Lemma 2.6.1 (Lemma 5.1 & 5.2,[75],[74]) *Let* $\widehat{\mathcal{H}}(s,p)$ interpolate *the data set (2.6.10). The* $\widehat{\mathcal{H}}(s,p)$ *can be written in state-space form with a parametric dependence in* $\widehat{\mathbf{c}}(p)$ *and* $\widehat{\mathbf{A}}(p)$ *as*

$$\widehat{\mathcal{H}}(s,p) = \widehat{\mathbf{c}}^{\top}(p)\left(s\widehat{\mathbf{E}} - \widehat{\mathbf{A}}(p)\right)^{-1}\widehat{\mathbf{b}}, \tag{2.6.11}$$

with system matrices

$$s\widehat{\mathbf{E}} - \widehat{\mathbf{A}}(p) = \begin{bmatrix} s-\lambda_1 & \lambda_2 - s & & \\ \vdots & & \ddots & \\ s-\lambda_1 & 0 & & \lambda_{k+1}-s \\ \alpha_1(p) & \dots & & \alpha_{k+1}(p) \end{bmatrix}, \quad \widehat{\mathbf{b}} = \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \widehat{\mathbf{c}}(p) = \begin{bmatrix} \beta_1(p) \\ \vdots \\ \beta_{k+1}(p) \end{bmatrix}. \tag{2.6.12}$$

We can restrict the parametric dependence to $\widehat{\mathbf{A}}(p)$ *and construct*

$$\widehat{\mathcal{H}}(s,p) = \widehat{\mathbf{c}}^{\top}\left(s\widehat{\mathbf{E}} - \widehat{\mathbf{A}}(p)\right)^{-1}\widehat{\mathbf{b}}, \tag{2.6.13}$$

with the system matrices (we use the same notation as in (2.6.12), although the matrices are different)

$$s\widehat{\mathbf{E}} - \widehat{\mathbf{A}}(p) = \begin{bmatrix} \mathbf{J}_{s,\lambda,k} & 0 & 0 \\ \mathbb{A} & \mathbf{J}^{*}_{p,\pi,q} & 0 \\ \mathbb{B} & 0 & [\mathbf{J}^{*}_{p,\pi,q}, \tau] \end{bmatrix} \quad \widehat{\mathbf{b}} = \begin{bmatrix} 0 \\ \tau \\ 0 \end{bmatrix}, \quad \widehat{\mathbf{c}} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, \tag{2.6.14}$$

and the factors $\mathbb{A}$, $\mathbb{D}$ *and* τ *as*

$$\mathbb{A}_i = \begin{bmatrix} c_{i,1} \\ \vdots \\ c_{i,q+1} \end{bmatrix}, \quad \mathbb{B}_i = \begin{bmatrix} c_{i,1} w_{i,1} \\ \vdots \\ c_{i,q+1} w_{i,q+1} \end{bmatrix}, \quad \tau_i = \left(\prod_{j=1, j \neq i}^{q+1} (\pi_i - \pi_j) \right)^{-1}. \tag{2.6.15}$$

We emphasize that our goal is a selection of frequency and parameter sampling points (σ_i, p_j) that yields an optimal reduced model with respect to a *joint* error norm. Now that all the necessary background has been introduced, we continue with our investigation of parametric rational interpolation in the following chapter.

Chapter 3

Jointly Optimal Approximation in Frequency and Parameter

The main focus of this chapter is on rational interpolation methods for parametric dynamical systems. Specifically we investigate linear, finite dimensional, time invariant, parametric dynamical systems. For a joint error measure with respect to frequency and parameter, interpolation conditions are derived that characterize an optimal reduced model. The problem of finding such a model is addressed via a gradient descent algorithm.

We start by introducing norms and notation to precisely formulate the parametric approximation problem. Then the appropriate function spaces in the two-variable setting are defined with connections to linear system theory in Section 3.2. In Section 3.3, we derive optimality conditions extending the Meier-Luenberger conditions [91] to the parametric case. Examples

and details on the implementation of our algorithm are provided in Section 3.4, followed by a summary of contributions and an outlook on future work in Section 3.5.

3.1 Problem Setting

In this chapter, we focus on a parametric, linear dynamical system of the form

$$\begin{aligned} \mathbf{E}(p)\dot{\mathbf{x}}(t;p) &= \mathbf{A}(p)\mathbf{x}(t;p) + \mathbf{b}(p)u(t), \\ y(t;p) &= \mathbf{c}^\top(p)\mathbf{x}(t;p), \end{aligned} \tag{3.1.1}$$

with bounded parameter domain $\mathcal{P} \subsetneq \mathbb{C}$ and analytic matrix functions $\mathbf{A}, \mathbf{E} : \mathcal{P} \to \mathbb{R}^{n\times n}$ and $\mathbf{b}, \mathbf{c} : \mathcal{P} \to \mathbb{R}^n$. For simplicity, we assume $\mathbf{E}$ to be invertible for all $p \in \mathcal{P}$. We restrict ourselves to the case that $p \in \mathcal{P}$ is a scalar but all the following results in this chapter can be generalized to multiple parameters, i.e., $\mathcal{P} \subseteq \mathbb{C}^k$. Note that we do not require any particular structure in the parametric dependency appearing in (3.1.1).

Remark 3.1.1 Dynamical systems commonly carry a *feed-through term* $d(p)$, so that $y(t;p)$ in (3.1.1) becomes

$$y(t;p) = \mathbf{c}^\top(p)\mathbf{x}(t;p) + d(p)u(t).$$

The transfer function then becomes

$$\mathcal{H}(s;p) = \mathbf{c}^\top(p)\left(s\mathbf{E}(p) - \mathbf{A}(p)\right)^{-1}\mathbf{b}(p) + d(p),$$

and for $\|\mathcal{H}\|_{\mathcal{H}_2} < \infty$ we need $d(p) = 0$ for each $p \in \mathcal{P}$. For a reduced model $\widehat{\mathcal{H}}(s;p)$, it is necessary that $\widehat{d}(p) = d(p)$ $(p \in \mathcal{P})$ to ensure that $\|\widehat{\mathcal{H}} - \mathcal{H}\|_{\mathcal{H}_2} < \infty$. Hence we omit $d(p)$ in our model reduction approach.

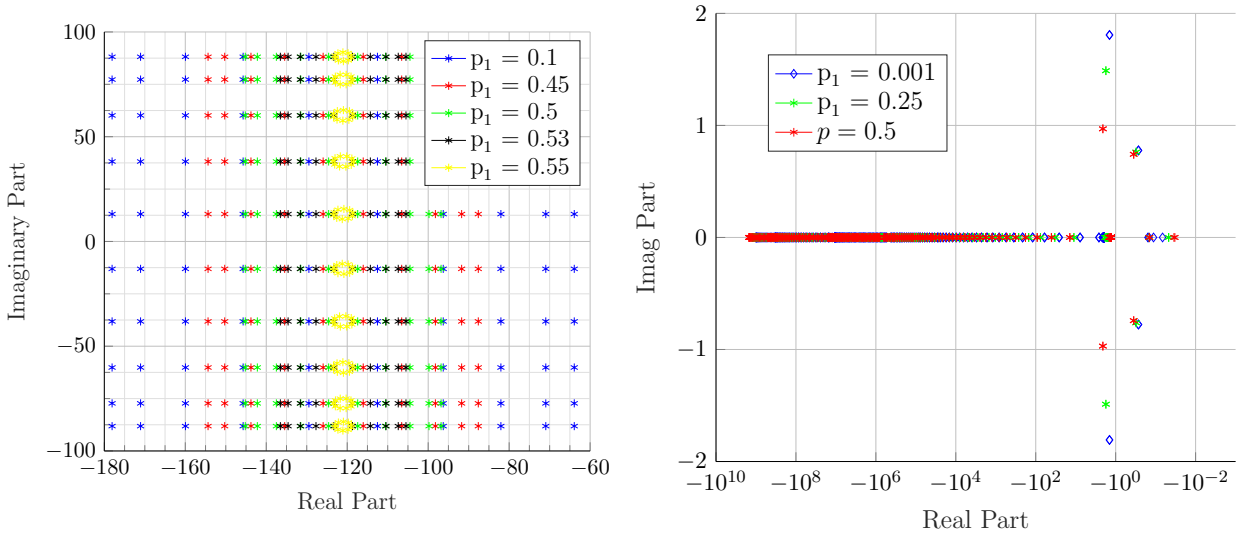
3.1.1 Parametric Dynamical Systems

Stability of non-parametric dynamical systems can be characterized through the eigenvalues of the matrix pencil involving the matrices $\mathbf{A}$ and $\mathbf{E}$. For parametric systems, an extension is provided in the following definition.

Definition 3.1.2 A parametric dynamical system $(\mathbf{E}(p), \mathbf{A}(p), \mathbf{b}(p), \mathbf{c}(p))$ with the parameter $p \in \mathcal{P}$, is said to be *uniformly stable* if the spectrum $\sigma(\mathbf{A}(p), \mathbf{E}(p)) \in \mathbb{C}_L$ for all $p \in \mathcal{P}$.

Recall that, for a fixed $p \in \mathcal{P}$, if $(\mathbf{A}(p), \mathbf{E}(p), \mathbf{b}(p), \mathbf{c}(p))$ is a minimal realization of $\mathcal{H}(s, p)$, the eigenvalues $\sigma(\mathbf{A}(p), \mathbf{E}(p))$ are precisely the poles of $\mathcal{H}(s, p)$. We illustrate how the eigenvalues of the pencil $(\mathbf{A}(p), \mathbf{E}(p))$ can change with p in the following two examples. Keep in mind that the eigenvalues move continuously with $p \in \mathcal{P}$, since we assume that $\mathbf{A}(p)$ and $\mathbf{E}(p)$ are analytic on $\mathcal{P}$ and $\mathbf{E}(p)$ is invertible for all $p \in \mathcal{P}$.

Consider the convection-diffusion example from Section 1.6.2 and the vibrating beam problem from Section 1.6.1. In Figure 3.1, we show the pole movement of the eigenvalues $\sigma(\mathbf{A}(p)\mathbf{E}(p))$ for selected values of $p \in \mathcal{P} = [0, 1]$.



(a) Convection Diffusion Model, $n = 100$

(b) Vibrating Beam Model, $n = 600$

Figure 3.1: Eigenvalues of $(\mathbf{A}(p), \mathbf{E}(p))$ for two example models with parameter $p \in [0, 1]$.

Observe that all poles displayed in Figure 3.1 are in the left half plane, suggesting the system is uniformly stable by Definition 3.1.2. Obviously Figure 3.1 only serves as an illustration; we need to verify Definition 3.1.2 analytically to claim that a system is, in fact, stable.

3.1.2 Topics to be Discussed

A brief review of common model reduction approaches for parametric dynamical systems is provided in Section 2.6. Most methods interpolate local reduced models (see [12, 50]) using various basis functions (Lagrange interpolation, for example). The selection of parameter values at which local models are constructed is done heuristically or based on measurement

data. In contrast, we introduce a joint frequency/parameter-based error measure that governs the choice of parameter interpolation points. Given $\mathcal{H}(s,p)$, the goal is to construct a parametric model $\widehat{\mathcal{H}}(s,p)$ that is easier to evaluate than $\mathcal{H}(s,p)$ and so that

$$\|\mathcal{H} - \widehat{\mathcal{H}}\| \to \min, \tag{3.1.2}$$

where $\|\cdot\|$ denotes a suitable two-variable norm that is introduced in the following section. We start by presenting the underlying spaces as a basis for a joint optimality measure leading to the parametric norm in (3.1.2).

3.2 Hardy Spaces in Several Variables

Transfer functions of non-parametric systems are contained in Hardy spaces; see Definition 2.2.3 and Definition 2.2.4. To make the approximation problem (3.1.2) precise, we need to clarify the underlying function spaces for the parametric case. Multi-variable Hardy spaces have been subject to investigation, see e.g., [32, 38, 49] for an overview. A natural extension of Hardy spaces to multiple variables, similar to [89], is the following definition, extending $\mathcal{H}_2(\mathbb{C}_R)$ using the unit disc as the parameter domain:

$$\begin{aligned}\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D}) = \Big\{ \mathcal{H} :\ \mathbb{C}^2 \to \mathbb{C} :\ \mathcal{H} \text{ analytic in } \mathbb{C}_R \times \mathbb{D}, \text{ and} \\ \sup_{0<s<1,\ x>0} \int_{-\infty}^{\infty} \int_{0}^{2\pi} \left|\mathcal{H}(x + \mathrm{i}y, s\mathrm{e}^{\mathrm{i}\theta})\right|^2 \mathrm{d}\theta \,\mathrm{d}y < \infty \Big\}.\end{aligned} \tag{3.2.1}$$

We think of the first variable as the frequency and the second variable as a parameter.

The inner product on $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ is defined as follows.

Definition 3.2.1 For two functions $\mathcal{G}$ and $\mathcal{H}$ in $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$, let

$$\langle \mathcal{G}, \mathcal{H} \rangle_{\otimes} := \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_0^{2\pi} \overline{\mathcal{G}(\mathrm{i}\omega, \mathrm{e}^{\mathrm{i}\theta})} \mathcal{H}(\mathrm{i}\omega, \mathrm{e}^{\mathrm{i}\theta}) \, \mathrm{d}\theta \, \mathrm{d}\omega. \tag{3.2.2}$$

Observe that this inner product makes $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ a Hilbert space. The corresponding induced norm is defined as

$$\|\mathcal{H}\|_{\otimes} := \sqrt{\langle \mathcal{H}, \mathcal{H} \rangle_{\otimes}} = \left(\frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_0^{2\pi} \left| \mathcal{H}(\mathrm{i}\omega, \mathrm{e}^{\mathrm{i}\theta}) \right|^2 \, \mathrm{d}\theta \, \mathrm{d}\omega \right)^{1/2}. \tag{3.2.3}$$

We make use of a standard tensor product construction [102, Chap. 2] to define $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$:

$$\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D}) := \mathcal{H}_2(\mathbb{C}_R) \otimes \mathcal{H}_2(\mathbb{D}). \tag{3.2.4}$$

Let $\left\{ F_k^{(\mathbb{C}_R)} \right\}$ be a countable dense subset for $\mathcal{H}_2(\mathbb{C}_R)$ and $\left\{ F_k^{(\mathbb{D})} \right\}$ one for $\mathcal{H}_2(\mathbb{D})$ (which exist since both spaces are separable). From (3.2.4) it is obvious that

$$\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D}) = \overline{\mathrm{Span}_{\substack{i=1,2,\ldots \\ j=1,2,\ldots}} \left(F_i^{(\mathbb{C}_R)} \cdot F_j^{(\mathbb{D})} \right)}, \tag{3.2.5}$$

so the product of the individual bases is dense in $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$. Note that $\mathcal{H}_2(\mathbb{D})$ is a closed subspace of $L_2(\partial\mathbb{D})$. More precisely $L_2(\partial\mathbb{D}) = \mathcal{H}_2(\mathbb{D}) \oplus \mathcal{H}_2(\mathbb{C} \backslash \mathbb{D})$ [80], with the corresponding inclusion $\mathcal{H}_2(\mathbb{D}) \hookrightarrow L_2(\partial\mathbb{D})$. To emphasize the distinction between frequency and parameter variable, let us introduce the short notation $\mathcal{H}_2 \otimes L_2$ for $\mathcal{H}_2(\mathbb{C}_R) \otimes \mathcal{H}_2(\mathbb{D})$.

Let us further introduce the notion of *rational Hardy spaces* in two variables:

$$\mathcal{RH}_2(\mathbb{C}_R \times \mathbb{D}) := \{ \mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D}) : \quad \mathcal{H} \ \text{ is rational} \}. \tag{3.2.6}$$

In this chapter, we focus on $\mathcal{P} = \mathbb{D}$, giving us the advantage of a compact parameter domain. Although many examples give no rise to a physical meaning of a complex value of

the parameter away from $[-1,1]$, this approach enables the use of contour integrals and the residue theorem and yields good approximation results also for purely real parameters, e.g. $\mathcal{P} = [-1,1]$.

We approximate a given two variable rational function $\mathcal{H}(s,p)$ with

$$\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}, \qquad \begin{array}{l} \lambda_i \in \mathbb{C}_L \ (i = 1,\ldots,r_s), \\ \pi_j \in \mathbb{C}\backslash\overline{\mathbb{D}} \ (j = 1,\ldots,r_p). \end{array} \tag{3.2.7}$$

The inner product in (3.2.2) can be expensive to evaluate for general functions $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$, since a suitable discretization of the integration domains in (3.2.3) is needed, resulting in a large number of function evaluations of $\mathcal{H}(s,p)$. For functions of the form (3.2.7) we can find an explicit formula similar to the nonparametric case in Lemma 2.2.8, which aids in deriving the joint optimality conditions for parametric reduced models in the following sections.

3.2.1 Basis Functions and Approximation

To illustrate Definition 3.2.1, consider

$$\mathcal{B}_{ij}(s,p) := \frac{1}{(s-\lambda_i)(p-\pi_j)}, \qquad \lambda_i \in \mathbb{C}_L, \ \pi_j \in \mathbb{C}\backslash\mathbb{D}, \tag{3.2.8}$$

where the indices i and j are introduced here, since we aim to consider functions in the span of such $\mathcal{B}_{ij}$. The following lemma derives closed form expressions for the $\mathcal{H}_2 \otimes L_2$ inner product and norm.

Lemma 3.2.2 *Let $\mathcal{B}_{ij}$ and $\mathcal{B}_{k\ell}$ be as in (3.2.8). The $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ inner product and norm are given by*

$$\langle \mathcal{B}_{ij},\ \mathcal{B}_{k\ell} \rangle_{\otimes} = \frac{1}{(1 - \overline{\pi}_j \pi_\ell)(\overline{\lambda_i} + \lambda_k)}, \quad \textit{and} \quad \|\mathcal{B}_{ij}\|_{\otimes} = \frac{1}{\sqrt{(1 - |\pi_j|^2) \cdot 2\Re(\lambda_i)}}, \tag{3.2.9}$$

where $\Re(\cdot)$ denotes the real part.

Proof The proof relies on the residue theorem (see, for example, [48]). Since the denominator is separable in s and p, we have

$$\begin{aligned} \langle \mathcal{B}_{ij},\ \mathcal{B}_{k\ell} \rangle_{\otimes} &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_0^{2\pi} \frac{1}{\overline{\mathrm{e}^{\mathrm{i}\theta} - \pi_j}} \frac{1}{\mathrm{e}^{\mathrm{i}\theta} - \pi_\ell} \frac{1}{\overline{(\mathrm{i}\omega - \lambda_i)}} \frac{1}{(\mathrm{i}\omega - \lambda_k)} \,\mathrm{d}\theta \,\mathrm{d}\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\overline{(\mathrm{i}\omega - \lambda_i)}} \frac{1}{(\mathrm{i}\omega - \lambda_k)} \,\mathrm{d}\omega \cdot \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{\overline{\mathrm{e}^{\mathrm{i}\theta} - \pi_j}} \frac{1}{\mathrm{e}^{\mathrm{i}\theta} - \pi_\ell} \,\mathrm{d}\theta. \end{aligned} \tag{3.2.10}$$

Beginning with the left integral, let Γ_R be a semi-circular contour in the right half plane with radius R, centered at 0 and oriented so that the imaginary axis is traced upwards in positive imaginary direction. Then the integral in ω is

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\overline{(\mathrm{i}\omega - \lambda_i)}} \frac{1}{(\mathrm{i}\omega - \lambda_k)} \,\mathrm{d}\omega &= \frac{1}{2\pi\mathrm{i}} \int_{-\infty}^{\infty} \frac{1}{-w - \overline{\lambda_i}} \frac{1}{(w - \lambda_k)} \,\mathrm{d}w \\ &= \lim_{R \to \infty} \frac{1}{2\pi\mathrm{i}} \int_{\Gamma_R} \frac{1}{-w - \overline{\lambda_i}} \frac{1}{(w - \lambda_k)} \,\mathrm{d}w \\ &= -\operatorname{Res}\left[\frac{1}{-w - \overline{\lambda_i}} \frac{1}{(w - \lambda_k)},\ w = -\overline{\lambda_i} \right] \\ &= -\lim_{w \to -\overline{\lambda_i}} (w + \overline{\lambda_i}) \frac{1}{-w - \overline{\lambda_i}} \frac{1}{(w - \lambda_k)} \\ &= \frac{-1}{\overline{\lambda_i} + \lambda_k}. \end{aligned} \tag{3.2.11}$$

Note that we gather a minus sign since the orientation of Γ_R is opposed to the standard definition for residue contours, which are traversed in mathematically positive direction

(exterior of the contour to the right). For the integral in θ, we note that $\overline{\mathrm{e}^{\mathrm{i}\theta}} = e^{-\mathrm{i}\theta}$, so that for $z = \mathrm{e}^{\mathrm{i}\theta}$, $\overline{z} = z^{-1}$. Then

$$\begin{aligned}
\frac{1}{2\pi}\int_0^{2\pi} \frac{1}{\overline{\mathrm{e}^{\mathrm{i}\theta} - \pi_j}} \frac{1}{\mathrm{e}^{\mathrm{i}\theta} - \pi_\ell}\,\mathrm{d}\theta &= \frac{1}{2\pi\mathrm{i}}\int_{\partial\mathbb{D}} \frac{1}{z}\frac{1}{z^{-1} - \overline{\pi_j}}\frac{1}{z - \pi_\ell}\,\mathrm{d}z \\
&= \mathrm{Res}\left[\frac{1}{1 - z\overline{\pi_j}}\frac{1}{z - \pi_\ell},\ z = \overline{\pi_j}^{-1}\right] \\
&= \lim_{z\to\overline{\pi_j}^{-1}} (z - \overline{\pi_j}^{-1})\frac{1}{1 - z\overline{\pi_j}}\frac{1}{z - \pi_\ell} \qquad (3.2.12) \\
&= \lim_{z\to\overline{\pi_j}^{-1}} \frac{\overline{\pi_j}z - 1}{\overline{\pi_j}}\frac{1}{1 - z\overline{\pi_j}}\frac{1}{z - \pi_\ell} \\
&= \frac{-1}{\overline{\pi_j}}\frac{1}{\overline{\pi_j}^{-1} - \pi_\ell} = \frac{1}{\overline{\pi_j}\pi_\ell - 1}.
\end{aligned}$$

Combining (3.2.11) and (3.2.12) in (3.2.10), we get

$$\langle \mathcal{B}_{ij}, \mathcal{B}_{k\ell}\rangle = \frac{1}{(1 - \overline{\pi}_j\pi_\ell)(\overline{\lambda_i} + \lambda_k)},$$

which is the desired formula. Further, for the norm:

$$\|\mathcal{B}_{ij}\|_\otimes^2 = \left\langle \frac{1}{(s - \lambda_i)(p - \pi_j)}\ ,\ \frac{1}{(s - \lambda_i)(p - \pi_j)}\right\rangle_\otimes \qquad (3.2.13)$$

$$= \frac{1}{(1 - \overline{\pi}_j\pi_j)(\overline{\lambda_i} + \lambda_i)} = \frac{1}{(1 - |\pi_j|^2)\cdot 2\Re(\lambda_i)}. \qquad (3.2.14)$$

■

By conjugate linearity of the inner product, the explicit formula for inner products on the elementary functions $\mathcal{B}_{i,j}(s,p)$ in (3.2.9) can be extended to determine the inner product between functions of the form in (3.2.7).

Corollary 3.2.3 *Let $\mathcal{G}, \mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R)\otimes\mathcal{H}_2(\mathbb{D})$ be of the form*

$$\mathcal{G}(s,p) = \sum_{i=1}^{n_s}\sum_{j=1}^{n_p} \frac{\phi_{i,j}^{(\mathcal{G})}}{\left(s - \lambda_i^{(\mathcal{G})}\right)\left(p - \pi_j^{(\mathcal{G})}\right)}, \quad \textit{and} \quad \mathcal{H}(s,p) = \sum_{k=1}^{r_s}\sum_{\ell=1}^{r_p} \frac{\phi_{k,\ell}^{(\mathcal{H})}}{\left(s - \lambda_k^{(\mathcal{H})}\right)\left(p - \pi_\ell^{(\mathcal{H})}\right)}, \qquad (3.2.15)$$

with simple poles $\lambda_i^{(\mathcal{G})}, \lambda_k^{(\mathcal{H})}, \pi_j^{(\mathcal{G})}$ *and* $\pi_\ell^{(\mathcal{H})}$*. The* $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ *inner product then is*

$$\langle \mathcal{G},\ \mathcal{H} \rangle_\otimes = \sum_{i=1}^{n_s} \sum_{j=1}^{n_p} \sum_{k=1}^{r_s} \sum_{\ell=1}^{r_p} \frac{\overline{\phi_{i,j}^{(\mathcal{G})}}\, \phi_{k,\ell}^{(\mathcal{H})}}{\left(\left(\overline{\pi_j^{(\mathcal{G})}} \right)^{-1} \pi_\ell^{(\mathcal{H})} \right) \left(\overline{\lambda_i^{(\mathcal{G})}} + \lambda_k^{(\mathcal{H})} \right)}. \tag{3.2.16}$$

Further, the norm is

$$\|\mathcal{G}(s,p)\|_\otimes = \left(\sum_{i=1}^{n_s} \sum_{j=1}^{n_p} \sum_{k=1}^{r_s} \sum_{\ell=1}^{r_p} \frac{\overline{\phi_{i,j}^{(\mathcal{G})}}\, \phi_{k,\ell}^{(\mathcal{G})}}{\left(\left(\overline{\pi_j^{(\mathcal{G})}} \right)^{-1} \pi_\ell^{(\mathcal{G})} \right) \left(\overline{\lambda_i^{(\mathcal{G})}} + \lambda_k^{(\mathcal{G})} \right)} \right)^{1/2}. \tag{3.2.17}$$

While such functions span the space we want to use for the reduced (target) model, it is a subspace of $\mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ (see (3.2.4)), the space of transfer functions we wish to approximate. To proceed, we require the inner product between a reduced transfer functions in $\operatorname{span}(\mathcal{B}_{i,j})$ and a transfer function $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$. For $\mathcal{B}_{ij}(s,p)$ from (3.2.8), the $\mathcal{H}_2 \otimes L_2$ inner product has a simple expression, derived in the following lemma.

Lemma 3.2.4 *Let* $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ *and* $\mathcal{B}_{ij}(s,p)$ *as in (3.2.8). Then it holds*

$$\langle \mathcal{B}_{ij},\ \mathcal{H} \rangle_\otimes = \frac{-1}{\overline{\pi_j}} \mathcal{H}\left(-\overline{\lambda}_i, \frac{1}{\overline{\pi_j}} \right). \tag{3.2.18}$$

Proof The proof is performed, again, by using the residue theorem. We note that all poles of $\mathcal{H}(s,p)$ lie in the left half plane in s and outside the unit disc in p. By the definition of the $\mathcal{H}_2 \otimes L_2$ inner product, we need to compute

$$\langle \mathcal{B}_{ij},\ \mathcal{H} \rangle_\otimes = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_0^{2\pi} \frac{1}{(\mathrm{i}\omega - \lambda_i)(\mathrm{e}^{\mathrm{i}\theta} - \pi_j)} \mathcal{H}\left(\mathrm{i}\omega, \mathrm{e}^{\mathrm{i}\theta} \right) \mathrm{d}\theta\, \mathrm{d}\omega. \tag{3.2.19}$$

First, for the integral in p, let ω be fixed. Then

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{\overline{(\mathrm{e}^{\mathrm{i}\theta}-\pi_j)}}\mathcal{H}\left(\mathrm{i}\omega,\mathrm{e}^{\mathrm{i}\theta}\right)\,\mathrm{d}\theta &= \frac{1}{2\pi\mathrm{i}}\int_{\partial\mathbb{D}}\frac{1}{1-z\overline{\pi_j}}\mathcal{H}(\mathrm{i}\omega,z)\,\mathrm{d}z\\ &= \mathrm{Res}\left[\frac{1}{1-z\overline{\pi_j}}\mathcal{H}(\mathrm{i}\omega,z),\ z=\overline{\pi_j}^{-1}\right]\\ &= \lim_{z\to\overline{\pi_j}^{-1}}\frac{z-\overline{\pi_j}^{-1}}{1-z\overline{\pi_j}}\mathcal{H}(\mathrm{i}\omega,z)\\ &= \frac{-1}{\overline{\pi_j}}\mathcal{H}(\mathrm{i}\omega,\overline{\pi_j}^{-1}).\end{aligned} \tag{3.2.20}$$

Let Γ_R be a semi-circular contour in the right half plane. Substituting (3.2.20) into (3.2.19) we get:

$$\begin{aligned}\langle\mathcal{B}_{ij},\ \mathcal{H}\rangle_{\otimes} &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_0^{2\pi}\frac{1}{\overline{(\mathrm{i}\omega-\lambda_i)(\mathrm{e}^{\mathrm{i}\theta}-\pi_j)}}\mathcal{H}\left(\mathrm{i}\omega,\mathrm{e}^{\mathrm{i}\theta}\right)\,\mathrm{d}\theta\,\mathrm{d}\omega\\ &= \frac{-1}{\overline{\pi_j}}\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{\overline{\mathrm{i}\omega-\lambda_i}}\mathcal{H}(\mathrm{i}\omega,\overline{\pi_j}^{-1})\,\mathrm{d}\omega\\ &= \frac{-1}{\overline{\pi_j}}\frac{1}{2\pi\mathrm{i}}\int_{\Gamma_R}\frac{1}{-w-\overline{\lambda_i}}\mathcal{H}(w,\overline{\pi_j}^{-1})\,\mathrm{d}w\\ &= \frac{1}{\overline{\pi_j}}\,\mathrm{Res}\left[\frac{1}{-w-\overline{\lambda_i}}\mathcal{H}(w,\overline{\pi_j}^{-1}),\ w=-\overline{\lambda_i}\right]\\ &= \frac{1}{\overline{\pi_j}}\lim_{w\to-\overline{\lambda_i}}\frac{w+\overline{\lambda_i}}{-w-\overline{\lambda_i}}\mathcal{H}(w,\overline{\pi_j}^{-1}) = \frac{-1}{\overline{\pi_j}}\mathcal{H}(-\overline{\lambda_i},\overline{\pi_j}^{-1}),\end{aligned} \tag{3.2.21}$$

which is the desired expression in (3.2.18).

■

Using the conjugate linearity of the $\mathcal{H}_2\otimes L_2$ inner product, a direct consequence of the previous lemma is the following corollary.

Corollary 3.2.5 *Let* $\mathcal{H}\in\mathcal{H}_2(\mathbb{C}_R\otimes L_2)$ *and* $\mathcal{G}(s,p)=\sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\dfrac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}$ *with*

simple poles λ_i, $i = 1, \ldots, r_s$, π_j, $j = 1, \ldots, r_p$. *Then*

$$\langle \mathcal{G},\ \mathcal{H} \rangle_{\otimes} = \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{-\overline{\phi_{i,j}}}{\overline{\pi_j}} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}). \tag{3.2.22}$$

Since we aim to extend the classical $\mathcal{H}_2(\mathbb{C}_R)$ case, we recover the classic $\mathcal{H}_2$ norm if $\mathcal{H}(s,p)$ does not depend on p. Let $\mathcal{H}(s,p) = \mathcal{H}(s)$, then

$$\|\mathcal{H}\|_{\otimes}^2 = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_0^{2\pi} |\mathcal{H}(\mathrm{i}\omega)|^2 \,\mathrm{d}\theta \,\mathrm{d}\omega = \frac{2\pi}{4\pi^2} \int_{-\infty}^{\infty} |\mathcal{H}(\mathrm{i}\omega)|^2 \,\mathrm{d}\omega = \|\mathcal{H}\|_{\mathcal{H}_2(\mathbb{C}_R)}^2. \tag{3.2.23}$$

The following corollary is useful in deriving the optimality conditions in Section 3.3.

Corollary 3.2.6 *Let* $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$. *Then*

$$\begin{aligned}
\left\langle \frac{1}{(s-\lambda_i)^2(p-\pi_j)}, \mathcal{H}(s,p) \right\rangle_{\otimes} &= \frac{-1}{\overline{\pi}_j} \frac{\partial}{\partial s} \mathcal{H}(s, \overline{\pi}_j^{-1}) \Big|_{s=-\overline{\lambda}_i}, \quad \textit{and} \\
\left\langle \frac{1}{(s-\lambda_i)(p-\pi_j)^2}, \mathcal{H}(s,p) \right\rangle_{\otimes} &= \frac{-1}{\overline{\pi}_j^2} \mathcal{H}(-\overline{\lambda}_i, \overline{\pi}_j^{-1}) + \frac{1}{\overline{\pi}_j^3} \frac{\partial}{\partial p} \mathcal{H}(-\overline{\lambda}_i, p) \Big|_{p=\overline{\pi}_j^{-1}}.
\end{aligned} \tag{3.2.24}$$

Proof The residue for double poles can be computed as follows. If λ_i is a double pole of a meromorphic function $H(s)$, then

$$\mathrm{Res}\left[H(z),\ z=\lambda\right] = \lim_{z \to \lambda} \frac{d}{dz} \left((z-\lambda)^2 H(z) \right). \tag{3.2.25}$$

For the inner product in Corollary 3.2.6 with a double pole in p, we compute

$$\begin{aligned}
\left\langle \frac{1}{(s-\lambda_i)(p-\pi_j)^2}, \mathcal{H}(s,p) \right\rangle_{\otimes} &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_0^{2\pi} \frac{1}{(\mathrm{i}\omega - \lambda_i)(\mathrm{e}^{\mathrm{i}\theta} - \pi_j)^2} \mathcal{H}(\mathrm{i}\omega, \mathrm{e}^{\mathrm{i}\theta}) \,\mathrm{d}\theta \,\mathrm{d}\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\mathrm{i}\omega - \lambda_i} \underbrace{\frac{1}{2\pi} \int_0^{2\pi} \frac{1}{(\mathrm{e}^{\mathrm{i}\theta} - \pi_j)^2} \mathcal{H}(\mathrm{i}\omega, \mathrm{e}^{\mathrm{i}\theta}) \,\mathrm{d}\theta}_{=: \mathcal{I}_p} \,\mathrm{d}\omega
\end{aligned} \tag{3.2.26}$$

For the inner integral $\mathcal{I}_p$ in p, we compute

$$\begin{aligned}
\mathcal{I}_p &= \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{\overline{(\mathrm{e}^{\mathrm{i}\theta}-\pi_j)^2}}\mathcal{H}(\mathrm{i}\omega,\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\theta \\
&= \frac{1}{2\pi\mathrm{i}}\int_{\Gamma_R} \frac{1}{z}\frac{1}{(z^{-1}-\overline{\pi_j})^2}\mathcal{H}(\mathrm{i}\omega,z)\,\mathrm{d}z \\
&= \operatorname{Res}\left[\frac{1}{z}\frac{1}{(z^{-1}-\overline{\pi_j})^2}\mathcal{H}(\mathrm{i}\omega,z),\ z=-\overline{\pi_j}^{-1}\right] + \operatorname{Res}\left[\frac{1}{z}\frac{1}{(z^{-1}-\overline{\pi_j})^2}\mathcal{H}(\mathrm{i}\omega,z),\ z=0\right] \\
&= \lim_{z\to\overline{\pi_j}^{-1}}\frac{\partial}{\partial z}\left[\frac{1}{z}\frac{(z-\overline{\pi_j}^{-1})^2}{(z^{-1}-\overline{\pi_j})^2}\mathcal{H}(\mathrm{i}\omega,z)\right] + \lim_{z\to 0}\left[\frac{z}{z}\frac{1}{(z^{-1}-\overline{\pi_j})^2}\mathcal{H}(\mathrm{i}\omega,z)\right] \\
&= \lim_{z\to\overline{\pi_j}^{-1}}\frac{\partial}{\partial z}\left[\frac{z}{\overline{\pi_j}^2}\mathcal{H}(\mathrm{i}\omega,z)\right] + \lim_{z\to 0}\left[\frac{z^2}{(1-z\overline{\pi_j})^2}\mathcal{H}(\mathrm{i}\omega,z)\right] \\
&= \lim_{z\to\overline{\pi_j}^{-1}}\left[\frac{1}{\overline{\pi_j}^2}\mathcal{H}(\mathrm{i}\omega,z) - \frac{z}{\overline{\pi_j}^2}\frac{\partial}{\partial p}\mathcal{H}(\mathrm{i}\omega,z)\right] + 0 \\
&= \frac{1}{\overline{\pi_j}^2}\mathcal{H}(\mathrm{i}\omega,\overline{\pi_j}^{-1}) - \frac{1}{\overline{\pi_j}^3}\frac{\partial}{\partial p}\mathcal{H}(\mathrm{i}\omega,\overline{\pi_j}^{-1}).
\end{aligned} \tag{3.2.27}$$

Substituting (3.2.27) into (3.2.26), we get

$$\begin{aligned}
&\frac{1}{2\pi\mathrm{i}}\int_{\Gamma_R}\frac{1}{-w-\overline{\lambda_i}}\left[\frac{1}{\overline{\pi_j}^2}\mathcal{H}(\omega,\overline{\pi_j}^{-1}) - \frac{1}{\overline{\pi_j}^3}\frac{\partial}{\partial p}\mathcal{H}(w,\overline{\pi_j}^{-1})\right]\mathrm{d}w \\
&= \frac{1}{\overline{\pi_j}^2}\operatorname{Res}\left[\frac{1}{-w-\overline{\lambda_i}}\mathcal{H}(w,\overline{\pi_j}^{-1}),\ w=-\overline{\lambda_i}\right] - \frac{1}{\overline{\pi_j}^3}\operatorname{Res}\left[\frac{1}{-w-\overline{\lambda_i}}\frac{\partial}{\partial p}\mathcal{H}(w,\overline{\pi_j}^{-1}),\ w=-\overline{\lambda_i}\right] \\
&= \frac{1}{\overline{\pi_j}^2}\lim_{w\to-\overline{\lambda_i}}\left[\frac{w+\overline{\lambda_i}}{-w-\overline{\lambda_i}}\mathcal{H}(w,\overline{\pi_j}^{-1})\right] - \frac{1}{\overline{\pi_j}^3}\lim_{w\to-\overline{\lambda_i}}\left[\frac{w+\overline{\lambda_i}}{-w-\overline{\lambda_i}}\frac{\partial}{\partial p}\mathcal{H}(w,\overline{\pi_j}^{-1})\right] \\
&= \frac{-1}{\overline{\pi_j}^2}\mathcal{H}(-\overline{\lambda_i},\overline{\pi_j}^{-1}) + \frac{1}{\overline{\pi_j}^3}\frac{\partial}{\partial p}\mathcal{H}(-\overline{\lambda_i},\overline{\pi_j}^{-1}).
\end{aligned} \tag{3.2.28}$$

For the first integral in (3.2.24):

$$\left\langle \frac{1}{(s-\lambda_i)^2(p-\pi_j)}, \mathcal{H}(s,p)\right\rangle_{\otimes} = \frac{1}{2\pi}\int_0^{2\pi}\frac{1}{\overline{\mathrm{e}^{\mathrm{i}\theta}-\pi_j}}\underbrace{\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{\overline{\mathrm{i}\omega-\lambda_i}}\mathcal{H}(\mathrm{i}\omega,\mathrm{e}^{\mathrm{i}\theta})\,\mathrm{d}\omega}_{=:\mathcal{I}_\lambda}\,\mathrm{d}\theta \tag{3.2.29}$$

For the inner integral $\mathcal{I}_\lambda$, we compute

$$\begin{aligned}\mathcal{I}_\lambda &= \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_R} \frac{1}{(-w-\overline{\lambda_i})^2} \mathcal{H}(w, \mathrm{e}^{\mathrm{i}\theta})\, \mathrm{d}w \\ &= \operatorname{Res}\left[\frac{1}{(-w-\overline{\lambda_i})^2} \mathcal{H}(w, \mathrm{e}^{\mathrm{i}\theta}),\ w = -\overline{\lambda_i}\right] \\ &= \lim_{w \to -\overline{\lambda_i}} \frac{\partial}{\partial w}\left[\frac{(w+\overline{\lambda_i})^2}{(-w-\overline{\lambda_i})^2} \mathcal{H}(w, \mathrm{e}^{\mathrm{i}\theta})\right] \\ &= \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, \mathrm{e}^{\mathrm{i}\theta}).\end{aligned} \tag{3.2.30}$$

Substituting (3.2.30) into (3.2.29), we get

$$\begin{aligned}\left\langle \frac{1}{(s-\lambda_i)^2 (p-\pi_j)}, \mathcal{H}(s,p) \right\rangle_{\otimes} &= \frac{1}{2\pi \mathrm{i}} \int_{\Gamma_R} \frac{1}{1 - z\overline{\pi_j}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, z)\, \mathrm{d}z \\ &= \operatorname{Res}\left[\frac{1}{1 - z\overline{\pi_j}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, z),\ z = \overline{\pi_j}^{-1}\right] \\ &= \lim_{z \to \overline{\pi_j}^{-1}} \left[\frac{z - \overline{\pi_j}^{-1}}{1 - z\overline{\pi_j}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, z)\right] \\ &= \frac{-1}{\overline{\pi_j}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}).\end{aligned} \tag{3.2.31}$$

■

3.2.2 Real Transfer Functions

For applications, it is important that the system response to real inputs yields a real output, i.e., $\mathcal{H}(s)$ maps $\mathbb{R}$ to $\mathbb{R}$. An internal description $(\mathbf{E}, \mathbf{A}, \mathbf{b}, \mathbf{c})$ with real matrices guarantees that $\mathcal{H}(\mathbb{R}) \subseteq \mathbb{R}$.

Definition 3.2.7 A dynamical system $\mathcal{H}(s)$ is called *real* if $\overline{\mathcal{H}(s)} = \mathcal{H}(\overline{s})$ for all $s \in \mathbb{C}_R$ and a parametric dynamical system $\mathcal{H}(s,p)$ is called *real* if $\overline{\mathcal{H}(s,p)} = \mathcal{H}(\overline{s}, \overline{p})$ for all $(s,p) \in \mathbb{C}_R \times \mathbb{D}$.

A real system automatically satisfies $\mathcal{H}(z) \in \mathbb{R}$ for $z \in \mathbb{R}$ and similarly for the parametric case $\mathcal{H}(s,p) \in \mathbb{R}$ if $s \in \mathbb{R}$, $p \in [-1,1]$. Since the Hardy spaces $\mathcal{H}_2$ are defined over $\mathbb{C}$, the inner product is, in general, complex. To ensure the inner product $\langle \mathcal{H},\ \mathcal{G} \rangle_{\otimes}$ is real for real systems, we have the following lemma.

Lemma 3.2.8 *Let $\mathcal{H} \in \mathcal{RH}_2(\mathbb{C}_R)$ be as $\mathcal{H}(s) = \sum_{i=1}^{r} \dfrac{\phi_i}{s-\lambda_i}$. Further let $\mathcal{I} := \{1,\ldots,r\}$ and $\mathbf{P} : \mathcal{I} \to \mathcal{I}$ be a permutation so that $\lambda_{\mathbf{P}(i)} = \overline{\lambda_i}$, $i \in \mathcal{I}$. If $\phi_{\mathbf{P}(i)} = \overline{\phi_i}$ for $i \in \mathcal{I}$, then $\mathcal{H}(\overline{s}) = \overline{\mathcal{H}(s)}$, i.e., the system is real.*

Proof Let $\mathcal{I}_R := \{i \in \mathcal{I} \mid \lambda_i \in \mathbb{R}\}$ and $\mathcal{I}_*$ so that

$$\mathcal{I}_* \cup \mathbf{P}(\mathcal{I}_*) = \mathcal{I} \backslash \mathcal{I}_R, \qquad \text{with} \qquad \mathcal{I}_* \cap \mathbf{P}(\mathcal{I}_*) = \emptyset. \tag{3.2.32}$$

In particular, $\lambda_i, \phi_i \in \mathbb{R}$ for $i \in \mathcal{I}_R$. By expanding $\mathcal{H}(s)$ in pole-residue form, we get

$$\begin{aligned} \mathcal{H}(\overline{s}) &= \sum_{i \in \mathcal{I}} \frac{\phi_i}{\overline{s}-\lambda_i} = \sum_{j \in \mathcal{I}_R} \frac{\phi_j}{\overline{s}-\lambda_j} + \sum_{j \in \mathcal{J}} \left(\frac{\phi_j}{\overline{s}-\lambda_j} + \frac{\overline{\phi}_j}{\overline{s}-\overline{\lambda}_j} \right) \\ &= \sum_{j \in \mathcal{I}_R} \overline{\frac{\phi_j}{s-\lambda_j}} + \sum_{j \in \mathcal{J}} \overline{\left(\frac{\overline{\phi}_j}{s-\overline{\lambda}_j} + \frac{\phi_j}{s-\lambda_j} \right)} = \overline{\mathcal{H}(s)}. \end{aligned} \tag{3.2.33}$$

■

The following lemma generalizes this concept to the two-variable case.

Lemma 3.2.9 *Let $\mathcal{G}(s,p) = \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \dfrac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}$ with poles $\{\lambda_i\}_{i=1}^{r_s}$ in s and $\{\pi_j\}_{j=1}^{r_p}$ in p. Let $\mathcal{I} := \{1,\ldots,r_s\}$, $\mathcal{J} := \{1,\ldots,r_p\}$ and $\mathbf{P}_\lambda : \mathcal{I} \to \mathcal{I}$ and $\mathbf{P}_\pi : \mathcal{J} \to \mathcal{J}$ be permutations, so that*

$$\overline{\lambda_k} = \lambda_{\mathbf{P}_\lambda(k)}, \qquad k \in \mathcal{I}_* \qquad \text{and} \qquad \overline{\pi_k} = \pi_{\mathbf{P}_\pi(k)}, \qquad k \in \mathcal{J}_*, \tag{3.2.34}$$

where we define

$$\mathcal{I}_R := \{i \in \mathcal{I} \mid \lambda_i \in \mathbb{R}\}, \quad \textit{and} \quad \mathcal{J}_R := \{j \in \mathcal{J} \mid \pi_j \in \mathbb{R}\}, \tag{3.2.35}$$

and $\mathcal{I}_* \subset \mathcal{I}$ *and* $\mathcal{J}_* \subset \mathcal{J}$ *so that*

$$\begin{aligned} &\mathcal{I}_* \cup \mathbf{P}_\lambda(\mathcal{I}_*) = \mathcal{I}\backslash\mathcal{I}_R, \qquad \textit{where} \qquad \mathcal{I}_* \cap \mathbf{P}_\lambda(\mathcal{I}_*) = \emptyset, \quad \textit{and} \\ &\mathcal{J}_* \cup \mathbf{P}_\pi(\mathcal{J}_*) = \mathcal{J}\backslash\mathcal{J}_R \qquad \textit{where} \qquad \mathcal{J}_* \cap \mathbf{P}_\pi(\mathcal{J}_*) = \emptyset. \end{aligned} \tag{3.2.36}$$

Furthermore, we assume for the numerator values the following relations:

$$\begin{aligned} \overline{\phi_{\mathbf{P}_\lambda(i),j}} &= \phi_{i,\mathbf{P}_\pi(j)} && i \in \mathcal{I}_*,\ j \in \mathcal{J}_*, \\ \overline{\phi_{i,j}} &= \phi_{i,\mathbf{P}_\pi(j)} && i \in \mathcal{J}_R,\ j \in \mathcal{J}_*, \\ \overline{\phi_{i,j}} &= \phi_{\mathbf{P}_\lambda(i),j} && i \in \mathcal{J}_*,\ j \in \mathcal{J}_R, \\ \overline{\phi_{i,j}} &= \phi_{i,j} && i \in \mathcal{J}_R,\ j \in \mathcal{J}_R. \end{aligned} \tag{3.2.37}$$

Then $\mathcal{H}(\overline{s},\overline{p}) = \overline{\mathcal{H}(s,p)}$*, i.e.,* $\mathcal{H}(s,p)$ *is real.*

Proof Expand $\mathcal{H}(\overline{s},\overline{p})$ as

$$\begin{aligned} \mathcal{H}(\overline{s},\overline{p}) &= \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \frac{\phi_{i,j}}{(\overline{s}-\lambda_i)(\overline{p}-\pi_j)} \\ &= \sum_{i\in\mathcal{I}_R} \frac{1}{\overline{s}-\lambda_i}\left(\sum_{j\in\mathcal{J}_R}\frac{\phi_{i,j}}{\overline{p}-\pi_j} + \sum_{j\in\mathcal{J}_*}\left(\frac{\phi_{i,j}}{\overline{p}-\pi_j} + \frac{\phi_{i,\mathbf{P}_\pi(j)}}{\overline{p}-\overline{\pi_j}}\right)\right) \\ &\quad + \sum_{i\in\mathcal{I}_*}\left[\frac{1}{\overline{s}-\lambda_i}\left(\sum_{j\in\mathcal{J}_R}\frac{\phi_{i,j}}{\overline{p}-\pi_j} + \sum_{j\in\mathcal{J}_*}\left(\frac{\phi_{i,j}}{\overline{p}-\pi_j} + \frac{\phi_{i,\mathbf{P}_\pi(j)}}{\overline{p}-\overline{\pi_j}}\right)\right)\right] \\ &\quad + \sum_{i\in\mathcal{I}_*}\left[\frac{1}{\overline{s}-\overline{\lambda_i}}\left(\sum_{j\in\mathcal{J}_R}\frac{\phi_{\mathbf{P}_\lambda(i),j}}{\overline{p}-\pi_j} + \sum_{j\in\mathcal{J}_*}\left(\frac{\phi_{\mathbf{P}_\lambda(i),j}}{\overline{p}-\pi_j} + \frac{\phi_{\mathbf{P}_\lambda(i),\mathbf{P}_\pi(j)}}{\overline{p}-\overline{\pi_j}}\right)\right)\right]. \end{aligned} \tag{3.2.38}$$

We pull out the complex conjugation from each term:

$$\begin{aligned}\mathcal{H}(\overline{s},\overline{p}) = &\sum_{i\in\mathcal{I}_R} \overline{\frac{1}{s-\lambda_i}} \left(\sum_{j\in\mathcal{J}_R} \overline{\left(\frac{\overline{\phi_{i,j}}}{p-\pi_j}\right)} + \sum_{j\in\mathcal{J}_*} \overline{\left(\frac{\overline{\phi_{i,j}}}{p-\overline{\pi_j}} + \frac{\overline{\phi_{i,\mathbf{P}_\pi(j)}}}{p-\pi_j}\right)} \right) \\ &+ \underbrace{\sum_{i\in\mathcal{I}_*} \left[\overline{\frac{1}{s-\overline{\lambda_i}}} \right]}_{\lozenge} \left(\sum_{j\in\mathcal{J}_R} \overline{\left(\frac{\overline{\phi_{i,j}}}{p-\pi_j}\right)} + \underbrace{\sum_{j\in\mathcal{J}_*}}_{\lozenge} \overline{\left(\frac{\overline{\phi_{i,j}}}{p-\overline{\pi_j}} + \underbrace{\frac{\overline{\phi_{i,\mathbf{P}_\pi(j)}}}{p-\pi_j}}_{\lozenge}\right)} \right) \Bigg] \\ &+ \underbrace{\sum_{i\in\mathcal{I}_*} \left[\overline{\frac{1}{s-\lambda_i}} \right]}_{\lozenge} \left(\sum_{j\in\mathcal{J}_R} \overline{\left(\frac{\overline{\phi_{\mathbf{P}_\lambda(i),j}}}{p-\pi_j}\right)} + \underbrace{\sum_{j\in\mathcal{J}_*}}_{\lozenge} \overline{\left(\underbrace{\frac{\overline{\phi_{\mathbf{P}_\lambda(i),j}}}{p-\overline{\pi_j}}}_{\lozenge} + \frac{\overline{\phi_{\mathbf{P}_\lambda(i),\mathbf{P}_\pi(j)}}}{p-\pi_j}\right)} \right) \Bigg] \\ = &\ \overline{\mathcal{H}(s,p)}. \end{aligned} \tag{3.2.39}$$

The last equality in (3.2.39) comes from applying (3.2.37) to identify complex conjugate pairs. We show an example for the terms marked with $\lozenge$:

$$\sum_{i\in\mathcal{I}_*} \overline{\frac{1}{s-\overline{\lambda_i}}} \sum_{j\in\mathcal{J}_*} \overline{\frac{\overline{\phi_{i,\mathbf{P}_\pi(j)}}}{p-\pi_j}} = \sum_{i\in\mathcal{I}_*} \frac{1}{\overline{s}-\lambda_i} \sum_{j\in\mathcal{J}_*} \overline{\frac{\overline{\phi_{\mathbf{P}_\lambda(i),j}}}{p-\pi_j}}. \tag{3.2.40}$$

The other terms in (3.2.39) can be paired similarly. ■

Corollary 3.2.10 *For $s \in \mathbb{R}$ and $p \in \mathbb{R}$, the previous theorem implies $\mathcal{H}(s,p) \in \mathbb{R}$.*

With Lemma 3.2.9, we now address the inner product $\langle \cdot,\ \cdot \rangle_\otimes$.

Lemma 3.2.11 *Let $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ be real and $\mathcal{G}(s,p) = \sum_{i=1}^{r_s} \sum_{k=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ be real. Then $\langle \mathcal{G},\ \mathcal{H} \rangle_\otimes \in \mathbb{R}$.*

Proof We expand the inner product $\langle \mathcal{G},\ \mathcal{H} \rangle_{\otimes}$ using the index sets in (3.2.36):

$$\begin{aligned}
\langle \mathcal{G},\ \mathcal{H} \rangle_{\otimes} &= \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}} \mathcal{H}\left(-\overline{\lambda_i}, \overline{\pi_j}^{-1}\right) \\
&= \sum_{i \in \mathcal{I}_R} \left[\sum_{j \in \mathcal{J}_R} \underbrace{\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, {\pi_j}^{-1}\right)}{\pi_j}}_{\in \mathbb{R}} + \sum_{j \in \mathcal{J}_*} \left(\underbrace{\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, \overline{\pi_j}^{-1}\right)}{\overline{\pi_j}}}_{\diamondsuit} + \underbrace{\frac{\overline{\phi_{i,\mathbf{P}_\pi(j)}} \mathcal{H}\left(-\lambda_i, {\pi_j}^{-1}\right)}{\pi_j}}_{\diamondsuit} \right) \right] \\
&+ \sum_{i \in \mathcal{I}_*} \left[\sum_{j \in \mathcal{J}_R} \underbrace{\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, {\pi_j}^{-1}\right)}{\pi_j}}_{\sharp} + \sum_{j \in \mathcal{J}_*} \left(\underbrace{\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, \overline{\pi_j}^{-1}\right)}{\overline{\pi_j}}}_{\clubsuit} + \underbrace{\frac{\overline{\phi_{i,\mathbf{P}_\pi(j)}} \mathcal{H}\left(-\lambda_i, {\pi_j}^{-1}\right)}{\pi_j}}_{\spadesuit} \right) \right] \\
&+ \sum_{i \in \mathcal{I}_*} \left[\sum_{j \in \mathcal{J}_R} \underbrace{\frac{\overline{\phi_{\mathbf{P}_\lambda(i),j}} \mathcal{H}\left(-\overline{\lambda_i}, {\pi_j}^{-1}\right)}{\pi_j}}_{\sharp} \right. \\
&\qquad \left. + \sum_{j \in \mathcal{J}_*} \left(\underbrace{\frac{\overline{\phi_{\mathbf{P}_\lambda(i),j}} \mathcal{H}\left(-\overline{\lambda_i}, \overline{\pi_j}^{-1}\right)}{\overline{\pi_j}}}_{\spadesuit} + \underbrace{\frac{\overline{\phi_{\mathbf{P}_\lambda(i),\mathbf{P}_\pi(j)}} \mathcal{H}\left(-\overline{\lambda_i}, {\pi_j}^{-1}\right)}{\pi_j}}_{\clubsuit} \right) \right].
\end{aligned} \tag{3.2.41}$$

Here the symbols indicate terms that are complex conjugates of each other and will lead a real number when added up. Recall that $\mathcal{G}(s,p)$ satisfies $\mathcal{G}(\overline{s},\overline{p}) = \overline{\mathcal{G}(s,p)}$. Further observe that $\mathcal{H}(s,\overline{p}) = \overline{\mathcal{H}(\overline{s},p)}$. Using the assumptions on the numerator values $\phi_{i,j}$ from (3.2.37), we demonstrate this for the terms marked by $\clubsuit$:

$$\begin{aligned}
&\sum_{i \in \mathcal{I}_*} \sum_{j \in \mathcal{J}_*} \left[\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, \overline{\pi_j}^{-1}\right)}{\overline{\pi_j}} + \frac{\overline{\phi_{\mathbf{P}_\lambda(i),\mathbf{P}_\pi(j)}} \mathcal{H}\left(-\overline{\lambda_i}, {\pi_j}^{-1}\right)}{\pi_j} \right] \\
&= \sum_{i \in \mathcal{I}_*} \sum_{j \in \mathcal{J}_*} \left[\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, \overline{\pi_j}^{-1}\right)}{\overline{\pi_j}} + \frac{\phi_{i,j} \overline{\mathcal{H}\left(-\lambda_i, \overline{\pi_j}^{-1}\right)}}{\pi_j} \right] \\
&= \sum_{i \in \mathcal{I}_*} \sum_{j \in \mathcal{J}_*} 2\Re \left[\frac{\overline{\phi_{i,j}} \mathcal{H}\left(-\lambda_i, \overline{\pi_j}^{-1}\right)}{\overline{\pi_j}} \right] \in \mathbb{R}.
\end{aligned} \tag{3.2.42}$$

Similarly, the other terms are paired up. ■

Throughout this work, we assume real dynamical systems $\mathcal{H}(s,p)$ and $\widehat{\mathcal{H}}(s,p)$ that satisfy the complex conjugate symmetry conditions from Lemma 3.2.9.

3.2.3 Gramians for the Parametric Case

First note that the transfer function $\mathcal{H}(s) = \sum_{i=1}^{n} \frac{\phi_i}{s-\lambda_i} \in \mathcal{H}_2(\mathbb{C}_R)$ admits the realization

$$\mathbf{A} = \operatorname{diag}[\lambda_1, \ldots, \lambda_r], \qquad \mathbf{b} = [1, \ldots, 1]^\top \qquad \text{and} \qquad \mathbf{c} = [\phi_1, \ldots, \phi_r]^\top, \tag{3.2.43}$$

so that $\mathcal{H}(s) = \mathbf{c}^\top (s\mathbf{I} - \mathbf{A})^{-1} \mathbf{b}$. Then the $\mathcal{H}_2$ norm of $\mathcal{H}(s)$ can be computed via the system Gramians, introduced in Section 2.3.1 (see, e.g., [3, Chap. 5.5]) as

$$\|\mathcal{H}\|_{\mathcal{H}_2}^2 = \mathbf{c}^* \mathcal{P} \mathbf{c} = \mathbf{b}^* \mathcal{Q} \mathbf{b}. \tag{3.2.44}$$

In the following, we derive similar expressions for the $\mathcal{H}_2 \otimes L_2$ norm.

To derive an expression like (3.2.44) in the parametric case, we start by finding a particular internal description from the pole-residue form similar to (3.2.43). Recall the form for $\widehat{\mathcal{H}}(s,p)$ from (3.2.7) as

$$\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}, \qquad \begin{array}{l} \lambda_i \in \mathbb{C}_L,\ i = 1, \ldots, r_s; \\ \pi_j \in \mathbb{C} \backslash \overline{\mathbb{D}},\ j = 1, \ldots, r_p. \end{array}$$

For such a separable, two-variable transfer function $\widehat{\mathcal{H}}(s,p)$, consider the following realizations:

$$\begin{aligned} \widetilde{\mathcal{H}}_1(s,p) &:= \widetilde{\mathbf{C}}_s^\top(p) \left(s\mathbf{I}_{r_s} - \widetilde{\mathbf{A}}_s\right)^{-1} \widetilde{\mathbf{B}}_s, \qquad \text{with} \qquad [\widetilde{\mathbf{C}}_s(p)]_i = \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{p-\pi_j}, \quad \text{and} \\ \widetilde{\mathcal{H}}_2(s,p) &:= \widetilde{\mathbf{C}}_p^\top(s) \left(p\mathbf{I}_{r_p} - \widetilde{\mathbf{A}}_p\right)^{-1} \widetilde{\mathbf{B}}_p, \qquad \text{with} \qquad [\widetilde{\mathbf{C}}_p(s)]_j = \sum_{i=1}^{r_s} \frac{\phi_{i,j}}{s-\lambda_i}. \end{aligned} \tag{3.2.45}$$

Explicitely, $\widetilde{\mathbf{A}}_s = \operatorname{diag}[\lambda_1, \ldots, \lambda_{r_s}] \in \mathbb{C}^{r_s \times r_s}$, $\widetilde{\mathbf{A}}_p = \operatorname{diag}[\pi_1, \ldots, \pi_{r_p}] \in \mathbb{C}^{r_p \times r_p}$ as well as $\widetilde{\mathbf{B}}_s = [1, \ldots, 1]^\top \in \mathbb{C}^{r_s}$ and $\widetilde{\mathbf{B}}_p = [1, \ldots, 1]^\top \in \mathbb{C}^{r_p}$. Further, let

$$\begin{aligned} \widetilde{\mathcal{H}}_3(s,p) &:= \widetilde{\mathbf{C}}_s^\top \left(s\mathbf{I}_{r_s} - \widetilde{\mathbf{A}}_s \right)^{-1} \widetilde{\mathbf{B}}_s(p), \qquad \text{with} \qquad [\widetilde{\mathbf{B}}_s(p)]_i = \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{p - \pi_j}, \quad \text{and} \\ \widetilde{\mathcal{H}}_4(s,p) &:= \widetilde{\mathbf{C}}_p^\top \left(p\mathbf{I}_{r_p} - \widetilde{\mathbf{A}}_p \right)^{-1} \widetilde{\mathbf{B}}_p(s), \qquad \text{with} \qquad [\widetilde{\mathbf{B}}_p(s)]_j = \sum_{i=1}^{r_s} \frac{\phi_{i,j}}{s - \lambda_i}, \end{aligned} \tag{3.2.46}$$

where $\widetilde{\mathbf{C}}_s = [1, \ldots, 1]^\top \in \mathbb{C}^{r_s}$ and $\widetilde{\mathbf{C}}_p = [1, \ldots, 1]^\top \in \mathbb{C}^{r_p}$. Observe that the expressions in (3.2.45) are merely different representations of $\widehat{\mathcal{H}}(s,p)$, so that

$$\widehat{\mathcal{H}}(s,p) = \widetilde{\mathcal{H}}_1(s,p) = \widetilde{\mathcal{H}}_2(s,p) = \widetilde{\mathcal{H}}_3(s,p) = \widetilde{\mathcal{H}}_4(s,p), \qquad \text{for all} \quad (s,p) \in \mathbb{C}_R \times \mathbb{D}. \tag{3.2.47}$$

Since all representations in (3.2.45) and (3.2.46) are equivalent, we are free to choose one that leads to a simple expression of the $\mathcal{H}_2 \otimes L_2$ norm.

Lemma 3.2.12 *Let $\widehat{\mathcal{H}}(s,p)$ be a two-variable transfer function as in (3.2.45). With the matrices defined in (3.2.45), let $\mathcal{P}_\lambda \in \mathbb{C}^{r_s \times r_s}$ (λ-*reachability Gramian*) be the unique solution of the Lyapunov equation*

$$\widetilde{\mathbf{A}_s}\mathcal{P}_\lambda + \mathcal{P}_\lambda \widetilde{\mathbf{A}_s}^* + \widetilde{\mathbf{B}_s}\widetilde{\mathbf{B}_s}^* = 0. \tag{3.2.48}$$

*Further let $\mathcal{P}_\pi \in \mathbb{C}^{r_p \times r_p}$ (π-*reachability Gramian*) be the solution of the Stein equation*

$$\widetilde{\mathbf{A}_p}\mathcal{P}_\pi \widetilde{\mathbf{A}_p}^* + \widetilde{\mathbf{B}_p}\widetilde{\mathbf{B}_p}^* = \mathcal{P}_\pi. \tag{3.2.49}$$

Then the elements of $\mathcal{P}_\lambda$ and $\mathcal{P}_\pi$ can be expressed as

$$[\mathcal{P}_\lambda]_{i,j} = \frac{-1}{\lambda_i + \overline{\lambda_j}}, \qquad i,j = 1, \ldots, r_s \qquad \textit{and} \qquad [\mathcal{P}_\pi]_{i,j} = \frac{1}{\overline{\pi_j}\pi_i - 1}, \qquad i,j = 1, \ldots, r_p. \tag{3.2.50}$$

Proof Bartels and Stewart [8] noted that (3.2.48) has a unique solution if and only if the eigenvalues of $\widetilde{\mathbf{A}_s}$ and $\widetilde{\mathbf{A}_s}^*$ never sum up to 0. With the eigenvalues $\lambda_1, \ldots, \lambda_{r_s}$ of $\widetilde{\mathbf{A}_s}$ and $\overline{\lambda_1}, \ldots, \overline{\lambda_{r_s}}$ of $\widetilde{\mathbf{A}_s}^*$, it is clear that $\lambda_i + \overline{\lambda_j} \neq 0$ for any $i, j = 1, \ldots, r_s$.

Observe that

$$\begin{aligned}
\widetilde{\mathbf{A}}_s\mathcal{P}_\lambda + \mathcal{P}_\lambda\widetilde{\mathbf{A}}_s^* &= \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{r_s} \end{bmatrix} \begin{bmatrix} \frac{-1}{\lambda_1+\overline{\lambda_1}} & \cdots & \frac{-1}{\lambda_1+\overline{\lambda_{r_s}}} \\ \vdots & \ddots & \vdots \\ \frac{-1}{\lambda_{r_s}+\overline{\lambda_1}} & \cdots & \frac{-1}{\lambda_{r_s}+\overline{\lambda_{r_s}}} \end{bmatrix} \\
&\quad + \begin{bmatrix} \frac{-1}{\lambda_1+\overline{\lambda_1}} & \cdots & \frac{-1}{\lambda_1+\overline{\lambda_{r_s}}} \\ \vdots & \ddots & \vdots \\ \frac{-1}{\lambda_{r_s}+\overline{\lambda_1}} & \cdots & \frac{-1}{\lambda_{r_s}+\overline{\lambda_{r_s}}} \end{bmatrix} \begin{bmatrix} \overline{\lambda_1} & & \\ & \ddots & \\ & & \overline{\lambda_{r_s}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{-\lambda_1}{\lambda_1+\overline{\lambda_1}} & \cdots & \frac{-\lambda_1}{\lambda_1+\overline{\lambda_{r_s}}} \\ \vdots & \ddots & \vdots \\ \frac{-\lambda_{r_s}}{\lambda_{r_s}+\overline{\lambda_1}} & \cdots & \frac{-\lambda_{r_s}}{\lambda_{r_s}+\overline{\lambda_{r_s}}} \end{bmatrix} + \begin{bmatrix} \frac{-\overline{\lambda_1}}{\lambda_1+\overline{\lambda_1}} & \cdots & \frac{-\overline{\lambda_{r_s}}}{\lambda_1+\overline{\lambda_{r_s}}} \\ \vdots & \ddots & \vdots \\ \frac{-\overline{\lambda_1}}{\lambda_{r_s}+\overline{\lambda_1}} & \cdots & \frac{-\overline{\lambda_{r_s}}}{\lambda_{r_s}+\overline{\lambda_{r_s}}} \end{bmatrix} \\
&= -\begin{bmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{bmatrix} = \widetilde{\mathbf{B}_s}\widetilde{\mathbf{B}_s}^*.
\end{aligned} \tag{3.2.51}$$

Further it is well known that the Stein equation (3.2.49) has a unique solution if and only if the eigenvalues of $\widetilde{\mathbf{A}_p}$ and $\widetilde{\mathbf{A}_p}^*$ never multiply to 1. In our case, it is clear that for $\pi_1, \ldots, \pi_{r_p}$ the eigenvalues of $\widetilde{\mathbf{A}_p}$ and $\overline{\pi_1}, \ldots, \overline{\pi_{r_p}}$ those of $\widetilde{\mathbf{A}_p}^*$, it holds that $\pi_i\overline{\pi_j} > 1$ for all

$i, j = 1, \ldots, r_p$. For $\mathcal{P}_\pi$, we compute

$$
\begin{aligned}
&\widetilde{\mathbf{A}}_p \mathcal{P}_\pi \widetilde{\mathbf{A}}_p^* + \widetilde{\mathbf{B}}_p \widetilde{\mathbf{B}}_p^* \\
&= \begin{bmatrix} \pi_1 & & \\ & \ddots & \\ & & \pi_{r_s} \end{bmatrix} \begin{bmatrix} \frac{1}{\pi_1\overline{\pi}_1 - 1} & \cdots & \frac{1}{\pi_1\overline{\pi}_{r_s} - 1} \\ & \ddots & \\ \frac{1}{\pi_{r_s}\overline{\pi}_1 - 1} & \cdots & \frac{1}{\pi_{r_s}\overline{\pi}_{r_s} - 1} \end{bmatrix} \begin{bmatrix} \overline{\pi}_1 & & \\ & \ddots & \\ & & \overline{\pi}_{r_s} \end{bmatrix} - \begin{bmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{bmatrix} \qquad (3.2.52) \\
&= \begin{bmatrix} \frac{\pi_1\overline{\pi}_1}{\pi_1\overline{\pi}_1 - 1} - 1 & \cdots & \frac{\pi_1\overline{\pi}_{r_s}}{\pi_1\overline{\pi}_{r_s} - 1} - 1 \\ & \ddots & \\ \frac{\pi_{r_s}\overline{\pi}_1}{\pi_{r_s}\overline{\pi}_1 - 1} - 1 & \cdots & \frac{\pi_{r_s}\overline{\pi}_{r_s}}{\pi_{r_s}\overline{\pi}_{r_s} - 1} - 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_1\overline{\pi}_1 - 1} & \cdots & \frac{1}{\pi_1\overline{\pi}_{r_s} - 1} \\ & \vdots & \\ \frac{1}{\pi_{r_s}\overline{\pi}_1 - 1} & \cdots & \frac{1}{\pi_{r_s}\overline{\pi}_{r_s} - 1} \end{bmatrix} = \mathcal{P}_\pi.
\end{aligned}
$$

■

The matrices $\mathcal{P}_\lambda$ and $\mathcal{P}_\pi$ in (3.2.50) are reachability Gramians as in the single-variable case with respect to the domain of the parameter, leading to Lyapunov and Stein equations since $\mathcal{H}(\cdot, p) \in \mathcal{H}_2(\mathbb{C}_R)$ and $\mathcal{H}(s, \cdot) \in \mathcal{H}_2(\mathbb{D})$. Let $\mathcal{P} \in \mathbb{C}^{r_s r_p \times r_s r_p}$ be defined as

$$\mathcal{P} := (\mathbf{I}_{r_s} \otimes \mathcal{P}_\pi) \odot (\mathcal{P}_\lambda \otimes \mathbf{I}_{r_p}), \qquad (3.2.53)$$

where $\odot$ represents the Hadamard product of point wise multiplication of matrix entries.

We can write the entries of $\mathcal{P}$ (we use the same notation $\mathcal{P}$ here as for the non-parametric case) as

$$\mathcal{P}_{k,\ell} = \frac{-1}{(\underline{\boldsymbol{\lambda}_k} + \overline{\boldsymbol{\lambda}_\ell})(\overline{\boldsymbol{\pi}_k}\boldsymbol{\pi}_\ell - 1)} \quad \text{for} \quad k, \ell = 1, \ldots, r_s r_p. \qquad (3.2.54)$$

In a slight abuse of notation, we use a tensor notation for the indices for λ and π in (3.2.54):

$$\begin{aligned}\underline{\boldsymbol{\lambda}} &= [\underbrace{\lambda_1, \lambda_1, \ldots, \lambda_1}_{r_p \text{ times}}, \ldots, \underbrace{\lambda_{r_s}, \ldots, \lambda_{r_s}}_{r_p \text{ times}}], \quad \text{and} \\ \underline{\boldsymbol{\pi}} &= [\underbrace{\pi_1, \pi_2, \ldots, \pi_{r_p}, \pi_1, \pi_2, \ldots, \pi_{r_p}}_{r_s \text{ times repeated}}].\end{aligned} \tag{3.2.55}$$

The next lemma extends (3.2.44) to the two-variable case.

Lemma 3.2.13 *Let* $\widehat{\mathcal{H}}(s,p) = \displaystyle\sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ *and* $\mathcal{P}$ *as in (3.2.54). Then the* $\mathcal{H}_2 \otimes L_2$ *norm can be expressed as*

$$\|\widehat{\mathcal{H}}\|_{\otimes}^2 = \mathbf{C}_0^* \mathcal{P} \mathbf{C}_0, \tag{3.2.56}$$

with $\mathbf{C}_0 = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \ldots & \phi_{r_s,r_p} \end{bmatrix}^\top = \operatorname{vec}(\boldsymbol{\phi})$.

Proof Recall from Corollary 3.2.5 that the $\mathcal{H}_2 \otimes L_2$ inner norm can be written as

$$\begin{aligned}\|\widehat{\mathcal{H}}\|_{\otimes}^2 &= \sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{-\overline{\phi_{i,j}}}{\overline{\pi_j}} \widehat{\mathcal{H}}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) = \sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{-\overline{\phi_{i,j}}}{\overline{\pi_j}} \sum_{k,\ell} \frac{\phi_{k,\ell}}{(-\overline{\lambda_i} - \lambda_k)(\overline{\pi_j}^{-1} - \pi_\ell)} \\ &= \sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\sum_{k=1}^{r_s}\sum_{\ell=1}^{r_p} \frac{-\overline{\phi_{i,j}}\phi_{k,\ell}}{(-\overline{\lambda_i} - \lambda_k)(1 - \overline{\pi_j}\pi_\ell)} = \sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\sum_{k=1}^{r_s}\sum_{\ell=1}^{r_p} \frac{\overline{\phi_{i,j}}\phi_{k,\ell}}{(\overline{\lambda_i} + \lambda_k)(1 - \overline{\pi_j}\pi_\ell)}\end{aligned} \tag{3.2.57}$$

Then (3.2.57) implies

$$\|\mathcal{H}\|_{\otimes}^2 = \mathbf{C}_0^* \mathcal{P} \mathbf{C}_0. \tag{3.2.58}$$

■

After expressing a parametric analogue to the (infinite) reachability Gramian using the pole-residue formulation of a parametric dynamical system, we obtain the following expression

based on the observability Gramian. This is made precise in the following Lemma. As for the reachability Gramians, the two-variable setting decomposes into Gramians for each variable separately which is a direct result of the separable form of $\widehat{\mathcal{H}}(s)$.

Theorem 3.2.14 *Let* $\widetilde{\mathbf{A}}$ *and* $\widetilde{\mathbf{C}}$ *be as in (3.2.46). Let* $\mathcal{Q}_\pi$ *(*π-observability Gramian*) be the solution of the Stein equation*

$$\widetilde{\mathbf{A}}^* \mathcal{Q}_\pi \widetilde{\mathbf{A}} + \widetilde{\mathbf{C}_p}^* \widetilde{\mathbf{C}_p} = \mathcal{Q}_\pi. \tag{3.2.59}$$

Further let $\mathcal{Q}_\lambda$ *(*λ-observability Gramian*) be the solution of the Lyapunov equation*

$$\widetilde{\mathbf{A}}^* \mathcal{Q}_\lambda + \mathcal{Q}_\lambda \widetilde{\mathbf{A}} + \widetilde{\mathbf{C}_s}^* \widetilde{\mathbf{C}_s} = 0. \tag{3.2.60}$$

Define $\mathcal{Q}$ *as*

$$\mathcal{Q} := (\mathbf{I}_{r_s} \otimes \mathcal{Q}_\pi) \odot (\mathcal{Q}_\lambda \otimes \mathbf{I}_{r_p}). \tag{3.2.61}$$

Then the $\mathcal{H}_2 \otimes L_2$ *norm of* $\widehat{\mathcal{H}}(s,p)$ *can be expressed as*

$$\|\widehat{\mathcal{H}}\|_\otimes^2 = \mathbf{B}_0^* \mathcal{Q} \mathbf{B}_0, \tag{3.2.62}$$

with the vector $\mathbf{B}_0 = \begin{bmatrix} \phi_{1,1}, \phi_{1,2} & \dots & \phi_{r_s,r_p} \end{bmatrix}^\top = \operatorname{vec}(\boldsymbol{\phi})$.

Proof The proof is analogous to the proof of Lemma 3.2.13 with obvious modifications for the representations in (3.2.46). ■

After the system theoretic results above, we turn our attention to approximation results for reduced order models in the next subsection.

3.3 Optimality Conditions

In this section, we develop conditions for $\widehat{\mathcal{H}}(s,p)$ to be an optimal approximant to $\mathcal{H}(s,p)$ with respect to the combined $\mathcal{H}_2 \otimes L_2$ norm in frequency and parameter.

Recall the form of the approximating function from (3.2.7):

$$\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}, \qquad \Re(\lambda_i) < 0, |\pi_j| > 1. \tag{3.3.1}$$

We present two approaches to derive conditions for $\widehat{\mathcal{H}}(s,p)$ to be an $\mathcal{H}_2 \otimes L_2$ optimal approximation of $\mathcal{H}(s,p)$. First in Section 3.3.1, by differentiation of the error functional, and in Section 3.3.2, with variational arguments.

3.3.1 Gradient Based Optimality Conditions

For $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$, recall the inner product formula from Lemma 3.2.4:

$$\left\langle \frac{1}{(s-\lambda_i)(p-\pi_j)}, \mathcal{H} \right\rangle_{\otimes} = \frac{-1}{\overline{\pi_j}} \mathcal{H}\left(-\overline{\lambda_i}, \frac{1}{\overline{\pi_j}}\right), |\pi_j| > 1,\ \lambda_i \in \mathbb{C}_L.$$

The next theorem is the first central piece to approximate functions with separable poles as in (3.3.1). First consider the case for fixed poles.

Theorem 3.3.1 (Lagrange Optimality Conditions) *Let $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ and $\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}$ with fixed poles $\lambda_i \in \mathbb{C}_L$ and $\pi_j \in \mathbb{C}\backslash\overline{\mathbb{D}}$. Then $\widehat{\mathcal{H}}(s,p)$ minimizes $\|\widehat{\mathcal{H}} - \mathcal{H}\|_{\otimes}$ if and only if*

$$\mathcal{H}\left(-\overline{\lambda_i}, \frac{1}{\overline{\pi_j}}\right) = \widehat{\mathcal{H}}\left(-\overline{\lambda_i}, \frac{1}{\overline{\pi_j}}\right), \qquad \forall i = 1,\ldots,r_s,\ j = 1,\ldots,r_p. \tag{3.3.2}$$

Proof With $\widehat{\mathcal{H}} \in \operatorname{span}(\mathcal{B}_{i,j})$ the Hilbert space projection theorem yields that the residual $r(s,p) := \mathcal{H}(s,p) - \widehat{\mathcal{H}}(s,p)$ is minimal with respect to the $\|\cdot\|_\otimes$ norm if and only if r is orthogonal to $\mathcal{B}_{i,j}$. Using Lemma 3.2.4, we observe

$$\left\langle \frac{1}{(s-\lambda_i)(p-\pi_j)}, \mathcal{H} - \widehat{\mathcal{H}} \right\rangle_\otimes = \frac{-1}{\overline{\pi_j}} \left(\mathcal{H}\left(-\overline{\lambda_i}, \frac{1}{\overline{\pi_j}}\right) - \widehat{\mathcal{H}}\left(-\overline{\lambda_i}, \frac{1}{\overline{\pi_j}}\right)\right) = 0, \tag{3.3.3}$$

for all $i = 1, \ldots, r_s$ and $j = 1, \ldots, r_p$. Rearranging (3.3.3) directly yields (3.3.2). ■

Observe that the conditions (3.3.2) yield exactly $r_s r_p$ equations for the $r_s r_p$ unknowns $\phi_{i,j}$, $i = 1, \ldots, r_s$, $j = 1, \ldots, r_p$. Finding numerator values $\phi_{i,j}$ for a fixed set of poles (λ_i, π_j) can be implemented as a single linear system solve using Theorem 3.3.1:

$$\mathcal{A}\operatorname{vec}(\boldsymbol{\phi}) = \mathfrak{b}, \quad \text{where} \quad \mathcal{A} = \mathcal{A}_{k,\ell} = \mathcal{B}_{i,j}(-\overline{\lambda_k}, \overline{\pi_\ell}^{-1}), \qquad \mathfrak{b} = \operatorname{vec}(\mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1})). \tag{3.3.4}$$

In matrix form, that yields

$$\underbrace{\begin{bmatrix} \frac{1}{(-\overline{\lambda_1} - \lambda_1)(\overline{\pi_1}^{-1} - \pi_1)} & \cdots & \frac{1}{(-\overline{\lambda_1} - \lambda_{r_s})(\overline{\pi_1}^{-1} - \pi_{r_p})} \\ & \vdots & \\ \frac{1}{(-\overline{\lambda_{r_s}} - \lambda_1)(\overline{\pi_{r_p}}^{-1} - \pi_1)} & \cdots & \frac{1}{(-\overline{\lambda_{r_s}} - \lambda_{r_s})(\overline{\pi_{r_p}}^{-1} - \pi_{r_p})} \end{bmatrix}}_{=\mathcal{A}} \begin{bmatrix} \phi_{1,1} \\ \phi_{1,2} \\ \vdots \\ \phi_{r_s,r_p} \end{bmatrix} = \begin{bmatrix} \mathcal{H}(-\overline{\lambda_1}, \overline{\pi_1}^{-1}) \\ \mathcal{H}(-\overline{\lambda_1}, \overline{\pi_2}^{-1}) \\ \vdots \\ \mathcal{H}(-\overline{\lambda_{r_s}}, \overline{\pi_{r_p}}^{-1}) \end{bmatrix}. \tag{3.3.5}$$

Notice that $\mathcal{A}$ has the structure of a two-variable Cauchy matrix.

The Lagrange optimality condition in Theorem 3.3.1 gives information about the choice of numerator values or residues. To choose the poles λ_i and π_j optimally, we differentiate $\|\widehat{\mathcal{H}} - \mathcal{H}\|_\otimes$ with respect to the poles. We first make the following observation.

Proposition 3.3.2 *Let $\widehat{\mathcal{H}}(s,p)$ be as in (3.3.1). Then,*

$$\begin{aligned}
\frac{\partial}{\partial \phi_{i,j}}\widehat{\mathcal{H}}(s,p) &= \frac{\partial}{\partial \phi_{i,j}}\sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)} = \frac{1}{(s-\lambda_i)(p-\pi_j)},\\
\frac{\partial}{\partial \lambda_i}\widehat{\mathcal{H}}(s,p) &= \frac{\partial}{\partial \lambda_i}\sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)} = \sum_{j=1}^{r_p}\frac{\phi_{i,j}}{(s-\lambda_i)^2(p-\pi_j)}, \quad \textit{and}\\
\frac{\partial}{\partial \pi_j}\widehat{\mathcal{H}}(s,p) &= \frac{\partial}{\partial \pi_j}\sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)} = \sum_{i=1}^{r_s}\frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)^2}.
\end{aligned} \tag{3.3.6}$$

Proof For the first line in (3.3.6), note that there is only one term in $\widehat{\mathcal{H}}(s,p)$ that contains $\phi_{i,j}$. For the derivatives in λ_i and π_j, we apply the chain rule. ■

For ease of presentation, we first consider the case $(r_s, r_p) = (1,1)$, i.e., a simple pole in each variable.

Lemma 3.3.3 *Let $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ be a real dynamical system. A (locally) $\mathcal{H}_2 \otimes L_2$ optimal approximant $\widehat{\mathcal{H}}(s,p) = \dfrac{\phi_*}{(s-\lambda_*)(p-\pi_*)}$ with $\lambda_* < 0$, $\pi_* \in \mathbb{R}\backslash[-1,1]$ satisfies the necessary conditions:*

$$\begin{aligned}
\mathcal{H}\left(-\lambda_*, \pi_*^{-1}\right) &= \widehat{\mathcal{H}}\left(-\lambda_*, \pi_*^{-1}\right),\\
\frac{\partial}{\partial s}\mathcal{H}\left(-\lambda_*, \pi_*^{-1}\right) &= \frac{\partial}{\partial s}\widehat{\mathcal{H}}\left(-\lambda_*, \pi_*^{-1}\right), \quad \textit{and}\\
\frac{\partial}{\partial p}\mathcal{H}\left(-\lambda_*, \pi_*^{-1}\right) &= \frac{\partial}{\partial p}\widehat{\mathcal{H}}\left(-\lambda_*, \pi_*^{-1}\right).
\end{aligned} \tag{3.3.7}$$

Note that we require $\lambda_*, \pi_*, \phi_* \in \mathbb{R}$ to ensure that $\widehat{\mathcal{H}}(s,p)$ is a *real dynamical system*. The Lagrange optimality conditions (3.3.2) automatically imply $\phi_* \in \mathbb{R}$.

Proof Define the error functional

$$E(\phi_*, \lambda_*, \pi_*) := \|\mathcal{H} - \widehat{\mathcal{H}}\|_\otimes^2 = \langle \mathcal{H},\ \mathcal{H}\rangle_\otimes - 2\Re\left\langle \widehat{\mathcal{H}},\ \mathcal{H}\right\rangle_\otimes + \left\langle \widehat{\mathcal{H}},\ \widehat{\mathcal{H}}\right\rangle_\otimes, \tag{3.3.8}$$

where we observe that the first term $\langle \mathcal{H},\ \mathcal{H} \rangle_{\otimes}$ is independent of the reduced model and any of its parameters, hence its derivative with respect to the model parameters $(\phi_*, \lambda_*, \pi_*)$ is zero. The real part $\Re \left\langle \widehat{\mathcal{H}},\ \mathcal{H} \right\rangle_{\otimes}$ in (3.3.8) can be replaced by the full value $\left\langle \widehat{\mathcal{H}},\ \mathcal{H} \right\rangle_{\otimes}$, since we assumed $\mathcal{H}(s,p)$ is a real system and $\phi_*, \lambda_*, \pi_* \in \mathbb{R}$. Using Proposition 3.3.2 and the explicit form of the inner product from Corollary 3.2.6, we compute

$$\begin{aligned}
\frac{\partial}{\partial \phi_*} E(\phi_*, \lambda_*, \pi_*) &= 2 \left\langle \frac{\partial \widehat{\mathcal{H}}}{\partial \phi_*}, \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2 \left\langle \frac{\partial \widehat{\mathcal{H}}}{\partial \phi_*}, \mathcal{H} \right\rangle_{\otimes} \\
&= 2 \left\langle \frac{1}{(s-\lambda_*)(p-\pi_*)}, \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2 \left\langle \frac{1}{(s-\lambda_*)(p-\pi_*)}, \mathcal{H} \right\rangle_{\otimes} \\
&= -2 \frac{1}{\overline{\pi_*}} \widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) + 2 \frac{1}{\overline{\pi_*}} \mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) = 0 \\
\Rightarrow \qquad & \widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) = \mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}).
\end{aligned} \tag{3.3.9}$$

For the derivative in λ, we get

$$\begin{aligned}
\frac{\partial}{\partial \lambda_*} E(\phi_*, \lambda_*, \pi_*) &= 2 \left\langle \frac{\partial \widehat{\mathcal{H}}}{\partial \lambda_*}, \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2 \left\langle \frac{\partial \widehat{\mathcal{H}}}{\partial \lambda_*}, \mathcal{H} \right\rangle_{\otimes} \\
&= 2 \left\langle \frac{\phi_*}{(s-\lambda_*)^2(p-\pi_*)}, \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2 \left\langle \frac{\phi_*}{(s-\lambda_*)^2(p-\pi_*)}, \mathcal{H} \right\rangle_{\otimes} \\
&= -2 \frac{\overline{\phi_*}}{\overline{\pi_*}} \frac{\partial}{\partial s} \widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) + 2 \frac{\overline{\phi_*}}{\overline{\pi_*}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) = 0 \\
\Rightarrow \qquad & \frac{\partial}{\partial s} \widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) = \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}).
\end{aligned} \tag{3.3.10}$$

Differentiating in π yields

$$\begin{aligned}
\frac{\partial}{\partial \pi_*} E(\phi_*, \lambda_*, \pi_*) &= 2\left\langle \frac{\partial \widehat{\mathcal{H}}}{\partial \pi_*}, \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2\left\langle \frac{\partial \widehat{\mathcal{H}}}{\partial \pi_*}, \mathcal{H} \right\rangle_{\otimes} \\
&= 2\left\langle \frac{\phi_*}{(s-\lambda_*)(p-\pi_*)^2}, \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2\left\langle \frac{\phi_*}{(s-\lambda_*)(p-\pi_*)^2}, \mathcal{H} \right\rangle_{\otimes} \\
&= -4\frac{\overline{\phi_*}}{\overline{\pi_*}^2}\widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) - 2\frac{\overline{\phi_*}}{\overline{\pi_*}^3}\frac{\partial}{\partial p}\widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) \\
&\qquad + 4\frac{\overline{\phi_*}}{\overline{\pi_*}^2}\mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) + 2\frac{\overline{\phi_*}}{\overline{\pi_*}^3}\frac{\partial}{\partial p}\mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) = 0 \\
\Rightarrow \qquad \frac{\partial}{\partial p}\widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) &= \frac{\partial}{\partial p}\mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}),
\end{aligned} \tag{3.3.11}$$

where the first term cancels since $\widehat{\mathcal{H}}(-\overline{\lambda_*}, \overline{\pi_*}^{-1}) = \mathcal{H}(-\overline{\lambda_*}, \overline{\pi_*}^{-1})$ by (3.3.9). ■

Observe that, as in the regular $\mathcal{H}_2$ case, the conditions from Lemma 3.3.3 are *necessary*, not in general sufficient. The following theorem generalizes Lemma 3.3.3 to a function $\widehat{\mathcal{H}}(s,p)$ as in (3.3.1).

Theorem 3.3.4 *Let $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ be real and $\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)} \in \mathcal{H}_2(\mathbb{C}_R \times \mathbb{D})$ be real. If $\widehat{\mathcal{H}}(s,p)$ is an $\mathcal{H}_2 \otimes L_2$ optimal approximation of $\mathcal{H}(s,p)$, then*

$$\begin{aligned}
\widehat{\mathcal{H}}\left(-\lambda_i, \pi_j^{-1}\right) &= \mathcal{H}\left(-\lambda_i, \pi_j^{-1}\right), && i = 1, \ldots, r_s,\ j = 1, \ldots, r_p, \\
\sum_{j=1}^{r_p} \frac{\phi_{i,j}}{\pi_j}\frac{\partial}{\partial s}\widehat{\mathcal{H}}(-\lambda_i, \pi_j^{-1}) &= \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{\pi_j}\frac{\partial}{\partial s}\mathcal{H}(-\lambda_i, \pi_j^{-1}), && i = 1, \ldots, r_s, \\
\sum_{i=1}^{r_s} \frac{\phi_{i,j}}{{\pi_j}^3}\frac{\partial}{\partial p}\widehat{\mathcal{H}}(-\lambda_i, \pi_j^{-1}) &= \sum_{i=1}^{r_s} \frac{\phi_{i,j}}{{\pi_j}^3}\frac{\partial}{\partial p}\mathcal{H}(-\lambda_i, \pi_j^{-1}), && j = 1, \ldots, r_p.
\end{aligned} \tag{3.3.12}$$

Proof The first part, the Lagrange conditions, have already been shown in Theorem 3.3.1, based on the Hilbert space projection:

$$\left\langle \mathcal{H} - \widehat{\mathcal{H}},\ \frac{1}{(s-\lambda_i)(p-\pi_j)} \right\rangle_{\otimes} = 0, \qquad \forall i = 1, \ldots, r_s,\ j = 1, \ldots, r_p. \tag{3.3.13}$$

We proceed in an analogous fashion as in the proof of Lemma 3.3.3. Consider again the error functional from (3.3.8) in the preceding proof:

$$E(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}) := \|\mathcal{H} - \widehat{\mathcal{H}}\|_{\otimes}^2 = \langle \mathcal{H},\ \mathcal{H} \rangle_{\otimes} - 2\Re \left\langle \widehat{\mathcal{H}},\ \mathcal{H} \right\rangle_{\otimes} + \left\langle \widehat{\mathcal{H}},\ \widehat{\mathcal{H}} \right\rangle_{\otimes}. \tag{3.3.14}$$

We make use of the explicit form of the $\mathcal{H}_2 \otimes L_2$ inner product from Corollary 3.2.6. The real part $\Re \left\langle \widehat{\mathcal{H}},\ \mathcal{H} \right\rangle_{\otimes}$ in (3.3.14) can be replaced by the full value $\left\langle \widehat{\mathcal{H}},\ \mathcal{H} \right\rangle_{\otimes}$, since we assumed both $\mathcal{H}(s,p)$ and $\widehat{\mathcal{H}}(s,p)$ to be real as in Lemma 3.2.11. For the partial derivatives, we compute

$$\begin{aligned}
\frac{\partial}{\partial \lambda_i} E(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}) &= 2 \left\langle \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)^2 (p-\pi_j)},\ \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2 \left\langle \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)^2 (p-\pi_j)},\ \mathcal{H} \right\rangle_{\otimes} \\
&= -2 \sum_{j=1}^{r_p} \overline{\phi_{i,j}} \frac{1}{\overline{\pi_j}} \frac{\partial}{\partial s} \widehat{\mathcal{H}}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) + 2 \sum_{j=1}^{r_p} \overline{\phi_{i,j}} \frac{1}{\overline{\pi_j}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) = 0 \\
\Rightarrow \qquad & \sum_{j=1}^{r_p} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}} \frac{\partial}{\partial s} \widehat{\mathcal{H}}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) = \sum_{j=1}^{r_p} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}} \frac{\partial}{\partial s} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}).
\end{aligned} \tag{3.3.15}$$

And for the derivative in π_j, we get

$$\begin{aligned}
\frac{\partial}{\partial \pi_j} E(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}) &= 2 \left\langle \sum_{i=1}^{r_s} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)^2},\ \widehat{\mathcal{H}} \right\rangle_{\otimes} - 2 \left\langle \sum_{i=1}^{r_s} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)^2},\ \mathcal{H} \right\rangle_{\otimes} \\
&= -4 \sum_{i=1}^{r_s} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}^2} \widehat{\mathcal{H}}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) - 2 \sum_{i=1}^{r_s} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}^3} \frac{\partial}{\partial p} \widehat{\mathcal{H}}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) \\
&\qquad + 4 \sum_{i=1}^{r_s} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}^2} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) + 2 \sum_{i=1}^{r_s} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}^3} \frac{\partial}{\partial p} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) = 0 \\
\Rightarrow \qquad & \sum_{i=1}^{r_s} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}^3} \frac{\partial}{\partial p} \widehat{\mathcal{H}}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}) = \sum_{i=1}^{r_s} \frac{\overline{\phi_{i,j}}}{\overline{\pi_j}^3} \frac{\partial}{\partial p} \mathcal{H}(-\overline{\lambda_i}, \overline{\pi_j}^{-1}),
\end{aligned} \tag{3.3.16}$$

where we used that $\widehat{\mathcal{H}}(-\overline{\lambda}_i, \overline{\pi}_j^{-1}) = \mathcal{H}(-\overline{\lambda}_i, \overline{\pi}_j^{-1})$. Since the systems are real, we sum over complex conjugate pairs in λ, π and ϕ and thus omit the conjugation in (3.3.12). ■

Remark 3.3.5 With the Hilbert space projection theorem, we get geometric intuition of the optimality conditions (3.3.12) as the error function $E(s,p) = \widehat{\mathcal{H}}(s,p) - \mathcal{H}(s,p)$ being orthogonal to the basis functions of $\widehat{\mathcal{H}}(s,p)$ (yielding the Lagrange conditions) and their derivatives:

$$\begin{aligned} 0 &= \left\langle \mathcal{H} - \widehat{\mathcal{H}},\ \frac{1}{(s-\lambda_i)^2(p-\pi_j)} \right\rangle_{\otimes} \quad \forall i = 1,\dots,r_s,\ j = 1,\dots,r_p, \quad \text{and} \\ 0 &= \left\langle \mathcal{H} - \widehat{\mathcal{H}},\ \frac{1}{(s-\lambda_i)(p-\pi_j)^2} \right\rangle_{\otimes} \quad \forall i = 1,\dots,r_s,\ j = 1,\dots,r_p. \end{aligned} \tag{3.3.17}$$

This leads to the interpolation conditions

$$\begin{aligned} \frac{\partial}{\partial p}\mathcal{H}\left(-\lambda_i, {\pi_j}^{-1}\right) &= \frac{\partial}{\partial p}\widehat{\mathcal{H}}\left(-\lambda_i, {\pi_j}^{-1}\right), \qquad i = 1,\dots,r_s,\ j = 1\dots,r_p, \quad \text{and} \\ \frac{\partial}{\partial s}\mathcal{H}\left(-\lambda_i, {\pi_j}^{-1}\right) &= \frac{\partial}{\partial s}\widehat{\mathcal{H}}\left(-\lambda_i, {\pi_j}^{-1}\right), \qquad i = 1,\dots,r_s,\ j = 1\dots,r_p. \end{aligned} \tag{3.3.18}$$

Observe that (3.3.18) implies the necessary conditions in (3.3.12).

The proof of Theorem 3.3.4 is based on differentiation of the error functional. In the following subsection, we present an alternative derivation using a variational approach.

3.3.2 Variational Derivation of Optimality Conditions

Using a variational approach rather then differentiating may provide additional insight into the parametric optimality conditions from Theorem 3.3.4. Consider an optimal approximant $\widehat{\mathcal{H}}(s,p)$ of order (r_s, r_p) of the form

$$\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s}\sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}. \tag{3.3.19}$$

In order for $\widehat{\mathcal{H}}(s,p)$ to be locally optimal, we require $\|\mathcal{H}-\widehat{\mathcal{H}}\|_\otimes \le \|\mathcal{H}-\delta\widehat{\mathcal{H}}\|_\otimes$ for a stable perturbed system $\delta\widehat{\mathcal{H}}(s,p)$ close to $\widehat{\mathcal{H}}(s,p)$, i.e., $\|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_\otimes$ small. Observe that

$$\begin{aligned}\|\mathcal{H}-\widehat{\mathcal{H}}\|_\otimes^2 &\le \|\mathcal{H}-\delta\widehat{\mathcal{H}}\|_\otimes^2\\ &= \|\mathcal{H}-\widehat{\mathcal{H}}+\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_\otimes^2\\ &= \|\mathcal{H}-\widehat{\mathcal{H}}\|_\otimes^2+2\Re\langle\mathcal{H}-\widehat{\mathcal{H}},\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\rangle_\otimes+\|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_\otimes^2\\ \Rightarrow 0 &\le 2\Re\langle\mathcal{H}-\widehat{\mathcal{H}},\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\rangle_\otimes+\|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_\otimes^2.\end{aligned} \tag{3.3.20}$$

Note that $\delta\widehat{\mathcal{H}}(s,p)$ may vary in both the numerators and the poles of $\widehat{\mathcal{H}}(s,p)$. Varying the numerator values $\widetilde{\phi_{i,j}}$ to $\widetilde{\phi_{i,j}}+\delta_{i,j}$ yields the familiar Lagrange conditions, so we focus on varying the poles in s and p. Let us start by looking at the poles in s and pick a $k\in\{1,\dots,r_s\}$ fixed but arbitrary. Denote the complex perturbation in polar form $\delta_{i,j}=\epsilon e^{i\theta}\in\mathbb{C}$ and define $\delta\widehat{\mathcal{H}}(s,p)$ as

$$\delta\widehat{\mathcal{H}}(s,p)=\sum_{j=1}^{r_p}\frac{\widetilde{\phi_{k,j}}}{(s-(\widetilde{\lambda}_k+\epsilon e^{i\theta}))(p-\widetilde{\pi}_j)}+\sum_{i\neq k}^{r_s}\sum_{j=1}^{r_p}\frac{\widetilde{\phi_{i,j}}}{(s-\widetilde{\lambda}_i)(p-\widetilde{\pi}_j)}. \tag{3.3.21}$$

The difference $\widehat{\mathcal{H}}(s,p)-\delta\widehat{\mathcal{H}}(s,p)$ can be simplified as

$$\begin{aligned}\widehat{\mathcal{H}}(s,p)-\delta\widehat{\mathcal{H}}(s,p) &= \sum_{j=1}^{r_p}\frac{\widetilde{\phi}_{k,j}}{(s-(\widetilde{\lambda}_k+\epsilon e^{i\theta}))(p-\widetilde{\pi}_j)}-\sum_{j=1}^{r_p}\frac{\widetilde{\phi}_{k,j}}{(s-\widetilde{\lambda}_k)(p-\widetilde{\pi}_j)}\\ &= \left(\frac{1}{(s-\widetilde{\lambda}_k-\epsilon e^{i\theta})}-\frac{1}{(s-\widetilde{\lambda}_k)}\right)\sum_{j=1}^{r_p}\frac{\widetilde{\phi}_{k,j}}{(p-\widetilde{\pi}_j)}\\ &= \frac{\epsilon e^{i\theta}}{(s-\widetilde{\lambda}_k)(s-\widetilde{\lambda}_k-\epsilon e^{i\theta})}\sum_{j=1}^{r_p}\frac{\widetilde{\phi}_{k,j}}{(p-\widetilde{\pi}_j)}.\end{aligned} \tag{3.3.22}$$

Substituting (3.3.22) into (3.3.20), we estimate

$$\begin{aligned} 0 &\le 2\Re\left\langle \mathcal{H}-\widehat{\mathcal{H}}, \frac{\epsilon \mathrm{e}^{\mathrm{i}\theta}}{(s-\widetilde{\lambda}_k)(s-\widetilde{\lambda}_k-\epsilon\mathrm{e}^{\mathrm{i}\theta})}\sum_{j=1}^{r_p}\frac{\widetilde{\phi}_{k,j}}{(p-\widetilde{\pi}_j)}\right\rangle_{\otimes} + \|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_{\otimes}^2 \\ &= 2\Re\left(\epsilon\mathrm{e}^{\mathrm{i}\theta}\sum_j \widetilde{\phi}_{k,j}\left\langle \mathcal{H}-\widehat{\mathcal{H}}, \frac{1}{\left((s-\widetilde{\lambda}_k)^2-\epsilon\mathrm{e}^{\mathrm{i}\theta}(s-\widetilde{\lambda}_k)\right)(p-\widetilde{\pi}_j)}\right\rangle_{\otimes}\right) + \|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_{\otimes}^2. \end{aligned} \tag{3.3.23}$$

We now turn our attention to the term $\|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_{\otimes}^2$ and observe that

$$\begin{aligned} \|\widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\|_{\otimes}^2 &= \langle \widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}, \widehat{\mathcal{H}}-\delta\widehat{\mathcal{H}}\rangle_{\otimes} \\ &= \left\langle \frac{\epsilon\mathrm{e}^{\mathrm{i}\theta}}{(s-\widetilde{\lambda}_k)(s-\widetilde{\lambda}_k-\epsilon\mathrm{e}^{\mathrm{i}\theta})}\sum_{j_1=1}^{r_p}\frac{\widetilde{\phi}_{k,j_1}}{(p-\widetilde{\pi}_{j_1})}, \frac{\epsilon\mathrm{e}^{\mathrm{i}\theta}}{(s-\widetilde{\lambda}_k)(s-\widetilde{\lambda}_k-\epsilon\mathrm{e}^{\mathrm{i}\theta})}\sum_{j_2=1}^{r_p}\frac{\widetilde{\phi}_{k,j_2}}{(p-\widetilde{\pi}_{j_2})}\right\rangle_{\otimes} \\ &= \sum_{j_1,j_2}\epsilon^2|\mathrm{e}^{\mathrm{i}\theta}|^2\overline{\widetilde{\phi}_{k,j_1}}\widetilde{\phi}_{k,j_2} \\ &\qquad \left\langle \frac{1}{(s-\widetilde{\lambda}_k)(s-\widetilde{\lambda}_k-\epsilon\mathrm{e}^{\mathrm{i}\theta})(p-\widetilde{\pi}_{j_1})}, \frac{1}{(s-\widetilde{\lambda}_k)(s-\widetilde{\lambda}_k-\epsilon\mathrm{e}^{\mathrm{i}\theta})(p-\widetilde{\pi}_{j_2})}\right\rangle_{\otimes} \\ &= O(\epsilon^2), \qquad (\epsilon\to 0). \end{aligned} \tag{3.3.24}$$

Combining (3.3.24) and (3.3.23), we get

$$0 \le 2\epsilon\Re\left(\mathrm{e}^{\mathrm{i}\theta}\sum_j \widetilde{\phi}_{k,j}\left\langle \mathcal{H}-\widehat{\mathcal{H}}, \frac{1}{\left((s-\widetilde{\lambda}_k)^2-\epsilon\mathrm{e}^{\mathrm{i}\theta}(s-\widetilde{\lambda}_k)\right)(p-\widetilde{\pi}_j)}\right\rangle_{\otimes}\right) + O(\epsilon^2). \tag{3.3.25}$$

Since $\epsilon > 0$, we can divide by ϵ and arrive at

$$0 \le 2\Re\left(\mathrm{e}^{\mathrm{i}\theta}\sum_j \widetilde{\phi}_{k,j}\left\langle \mathcal{H}-\widehat{\mathcal{H}}, \frac{1}{\left((s-\widetilde{\lambda}_k)^2-\epsilon\mathrm{e}^{\mathrm{i}\theta}(s-\widetilde{\lambda}_k)\right)(p-\widetilde{\pi}_j)}\right\rangle_{\otimes}\right) + O(\epsilon). \tag{3.3.26}$$

We are still free to choose the angle $\theta \in [0, 2\pi)$. Let θ be such that the inner product in

(3.3.26) is strictly positive and real. Then, for small enough $\epsilon > 0$, it follows that

$$0 \leq \left(\sum_j \widetilde{\phi}_{k,j} \left\langle \mathcal{H} - \widehat{\mathcal{H}}, \frac{1}{\left((s - \widetilde{\lambda}_k)^2 - \epsilon \mathrm{e}^{\mathrm{i}\theta} (s - \widetilde{\lambda}_k) \right) (p - \widetilde{\pi}_j)} \right\rangle_{\otimes} \right) = o(\epsilon). \tag{3.3.27}$$

Taking the limit $\epsilon \to 0$, we arrive at a contradiction unless

$$0 = \sum_j \widetilde{\phi}_{k,j} \left\langle \mathcal{H} - \widehat{\mathcal{H}}, \frac{1}{(s - \widetilde{\lambda}_k)^2 (p - \widetilde{\pi}_j)} \right\rangle_{\otimes} = \left\langle \mathcal{H} - \widehat{\mathcal{H}}, \sum_j \frac{\widetilde{\phi}_{k,j}}{(s - \widetilde{\lambda}_k)^2 (p - \widetilde{\pi}_j)} \right\rangle_{\otimes}. \tag{3.3.28}$$

Since k was chosen arbitrarily among the indices of λ, the condition in (3.3.28) has to hold for all indices $k = 1, \ldots, r_s$.

With the same argument for the poles in p, we arrive at a similar conclusion in the second argument as

$$0 = \sum_j \widetilde{\phi}_{i,k} \left\langle \mathcal{H} - \widehat{\mathcal{H}}, \frac{1}{(s - \widetilde{\lambda}_i)(p - \widetilde{\pi}_k)^2} \right\rangle_{\otimes} = \left\langle \mathcal{H} - \widehat{\mathcal{H}}, \sum_i \frac{\widetilde{\phi}_{i,k}}{(p - \widetilde{\pi}_k)^2 (s - \widetilde{\lambda}_i)} \right\rangle_{\otimes}. \tag{3.3.29}$$

Combining (3.3.28) and (3.3.29), we arrive at the same optimality conditions as in Theorem 3.3.4.

In the following section, we illustrate our approach to implement the optimality conditions in Theorem 3.3.4 or Theorem 3.3.4 on several examples with increasing complexity level.

3.4 Implementation and Numerical Examples

For a practical implementation of Theorem 3.3.4, we consider several options:

- Solve/approximate the nonlinear system of equations from the optimality conditions from Theorem 3.3.4;

- Minimize $\|\mathcal{H} - \widehat{\mathcal{H}}\|_{\otimes}^2$ as a function of the reduced model configuration (poles and residues). The objective function F then is

$$F(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}) := \|\widehat{\mathcal{H}} - \mathcal{H}\|_{\otimes}^2, \qquad \widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}, \tag{3.4.1}$$

where $\boldsymbol{\phi} = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \dots & \phi_{r_s,r_p} \end{bmatrix}^\top$, $\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 & \dots & \lambda_{r_s} \end{bmatrix}^\top$ and $\boldsymbol{\pi} = \begin{bmatrix} \pi_1 & \dots & \pi_{r_p} \end{bmatrix}^\top$.

Both approaches result in a valid reduced model configuration. Our choice of implementation is a *descent algorithm* to optimize over the pole locations $[\boldsymbol{\lambda}, \boldsymbol{\pi}]^\top \in \mathbb{C}^{r_s+r_p}$, where the proof of Theorem 3.3.4 yields a direct expression of the necessary gradients of $F(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ in (3.4.1) with respect to λ_i, $i = 1, \dots, r_s$ and π_j, $j = 1, \dots, r_p$. The numerator values are updated at every step by solving (3.3.4). This corresponds to using the objective function

$$F_{\boldsymbol{\lambda},\boldsymbol{\pi}}(\boldsymbol{\lambda}, \boldsymbol{\pi}) = F(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}), \qquad \text{with } \boldsymbol{\phi} \text{ via (3.3.4).} \tag{3.4.2}$$

We summarize this procedure in Algorithm 3.4.1.

Algorithm 3.4.1 $\mathcal{H}_2 \otimes L_2$ Gradient Descent

INPUT: Real transfer function $\mathcal{H} \in \mathcal{H}_2(\mathbb{C}_R \otimes \mathbb{D})$, order of approximation (r_s, r_p).

OUTPUT: Rational function $\widehat{\mathcal{H}}(s,p)$.

1. Chose initial selection of poles $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$.

2. Perform Gauss-Newton optimization of $\boldsymbol{\lambda}$ and $\boldsymbol{\pi}$ using Jacobian J_F of $F_{\boldsymbol{\lambda},\boldsymbol{\pi}}$ from (3.4.2) (computed by (3.3.15) and (3.3.16)).

3. At every step of the Gauss-Newton optimization, update numerator values $\boldsymbol{\phi}$ by solving (3.3.4).

4. The final reduced model is

$$\widehat{\mathcal{H}}(s,p) = \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}. \tag{3.4.3}$$

We emphasize that the model resulting from Algorithm 3.4.1 is, in fact, a locally optimal approximation to $\mathcal{H}(s,p)$ with respect to $\|\cdot\|_{\otimes}$, since the nonlinear optimization method used in Algorithm 3.4.1, upon convergence, finds a locally optimal subspace of $\mathcal{H}_2(\mathbb{C}_R) \otimes \mathcal{H}_2(\mathbb{D})$.

Note that, in practice, the ranks of the reduced model (r_s, r_p) are chosen via numerical rank of the corresponding (parametric) Loewner matrices; see [86]. For the initialization of the poles, $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$, we use optimal rational interpolation via IRKA in the following way. Let p_0 be fixed, then $\mathcal{H}_{p_0}(s) := \mathcal{H}(s,p_0) \in \mathcal{H}_2(\mathbb{C}_R)$ is a single variable transfer function. We can apply $\mathcal{H}_2$ optimal rational approximation using TF-IRKA [14] to find (locally) $\mathcal{H}_2$

optimal poles $\lambda_1, \ldots, \lambda_{r_s}$, which depend on p_0. In the same way

We proceed with some examples that illustrate the implementation of Algorithm 3.4.1.

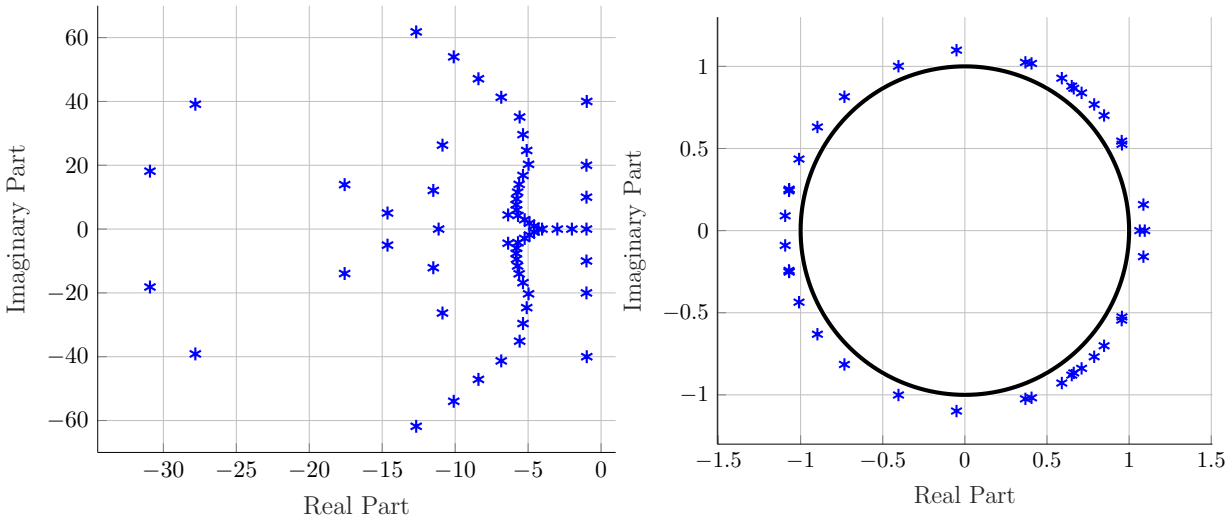
3.4.1 Small Synthetic Example

To illustrate our procedure, we start with a model $\mathcal{H}(s,p)$ of order $(n_s, n_p) = (100, 42)$ of the form

$$\mathcal{H}(s,p) = \sum_{i=1}^{n_s} \sum_{j=1}^{n_p} \frac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}, \tag{3.4.4}$$

which has the same structure as the reduced model. We aim for a reduced model that has fewer terms than $\mathcal{H}(s,p)$ to illustrate the reduction of complexity.

In particular, we chose a variation of [101, Example 3] with added parametric dependence. Poles in p are chosen at random with stability restrictions, i.e., boundedness of the $\mathcal{H}_2 \otimes L_2$ norm (the π_j that are added lie outside the unit disc) and closed under conjugation, so $\mathcal{H}(s,p)$ is real. We chose the order $n_s = 100$ (rather than the usual $n = 1006$). The position of the poles in s and p are shown in Figure 3.3 on the left and right, respectively.

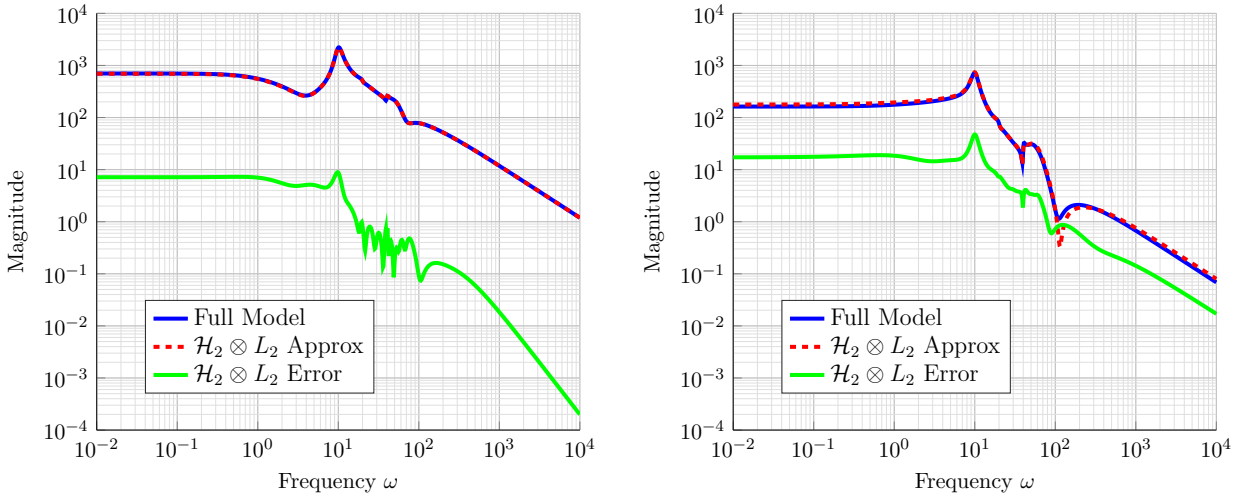


(a) Poles in s

(b) Poles in p

Figure 3.2: Pole configuration in s and p for full order model.

The reduced model $\widehat{\mathcal{H}}(s,p)$ is then constructed using Algorithm 3.4.1. We compare Bode plots of $\mathcal{H}(\cdot,p^*)$ and $\widehat{\mathcal{H}}(\cdot,p^*)$ at certain, fixed parameter values p^*, one chosen as the best approximation over $p \in \mathbb{D}$ and one chosen from among the worst. Those, together with absolute error curves (in dashed lines) are shown in Figure 3.3.



(a) $p = 0.8$ (b) $p = 0.83 + 0.29i$

Figure 3.3: Local frequency approximation quality for selected parameter values, comparison in Bode plot. Left side parameter for best frequency approximation, right side: worst frequency approximation.

In Figure 3.3, we also plot the locally best approximation, computed using non-parametric rational interpolation, IRKA. Note that the locally best $\mathcal{H}_2$ approximation for fixed parameter $p = p^*$ of the same order is better than the joint $\mathcal{H}_2 \otimes L_2$ approximation.

This is not surprising, since the full degrees of freedom r_s are available for a particular parameter choice. In the parametric approximation $\widehat{\mathcal{H}}(s, p)$, the poles are used across the parameter range, hence at every given point p^*, IRKA performs better but $\widehat{\mathcal{H}}(s, p)$ approximates $\mathcal{H}(s, p)$ over all $p \in \mathbb{D}$ with respect to the joint $\mathcal{H}_2 \otimes L_2$ norm.

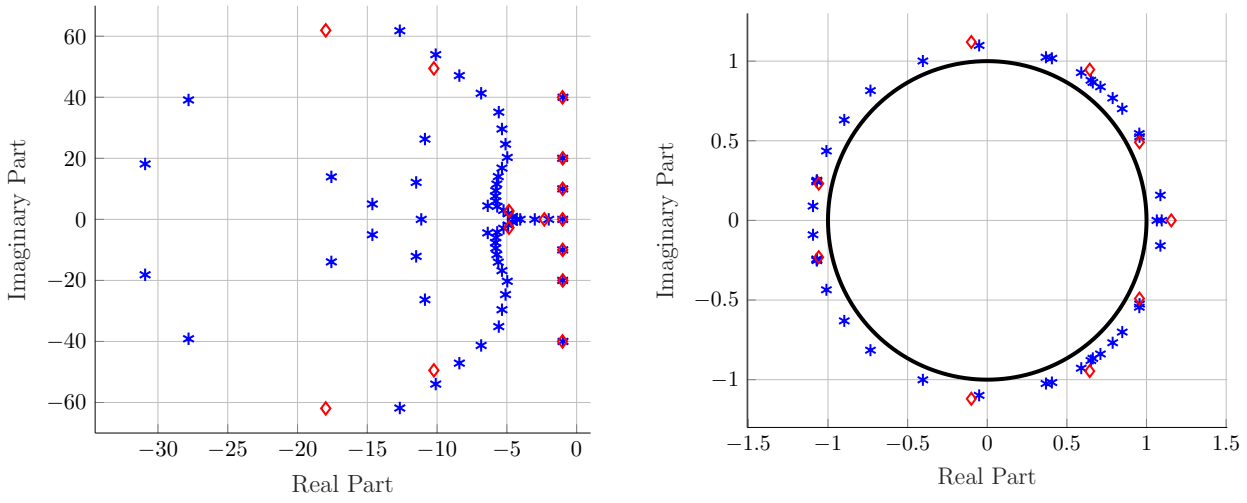


Figure 3.4: Comparison of pole configuration in s and p between full order model (blue) and $\mathcal{H}_2 \otimes L_2$ optimal reduced model (red).

Since the full model and reduced model are of the same structure, we are able to compare the poles in s and p directly; see Figure 3.4. Algorithm 3.4.1 tends to best match poles closest to the imaginary axis and the unit circle. To illustrate the approximation quality over the entire parameter range $p \in \mathbb{D}$, we show the pointwise error in Figure 3.5.

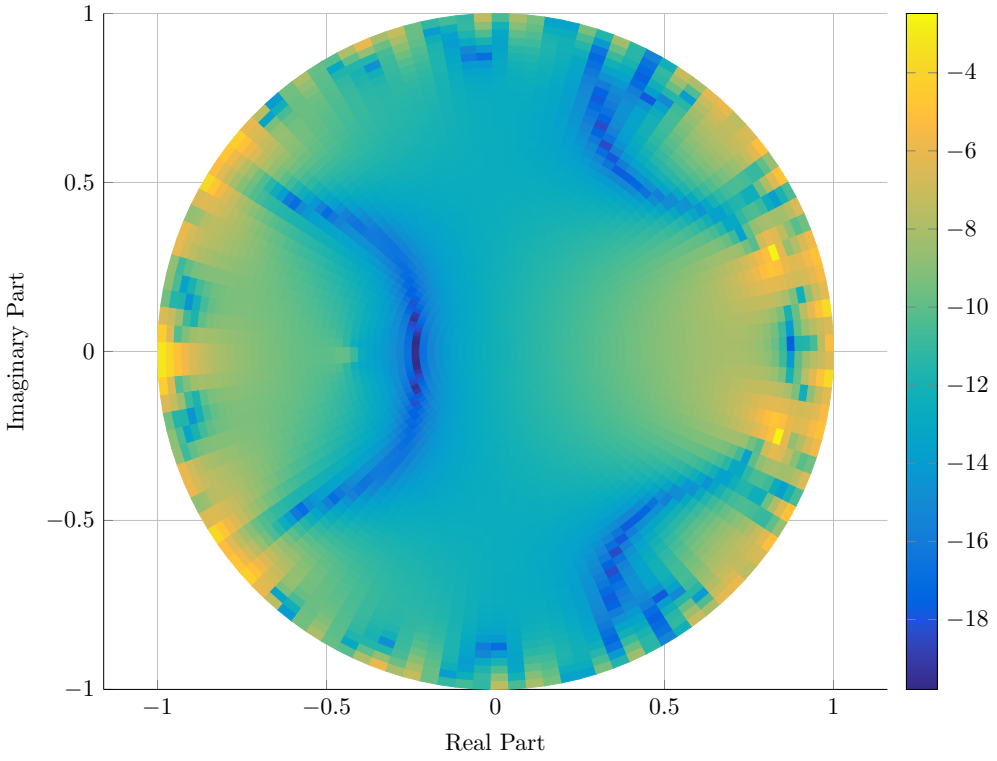


Figure 3.5: Approximation Quality in $\mathcal{H}_2$ norm for fixed $p \in \mathbb{D}$ on a logarithmic scale.

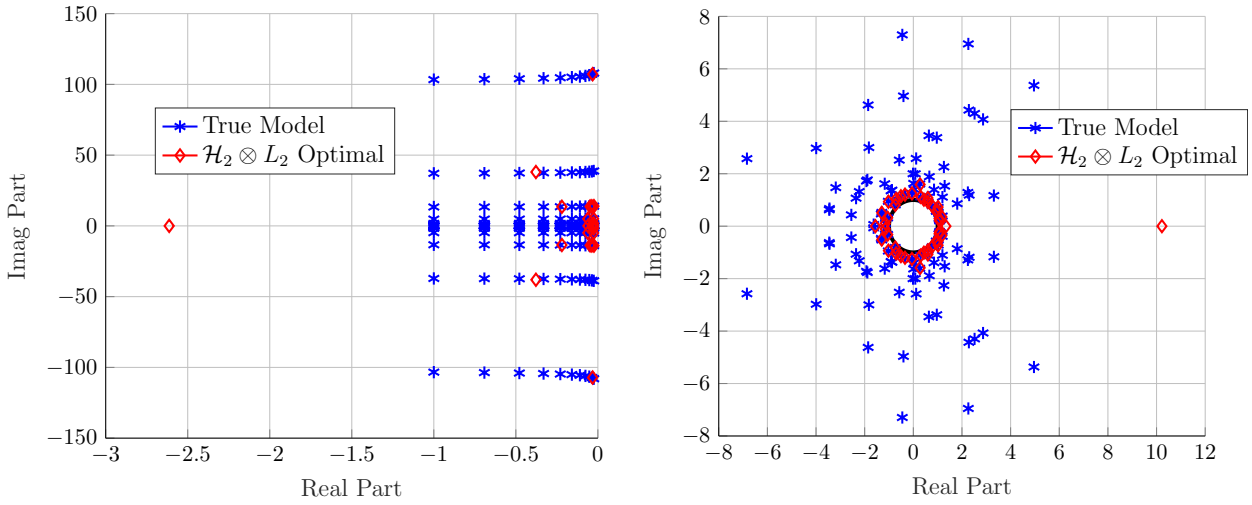
We observe that the reduced model is a good approximation over the whole parameter range $p \in \mathbb{D}$ (note the logarithmic scale in Figure 3.5). Even the lowest quality approximation, found at $p \approx 0.83 + 0.29\mathrm{i}$ (shown in Figure 3.3b) still captures the dominant characteristics of $\mathcal{H}(s, p)$. The joint $\mathcal{H}_2 \otimes L_2$ approximation error is

$$\frac{\|\mathcal{H} - \widehat{\mathcal{H}}\|_{\otimes}}{\|\mathcal{H}\|_{\otimes}} \approx 2.73710^{-6}. \tag{3.4.5}$$

3.4.2 Larger Synthetic Example

Similar to the previous example, we construct a function $\mathcal{H}(s,p)$ with separable poles of order $(n_s, n_p) = (220, 130)$ in s and p respectively. The numerator values of $\mathcal{H}(s,p)$ are chosen at random, under the constraint that $\mathcal{H}(s,p)$ is real.

We determine the order of the reduced model in s and p using the rank of Loewner matrices for a sufficient number of (s,p) samples. Our choice of tolerance $\epsilon = 10^{-4}$ to truncate the singular values of the corresponding (parametric) Loewner matrices yields rank $(r_s, r_p) = (26, 16)$. Observe that the order in s is decreased more than the order in r_p compared to the full model.



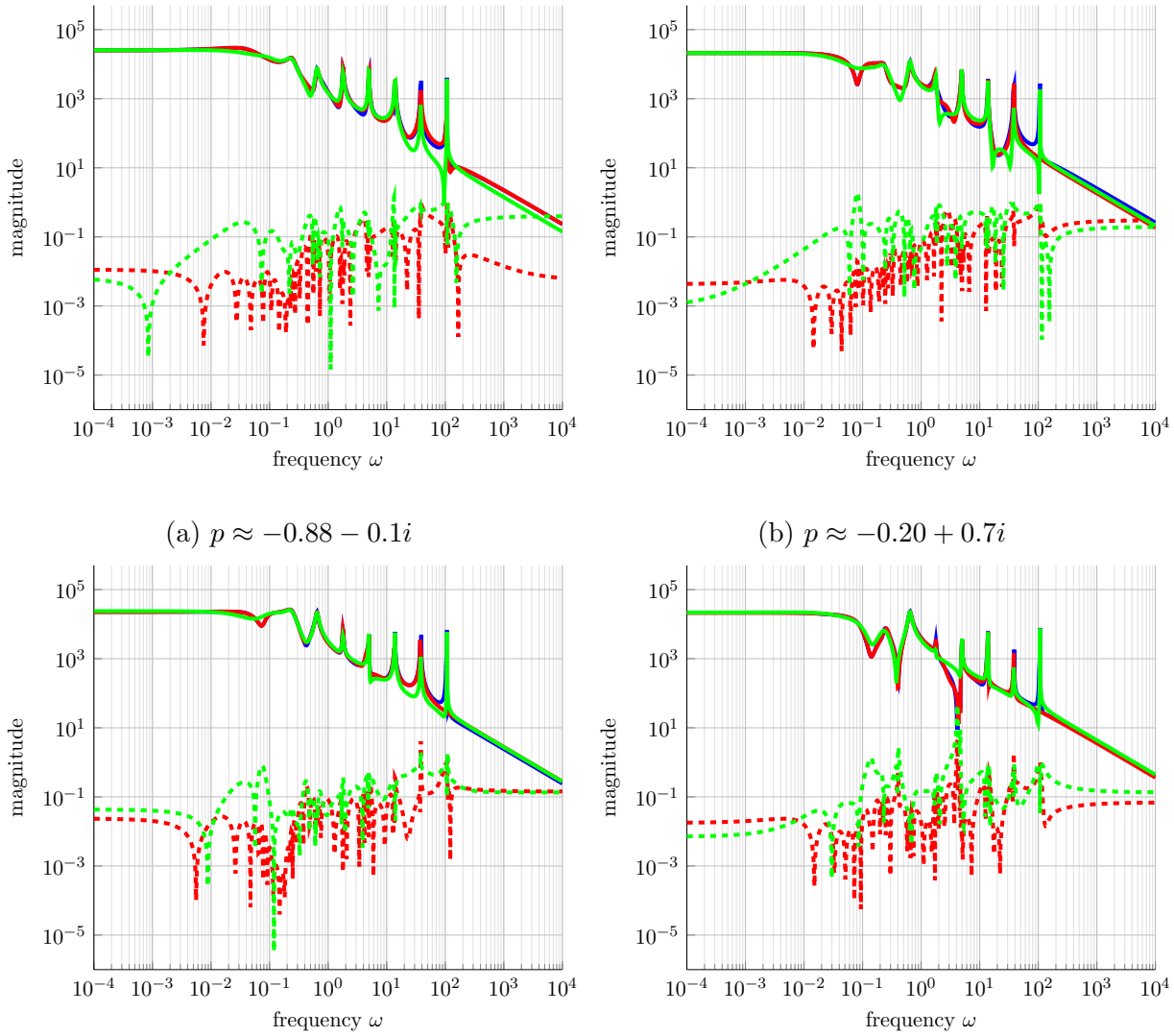
(a) Poles in s

(b) Poles in p

Figure 3.6: Poles of the synthetic transfer function $\mathcal{H}(s,p)$ and $\mathcal{H}_2 \otimes L_2$ optimal pole selection.

The poles in s of the full model $\mathcal{H}(s,p)$ (see Figure 3.6) are deliberately chosen with a

structure (not parallel to the real axis). The poles in p have been chosen at several radii at randomly generated angles θ, closed under conjugation. The initial pole position for the gradient descent algorithm is chosen by (non-parametric) IRKA for fixed s and p values. Performing Algorithm 3.4.1 yields the approximation result shown in Figure 3.7 at some representative parameter values.



(a) $p \approx -0.88 - 0.1i$

(b) $p \approx -0.20 + 0.7i$

(c) $p \approx -0.3 - 0.8i$

(d) $p \approx 0.76$

Figure 3.7: Approximation quality for the synthetic example of order $(220, 130)$, reduced order $(26, 16)$ at representative parameter values $p \in \mathbb{D}$. Full model in blue, IRKA in green, $\mathcal{H}_2 \otimes L_2$ approximation in red, absolute errors in dashed lines.

We observe that the optimal $\mathcal{H}_2 \otimes L_2$ approximant matches the transfer function behavior of $\mathcal{H}(s, p)$ well (see Figure 3.7); it captures almost all peaks and the asymptotic behavior

for $\omega \to \infty$ and $\omega \to 0$. In Figure 3.7, we also compare the $\mathcal{H}_2 \otimes L_2$ best approximant to the locally (fixed p) non-parametric optimal $\mathcal{H}_2$ approximation, computed using IRKA. As one might expect, the $\mathcal{H}_2$ optimal approximation for fixed p performs better than the $\mathcal{H}_2 \otimes L_2$ approximation. However, this comparison is only for illustration, since we require a parametric reduced model $\widehat{\mathcal{H}}(s, p)$.

Since we consider the unit disc $\mathbb{D}$ as the range of parameters, we want to observe the quality of the approximant over the entire parameter range in Figure 3.8. In particular, for fixed p^*, we compare the pointwise $\mathcal{H}_2$ error, defined by

$$\mathcal{E}(p) := \frac{\|\mathcal{H}(\cdot, p) - \widehat{\mathcal{H}}(\cdot, p)\|_{\mathcal{H}_2}}{\|\mathcal{H}(\cdot, p)\|_{\mathcal{H}_2}}, \tag{3.4.6}$$

which is shown in Figure 3.8.

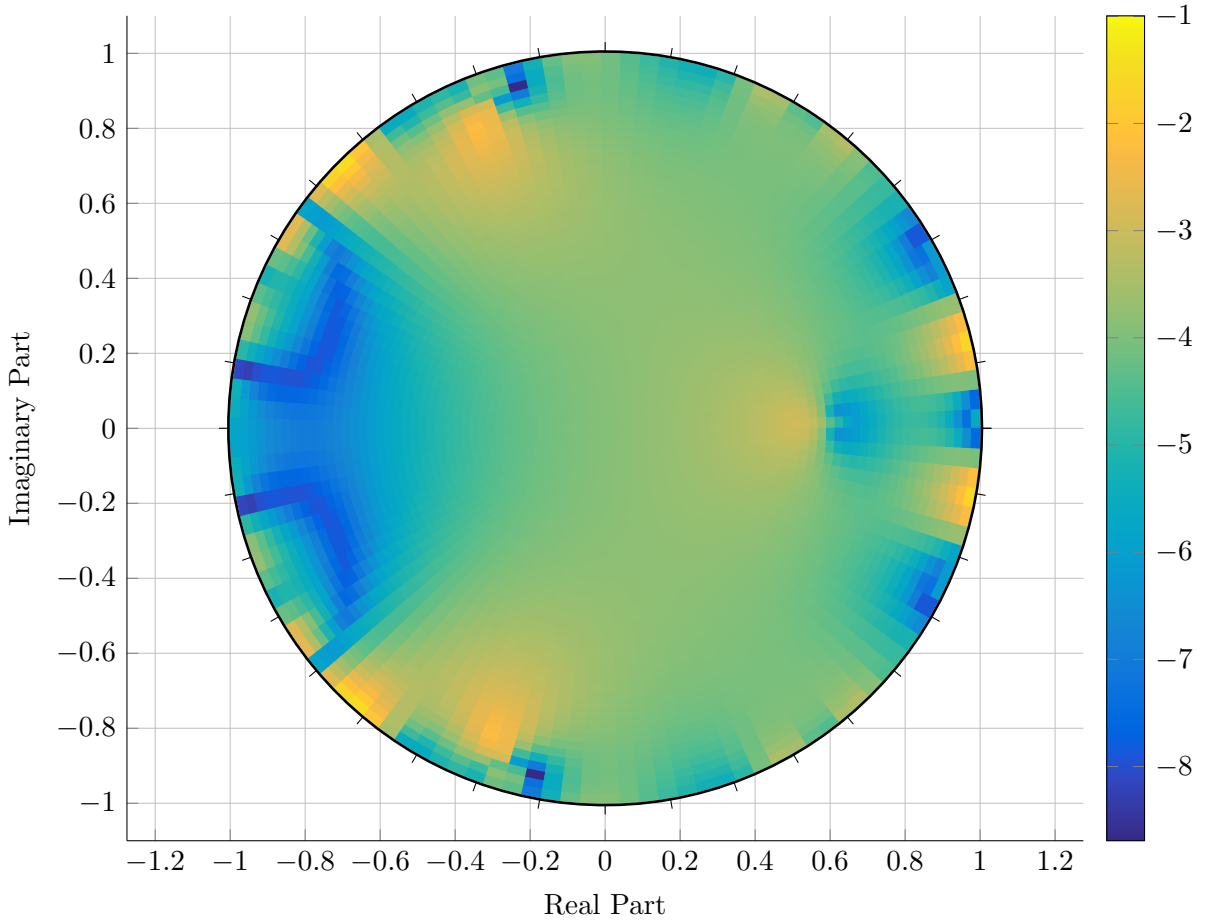
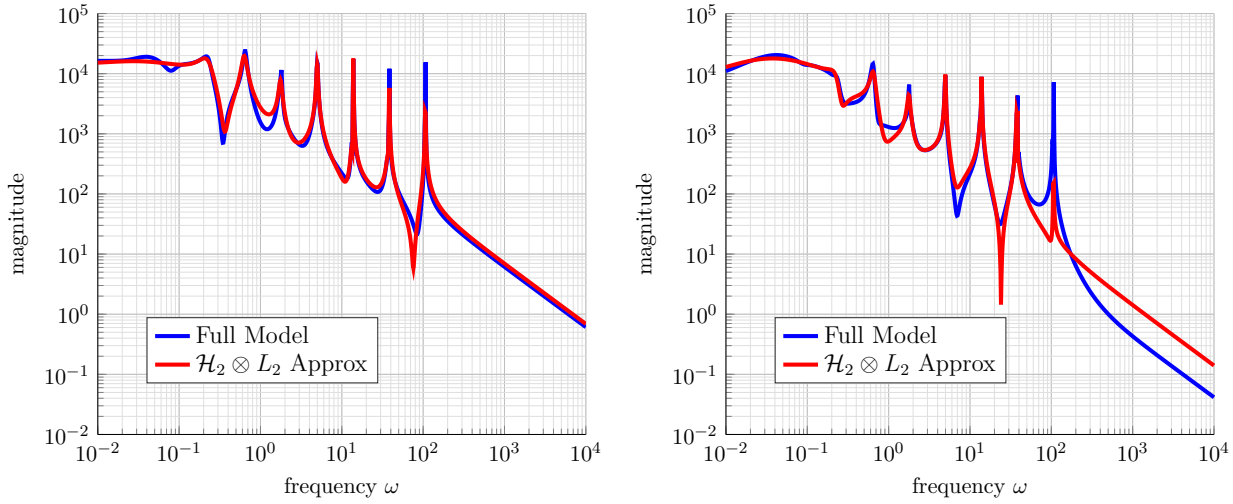


Figure 3.8: Approximation quality over the unit disc. Displayed are the *pointwise* relative errors from (3.4.6) on a logarithmic scale.

Observe that our approximation $\widehat{\mathcal{H}}(s, p^*)$ performs better at some parameter values $p^* \in \mathbb{D}$. We compare Bode plots of $\mathcal{H}(\cdot, p^*)$ and $\widehat{\mathcal{H}}(\cdot, p^*)$ in Figure 3.9 for representative selections of points $p^* \in \mathbb{D}$.



(a) Best local solution, $p \approx 0.97$

(b) Worst local solution, $p \approx -0.34 + 0.78i$

Figure 3.9: Comparison between best and worst solution for $\mathcal{H}_2 \otimes L_2$ approximation.

In Figure 3.9, we see that even the worst local approximation in Figure 3.8 still captures the peaks of the original transfer function $\mathcal{H}(s, p)$, while the best local approximation also captures the asymptotic behavior. The joint relative error in the $\mathcal{H}_2 \otimes L_2$ norm is

$$\frac{\|\mathcal{H} - \widehat{\mathcal{H}}\|_{\mathcal{H}_2 \otimes L_2}}{\|\mathcal{H}\|_{\mathcal{H}_2 \otimes L_2}} \approx 0.0743. \tag{3.4.7}$$

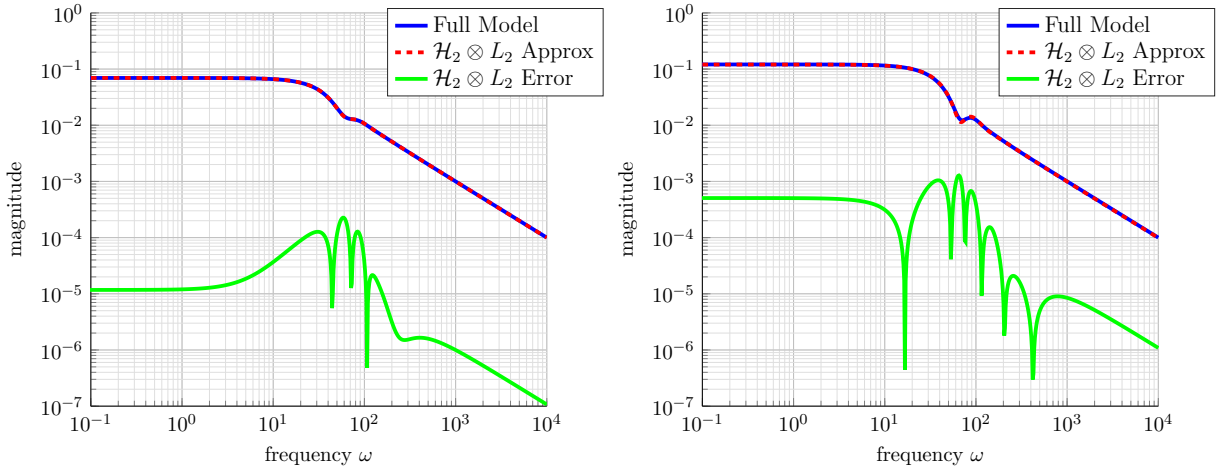
3.4.3 Convection-Diffusion Example

Recall the convection-diffusion example from Section 1.6.2 as

$$\begin{aligned} \dot{\mathbf{x}}(t) &= (\mathbf{A}_0 + p\mathbf{A}_1)\,\mathbf{x}(t) + \mathbf{B}u(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t). \end{aligned} \tag{3.4.8}$$

The parameter p represents the convection coefficient. Note that for this example, the main interest lies on *real* values of p, even though the $\mathcal{H}_2 \otimes L_2$ norm takes the entire unit disc under consideration. We fix the convection coefficient in the y-direction at $p_2 = 0.1$, the diffusion as $p_3 = 0.25$. The finite difference discretization has $N = 100$ degrees of freedom in both x and y direction, resulting in a matrix dimension of $10,000 \times 10,000$. For the reduced model, we choose the order $(r_s, r_p) = (8, 14)$.

Figure 3.10 shows the resulting approximation error in Bode plot form for selected real parameters $p \in [0, 1]$. Observe how both the asymptotic behavior for $\omega \to 0$ and $\omega \to \infty$ is captured well.



(a) $p = 0.09$ (b) $p = 0.6$

Figure 3.10: Bode plot comparison for the convection-diffusion example for representative parameter choices $p \in [0, 1]$

Similar to the previous experiments, we wish to see how the error behaves for arbitrary

$p \in \mathbb{D}$, not just at sampling points. In Figure 3.11, the $\mathcal{H}_2$ norm of the error $E(s,p) := \mathcal{H}(s,p) - \widehat{\mathcal{H}}(s,p)$ is displayed at fixed p values, $p^* \in \mathbb{D}$. The color code in Figure 3.11 represents the pointwise $\mathcal{H}_2$ error from (3.4.6). on a logarithmic scale.

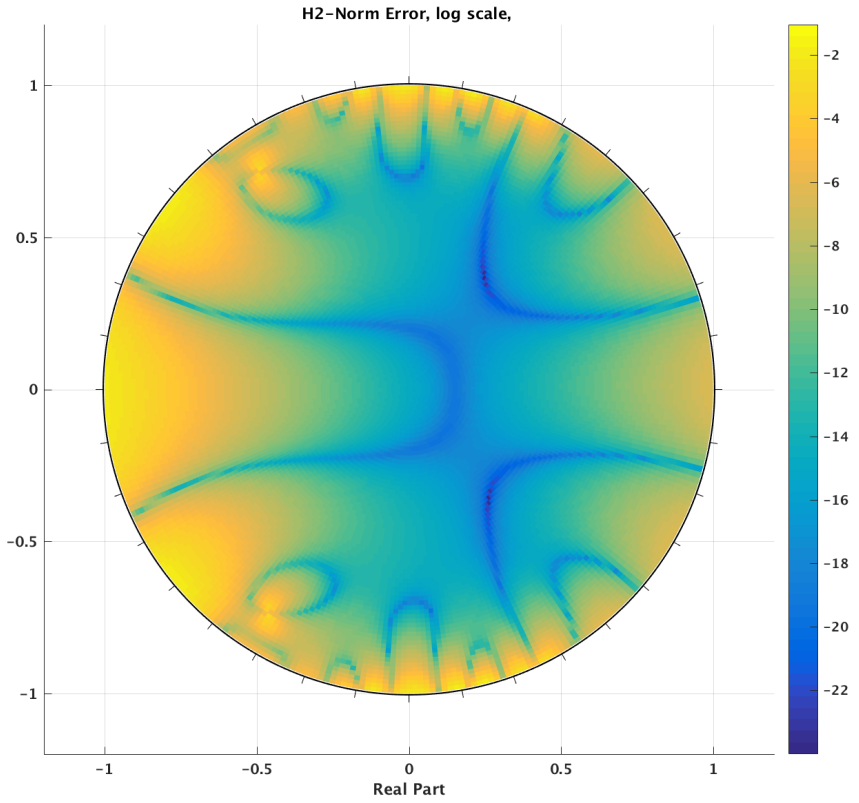
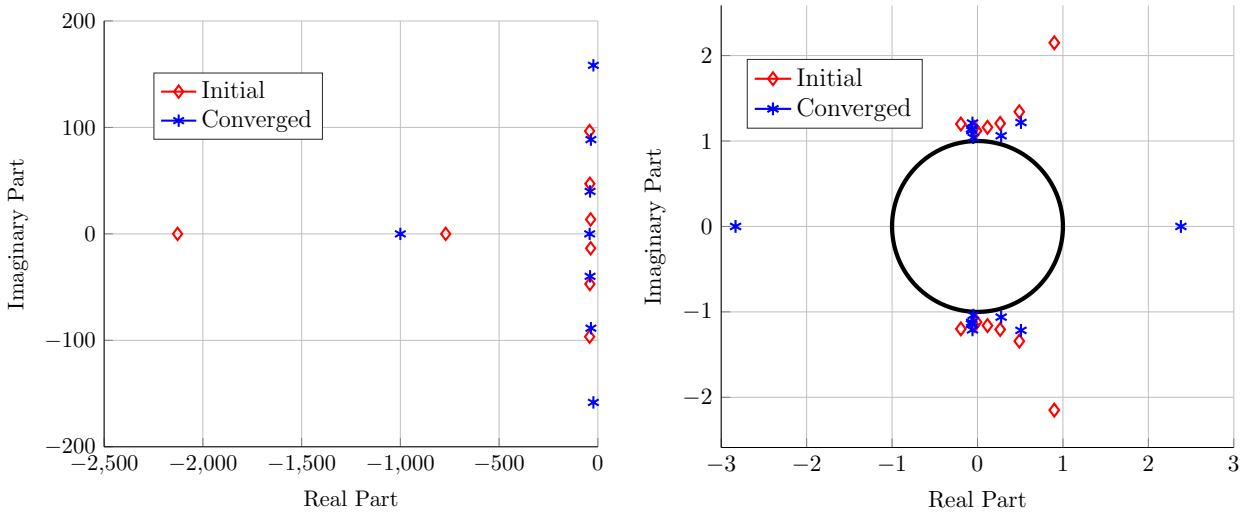


Figure 3.11: Approximation quality for the convection-diffusion model over the unit disc. Displayed is the *pointwise* relative $\mathcal{H}_2$ norm difference on a logarithmic scale.

We observe that the approximation quality over the real interval $[0,1]$ is better than on other points on the unit disc. However, values of $p_1 \in \mathbb{D}\backslash[0,1]$ carry no physical meaning. Also note the dark blue curves in Figure 3.11, which represent p^*-values where $\mathcal{H}(s,p^*)$ and

$\widehat{\mathcal{H}}(s, p^*)$ agree up to machine precision, with respect to the $\mathcal{H}_2$ norm in s.

In Figure 3.12, we compare the position of the poles λ_i, $i = 1, \ldots, 8$ and π_j, $j = 1, \ldots, 14$ between the initialization using IRKA and the converged $\mathcal{H}_2 \otimes L_2$ optimization algorithm.



(a) Poles in s

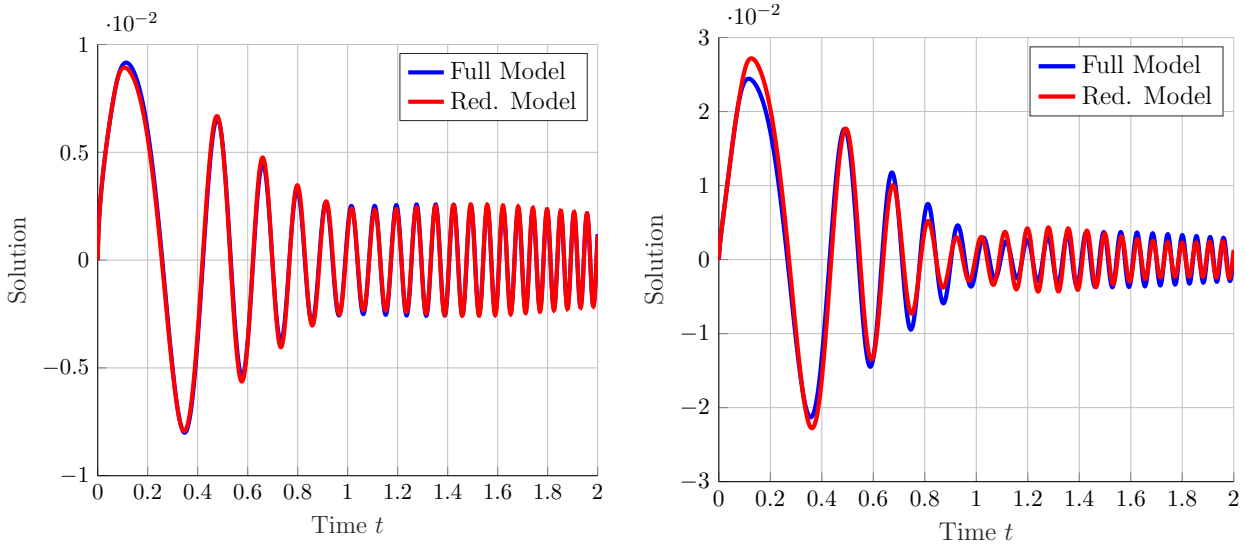
(b) Poles in p

Figure 3.12: Comparison of poles in s and p from initial conditions to converged poles of $\widehat{\mathcal{H}}(s, p)$

Observe in Figure 3.12 that initial poles, generated using IRKA with fixed parameter values are not optimal in the joint $\mathcal{H}_2 \otimes L_2$ norm. Our descent algorithm performs a pole reallocation with respect to the joint error measure.

Since (3.4.8) has physical meaning, we can compare the full model $\mathcal{H}(s, p)$ and the reduced model $\widehat{\mathcal{H}}(s, p)$ in the time domain. For simplicity, we chose a backward Euler discretization and the input $u(t)$ as the `chirp` signal from the standard MATLAB® library. Note that the `chirp` signal oscillates at different frequencies, enabling us to compare the system response

across a range of frequencies.



(a) $p_1 = 0.1$

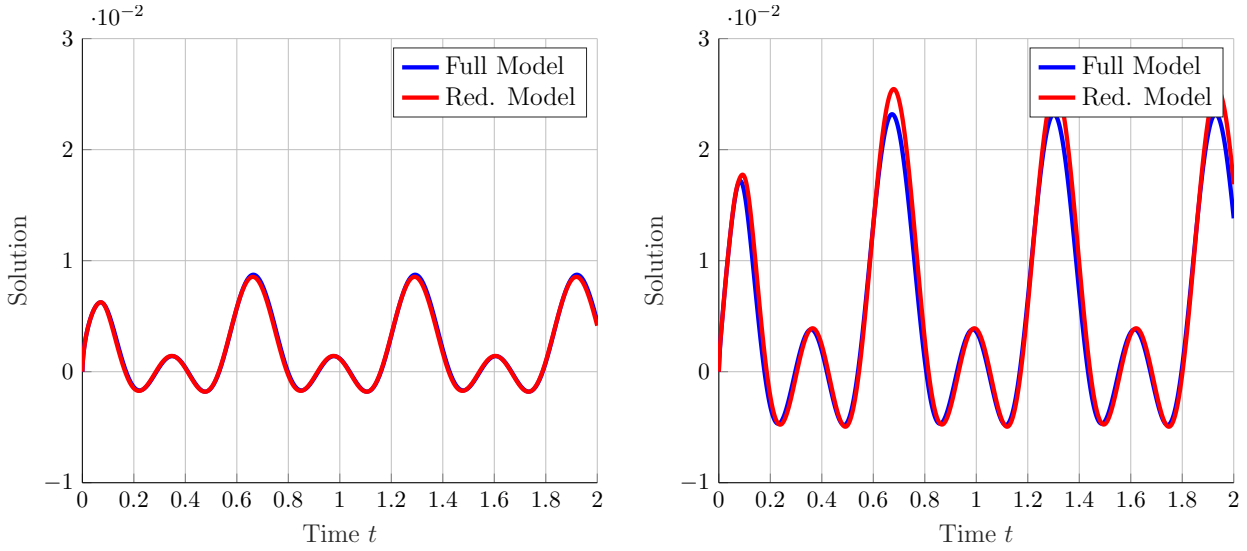
(b) $p_1 = 0.8$

Figure 3.13: Time domain comparison between full order and reduced model for the `chirp` input signal at selected parameter values.

To give another comparison, we choose a different input signal, the Dirichlet function

$$d(t) = \frac{\sin(Nt/2)}{N\sin(t/2)}, \qquad N \in \mathbb{N}, t > 0, \tag{3.4.9}$$

from the standard MATLAB® library `diric`. Here N represents the wavelength of the Dirichlet function $d(t)$. We pick $N = 5$ and scale the time samples by 10. In Figure 3.14, we compare the time domain output $y(t)$ and $\widehat{y}(t)$ in Figure 3.14.



(a) $p_1 = 0.1$ (b) $p_1 = 0.8$

Figure 3.14: Time domain comparison between full order and reduced model for the `diric` input signal at selected parameter values.

In Figure 3.13 and Figure 3.14, we observe that the reduced model matches the full order model in the time domain reasonably well, considering the reduction of order in $\widehat{\mathcal{H}}(s, p)$.

3.5 Summary of Contributions and Future Direction

We introduced a framework to compute optimal reduced models for parametric systems with respect to a joint $\mathcal{H}_2 \otimes L_2$ norm in frequency and parameter. Similar optimality conditions now with respect to both variables imply necessary optimality conditions. The theoretical framework relies on the structure of the inner product and the ansatz of a separable structure for the reduced model. Implementation details are discussed as well as illustrated by

numerical examples.

Future work includes a more structured way to determine the reduced order (r_s, r_p) similar to the McMillan degree [53] for single-variable systems for the case of separable pole-residue formulation.

Even though an $\mathcal{H}_2 \otimes L_2$ optimal reduced model considers p on the entire unit disc, the last example of a convection-diffusion problem already showed that this yields a good approximation on the real interval $[0, 1]$ which is of interest in the application. Future directions include a weighted version of $\mathcal{H}_2 \otimes L_2$ optimal model reduction focusing on subsets of physical relevance. This can be achieved by weighted two-variable Hardy spaces, an extension of the work in [1, 116] to the parametric case.

Extending optimal rational interpolation in two variables to the MIMO system case is another natural next step. We expect this to involve tangential interpolation conditions as in [59, 115].

Chapter 4

Parametric Vector Fitting

The Vector Fitting (VF) method, introduced in [63], targets the construction of a rational model in barycentric form measurement points in the frequency domain. Instead of interpolation, VF solves a least squares problem, making it more resistant to outliers in the measurement data as well as noise; see Section 2.5.2. In this chapter, we extend Vector Fitting to the parametric case by solving the least squares problem in the parameter and frequency domains via a special parametrization of the approximation. Several choices of such parametrizations are investigated and compared.

4.1 Goals and Problem Statement

In many physical situations, the measurements $\mathcal{H}(\xi_i)$ depend on a certain parameter configuration of the (physical) problem. It seems desirable to reflect the parameter dependency in

the realization $\widehat{\mathcal{H}}(s)$. Therefore, we would like to extend the SK iteration and Vector Fitting to the parametric setting.

Problem 4.1.1 *The problem for the parametric case is the following.*

Given a data set $\{\xi_i, \mu_j, \mathcal{H}(\xi_i, \mu_j)\}_{i=1,j=1}^{i=m_s,j=m_p} \subset \mathbb{C} \times \mathbb{C} \times \mathbb{C}$*, find a stable two-variable rational function that fits the data set in a least squares sense. More precisely, we want to construct* $\widehat{\mathcal{H}}(s,p)$ *that solves*

$$\sum_{i=1}^{m_s} \sum_{j=1}^{m_p} \left| \widehat{\mathcal{H}}(\xi_i, \mu_j) - \mathcal{H}(\xi_i, \mu_j) \right|^2 \to \min . \tag{4.1.1}$$

We focus on a one-dimensional parameter dependence, $p \in \mathbb{C}$ and extend the problem to several parameters in Section 4.3.

4.2 Parametric Vector Fitting

Following the process leading to Vector Fitting, we recall a parametric version of the SK-iteration with more details on the choice of basis functions as well as its extension to Vector Fitting in both polynomial and rational settings. The distinction is the allocation of the poles: for the parametric SK-iteration, a parametric least squares problem is solved with a fixed set of basis functions, the resulting coefficients are computed from a nonlinear least-squares problem, which results in an iterative procedure. For our extension of Vector Fitting, we allow for an adaptive choice of basis functions.

Our contributions can be summarized as follows

1. We show how to combine local models to solve a global least squares problem.

2. Expand Vector Fitting to the parametric case with a pole reallocation step, based on local models.

We emphasize that our goal is to construct a parametric model $\widehat{\mathcal{H}}$ solving

$$\sum_{i=1}^{m_s}\sum_{j=1}^{m_p}\left|\widehat{\mathcal{H}}(\xi_i,\mu_j)-\mathcal{H}(\xi_i,\mu_j)\right|^2\to\min\,. \tag{4.2.1}$$

We refer to this as the *global approximation problem*, since, in contrast to the minimization problem for local models (4.2.8), the minimization is performed simultaneously over the frequency and parameter samples.

4.2.1 Problem Setting

To formulate our approximation problem precisely, for the parameter domain $\mathcal{P}$, we consider the following space of functions:

$$\mathcal{R}_{\text{stab}}\mathcal{H}_2(\mathbb{C}_R\times\mathcal{P}):=\{\mathcal{H}:\mathbb{C}^2\to\mathbb{C}\mid\mathcal{H}(\cdot,p)\in\mathcal{RH}_2(\mathbb{C}_R)\ \text{ for all }\ p\in\mathcal{P}\} \tag{4.2.2}$$

The problem at hand is to find a two-variable model $\widehat{\mathcal{H}}(s,p)$ so that

$$\widehat{\mathcal{H}}=\underset{\mathcal{G}\in\mathcal{R}_{\text{stab}}\mathcal{H}_2(\mathbb{C}_R\times\mathcal{P})}{\arg\min}\ \sum_{i=1}^{m_s}\sum_{j=1}^{m_p}|\mathcal{G}(\xi_i,\mu_j)-\mathcal{H}(\xi_i,\mu_j)|^2 \tag{4.2.3}$$

The natural generalization of the barycentric form of $\widehat{\mathcal{H}}(s)$ used for VF is the two-variable barycentric form

$$\widehat{\mathcal{H}}(s,p)=\frac{n(s,p)}{d(s,p)}:=\frac{\sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\dfrac{\phi_{i,j}}{(s-\lambda_i)(p-\pi_j)}}{1+\sum_{i=1}^{r_s}\sum_{j=1}^{r_p}\dfrac{\varphi_{i,j}}{(s-\lambda_i)(p-\pi_j)}}, \tag{4.2.4}$$

for $\alpha_{i,j}, \beta_{i,j}, \lambda_i, \pi_j \in \mathbb{C}, \quad i = 1, \ldots, r_s,\ j = 1, \ldots, r_p$. Mimicking the pole-update step in VF, in the parametric setting requires to find the zeros of

$$d(s,p) = 1 + \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{\varphi_{i,j}}{(s-\lambda_i)(p-\pi_j)}. \tag{4.2.5}$$

Solving $d(s,p) = 0$ is equivalent to finding $(s^*, p^*) \in \mathbb{C}^2$ so that

$$1 + \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \frac{\varphi_{i,j}}{(s^*-\lambda_i)(p^*-\pi_j)} = 0. \tag{4.2.6}$$

Which, in turn, leads to the polynomials root finding problem

$$\prod_{k=1}^{r_s} \prod_{\ell=1}^{r_p} (s^*-\lambda_k)(p^*-\pi_\ell) + \sum_{i=1}^{r_s} \sum_{j=1}^{r_p} \varphi_{i,j} \prod_{\substack{k=1,\\ k\neq i}}^{k=r_s} \prod_{\substack{\ell=1\\ \ell\neq j}}^{\ell=r_p} (s^*-\lambda_k)(p^*-\pi_\ell) = 0 \tag{4.2.7}$$

The solutions are curves in $\mathbb{C}^2$ rather than discrete points as in the non-parametric case and thus cannot directly provide information for the pole update. Instead, we take a two-step approach to solving the least squares problem (4.2.3).

We define *local models* or *macro-models* as $\widehat{\mathcal{H}}_j(s)$ that solve

$$\widehat{\mathcal{H}}_j := \arg \min_{\mathcal{G} \in R\mathcal{H}_2(\mathbb{C}_R)} \sum_{i=1}^{m_s} |\mathcal{G}(\xi_i) - \mathcal{H}(\xi_i, \mu_j)|^2, \qquad \text{for} \qquad j = 1, \ldots, m_p. \tag{4.2.8}$$

Each $\widehat{\mathcal{H}}_j(s)$ can be constructed using classical VF on the data set $\{\xi_i, \mathcal{H}(\xi_i, \mu_j)\}_{i=1}^{m_s}$, resulting in a set of models models $\{\mu_j, \widehat{\mathcal{H}}_j\}_{j=1}^{m_p}$.

Our proposed two-step approach is organized as follows.

1. Construct local models $\{\mu_j, \widehat{\mathcal{H}}_j\}_{j=1}^{m_p}$

2. Combine the local models using a *joint* least squares approach in frequency and parameter

$$\widehat{\mathcal{H}}(s,p) := \sum_{k=1}^{m_p} \alpha_k(p)\widehat{\mathcal{H}}_k(s) = \sum_{k=1}^{m_p} \left(\sum_{\ell=1}^{r_p} \beta_{k,\ell} P_\ell(p) \right) \widehat{\mathcal{H}}_k(s), \tag{4.2.9}$$

with suitably chosen basis functions $P_\ell(p)$.

The specific choices of $P_\ell(p)$ and the computation of $\beta_{k,\ell}$ determines characteristic properties of the resulting model.

Let $\widehat{\mathcal{H}}_k(s) = \sum_{j=1}^{r_s} \frac{\phi_{j,k}}{s-\lambda_{j,k}}$ be a pole-residue form of the local models $\widehat{\mathcal{H}}_k(s)$ in (4.2.9). A simple computations reveals that

$$\widehat{\mathcal{H}}(s,p) = \sum_{k=1}^{m_p} \left(\sum_{\ell=1}^{r_p} \beta_{k,\ell} \frac{1}{p-\pi_\ell} \right) \sum_{j=1}^{r_s} \frac{\phi_{j,k}}{s-\lambda_{j,k}} = \sum_{k=1}^{m_p} \sum_{j=1}^{r_s} \sum_{\ell=1}^{r_p} \frac{\beta_{k,\ell}}{p-\pi_\ell} \frac{\phi_{j,k}}{s-\lambda_{j,k}}, \tag{4.2.10}$$

which we recognize as a two-variable pole-residue form, in constrast to the two-variable barycentric form of the generalized SK iteration in [54].

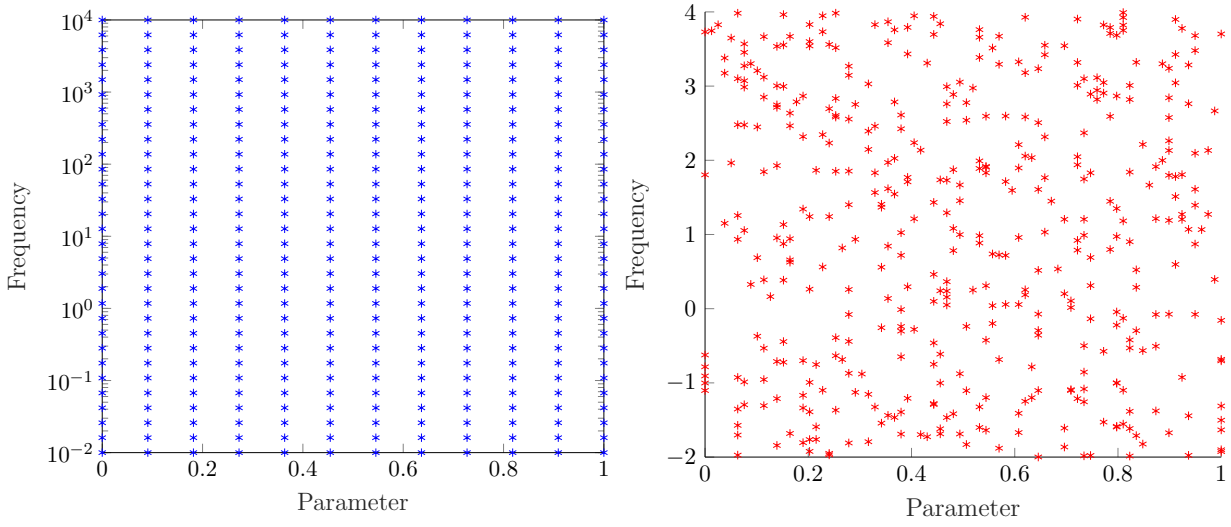
The main difference of our approach is that instead of using interpolation as in [10], for example, we chose the coefficients $\beta_{k,\ell}$ in (4.2.9) to solve a *global* least squares problem (4.2.3). In particular, we explore the choice of polynomial and rational functions. Prior knowledge, if available, can give rise to a suitable selection of functions $P_\ell(p)$ for the parametric dependence, for example sin/cos, rational functions or various polynomials [56]. Moreover, in the case of rational functions $P_\ell(p)$, we allow for an adaptive pole update in Section 4.2.4.

4.2.2 A Note on Sampling Points

Note that, in general, we do not require the (ξ_i, μ_j) grid of measurement points to be homogeneous. To clarify our definition of a homogeneous grid: Let $[\xi_1, \ldots, \xi_{m_s}]^\top \in \mathbb{C}^{m_s}$ be sampling points in s and $[\mu_1, \ldots, \mu_{m_p}]^\top \in \mathbb{C}^{m_p}$ sampling points in p. The tensor grid of those sampling points can be interpreted as ordered pairs

$$(\xi_1, \mu_1), (\xi_1, \mu_2), \ldots, (\xi_1, \mu_{m_p}), (\xi_2, \mu_1), \ldots, (\xi_{m_s}, \mu_{m_p}), \tag{4.2.11}$$

where all combinations of ξ_i and μ_j are listed. It may be possible, however, that the frequency sampling points ξ_i depend on the parameter μ_j (or vice versa). Then only a subset of the combinations in (4.2.11) are in the set of joint sampling points. This distinction is illustrated in Figure 4.1.



(a) Homogeneous grid

(b) Non-Homogeneous grid

Figure 4.1: Examples of homogeneous and non-homogeneous grid in frequency and parameter.

Observe how the grid in Figure 4.1a is a tensor product of vectors of sampling points in s and p, while no such structure can be found in Figure 4.1b. Since this distinction does not impact the presentation of our algorithm, we assume a homogeneous measurement grid for ease of exposition. In contrast to, for example, parametric Loewner interpolation [76], which requires tensor grids in frequency and parameter, our implementation remains valid for any choice of sampling grid.

4.2.3 Fixed Basis Functions $P_\ell(p)$

In this section, we solve the joint least-squares problem (4.2.3) using the proposed two-step approach with a fixed choice of coefficient functions $P_\ell(p)$. For illustration, we chose polynomial coefficient functions for $P_\ell(p)$. Assume the following form for $\widehat{\mathcal{H}}(s,p)$:

$$\widehat{\mathcal{H}}(s,p) = \sum_{k=1}^{m_p} \alpha_k(p)\widehat{\mathcal{H}}_k(s), \qquad \text{with} \qquad \alpha_k(p) = \sum_{\ell=1}^{r_p} \beta_{k,\ell} P_{k,\ell}(p), \tag{4.2.12}$$

for a choice of polynomials $P_{k,\ell}(p)$. $N(k)$ denotes the order of polynomials in $\alpha_k(p)$ for the k-th parameter μ_k, $k = 1, \ldots, m_p$.

Substituting (4.2.12) into (4.2.3), we have to solve the minimization problem

$$\sum_{i=1}^{m_s}\sum_{j=1}^{m_p} \left| \sum_{k=1}^{m_p}\sum_{\ell=1}^{r_p} \beta_{k,\ell} P_\ell(\mu_j)\widehat{\mathcal{H}}_k(\xi_i) - \mathcal{H}(\xi_i,\mu_j) \right|^2 \to \min, \tag{4.2.13}$$

for $\beta_{i,j}$, which we can recast as the linear least-squares problem $\mathcal{A}\boldsymbol{\beta} = \mathbf{b}$ (linear in the

coefficients $\beta_{i,j}$) as

$$\underbrace{\begin{bmatrix} \widehat{\mathcal{H}}_1(\xi_1)P_1(\mu_1) & \widehat{\mathcal{H}}_1(\xi_1)P_2(\mu_1) & \cdots & \widehat{\mathcal{H}}_{m_p}(\xi_1)P_{r_p}(\mu_1) \\ \widehat{\mathcal{H}}_1(\xi_1)P_1(\mu_2) & \widehat{\mathcal{H}}_1(\xi_1)P_2(\mu_2) & \cdots & \widehat{\mathcal{H}}_{m_p}(\xi_1)P_{r_p}(\mu_2) \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{\mathcal{H}}_1(\xi_{m_s})P_1(\mu_{m_p}) & \widehat{\mathcal{H}}_1(\xi_{m_s})P_2(\mu_{m_p}) & \cdots & \widehat{\mathcal{H}}_{m_p}(\xi_{m_s})P_{r_p}(\mu_{m_p}) \end{bmatrix}}_{=:\mathcal{A}} \underbrace{\begin{bmatrix} \beta_{1,1} \\ \beta_{1,2} \\ \vdots \\ \beta_{m_p,r_p} \end{bmatrix}}_{=:\mathrm{vec}(\boldsymbol{\beta})} = \underbrace{\begin{bmatrix} \mathcal{H}(\xi_1,\mu_1) \\ \mathcal{H}(\xi_1,\mu_2) \\ \vdots \\ \mathcal{H}(\xi_{m_s},\mu_{m_p}) \end{bmatrix}}_{=:\mathbf{b}} \tag{4.2.14}$$

The dimensions are $\mathcal{A} \in \mathbb{C}^{m_s m_p \times r_s r_p}$, $\mathrm{vec}(\boldsymbol{\beta}) \in \mathbb{C}^{r_s r_p}$ and $\mathbf{b} \in \mathbb{C}^{m_s m_p}$. The resulting parametric model can be evaluated at any parameter and frequency value and requires evaluation of the basis functions $P_\ell(p)$, matrix-multiplication with the coefficients $\boldsymbol{\beta}$ as well as frequency-evaluations of all local reduced models $\widehat{\mathcal{H}}_k(s)$. We summarize this procedure in Algorithm 4.2.1.

Algorithm 4.2.1 Parametric VF - Fixed Basis Functions $P_\ell(p)$

INPUT: Weights $w_{i,j} > 0,\ i = 0,\dots,m_s\ j = 1,\dots,m_p$, measurement data $\{\xi_i, \mu_j, \mathcal{H}(\xi_i,\mu_j)\}_{i=1,j=1}^{i=m_s,j=m_p}$, basis fuctions $\{P_\ell\}_{\ell=1}^{r_p}$

OUTPUT: Rational function $\widehat{\mathcal{H}}(s,p)$

1. Define $\boldsymbol{\Delta} := \mathrm{diag}[w_{1,1}, w_{1,2}, \dots, w_{m_s,m_p}]$ and the measurement vector $\mathbf{b}$ as

$$\mathbf{b} := [\mathcal{H}(\xi_1,\mu_1), \mathcal{H}(\xi_1,\mu_2), \dots, \mathcal{H}(\xi_{m_s},\mu_{m_p})]^\top \in \mathbb{C}^{m_s m_p}.$$

2. Construct *local* VF approximations of order r_j for each fixed parameter μ_j

$$\widehat{\mathcal{H}}_j = \underset{\mathcal{G}(s)\in R\mathcal{H}_2(\mathbb{C}_R)}{\arg\min} \sum_{i=1}^{m_s} |\mathcal{G}(\xi_i) - \mathcal{H}(\xi_i,\mu_j)|^2, \tag{4.2.15}$$

3. Construct the matrix $\mathcal{A}$ as in (4.2.14)

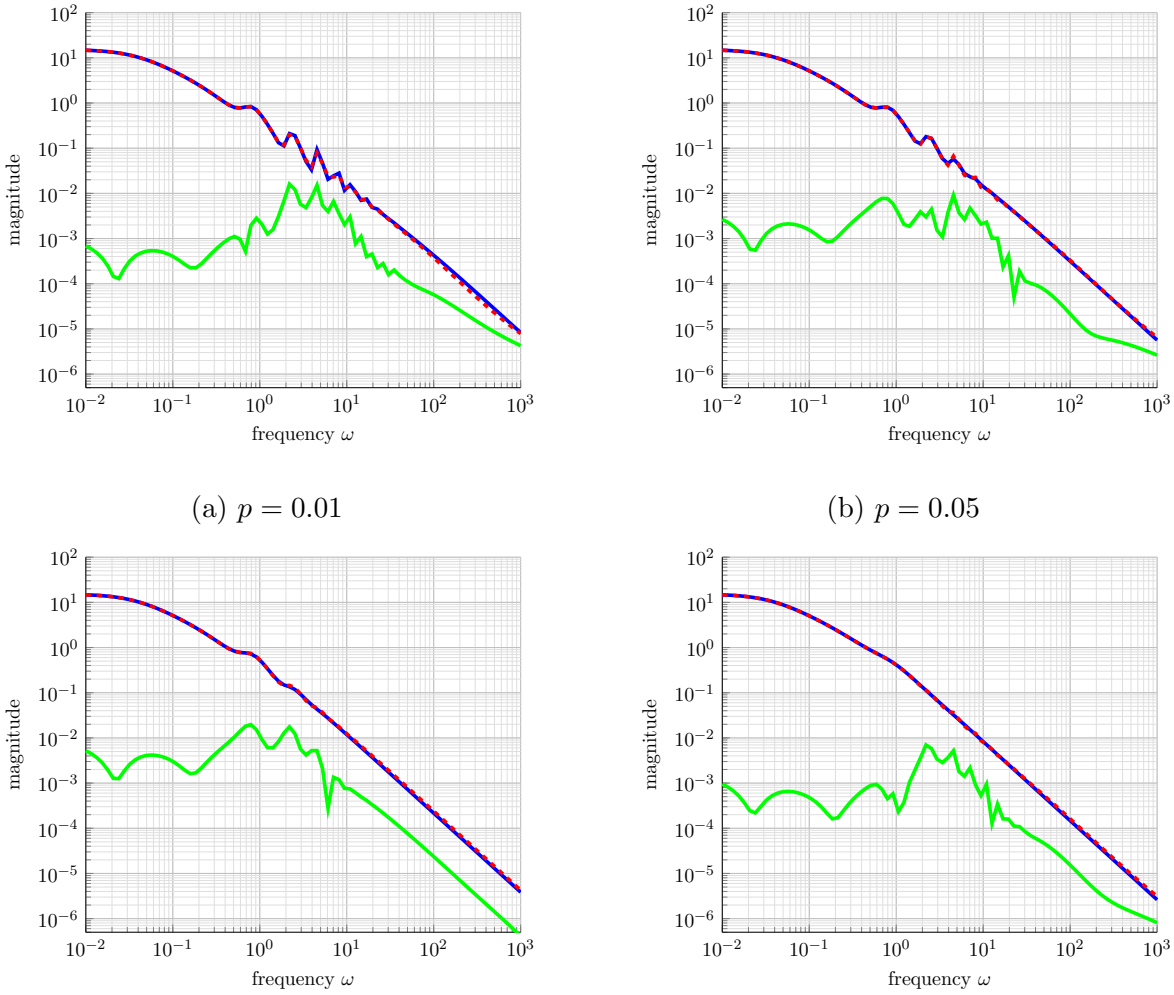
4. Solve the linear system $\|\boldsymbol{\Delta}(\mathcal{A}\boldsymbol{\beta} - \mathbf{b})\|_2^2 \to \min$

5. The final reduced model is given by

$$\widehat{\mathcal{H}}(s,p) = \sum_{k=1}^{m_p}\sum_{\ell=1}^{r_p} \beta_{k,\ell} P_\ell(p) \widehat{\mathcal{H}}_k(s) \tag{4.2.16}$$

The main concept in Algorithm 4.2.1 is the two-step process to first find reduced models in s with fixed μ_j, then combine those to a final two-variable model. We emphasize that the coefficients $\beta_{k,\ell}$ in (4.2.14) are constructed to solve the *joint* least squares problem (4.2.1), rather than interpolation.

Example 4.2.1 We illustrate Algorithm 4.2.1 on the beam example [99], introduced in Section 1.6.1. The state space formulation of the finite difference discretization has $n = 4000$ degrees of freedom. We sample at $n_s = 80$ logarithmically spaced frequency points between 10^{-4} and 10^{1} on the imaginary axis (the main dynamic range of the model) and $m_p = 7$ samples, chosen equally spaced on the interval $[0.01, 1]$. The orders for the local Vector Fitting models are chosen adaptively in p by the rank of the corresponding Loewner matrices. We use polynomials of order $r_p = 6$ for $P_\ell(p)$ for the approximation, in particular Berstein polynomials; see [5, 44].



(a) $p = 0.01$

(b) $p = 0.05$

(c) $p = 0.22$

(d) $p = 1$

Figure 4.2: Error plot of parametric vector fitting using polynomial basis functions. The original function is shown in blue, the approximation in dashed red lines and the (absolute) point wise error in green.

The error shown in Figure 4.2 at representative sampling points in $\mu_1, \ldots, \mu_7$ used to generate the approximation. We observe that the error is small across at the sampled points. Table 4.2 gives the numerical values of the least squares residual from (4.2.13) for several choices of

polynomial bases for $P_\ell(p)$. Note that our implementation chooses a particular set of basis functions with the lowest least squares error in (4.2.13), in this case Bernstein polynomials. Our algorithm chooses the set of functions $P_\ell(p)$ with the lowest least squares residue.

Choice of $P_\ell(p)$	Monomial	Bernstein	Legendre	Chebyshev
Rel. LS Residual	0.1047	0.0654	0.0822	0.0758

Table 4.1: Comparison of least squares error from (4.2.3) for various choices of $P_\ell(p)$ in (4.2.12)

For the functions $P_\ell(p)$ with the minimal least squares residual, chosen by the error from Table 4.1, we compare the approximation error of $\widehat{\mathcal{H}}(s,p)$ at sampled parameter values $\mu_1, \ldots, \mu_7$ in Table 4.2.

	Rel. Error	
Parameter	PVF $\|\cdot\|_2$	PVF $\|\cdot\|_\infty$
0.01	8.17e-04	1.22e-03
0.05	4.74e-04	7.04e-04
0.10	6.35e-04	9.39e-04
0.22	3.96e-04	4.24e-04
0.50	5.23e-04	6.35e-04
0.90	1.34e-04	1.43e-04
1.00	4.33e-04	7.27e-04

Table 4.2: Error in the polynomial parametric Vector Fitting approximation. Evaluated at original sample points μ_i, $j = 1, \ldots, 7$

While the approximation at sampled points μ_j, $j = 1, \ldots, m_p$ is low (see Table 4.2), it is appropriate to compare the approximation error on the continuous parameter interval $[0, 1]$.

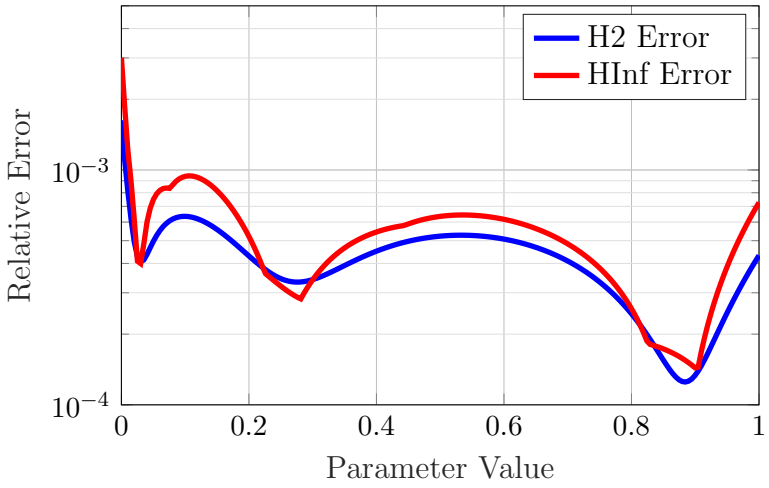


Figure 4.3: Error plot of parametric vector fitting using polynomial functions $P_\ell(p)$ over the paramter interval $[0.01, 0.8]$.

In Figure 4.3, we show the relative error between $\widehat{\mathcal{H}}(s,p)$ and the original model $\mathcal{H}(s,p)$ in the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norm. We observe that even outside of the sampled points, the approximation performs well.

So far, we assumed the basis function $P_\ell(p)$ to be fixed and compute coefficients to solve the global least squares problem (4.2.3). In the non-parametric setting, VF (see Section 2.5.4) dynamically updates the basis functions (in that case rational functions) to achieve a better approximation and improved conditioning of the least squares problem. We mimique this procedure in the parametric case in the following subsection.

4.2.4 Adaptive Basis Functions $P_\ell(p)$ and Variable Projection

Our choice of adaptive basis functions $P_\ell(p)$ in (4.2.12) are rational functions. Explicitly, consider

$$P_\ell(p) := \frac{1}{p-\pi_\ell}, \qquad \pi_\ell \in \mathbb{C}\backslash\mathcal{P},\ \ell = 1,\dots,r_p. \tag{4.2.17}$$

Denote $\boldsymbol{\pi} : \begin{bmatrix} \pi_1 & \dots & \pi_{r_p} \end{bmatrix}^\top \in \mathbb{C}^{r_p}$.

Our form of $\widehat{\mathcal{H}}(s,p)$ then becomes

$$\widehat{\mathcal{H}}(s,p) = \sum_{k=1}^{m_p} \alpha_k(p)\widehat{\mathcal{H}}_k(s) = \sum_{k=1}^{m_p}\sum_{\ell=1}^{r_p} \frac{\beta_{k,\ell}}{p-\pi_\ell}\widehat{\mathcal{H}}_k(s). \tag{4.2.18}$$

For fixed π_ℓ, $\ell = 1,\dots,r_p$, the coefficients $\beta_{k,\ell}$ can be computed via the least-squares problem

$$\|\mathcal{A}(\boldsymbol{\pi})\boldsymbol{\beta} - \mathbf{b}\|_2^2 \to \min, \tag{4.2.19}$$

where $\mathbf{b} = [\mathcal{H}(\xi_1,\mu_1), \mathcal{H}(\xi_1,\mu_2), \dots, \mathcal{H}(\xi_{m_s},\mu_{m_p})]^\top$ and

$$\mathcal{A}(\boldsymbol{\pi}) := \begin{bmatrix} \frac{\widehat{\mathcal{H}}_1(\xi_1)}{\mu_1-\pi_1} & \frac{\widehat{\mathcal{H}}_1(\xi_1)}{\mu_1-\pi_2} & \dots & \frac{\widehat{\mathcal{H}}_1(\xi_1)}{\mu_1-\pi_{r_p}} & \frac{\widehat{\mathcal{H}}_2(\xi_1)}{\mu_1-\pi_1} & \dots & \frac{\widehat{\mathcal{H}}_{m_p}(\xi_1)}{\mu_1-\pi_{r_p}} \\ \frac{\widehat{\mathcal{H}}_1(\xi_1)}{\mu_2-\pi_1} & \frac{\widehat{\mathcal{H}}_1(\xi_1)}{\mu_2-\pi_2} & \dots & \frac{\widehat{\mathcal{H}}_1(\xi_1)}{\mu_2-\pi_{r_p}} & \frac{\widehat{\mathcal{H}}_2(\xi_1)}{\mu_2-\pi_1} & \dots & \frac{\widehat{\mathcal{H}}_{m_p}(\xi_1)}{\mu_2-\pi_{r_p}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\widehat{\mathcal{H}}_1(\xi_{m_s})}{\mu_{m_p}-\pi_1} & \frac{\widehat{\mathcal{H}}_1(\xi_{m_s})}{\mu_{m_p}-\pi_2} & \dots & \frac{\widehat{\mathcal{H}}_1(\xi_{m_s})}{\mu_{m_p}-\pi_{r_p}} & \frac{\widehat{\mathcal{H}}_2(\xi_{m_s})}{\mu_{m_p}-\pi_1} & \dots & \frac{\widehat{\mathcal{H}}_{m_p}(\xi_{m_s})}{\mu_{m_p}-\pi_{r_p}} \end{bmatrix}. \tag{4.2.20}$$

Observe that the mapping $\pi \mapsto \mathcal{A}(\boldsymbol{\pi})$ in (4.2.20) is nonlinear. Hence an adaptive choice of poles $\boldsymbol{\pi}_\ell$, $\ell = 1,\dots,r_p$ introduces a nonlinear, nonconvex optimization problem.

A standard method to solve a problem like (4.2.19) for both $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ is variable projection [52, 72]. The main principle in variable projection is that the minimization problem (4.2.19)

can be solved for $\boldsymbol{\beta}$ directly using the pseudo-inverse $\mathcal{A}(\boldsymbol{\pi})^+$ when $\boldsymbol{\pi}$ is fixed. More precisely, we use the Moor-Pensore inverse for $\mathcal{A}(\boldsymbol{\pi})^+$ so that

$$\boldsymbol{\beta}(\boldsymbol{\pi}) := \mathcal{A}(\boldsymbol{\pi})^+ \mathbf{b}, \qquad \text{with} \qquad \mathcal{A}(\boldsymbol{\pi})^+ = \left(\mathcal{A}(\boldsymbol{\pi})^* \mathcal{A}(\boldsymbol{\pi})\right)^{-1} \mathcal{A}(\boldsymbol{\pi})^*. \tag{4.2.21}$$

Replacing $\boldsymbol{\beta}(\boldsymbol{\pi})$ with $\mathcal{A}(\boldsymbol{\pi})^+ \mathbf{b}$ in (4.2.19) leads to an optimization problem for $\boldsymbol{\pi}$:

$$\|\mathcal{A}(\boldsymbol{\pi})\mathcal{A}(\boldsymbol{\pi})^+ \mathbf{b} - \mathbf{b}\|_2^2 \to \min. \tag{4.2.22}$$

Observe that $\mathcal{A}(\boldsymbol{\pi})\mathcal{A}(\boldsymbol{\pi})^+$ is an orthogonal projection onto the range of $\mathcal{A}(\boldsymbol{\pi})$, hence the name *variable projection*.

To solve (4.2.22) for $\boldsymbol{\pi}$, one can apply any nonlinear optimization method. We choose a Gauss-Newton implementation from [31]. This approach is summarized in Algorithm 4.2.2.

Algorithm 4.2.2 Variable Projection - Implementation via Gauss-Newton

INPUT: Measurement vector $\mathbf{b}$ and map $\boldsymbol{\pi} \mapsto \mathcal{A}(\boldsymbol{\pi})$.

OUTPUT: Poles $\boldsymbol{\pi}$ and coefficients $\boldsymbol{\beta}$.

1. Chose initial $\boldsymbol{\pi}_0$

2. For $k = 1, 2, \ldots$ until converged

   (a) Solve the linear least-squares problem

$$\boldsymbol{\beta}^{(k)} = \arg\min_{\boldsymbol{\beta}} \left\| \mathbf{A}(\boldsymbol{\pi}^{(k)})\boldsymbol{\beta} - \mathbf{b} \right\|_2^2 \tag{4.2.23}$$

   (b) Compute the residue $\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}(\boldsymbol{\pi}^{(k)})\boldsymbol{\beta}^{(k)}$

   (c) Assemble Jacobian $\mathbf{J}^{(k)}$ by

$$\mathbf{J}^{(k)} := \left[\frac{\partial[\mathbf{A}(\boldsymbol{\pi}^{(k)})\boldsymbol{\beta}^{(k)}]}{\partial \boldsymbol{\pi}_1} \quad \cdots \quad \frac{\partial[\mathbf{A}(\boldsymbol{\pi}^{(k)})\boldsymbol{\beta}^{(k)}]}{\partial \boldsymbol{\pi}_{r_p}} \right] \tag{4.2.24}$$

   (d) Compute update step $\mathbf{d}^{(k)}$ by

$$\mathbf{d}^{(k)} = \arg\min_{\mathbf{d}} \left\| \mathbf{J}^{(k)}\mathbf{d} - \mathbf{r}^{(k)} \right\|_2^2 \tag{4.2.25}$$

   (e) Update $\boldsymbol{\pi}^{(k+1)} = \boldsymbol{\pi}^{(k)} + \mathbf{d}^{(k)}$.

Note that one may introduce a weight $\boldsymbol{\Delta} = \mathrm{diag}(w_{i,j})$ to the minimization problem (4.2.23) and instead solve

$$\|\boldsymbol{\Delta}(\mathbf{A}(\boldsymbol{\pi}^{(k)})\boldsymbol{\beta} - \mathbf{b})\|_2^2 \to \min\,. \tag{4.2.26}$$

To improve convergence rates, it may be necessary to include a line search for the update step $\boldsymbol{\pi}^{(k+1)} = \boldsymbol{\pi}^{(k)} + \mathbf{d}^{(k)}$ for some applications, which can easily be done with standard methods, e.g., using Armijo linesearch [97]. Further note that Algorithm 4.2.2 is not restricted to rational functions $P_\ell(p)$ but can be adapted to any basis functions $P_\ell(p)$ that involve a nonlinear parametrization.

We apply variable projection to the parametric Vector Fitting problem (4.2.13) by combining Algorithm 4.2.2 and Algorithm 4.2.1. The final algorithm is presented in Algorithm 4.2.3.

Algorithm 4.2.3 Parametric VF - Rational Coefficient Functions & Variable Projection

---

INPUT: Measurements $\{\xi_i, \mu_j, \mathcal{H}(\xi_i, \mu_j)\}_{i=1,j=1}^{i=m_s,j=m_p}$, map $\boldsymbol{\pi} \mapsto \mathcal{A}(\boldsymbol{\pi})$, order r_p

OUTPUT: Rational function $\widehat{\mathcal{H}}(s, p)$

1. Construct local Vector Fitting approximations of order r_p at each fixed parameter μ_j

$$\widehat{\mathcal{H}}_j = \underset{\mathcal{G} \in \mathcal{RH}_2(\mathbb{C}_R)}{\arg\min} \sum_{i=1}^{m_s} |\mathcal{G}(\xi_i) - \mathcal{H}(\xi_i, \mu_j)|^2. \tag{4.2.27}$$

2. Apply variable projection, Algorithm 4.2.2 to find $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$.

3. Assemble $\widehat{\mathcal{H}}(s, p)$ as

$$\widehat{\mathcal{H}}(s, p) = \sum_{k=1}^{m_s} \sum_{\ell=1}^{r_p} \frac{\beta_{k,\ell}}{p - \pi_\ell} \widehat{\mathcal{H}}_k(s). \tag{4.2.28}$$

First, we highlight the properties of the rational approximation by applying Algorithm 4.2.3 to a synthetic transfer function, that is constructed with explicit rational dependence in p.

Example 4.2.2 We consider the synthetic transfer function model

$$\mathcal{H}(s,p) = \sum_{k=1}^{6} \frac{\phi_k}{p-\pi_k} \widehat{\mathcal{H}}_k(s), \qquad \pi_k \in \mathbb{C}, \tag{4.2.29}$$

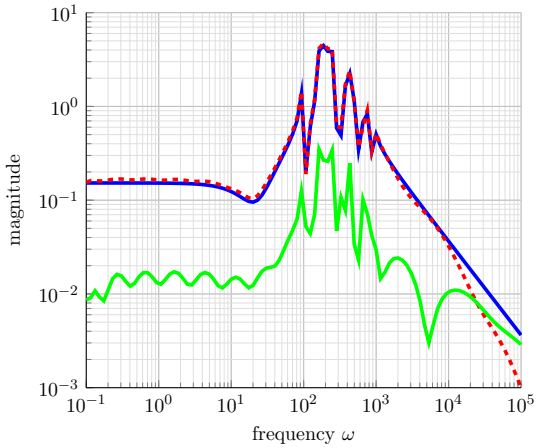
where the π_k are closed under conjugation. Explicitely, we chose

$$\boldsymbol{\pi} := \begin{bmatrix} 0.4 & 2 \pm 1.5\mathrm{i} & 4 \pm 0.8\mathrm{i} & 5.1 \end{bmatrix}^\top. \tag{4.2.30}$$

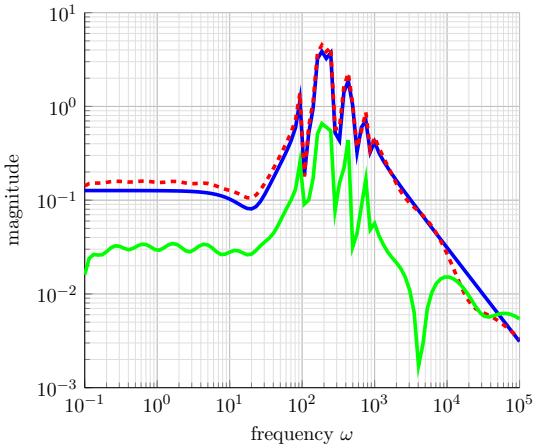
The local models $\widehat{\mathcal{H}}_k(s)$ are chosen as synthetic transfer functions based on [101, Ex. 3].

We sample at 100 frequency samples, logarithmically spaced between 10^{-1} and 10^5 and 8 parameter samples linearly spaced on $\mathcal{P} = [1, 5]$.

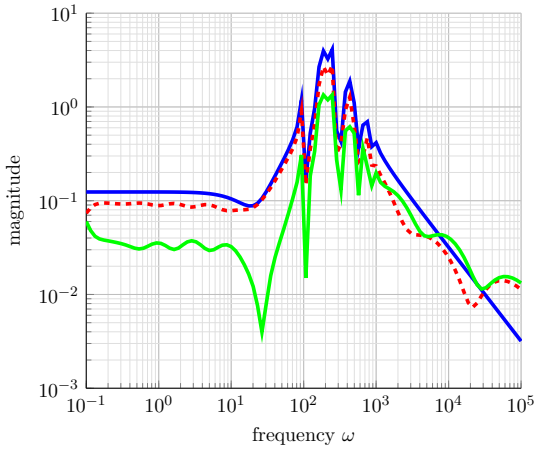
To illustrate the approximation quality, we compare the approximation for polynomial and rational $P_\ell(p)$ in over the frequency range via Bode plots in Figure 4.4 and Figure 4.5, respectively.



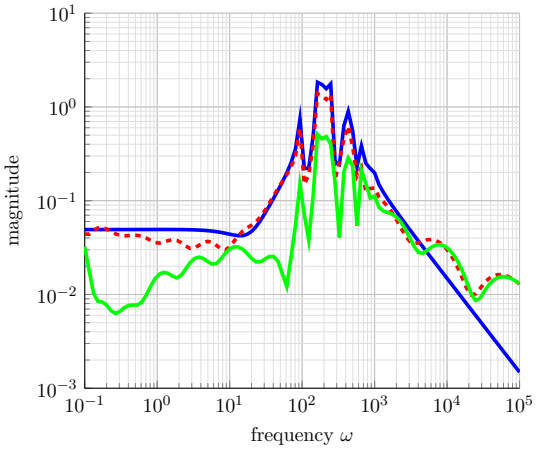
(a) $p = 1.57$



(b) $p = 2.14$

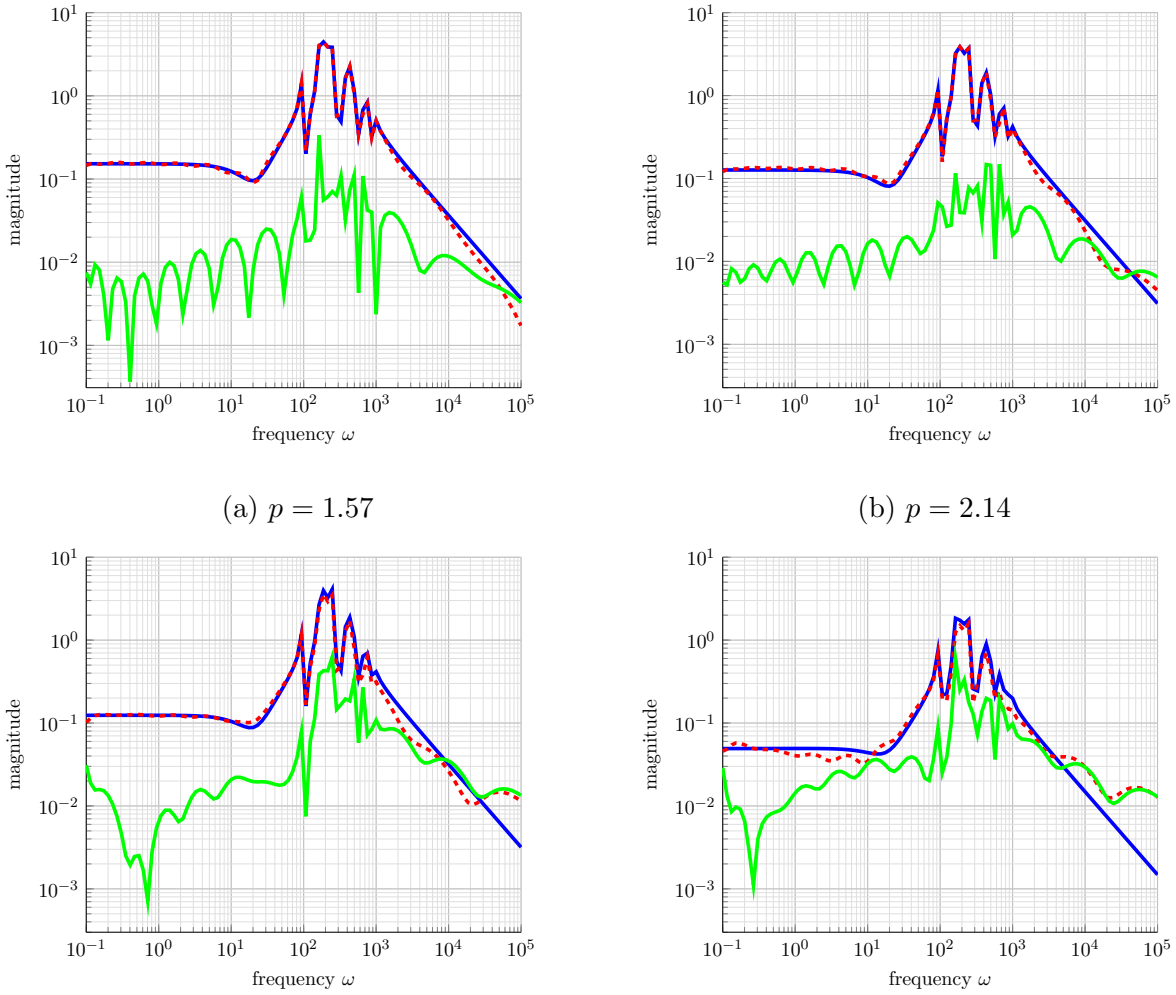


(c) $p = 3.28$



(d) $p = 3.86$

Figure 4.4: BODE plot comparison between *polynomial* basis functions in p the original model $\mathcal{H}(s, p)$.



(a) $p = 1.57$

(b) $p = 2.14$

(c) $p = 3.29$

(d) $p = 3.86$

Figure 4.5: Bode plot comparison between parametric VF result using *rational* $P_\ell(p)$ and the full model $\mathcal{H}(s, p)$

Comparing Figure 4.4 and Figure 4.5, we note that using rational $P_\ell(p)$ seems to perform better at the sampled points $\mu_1, \ldots, \mu_8$. For a more detailed comparison, we show the approximation error over the entire parameter domain $\mathcal{P} = [1, 5]$ in Figure 4.6.

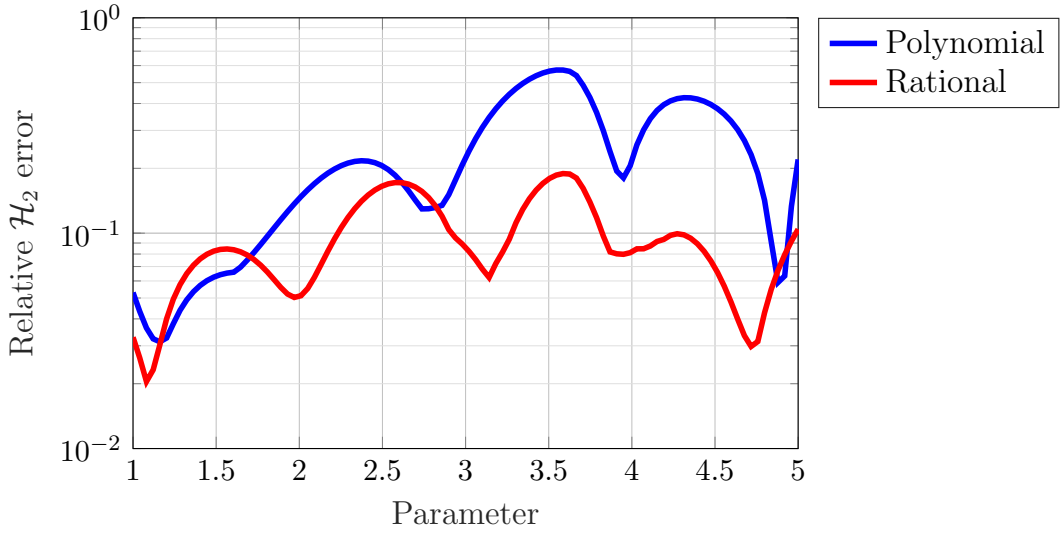


Figure 4.6: Error Comparison between polynomial and rational basis functions $P_\ell(p)$, computed by Algorithm 4.2.3

In Figure 4.6, we observe that, except for very few points, rational basis functions $P_\ell(p)$ yield superior approximation of $\widehat{\mathcal{H}}(s,p)$. The adaptive choice of poles for the rational functions $P_\ell(p)$ is show in Figure 4.7, where we compare an initial selection of poles and the converged poles from Algorithm 4.2.3.

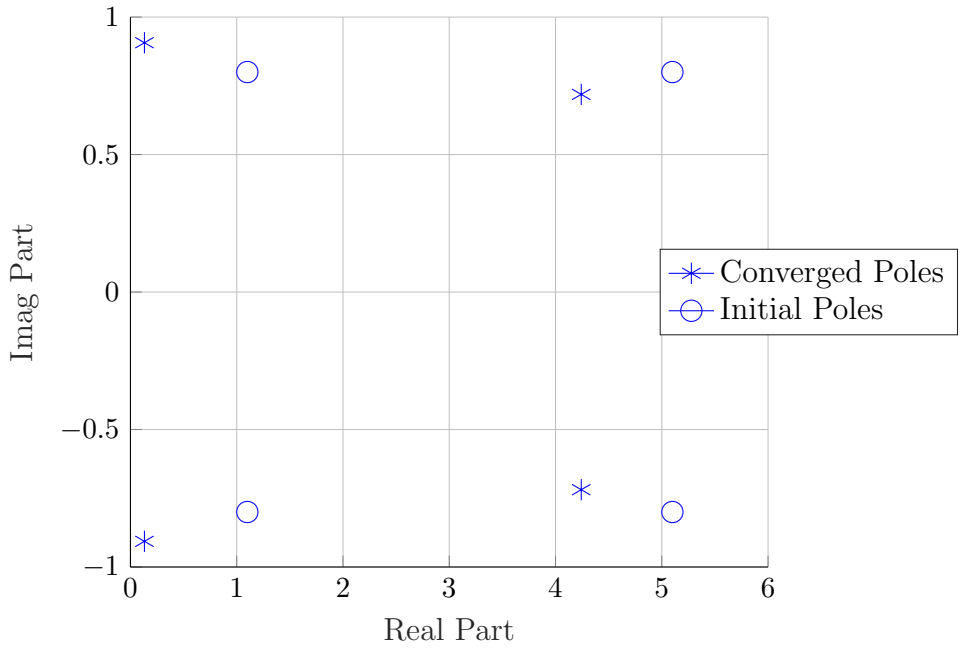
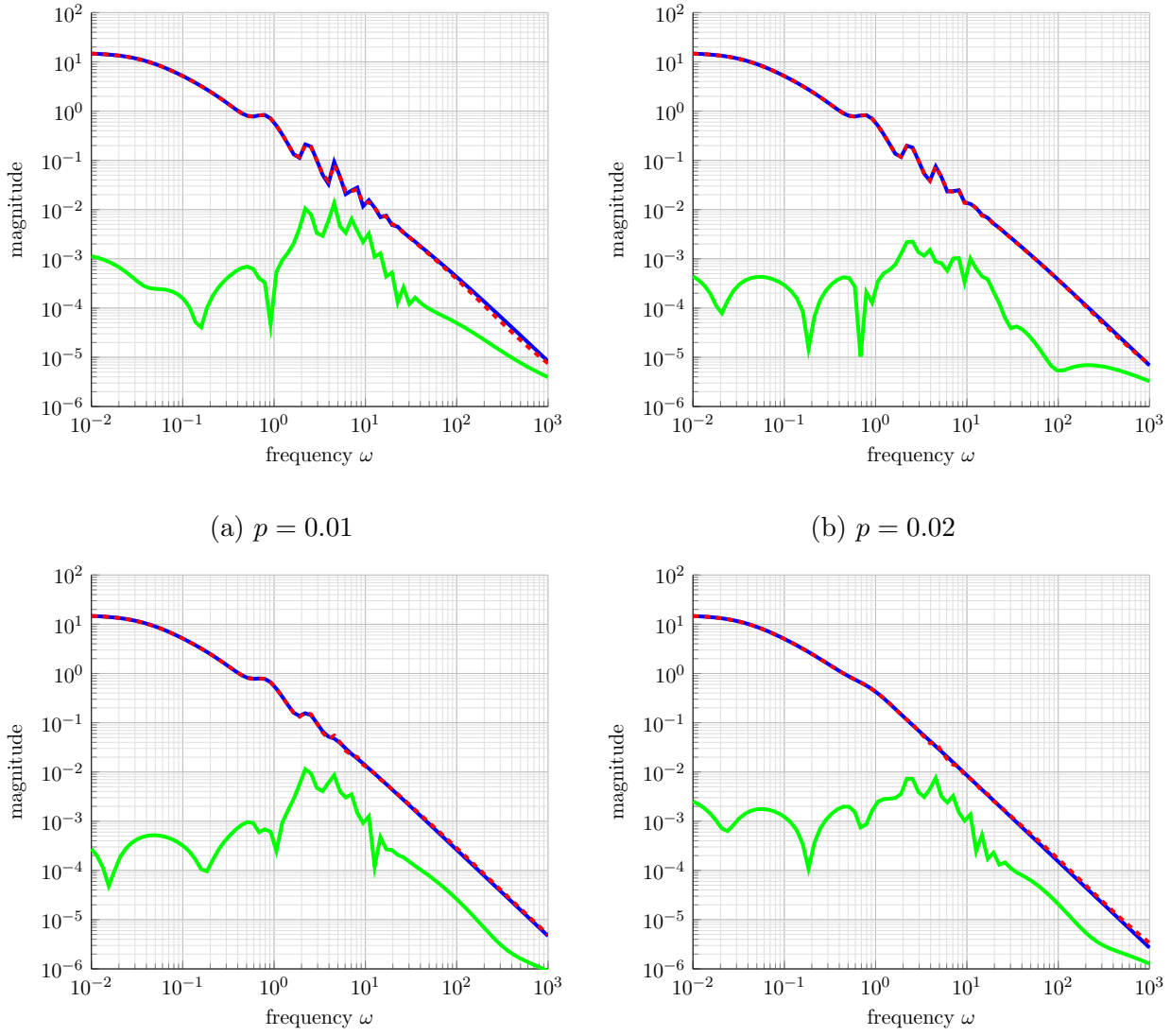


Figure 4.7: Initial and converged pole locations for rational $P_\ell(p)$ from Algorithm 4.2.3

The converged pole locations for the $r_p = 4$ rational poles of $P_\ell(p)$ are shown in Figure 4.7 above. We note that the algorithm converged in 38 iterations up to a pole movement tolerance of 10^{-5}.

Next, we illustrate Algorithm 4.2.3 on a more realistic example closer to engineering applications.

Example 4.2.3 Consider the vibrating Cantilever beam model from Section 1.6.1. The parameter p specified the amount of proportional damping. We pick 80 logarithmically spaced frequency samples in $[10^{-2}\mathrm{i}, 10^{3}\mathrm{i}]$, 10 linearly spaced parameter samples in $[0.01, 0.8]$. In contrast to Example 4.2.1, here we consider rational basis functions $P_\ell(p)$ of order $r_p = 10$.



(a) $p = 0.01$

(b) $p = 0.02$

(c) $p = 0.1$

(d) $p = 0.9$

Figure 4.8: Parametric Vector Fitting results using rational basis functions. Compared at the sampled points. Original model is displayed in blue, approximation dashed red and the (absolute) error in green.

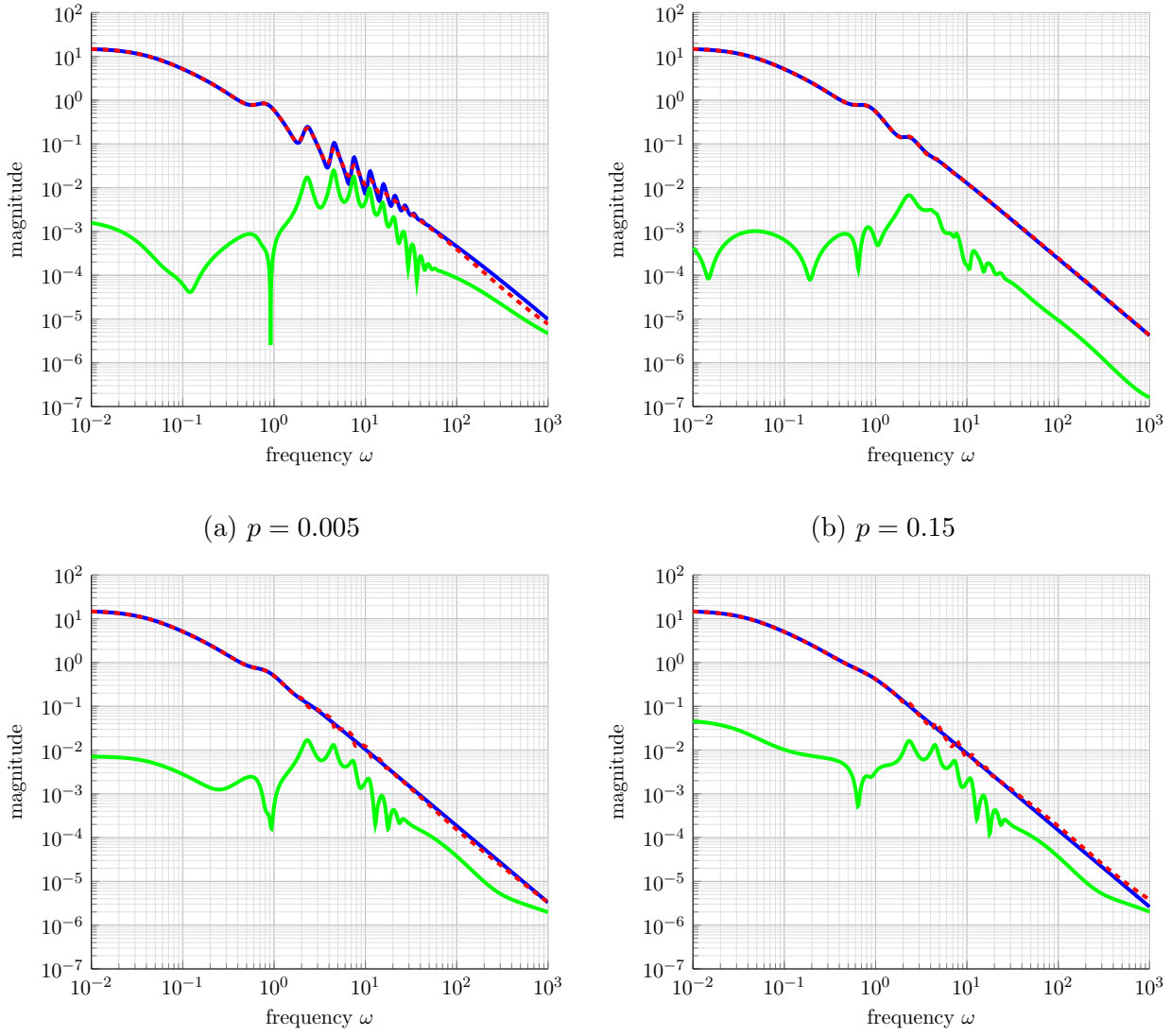
In Figure 4.8 we compare the original function $\mathcal{H}(s, p)$ to the approximation result from

Algorithm 4.2.3 using rational basis function $P_\ell(p)$. The corresponding numerical values are shown in Table 4.3.

	Rel. Error	
Parameter	PVF $\|\cdot\|_2$	PVF $\|\cdot\|_\infty$
0.01	2.97e-03	4.44e-03
0.0978	2.19e-03	2.14e-03
0.186	1.57e-03	1.46e-03
0.273	1.67e-03	1.58e-03
0.361	1.58e-03	1.29e-03
0.449	1.27e-03	1.08e-03
0.537	1.12e-03	9.51e-04
0.624	1.05e-03	9.86e-04
0.712	1.83e-03	1.47e-03
0.8	1.78e-03	1.89e-03

Table 4.3: Local relative error in rational parametric VF approximation. All errors are relative.

Observe in Table 4.3 that the rational basis yields a good approximation with $\approx 10^{-3}$ order of magnitude relative error.



(a) $p = 0.005$

(b) $p = 0.15$

(c) $p = 0.4$

(d) $p = 0.9$

Figure 4.9: Approximation quality at non-sampled points.

In Figure 4.9, we compare the approximation quality of $\widehat{\mathcal{H}}(s, p)$ at non-sampled points in the parameter domain.

Example 4.2.4 Consider the vibrating Cantilever beam example with the setup as in Ex-

ample 4.2.3. To illustrate the outlier resistance of least-squares approximation, we simulate a measurement error by choosing local vector fitting orders $r_p = 2$ for two sampling points, leading to poor approximation of the corresponding local models. We compare the result from Algorithm 4.2.1 and Algorithm 4.2.1 with polynomial and rational basis functions, respectively.

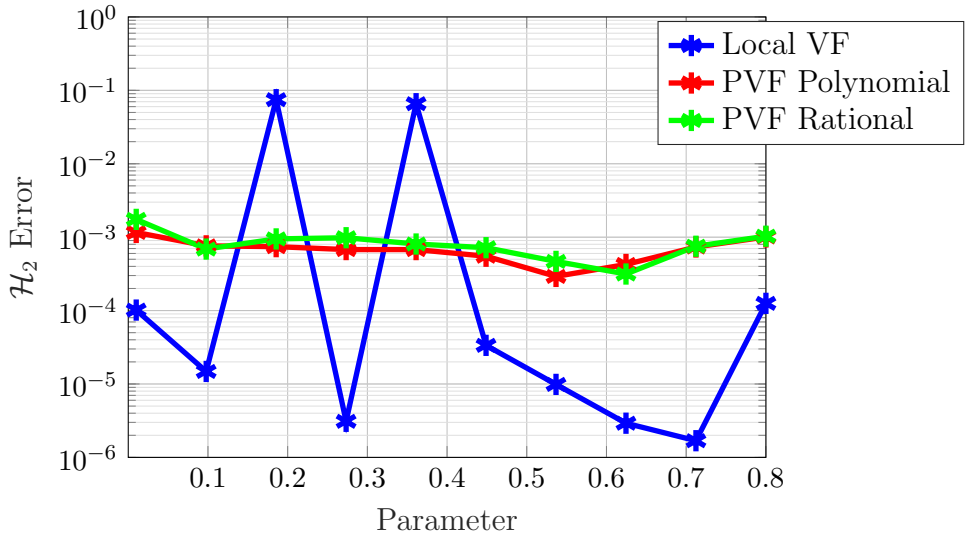


Figure 4.10: Comparison of discrete $\mathcal{H}_2$ approximation error at sampled parameter values.

Observe in Figure 4.11 how the parametric vector fitting approximation for both polynomial and rational functions has an almost uniform error across sampled parameter points. Solving the joint least squares problem compensates for outliers in the measurements at the simulated measurement outliers. We show the converged poles for the rational approximation in Figure 4.11.

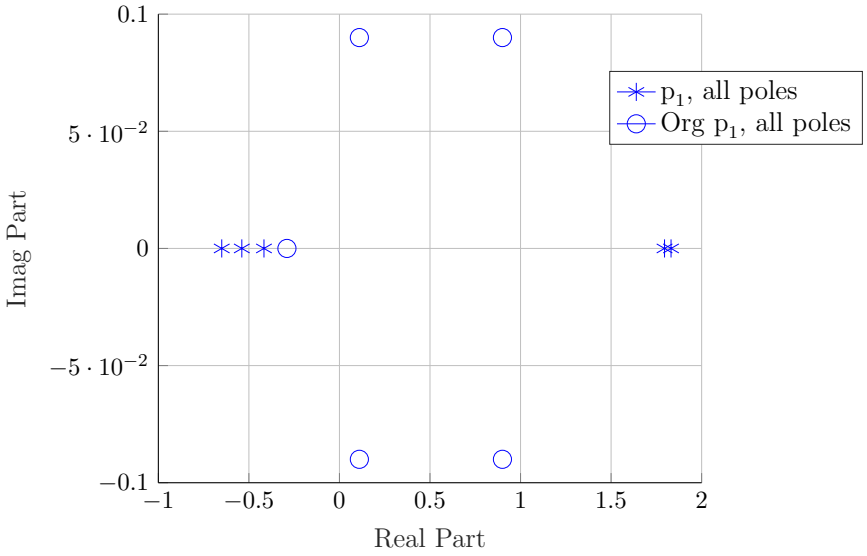


Figure 4.11: Initial (circle) and converged (star) positions of the reduced order poles in the rational basis functions for the parametric dependence.

It is interesting to remark that the poles for the rational basis functions in p converge to real poles, Figure 4.11, where we did not enforce such a constraint.

Our observations from the previous examples can be summarized as follows

1. Choosing rational functions $P_\ell(p)$ matches the approximation result for polynomial $P_\ell(p)$, independent of the structure of the original model, see Example 4.2.4.

2. If the underlying model has, in fact, a rational parametrization, using rational basis functions $P_\ell(p)$ perform better than polynomial $P_\ell(p)$, see Example 4.2.2.

4.3 Vector Fitting for Several Parameters

In this section, we assume that the parameter is no longer a scalar, i.e., $\mathbf{p} = \begin{bmatrix} p_1 & p_2 & \dots & p_d \end{bmatrix}^\top$. For ease of presentation, we restrict ourselves to two parameters ($d = 2$) and denote them $\mathbf{p} = \begin{bmatrix} p & q \end{bmatrix}^\top$. Further, we chose basis functions in p and q independently, following a standard tensor grid approach similar to [45]. Note that Algorithm 4.2.1 remains valid for other choices of basis functions.

Let $P_k(p)$, $k = 1, \dots, r_p$ be the basis functions corresponding to the first parameter, p of degree r_p. Similarly, let $Q_\ell(q)$, $\ell = 1, \dots, r_q$ the basis functions corresponding to the second parameter, q of degree r_q. Further denote the local Vector Fitting approximations at parameter samples (p_k, q_ℓ) by $\widehat{\mathcal{H}}_{k,\ell}(s)$. Then the function $\widehat{\mathcal{H}}(s, p, q)$ is defined as

$$\begin{aligned} \widehat{\mathcal{H}}(s,p,q) :&= \sum_{k_p=1}^{m_p} \sum_{k_q=1}^{m_q} \alpha_{k_p,k_q}(p,q) \widehat{\mathcal{H}}_{k_p,k_q}(s) \\ &= \sum_{k_p=1}^{m_p} \sum_{k_q=1}^{m_q} \sum_{\ell_p=1}^{r_p} \sum_{\ell_q=1}^{r_q} \beta_{k_p,\ell_p,k_q,\ell_q} P_{\ell_p}(p) Q_{\ell_q}(q) \widehat{\mathcal{H}}_{k_p,k_q}(s) \end{aligned} \tag{4.3.1}$$

As for the one-parameter case, we are free to chose polynomial or rational basis functions for $P_{\ell_p}(p)$ and $Q_{\ell_q}(q)$. In terms of implementing Algorithm 4.2.1 and Algorithm 4.2.3 for higher dimensional parameters, we remark that a proper ordering and storing of the variables is important.

Example 4.3.1 We consider the convection-diffusion example, introduced in Section 1.6.2. A finite difference discretization with orders $n_x = 100$ and $n_y = 100$ in x and y direction yields a full order model of dimension $n = 10.000$. The parameters p and q here represent

the convection in x and y direction on the domain $\Omega = [0, 1] \times [0, 1]$ (in Section 1.6.2 denoted by $p = p_1$ and $q = p_2$). For the polynomial approximation order, we chose $r_p = r_q = 12$.

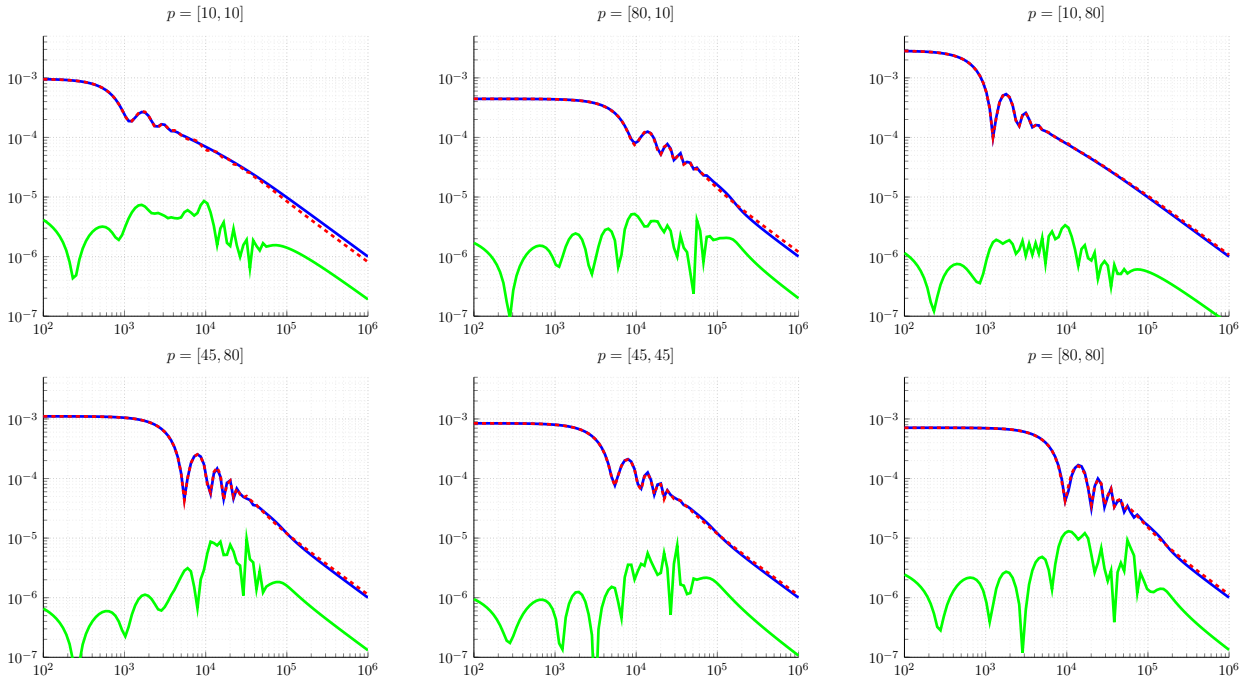


Figure 4.12: Approximation result for two-parameter Vector Fitting using polynomial basis functions. Original model in blue, approximation result dashed red, the (absolute) error in green.

The approximation quality of the two-variable parametric Vector Fitting approximation is shown in Figure 4.12. Our algorithm selected Chebyshev polynomials through the lowest least-squares residual.

4.4 Post-Processing for Parametric Vector Fitting

The previous section shows, how to construct a parametric model to approximate a given data set in a least squares sense. In the process, we chose local Vector Fitting models of certain orders r_p. One standard approach is to use parametric Loewner matrices (c.f. [86]) to gauge the required rank of each local model. However, the resulting model carries the combined orders of the local approximations. For a larger number of parameter samples this can be cumbersome and yield unnecessarily large reduced models. Assume $m_p = 20$ parameter samples and order $r_s = 50$ for the local models. The resulting total degree would be $d = m_p \cdot r_s = 1000$, even though the system dynamics may be similar on parts of the parameter domain $\mathcal{P}$.

This section aims to resolve this issue by combine parametric Vector Fitting with $\mathcal{H}_2$ optimal rational interpolation to approximate parametric models with a lower complexity. Our proposed approach has three steps:

1. Apply parametric Vector Fitting Algorithm 4.2.1 or Algorithm 4.2.3 to construct a surrogate model $\widehat{\mathcal{H}}(s,p) = \sum_{k=1}^{m_p} \alpha_k(p)\widehat{\mathcal{H}}_k(s)$.

2. Transform parametric SISO system into MIMO system with equivalent norm.

3. Use MIMO-IRKA, Algorithm 2.4.2 to find an optimal reduced model of fixed reduced degree $\underline{r}$.

To implement the last two steps of our post-processing, we review $\mathcal{H}_2$ optimal rational

interpolation for a special parametric dependance in the next subsection.

4.4.1 Review of IRKA for a Special Parametric Dependancy

Baur et all in [11] consider a system of the form

$$\mathcal{H}(s,\mathbf{p}) = (\mathbf{c}_0 + p_1\,\mathbf{c}_1)^T\,(s\,\mathbf{E}\,-\,\mathbf{A})^{-1}\,(\mathbf{b}_0 + p_2\,\mathbf{b}_1), \tag{4.4.1}$$

with parameter $\mathbf{p} = [p_1, p_2]^\top \in \mathcal{P} := [0,1]\times[0,1]$. Note that only the input and output matrices depend on the parameter and that the parameter dependency is assumed to be affine.

The error measure to approximate (4.4.1), $\|\cdot\|_{\mathcal{H}_2\otimes L_2(\mathcal{P})}$, is chosen in [11] as

$$\|\mathcal{H}\|_{\mathcal{H}_2\otimes L_2(\mathcal{P})} := \sqrt{\frac{1}{2\pi}\int_{-\infty}^{+\infty}\iint\limits_{\mathcal{P}} |\mathcal{H}(\imath\omega,\,\mathbf{p})|^2\;\mathrm{d}A(\mathbf{p})\;\mathrm{d}\omega}. \tag{4.4.2}$$

A key observation is that the $\mathcal{H}_2\otimes L_2(\mathcal{P})$ norm equals a weigthed $\mathcal{H}_2$ norm of a multi-input multi-output (MIMO) system. More precisely, let $\mathbf{L}$ be Cholesky factor of

$$\int_0^1 \begin{bmatrix}1\\ p\end{bmatrix}\begin{bmatrix}1 & p\end{bmatrix}\,\mathrm{d}p = \int_0^1 \begin{bmatrix}1\\ q\end{bmatrix}\begin{bmatrix}1 & q\end{bmatrix}\,\mathrm{d}q = \mathbf{L}\mathbf{L}^\top. \tag{4.4.3}$$

Then we have

$$\|\mathcal{H}\|_{\mathcal{H}_2\otimes L_2(\mathcal{P})} = \left\|\mathbf{L}^\top H\mathbf{L}\right\|_{\mathcal{H}_2}, \tag{4.4.4}$$

where $H(s) = \begin{bmatrix}\mathbf{c}_0 & \mathbf{c}_1\end{bmatrix}^\top (s\mathbf{E}-\mathbf{A})^{-1}\begin{bmatrix}\mathbf{b}_0 & \mathbf{b}_1\end{bmatrix}$.

4.4.2 SISO Parametric to MIMO Nonparametric

We need to investigate the form of the transformation in (4.4.4) in more detail for our application in parametric Vector Fitting. At parameter samples μ_j, $j = 1, \ldots, m_p$, the (local) Vector Fitting models $\widehat{\mathcal{H}}_k(s)$ from Algorithm 4.2.1 can be written as

$$\widehat{\mathcal{H}}_k(s) = \widehat{\mathbf{c}}_k^\top \left(s\mathbf{I} - \widehat{\mathbf{A}}_k\right)^{-1} \widehat{\mathbf{b}}_k, \qquad k = 1, \ldots, m_p, \tag{4.4.5}$$

with the matrices $\widehat{\mathbf{A}}_k \in \mathbb{C}^{r_p \times r_p}$ and $\widehat{\mathbf{b}}_k, \widehat{\mathbf{c}}_k \in \mathbb{C}^{r_p}$. The resulting parametric model then is

$$\widehat{\mathcal{H}}_{VF}(s,p) = \sum_{k=1}^{m_p} \alpha_k(p)\mathcal{H}_k(s) = \sum_{k=1}^{m_p} \alpha_k(p)\widehat{\mathbf{c}}_k^\top \left(s\mathbf{I} - \widehat{\mathbf{A}}_k\right)^{-1} \widehat{\mathbf{b}}_k. \tag{4.4.6}$$

In order to make use of (4.4.2), we need to express the summation in (4.4.6) in a direct state space form. Define

$$\mathfrak{A} := \begin{bmatrix} \widehat{\mathbf{A}}_1 & & \\ & \ddots & \\ & & \widehat{\mathbf{A}}_{m_p} \end{bmatrix} \in \mathbb{C}^{r_p m_p \times r_p m_p}, \qquad \text{and} \qquad \mathfrak{B} := \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \vdots \\ \widehat{\mathbf{B}}_{m_p} \end{bmatrix} \in \mathbb{C}^{r_p m_p}. \tag{4.4.7}$$

Using (4.4.7), we rewrite (4.4.6) as

$$\widehat{\mathcal{H}}_{VF}(s,p) = \begin{bmatrix} \alpha_1(p)\widehat{\mathbf{c}}_1^\top & \alpha_2(p)\widehat{\mathbf{c}}_2^\top & \ldots & \alpha_n(p)\widehat{\mathbf{c}}_{m_p}^\top \end{bmatrix} (s\mathbf{I} - \mathfrak{A})^{-1}\mathfrak{B} \tag{4.4.8}$$

Recall that the parametric coefficient functions $\alpha_k(p)$ have the form $\alpha_k(p) = \sum_{\ell=1}^{r_p} \beta_{k,\ell} P_\ell(p)$ with basis functions $\{P_\ell(p)\}$, yielding

$$\begin{bmatrix} \alpha_1(p)\widehat{\mathbf{c}}_1^\top & \ldots & \alpha_{m_p}(p)\widehat{\mathbf{c}}_{m_p}^\top \end{bmatrix} = \begin{bmatrix} P_1(p) & \ldots & P_{r_p}(p) \end{bmatrix} \begin{bmatrix} \beta_{1,1} & & \\ & \ddots & \\ & & \beta m_p, r_p \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{c}}_1^\top & & \\ & \ddots & \\ & & \widehat{\mathbf{c}}_{m_p}^\top \end{bmatrix}. \tag{4.4.9}$$

Using (4.4.9), we define the MIMO system matrix $\mathfrak{C}$ by

$$\mathfrak{C} := \begin{bmatrix} \beta_{1,1} & & \\ & \ddots & \\ & & \beta m_p, r_p \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{c}}_1^\top & & \\ & \ddots & \\ & & \widehat{\mathbf{c}}_{m_p}^\top \end{bmatrix} \tag{4.4.10}$$

Let $\mathbb{V} : \mathcal{P} \to \mathbb{C}^{r_p}$ be defined by

$$\mathbb{V}(p) := \begin{bmatrix} P_1(p) & \dots & P_{r_p}(p) \end{bmatrix}^\top, \tag{4.4.11}$$

then the reduced model from (4.4.6) can be written as

$$\widehat{\mathcal{H}}_{VF}(s,p) = \mathbb{V}(p)^\top \mathfrak{C}(s\mathbf{I} - \mathfrak{A})^{-1}\mathfrak{B}. \tag{4.4.12}$$

We emphasize that the parameter only enters in the observation matrix, here $\mathbb{V}(p)^\top\mathfrak{C}$, not in the system dynamics $\mathfrak{A}$.

Similar to [11], the $\mathcal{H}_2 \otimes L_2(\mathcal{P})$ norm can be recast as a weighted $\mathcal{H}_2$ norm of a MIMO system.

Theorem 4.4.1 *Let $\widehat{\mathcal{H}}_{VF}(s,p)$ be as in (4.4.12), $\mathcal{P} = [a,b] \subset \mathbb{R}$ bounded and define*

$$\widehat{\mathcal{H}}_L(s) := \mathbf{L}^\top \mathfrak{C}(s\mathbf{I} - \mathfrak{A})^{-1}\mathfrak{B}, \quad \textit{with} \qquad \mathbf{L}\mathbf{L}^\top = \int_a^b \mathbb{V}(p)\mathbb{V}(p)^\top \,\mathrm{d}p, \tag{4.4.13}$$

where $\mathbf{L}\mathbf{L}^\top$ is a Cholesky factorization and $\mathbb{V}(p)$ as in (4.4.11). Then

$$\|\widehat{\mathcal{H}}_L\|_{\mathcal{H}_2} = \|\widehat{\mathcal{H}}_{VF}\|_{\mathcal{H}_2 \otimes L_2(\mathcal{P})} \tag{4.4.14}$$

where $\|\cdot\|_{\mathcal{H}_2 \otimes L_2(\mathcal{P})}$ is defined as in (4.4.2).

Proof Recall the definition of the norm $\|\cdot\|_{\mathcal{H}_2\otimes L_2(\mathcal{P})}$ for our particular parameter domain $\mathcal{P} = [a, b]$:

$$\|\mathcal{H}\|^2_{\mathcal{H}_2\otimes L_2([a,b])} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_a^b |\mathcal{H}(i\omega, p)|^2 \,\mathrm{d}p\,\mathrm{d}\omega \tag{4.4.15}$$

Since we consider a one-dimensional parameter domain, the area integral $dA(\mathbf{p})$ from (4.4.2) becomes a scalar integral. We start by computing the following integral

$$\int_a^b \mathbb{V}(p)\mathbb{V}(p)^\top \,\mathrm{d}p = \int_a^b \begin{bmatrix} P_1(p) \\ \vdots \\ P_{r_p}(p) \end{bmatrix} \begin{bmatrix} P_1(p) & \dots & P_{r_p}(p) \end{bmatrix} \,\mathrm{d}p =: \mathbf{P} \tag{4.4.16}$$

with $P_\ell(p)$ from (4.4.6). Let $\mathbf{P} = \mathbf{L}\mathbf{L}^\top$ be the Cholesky factorization of $\mathbf{P}$. Also recall that it holds

$$\operatorname{trace}(\mathbf{A}\mathbf{B}) = \operatorname{trace}(\mathbf{B}\mathbf{A}), \tag{4.4.17}$$

for matrices $\mathbf{A}$, $\mathbf{B}$ of suitable dimensions. Then for $\widehat{\mathcal{H}}_{VF}(s, p)$ from (4.4.12), the norm (4.4.15)

becomes

$$\begin{aligned}
\left\|\widehat{\mathcal{H}}_{VF}\right\|^2_{\mathcal{H}_2\otimes L_2([a,b])} &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_a^b \left\|\widehat{\mathcal{H}}_{VF}(i\omega,p)\right\|_F^2 \,\mathrm{d}p\,\mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_a^b \operatorname{trace}\left(\mathfrak{B}^\top(i\omega\mathbf{I}-\mathfrak{A})^{-\top}\mathfrak{C}^\top\mathbb{V}(p)\mathbb{V}(p)^\top\mathfrak{C}(i\omega\mathbf{I}-\mathfrak{A})^{-1}\mathfrak{B}\right)\,\mathrm{d}p\,\mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\operatorname{trace}\left(\underbrace{\int_a^b \mathbb{V}(p)\mathbb{V}(p)^\top\,\mathrm{d}p}_{=\mathbf{L}^\top\mathbf{L}\ \text{ from (4.4.16)}}\mathfrak{C}(i\omega\mathbf{I}-\mathfrak{A})^{-1}\mathfrak{B}\mathfrak{B}^\top(i\omega\mathbf{I}-\mathfrak{A})^{-\top}\mathfrak{C}^\top\right)\,\mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\operatorname{trace}\left(\mathfrak{B}^\top(i\omega\mathbf{I}-\mathfrak{A})^{-\top}\mathfrak{C}^\top\mathbf{L}^\top\mathbf{L}\mathfrak{C}(i\omega\mathbf{I}-\mathfrak{A})^{-1}\mathfrak{B}\right)\,\mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\operatorname{trace}\left(\widehat{\mathcal{H}}_L^\top(i\omega)\widehat{\mathcal{H}}_L(i\omega)\right)\,\mathrm{d}\omega \\
&= \|\widehat{\mathcal{H}}_L\|^2_{\mathcal{H}_2},
\end{aligned} \tag{4.4.18}$$

with $\widehat{\mathcal{H}}_L(s)$ defined in (4.4.13). ■

Now we are in the position to apply $\mathcal{H}_2$ optimal model reduction (IRKA) to reduce $\widehat{\mathcal{H}}_L(s)$ from (4.4.13) to a chosen target size $\underline{r}$, resulting in a reduced model

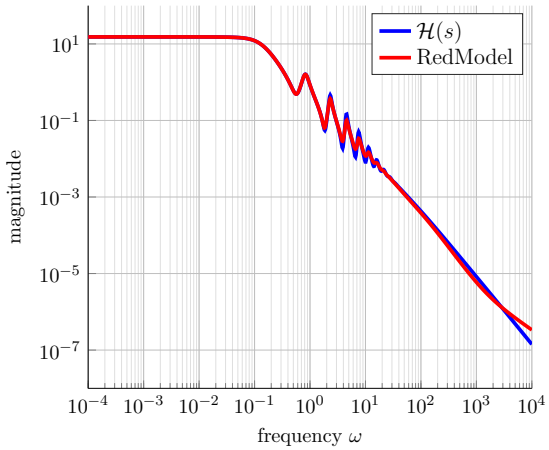
$$\widehat{\mathcal{H}}_F(s,p) := \mathbb{V}(p)^\top\widehat{\mathfrak{C}}(s\mathbf{I}-\widehat{\mathfrak{A}})^{-1}\widehat{\mathfrak{B}} = [P_1(p),\ldots,P_{r_p}(p)]\widehat{\mathfrak{C}}(s\mathbf{I}-\widehat{\mathfrak{A}})^{-1}\widehat{\mathfrak{B}} \tag{4.4.19}$$

Observe that the parametric dependence is unchanged with the transformation $\widehat{\mathcal{H}}_{VF}(s,p)$ to $\widehat{\mathcal{H}}_F(s,p)$.

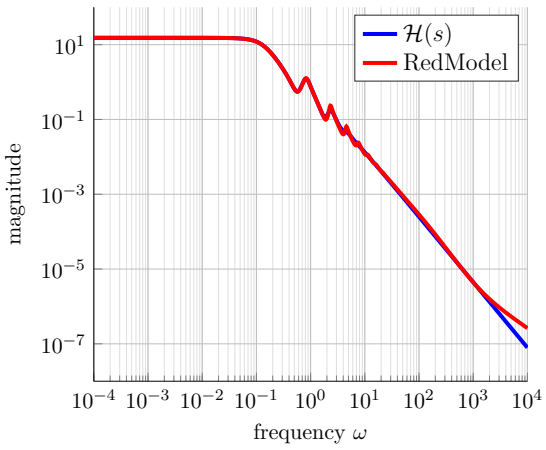
Function evaluations of the final reduced model then require evaluating the local models $\widehat{\mathcal{H}}_k(s)$ at given frequency samples and the evaluation of the basis functions $\{f_\ell(p)\}$ at the parameter value. We choose the order $\underline{r}$ of $\widehat{\mathcal{H}}_F(s,p)$ so that $\underline{r} < m_p r_p$, the sum of the local Vector Fitting orders.

4.4.3 Numerical Example

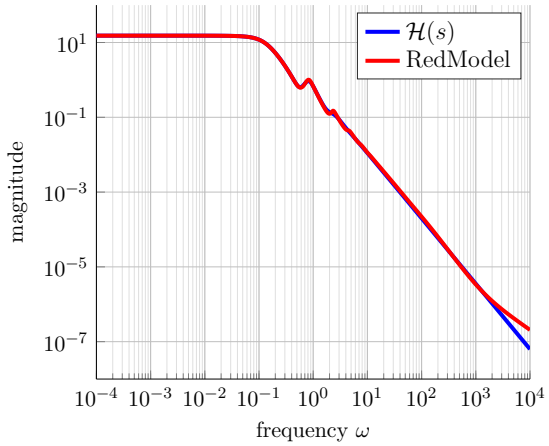
Again, we consider the vibrating Cantilever beam example from Section 1.6.1. The beam is discretized with a finite element model resulting in $n = 4600$ degrees of freedom. We chose 200 logarithmically spaced sampling points on the imaginary axis between 10^{-2} and 10^4 and $m_p = 30$ parameter samples, logarithmically spaced in $[0.0001, 1]$. The order of the local VF models is $r_s = 24$ and the polynomial order for $\alpha_k(p)$ chosen as $r_p = 8$. For the IRKA approximation, we chose $\underline{r} = 20$. In Figure 4.13, we show an error comparison for representative parameters in $[0, 1]$.



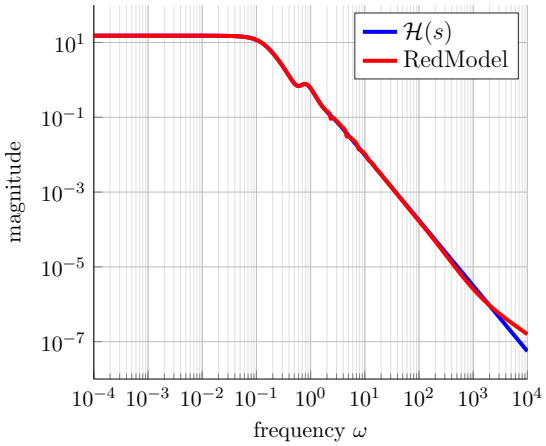
(a) $p = 0.01$



(b) $p = 0.12$



(c) $p = 0.3$



(d) $p = 0.5$

Figure 4.13: Comparison between original model $\mathcal{H}(s,p)$ (blue) and paramVF-IRKA model $\widehat{\mathcal{H}}(s,p)$ (red) at representative parameter values.

Observe that the approximation $\widehat{\mathcal{H}}(s,p)$, generated from parametric Vector Fitting & IRKA achieves a comparable approximation as parametric Vector Fitting at a lower order. In Figure 4.13, the dominant peaks of $\mathcal{H}(s,p)$ are matched well by $\widehat{\mathcal{H}}(s,p)$.

We expect this additional dimensionality reduction, since the choice of reduced model order r_p in (4.2.13) is made uniformally for each parameter without considering approximation properties over the whole parameter domain.

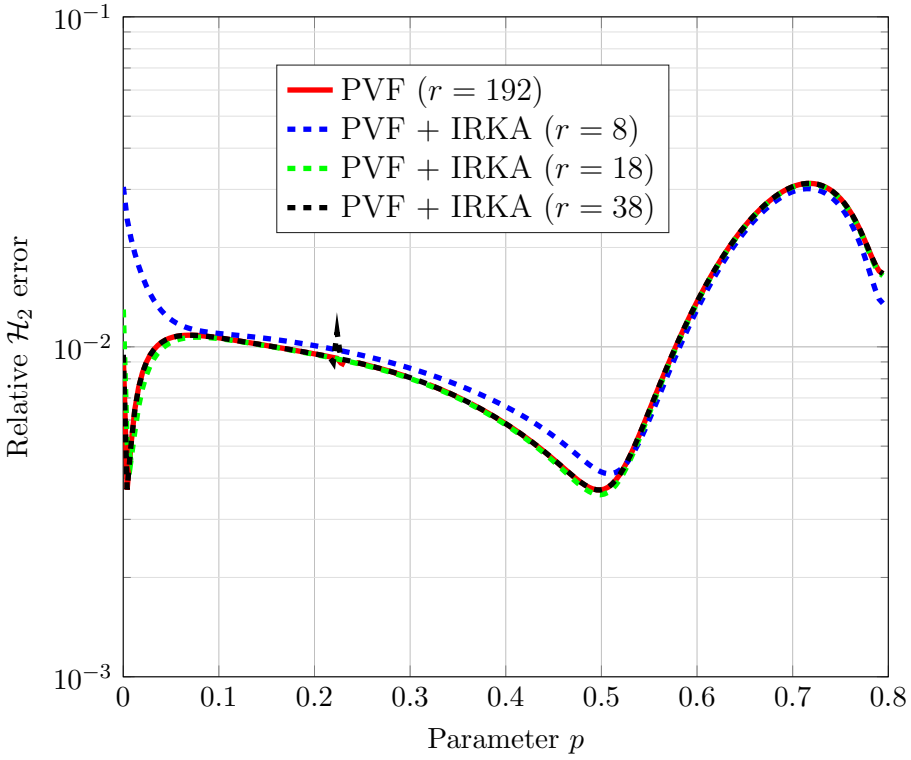


Figure 4.14: Local $\mathcal{H}_2$ approximation quality over sampled parameter range $[0, 0.8]$.

In Figure 4.14, we compare the local $\mathcal{H}_2$ error, meaning for fixed parameter values, continous in the frequency, between the parametric Vector Fitting model using $r_p = 12$ basis functions $P_\ell(p)$ and several post-processed PVF+IRKA models. We vary the final model order $\underline{r}$ and observe that the relative local $\mathcal{H}_2$ error between $\underline{r} = 18$ and $\underline{r} = 38$ is almost indistinguishable. The local relative error of parametric Vector Fitting is displayed in red. Since IRKA applies an additional reduction step, we do not expect to improve the approximation quality but

reduced the overall model order.

4.5 Summary of Contributions and Future Work

In this section, we presented an extension of Vector Fitting to the parametric case. For the parametric dependence both polynomials and rational basis functions are investigated. Adaptively chosen basis functions in p require nonlinear optimization methods. In the case of rational functions, we reallocate the poles from an initial guess to optimal positions. While the least squares approach has shown good performance and outlier resistance in several examples, some questions remain open. Possible future directions include an implementation of total-least squares-least squares Vector Fitting, as suggested in [40]. The extension of such an approach to the parametric case may greatly improve pole allocation properties in the adaptive choice of basis functions.

We observe that parametric vector fitting can lead to a moderate size surrogate model depending on the number of measurements and chosen orders for the local models. To further reduce the dimension of the final model, we combined the result of parametric Vector Fitting with $\mathcal{H}_2$ optimal rational interpolation via a state space transformation.

Chapter 5

Delay Differential Equations and Transient Analysis

In this chapter we consider a particular class of parametric dynamical systems, where the parameter represents a time delay in the state $x(t) \in \mathbb{C}$. More precisely, we consider variations of the scalar delay differential equation (DDE)

$$\begin{aligned} \dot{x}(t) &= ax(t) + bx(t-\tau), \quad \text{with} \quad a, b \in \mathbb{C}, \quad t \geq 0, \\ x(t) &= \phi(t+\tau), \qquad t \in [-\tau, 0], \end{aligned} \tag{5.0.1}$$

with initial history function $\phi(t) \in \mathcal{P}C\left([0,\tau], \mathbb{C}\right)$, the space of piecewise continuous functions on $[0,\tau]$. Note that now the information required to solve the equation is a history function instead of an initial state $x(0) \in \mathbb{C}$. It is well known [16] that the solution space to $\dot{x}(t) = ax(t) + bx(t-\tau)$ is infinite dimensional for any $\tau > 0$, which is in contrast to the case $\tau = 0$ corresponding to a regular ODE, where the solution space is finite dimensional (more

precisely, one dimensional). This distinction also surfaces in the spectral analysis of delay differential equations, where the generator of the solution semigroup has infinitely many eigenvalues, even though the solution $x(t) \in \mathbb{C}$ $(t \geq -\tau)$ evolves in a finite dimensional space.

5.1 Goals

ODEs in higher dimensions can exhibit transient growth [42, 114], but not so in the scalar case. Our investigations show that for delay differential equations, even the scalar case (5.0.1) can exhibit arbitrary transient growth. Even more, we present methods to

1. Identify examples for maximum transient growth;
2. Construct explicit parameter configurations that yield significant transient growth;
3. Examine transient behavior for several commensurate delays.

In short, the goal of our investigations is to answer the following question: Can an asymptotically stable solution to a delay equation grow larger than its initial condition? A more formal way to state this question is:

Is it possible for the solution of a scalar, asymptotically stable delay equation with $\|\phi\|_{L_\infty} \leq 1$ to have $\|x(t)\| \gg 1$ for some $t > 0$?

5.2 Motivation, Applications and Transients

Delay differential equations arise in multiple applications, including biological models and control problems in engineering. Interesting areas where delay differential equations play an essential role include biology [109], physics [6], chemistry [104], traffic flow analysis [98], population dynamics [82] and chaotic dynamics [111], to mention just a few areas. An extensive overview of applications for DDEs can also be found, for example, in [73].

While many of the references above are concerned with solution methods, stability analysis or control, our focus lies on transient behavior. More precisely, we are interested in asymptotically stable solutions, i.e., systems that satisfy $x(t) \to 0$ $(t \to \infty)$, but for which $\max_{t\geq 0} |x(t)|$ is large. Before analyzing transient behavior in delay equations, we present the standard methods for analyzing transient behavior in ODEs.

The solution to the ODE $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$ is given by

$$\mathbf{x}(t) = \mathrm{e}^{\mathbf{A}\mathrm{t}}\mathbf{x}(0), \qquad \mathrm{t} \geq 0. \tag{5.2.1}$$

In the scalar case, that means $x(t) = \mathrm{e}^{at}x(0)$. For stable systems, we require $\Re(a) < 0$ and the solution exhibits monotonic decay to 0. No transient behavior can occur for such systems in the scalar case.

For larger dimensions, transient behavior can be efficiently studied using the concept of pseudospectra [42, 112, 114].

Definition 5.2.1 Let $\mathbf{A} \in \mathbb{C}^{n\times n}$. For any $\epsilon > 0$, the *ϵ-pseudospectrum* of $\mathbf{A}$ is defined by

$$\sigma_\epsilon(\mathbf{A}) := \left\{ z \in \mathbb{C} \mid \| (z\mathbf{I} - \mathbf{A})^{-1} \| > \frac{1}{\epsilon} \right\}; \tag{5.2.2}$$

where we define $\| (z\mathbf{I} - \mathbf{A})^{-1} \| := \infty$ if $z \in \sigma(\mathbf{A})$. With the definition of the resolvent $R(\mathbf{A}, z) := (z\mathbf{I} - \mathbf{A})^{-1}$ for $z \notin \sigma(\mathbf{A})$, (5.2.2) can be rephrased as

$$z \in \sigma_\epsilon(\mathbf{A}) \qquad \Leftrightarrow \qquad \|R(\mathbf{A}, z)\| > \frac{1}{\epsilon}. \tag{5.2.3}$$

Another equivalent characterization of $\sigma_\epsilon(\mathbf{A})$ comes from perturbation theory:

$$\sigma_\epsilon(\mathbf{A}) = \left\{ z \in \sigma(\mathbf{A} + \mathbf{E}) : \quad \mathbf{E} \in \mathbb{C}^{n\times n},\ \|\mathbf{E}\| < \epsilon \right\}. \tag{5.2.4}$$

Definition 5.2.1 above can be easily extended from matrices to linear operators. Several quantities are of particular interest for our analysis of transient behavior.

Definition 5.2.2 For a matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ the *ϵ-pseudospectral abscissa* is

$$\alpha_\epsilon(\mathbf{A}) := \sup\{\Re(z) :\ z \in \sigma_\epsilon(\mathbf{A})\}. \tag{5.2.5}$$

With this definition, one can show [114] that

$$\sup_{t\geq 0} \|\mathrm{e}^{t\mathbf{A}}\| \geq \frac{\alpha_\epsilon(\mathbf{A})}{\epsilon}, \qquad \epsilon > 0. \tag{5.2.6}$$

Taking the supremum over all such $\epsilon > 0$ in (5.2.6) leads to the Kreiss constant.

Definition 5.2.3 For a matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, the *Kreiss constant* $\mathcal{K}(\mathbf{A})$ is defined by

$$\mathcal{K}(\mathbf{A}) := \sup_{\epsilon>0} \frac{\alpha_\epsilon(\mathbf{A})}{\epsilon}. \tag{5.2.7}$$

For discrete systems $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k$, we have analogous quantities. The *ϵ-pseudospectral radius* is defined by

$$\rho_\epsilon(\mathbf{A}) := \sup\{|z| : \ z \in \sigma_\epsilon(\mathbf{A})\}, \tag{5.2.8}$$

and the *discrete Kreiss constant* is

$$\mathcal{K}_d(\mathbf{A}) = \sup_{\epsilon>0} \frac{\rho_\epsilon(\mathbf{A}) - 1}{\epsilon}. \tag{5.2.9}$$

Note that the pseudospectra and derived quantities depend on the choice of norm. With these basic definitions, we continue with the spectral and pseudospectral analysis of delay equations.

5.2.1 Eigenvalue Analysis for Delay Differential Equations

Stability analysis amounts to understanding the eigenvalues of the governing equation, or, more precisely, the solution operator / semigroup [33, 43]. For the delay equation (5.0.1), we make the ansatz $x(t) = \mathrm{e}^{t\lambda}$ for some $\lambda \in \mathbb{C}$. Substituting this $x(t)$ into (5.0.1) leads to the *characteristic equation*

$$\Delta(\lambda) := (\lambda - a - b\,\mathrm{e}^{-\lambda\tau}) = 0. \tag{5.2.10}$$

This is a simple example of a nonlinear eigenvalue problem [64].

It is well known that solutions to (5.2.10) can be expressed through the Lambert W function [84, 119], which generally solves the equation

$$W(s)\,\mathrm{e}^{W(s)} = s, \qquad s \in \mathbb{C}. \tag{5.2.11}$$

Note that the solution $W(s)$ in (5.2.11) is not unique, but has branches $W_k(s)$, $k \in \mathbb{Z}$, every one of which solves (5.2.11). $W_0(s)$ is called the *principle branch* of the Lambert W function. The solutions of the characteristic equation (5.2.10) can be expressed as

$$\Delta(\lambda) = 0 \qquad \Leftrightarrow \qquad \lambda \in \left\{ \frac{1}{\tau} W_k(\tau b \, \mathrm{e}^{-a\tau}) + a : \ k \in \mathbb{Z} \right\}. \tag{5.2.12}$$

5.2.2 Pseudospectra for the Nonlinear EVP

The authors in [92, 93, 110] perform a thorough spectral analysis for understanding how perturbations of the coefficients affect stability of solutions to delay differential equations using the corresponding nonlinear eigenvalue problem (5.2.14). They introduce a notion of pseudospectra that can be used to gather information on sensitivity of the spectrum to perturbation of the coefficients a_k $(k = 1, \ldots, d)$ of the DDE and the delay value τ. For higher dimensions $n > 1$, pseudospectra are defined in Section 5.5. It suffices for our investigations to consider the scalar case.

Since the additional notation is minimal, we present the pseudospectral analysis for scalar delay equation with several, commensurate delays. Consider the equation

$$\dot{x}(t) = a_0 x(t) + a_1 x(t-\tau) + \cdots + a_d x(t - d\tau), \qquad a_0, \ldots, a_d \in \mathbb{C}. \tag{5.2.13}$$

The nonlinear eigenvalue problem for (5.2.13) is

$$F(\lambda) := \lambda - a_0 - \sum_{k=1}^{d} a_k \, \mathrm{e}^{-\lambda k \tau} = 0, \qquad \lambda \in \mathbb{C}. \tag{5.2.14}$$

The *spectrum* $\sigma(F)$ consists of all values $\lambda \in \mathbb{C}$ for which $F(\lambda) = 0$. For such $F(\lambda)$, Michiels and Niculescu [93] define a notion of pseudospectra using perturbation of the coefficients a_k:

$$\Lambda_{\epsilon,p}(F) = \left\{ \lambda \in \mathbb{C} : \ \lambda - \sum_{k=0}^{d} (a_k + \delta a_k)\, \mathrm{e}^{-k\lambda\tau} = 0, \quad \left\| \begin{bmatrix} \delta a_0 & \dots & \delta a_d \end{bmatrix}^\top \right\|_p < \epsilon \right\}. \quad (5.2.15)$$

We can consider $\Lambda_{\epsilon,p}(F)$ for any $p \in [1, \infty]$, corresponding to a different choice of norm on the perturbation vector $\begin{bmatrix} \delta a_0 & \dots & \delta a_d \end{bmatrix}^\top$. For some applications, it may be helpful to introduce a weight to the perturbations in (5.2.15). In our case, this will not be necessary.

However, for the nonlinear eigenvalue problem (5.2.14), define the following vector-valued *scaling function*:

$$w(\lambda) = \begin{bmatrix} 1 & \mathrm{e}^{-\lambda\tau} & \dots & \mathrm{e}^{-\lambda d\tau} \end{bmatrix}^\top \in \mathbb{C}^{d+1} \quad (5.2.16)$$

For easier computation, we define the scalar-valued function

$$f(\lambda; p) := \begin{cases} \left| \left(\lambda - a_0 - \sum_{k=1}^{d} a_k\, \mathrm{e}^{-\lambda k\tau} \right)^{-1} \right| \|w(\lambda)\|_p, & \lambda \notin \sigma(F); \\ +\infty, & \lambda \in \sigma(F). \end{cases} \quad (5.2.17)$$

Here $p \in [1, \infty]$ refers to the norm used for the scaling function $w(\lambda)$ from (5.2.16). With $f(\lambda; p)$ from (5.2.17), the ϵ-pseudospectrum is characterized [93, Thm. 3.2] as

$$\Lambda_{\epsilon,p}(F) = \left\{ \lambda \in \mathbb{C} : \quad f(\lambda; p) > \epsilon^{-1} \right\}, \qquad p \in [1, \infty]. \quad (5.2.18)$$

We note how the two equivalent characterizations in (5.2.15) and (5.2.18) parallel the characterizations of classic pseudospectra in Definition 5.2.1. Indeed, $\Lambda_{\epsilon,p}(F)$, generalizes linear pseudospectra from Definition 5.2.1. To see this, let $\tau = 0$, $d = 1$. Then $F(\lambda) = \lambda - a$ and

$w(\lambda) = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$, with $\|w\|_p = 1$ for any $p \in [1, \infty]$. Since $f(\lambda; p) = |(\lambda - a)^{-1}|$, we recover in this case $\Lambda_{\epsilon,p}(F) = \sigma_\epsilon(a)$.

The *distance to instability* is a quantity often of interest to engineers. In essence, it answers the question: How much can parameters vary before the system becomes unstable? Using $F(\lambda)$ from (5.2.14), the distance to instability can be expressed with a notion of *stability radius* r on a set $\Omega \subset \mathbb{C}$:

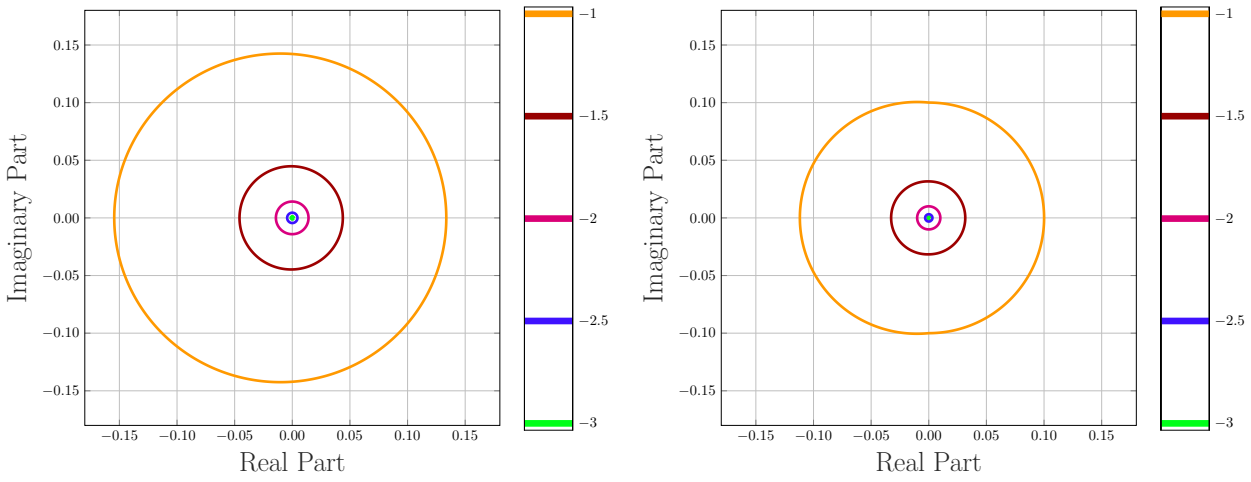
$$r(\Omega; p) := \inf_{\lambda \in \partial\Omega} \frac{1}{f(\lambda; p)} = \frac{1}{\sup_{\lambda \in \partial\Omega} f(\lambda; p)}. \tag{5.2.19}$$

While the notion of pseudospectra in (5.2.15) and (5.2.18) gives insight into the distance to instability, this approach is not designed to reveal transient behavior of a particular solution. We illustrate this point with three simple examples.

Example 5.2.4 Consider the trivial delay equation

$$\dot{x}(t) = ax(t) + bx(t-1), \qquad \text{with} \qquad (a, b) = (0, 0). \tag{5.2.20}$$

For simplicity, we choose $\tau = 1$. The choice of coefficients a and b makes the solution constant; no transient growth. Working from our intuition from the standard eigenvalue problem, we might expect the pseudospectra to reveal transient behavior through the level sets of the resolvent, shown in Figure 5.1. To visualize the pseudospectra, we graph the boundary $\partial\Lambda_{\epsilon,p}(F)$, or, equivalently, $\{\lambda \mid f(\lambda; p) = \epsilon^{-1}\}$, with different colors representing different ϵ values. The colorbar on the right shows $\log_{10}(\epsilon)$.



(a) $p = 2$ (b) $p = \infty$

Figure 5.1: Level set comparison of $f(\lambda; p)$ for (5.2.20) with structured perturbation from (5.2.17).

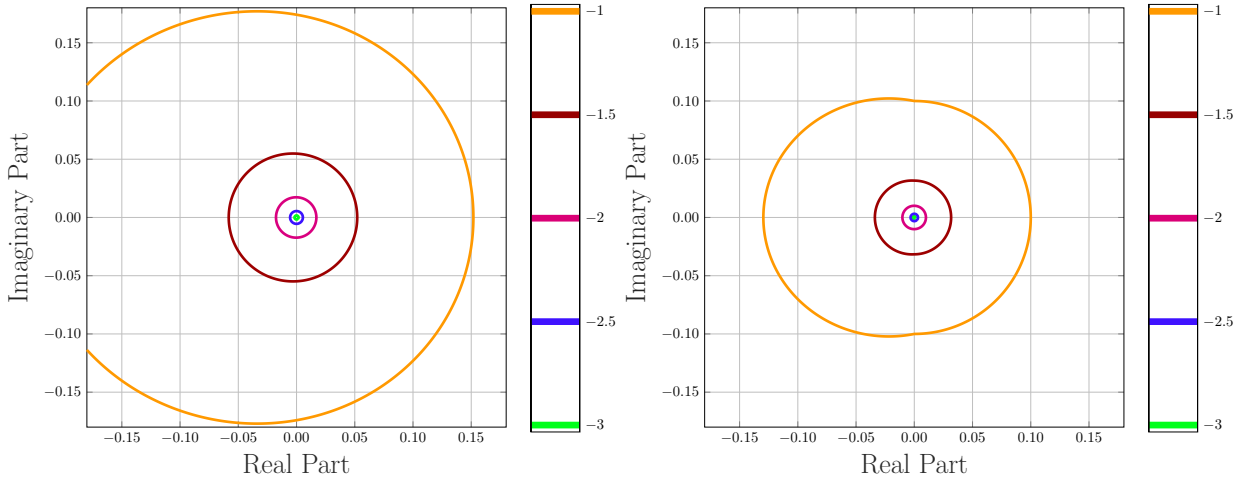
In particular, we might expect the natural generalization of the Kreiss constant to be never greater than 1, since the solution cannot grow for any initial history function $\phi(t)$. For the ∞ norm pseudospectra in Figure 5.1b, we observe that this seems to be the case: For example examine the level set line for $\epsilon = 0.1$, where $\max_{z \in \partial\Lambda_{0.1,p}} \Re(z) \approx 0.1$. Not so, however, for the 2-norm pseudospectrum in Figure 5.1a, where $\max_{z \in \partial\Lambda_{0.1,p}} \Re(z) > 0.1$. This highlights that the choice of norms is important, since different norms imply different scalings on the structured perturbation in (5.2.15).

Example 5.2.5 Next, we consider an equation similar to (5.2.20), but with $d = 2$:

$$\dot{x}(t) = ax(t) + bx(t-1) + cx(t-2), \qquad \text{with} \qquad (a, b, c) = (0, 0, 0). \tag{5.2.21}$$

This equation also leads to constant solutions, so we expect the pseudospectrum to behave

similar to that for (5.2.20). However the perturbation measure in (5.2.18) yields different pseudospectra, shown in Figure 5.2.



(a) $p = 2$ (b) $p = \infty$

Figure 5.2: Level set comparison of f for (5.2.21) with structured perturbation from (5.2.17).

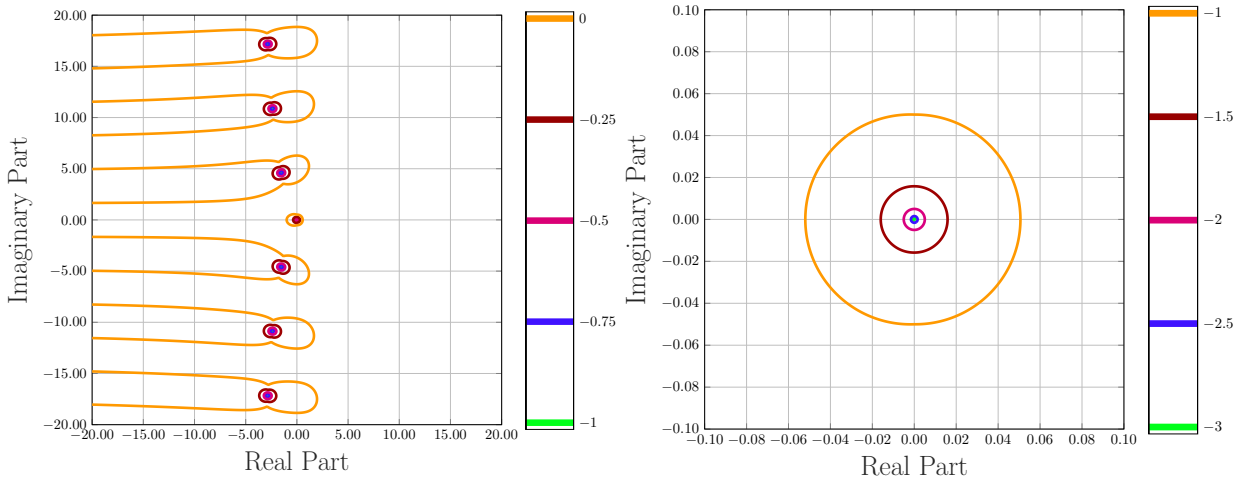
Again, the ∞-norm pseudospectra in Figure 5.2b suggest a Kreiss constant no greater than 1, but the level set lines for the 2-norm pseudospectra extend further into the complex plane, resulting in a larger Kreiss constant. Comparing Figure 5.2a and Figure 5.2b, the difference induced by the choice of norm is more prominent than in Example 5.2.4.

This example and the last illustrate how the same dynamical system can exhibit different pseudospectra if one allows for different perturbations to the coefficients a_k of the system, $k = 1, \ldots, d$.

Example 5.2.6 As a last example in this section, consider

$$\dot{x}(t) = ax(t) + bx(t-1), \qquad \text{with} \qquad (a, b) = (-1, 1). \tag{5.2.22}$$

Here the coefficients are nonzero, and the nonlinear eigenvalue problem (5.2.10) has infinitely many solutions, given by the branches of the Lambert W function from (5.2.12).



(a) $p = \infty$

(b) $p = \infty$, Zoom

Figure 5.3: Level set comparison of f with structured perturbation from (5.2.17) for the delay equation (5.2.22).

For (5.2.22), the contribution of eigenvalues corresponding to $k \neq 0$ branches of the Lambert W function significantly contribute to the Kreiss constant. In Figure 5.3a we observe how the pseudospectral level set lines corresponding to branches of the Lambert W function with $k \neq 0$ extend further into the right half plane than those for $k = 0$ (corresponding to the eigenvalue $\lambda = 0$).

The previous examples illustrate how perturbations to the coefficients a_k of the delay equation ($k = 1, \ldots, d$) shape the pseudospectra from (5.2.15). Information about transient behavior, however, does not coincide clearly with the perturbed eigenvalues, as in the (non-

delay) case, most notably in the first two examples with constant solutions $x(t)$ and $d = 1, 2$. Plischke [41] considers stability for delay equations on L_2 using the solution semigroup. To gain information on transient behavior, we investigate a different approach using the solution operator for the DDE. Stroh in [110] considers pseudospectra of the solution operator to assess transient growth. In the following, we extend this approach.

5.3 A Related Discrete Time Problem

The nonlinear eigenvalue problem (5.2.10) can be challenging to analyze, even for a single delay, since there are infinitely many eigenvalues. Understanding can be gained instead by studying the solution operator that advances the initial condition by one τ interval. We can then analyze a linear, discrete time dynamical system to gain insight into the transient behavior.

5.3.1 The Solution Operator

The *method of steps* is a typical way to solve The history function $\phi(t)$ can be considered an inhomogeneous contribution to (5.3.4) on $[0, \tau]$. The variation of parameters formula for $t \in [0, \tau]$ yields

$$x(t) = \mathrm{e}^{at}x(0) + \int_0^t \mathrm{e}^{a(t-s)}b\phi(s)\,\mathrm{d}s, \qquad t \in [0, \tau]. \tag{5.3.1}$$

The map $\phi(t) \mapsto x(t)$ in (5.3.1) defines an operator $\mathcal{M} : \mathcal{PC}([0,\tau];\mathbb{R}) \to \mathcal{PC}([0,\tau];\mathbb{R})$ by

$$(\mathcal{M}x)(t) = \mathrm{e}^{at}x(\tau) + \int_0^t \mathrm{e}^{a(t-s)}bx(s)\,\mathrm{d}s, \qquad t \in [0,\tau]. \tag{5.3.2}$$

The operator $\mathcal{M}$ acts on functions $x(t)$ on $[0,\tau]$ by advancing a function $x(t)$ according to the delay equation (5.3.4) by one τ interval. Note that $\mathcal{M}$ acts on functions defined on $[0,\tau]$, which can be thought of as a *reference interval*. Hence $\mathcal{M}$ to advances solutions from $[(k-1)\tau, k\tau]$ to the next interval $[k\tau,(k+1)\tau]$, $k \in \{0,1,\ldots\}$. This way, the solution to the DDE (5.3.4) can be computed for any $t > 0$ by iteratively applying $\mathcal{M}$ until the desired $t \in [(k-1)\tau, k\tau]$ is reached:

$$x\big|_{[(k-1)\tau,k\tau]} = \mathcal{M}^k\phi, \qquad k \in \mathbb{N}. \tag{5.3.3}$$

5.3.2 A Note on Norms

We first examine a simple case of (5.2.13) with $d = 1$ to present our approach, which we extend to $d > 1$ in Section 5.3.5. Consider

$$\begin{aligned} \dot{x}(t) &= ax(t) + bx(t-\tau), \qquad a, b \in \mathbb{R}, \\ x(t) &= \phi(t+\tau), \qquad t \in [-\tau, 0], \end{aligned} \tag{5.3.4}$$

with the initial history function $\phi(t) \in \mathcal{PC}([0,\tau],\mathbb{C})$.

Remark 5.3.1 Before continuing our analysis, let us clarify our choice $\phi \in \mathcal{PC}([0,\tau],\mathbb{C})$ in (5.3.4). In [43, 57, 111], the authors also consider $C^0([0,\tau],\mathbb{C})$ as the domain for generator of the solution semigroup. In contrast, [9] considers the Sobolev spaces $W^{1,p}([0,\tau],\mathbb{C})$ with

applications to control theory in mind. Piecewise continuous functions $\phi(t)$ are considered in [117] with applications to stability analysis. More general $L_p([0,\tau],\mathbb{C})$ spaces are discussed in [33] with a perspective on infinite dimensional dynamical systems.

The quest for the least required regularity of the initial history function $\phi(t)$ so that the solution to (5.3.4) is well defined is of tangential interest to our pursuit in this Chapter. We choose $\phi \in \mathcal{P}C([0,\tau],\mathbb{C})$ with the $\|\cdot\|_{L_\infty}$ norm, since we are interested in maximum transient growth of the solution. The following simple example illustrates why $L_2([0,\tau],\mathbb{C})$ is not a suitable choice for transient analysis.

Example 5.3.2 For $\tau = 1$, consider $\dot{x}(t) = -x(t) + \frac{1}{2}x(t-1)$. Let $\phi(t) = \mathrm{e}^{c(t-1)}$, for some $c > 0$ and $t \in [0,1]$ be the initial history function. Independent of c, $\phi(1) = 1$ but $\phi(t)$ is close to 0 for $t < 0$ and large values of c. In particular, the norm of ϕ is

$$\|\phi\|_2^2 = \int_0^1 |\phi(t)|^2 \,\mathrm{d}t = \int_0^1 \mathrm{e}^{2c(t-1)} \,\mathrm{d}t = \frac{1 - \mathrm{e}^{-2c}}{2c}. \tag{5.3.5}$$

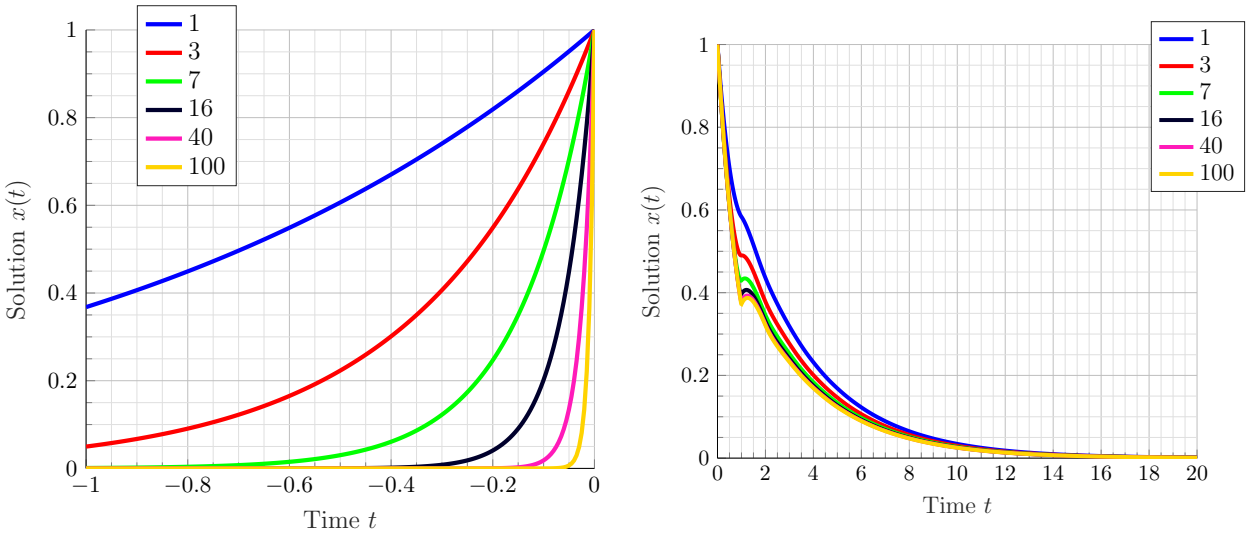
Observe that $\|\phi\|_{L_2}^2 \to 0$ for $c \to \infty$. A short computation reveals that

$$x(t) = \mathrm{e}^{-t} + \frac{1}{2(c+1)} \mathrm{e}^{-c} \left(\mathrm{e}^{ct} - \mathrm{e}^{-t} \right) \qquad t \in [0,1], \tag{5.3.6}$$

For the norm, another short computation shows that $\|x\|_{L_2}^2 \to \frac{1}{2}(1 - \mathrm{e}^{-2}) > 0$ for $c \to \infty$ and hence $\|x\|_{L_2}$ is bounded away from zero for any $c > 0$ and bounded above. On the interval $[0,1]$, it follows that

$$\frac{\|x\|_{L_2}}{\|\phi\|_{L_2}} \to \infty, \qquad c \to \infty. \tag{5.3.7}$$

To illustrate the computations above, Figure 5.4 shows the solution and the initial condition.



(a) Initial history function $\phi(t)$

(b) Solution on $[0, 20]$.

Figure 5.4: Initial history function $\phi(t)$ and solution for $\dot{x}(t) = -x(t) + \frac{1}{2}x(t-1)$, various values of c, indicated by the color of the curve.

The main point of this example is that while $|x(t)|$ does not grow beyond $1 = x(0)$ (see Figure 5.4b), the relative transient behavior in the $\|\cdot\|_{L_2}$-norm is large. This seemingly large transient growth is caused by the small $\|\cdot\|_{L_2}$ norm of the initial condition (compare Figure 5.4a) rather than growth in the solution.

To avoid such a phenomenon, we consider the $\|\cdot\|_{L_\infty}$ norm to assess transient behavior in the solution.

5.3.3 Discretizing the Solution Operator

The solution operator $\mathcal{M}$ in (5.3.2) leads to an infinite dimensional eigenvalue problem on $\mathcal{PC}([0,\tau],\mathbb{C})$. We follow [26, 78, 110] to discretize $\mathcal{M}$, to approximate function values of the exact solution $x(t)$ at discrete points in t. Chebyshev discretization is known to have superior approximation properties on finite intervals [113], so we represent $x(t)$ on $[0,\tau]$ by its values at Chebyshev points $t_k = \tau(1+\cos(k\pi/N))/2$, $k = 0,\ldots,N$. For simplicity, we use *Lagrange* basis functions $\ell_k(t)$ to construct the interpolant, defined by

$$\ell_k(t) = \prod_{\substack{j=0 \\ j\neq k}}^{N} \frac{t-t_k}{t_j-t_k}, \qquad k = 0,\ldots,N. \tag{5.3.8}$$

Then the solution $x(t)$ and initial history function $\phi(t)$ can be approximated on $[0,\tau]$ by

$$\phi(t) \approx \sum_{k=0}^{N} u_k^{(0)} \ell_k(t), \qquad \text{and} \qquad x(t) \approx \sum_{k=0}^{N} u_k^{(1)} \ell_k(t). \tag{5.3.9}$$

We write the coefficient vectors as $\mathbf{u}_N^{(k)} = [u_0^{(k)},\ldots,u_N^{(k)}]^\top$, $k = 0,1$, where the subscript on the vector notation emphasizes the discretization order N. Note that, by construction, $\phi(t_j) = u_j^{(0)}$. We can use vector norms on the coefficient vector $\mathbf{u}_N^{(k)}$ to approximate the L_∞ norm of $x(t)$:

$$\|\mathbf{u}_N^{(0)}\|_{L_\infty} \approx \|\phi\|_{L_\infty}, \qquad \text{and} \qquad \|\mathbf{u}_N^{(1)}\|_{L_\infty} \approx \|x\|_{L_\infty}. \tag{5.3.10}$$

As $N\to\infty$, we get equality in (5.3.10), by the convergence of the discretization. While we solve the DDE (5.3.4) on $\mathcal{PC}([0,\tau],\mathbb{R})$, we consider the L_∞ norm to examine the maximum transient behavior.

Substituting the approximations (5.3.9) into (5.3.2) leads to

$$\begin{aligned} x(t) &= \mathrm{e}^{at}x(0) + \int_0^t \mathrm{e}^{a(t-s)}b\phi(s)\,\mathrm{d}s \\ &\approx \mathrm{e}^{at}\phi(\tau) + \int_0^t b\,\mathrm{e}^{(t-s)a}\sum_{k=0}^{N}\phi(t)\ell_k(s)\,\mathrm{d}s \\ &= \mathrm{e}^{ta}\phi(\tau) + \sum_{k=1}^{N} b\,\mathrm{e}^{ta}\phi(t)\int_0^t \mathrm{e}^{-sa}\ell_k(s)\,\mathrm{d}s, \end{aligned} \tag{5.3.11}$$

for $t \in [0,\tau]$. Due to the interpolation at Chebyshev points in (5.3.9), at a point $t_j \in [0,\tau]$ (5.3.11) becomes

$$\begin{aligned} x(t_j) &\approx \mathrm{e}^{t_j a}\phi(t_0) + \sum_{k=0}^{N} b\phi(t_j)\,\mathrm{e}^{t_j a}\int_0^{t_j} \mathrm{e}^{-sa}\ell_k(s)\,\mathrm{d}s \\ &= \mathrm{e}^{t_j a}u_0^0 + \sum_{k=0}^{N} bu_j^{(0)}\,\underbrace{\mathrm{e}^{t_j a}\int_0^{t_j} \mathrm{e}^{-sa}\ell_k(s)\,\mathrm{d}s}_{=:\ m_{j,k}}. \end{aligned} \tag{5.3.12}$$

Using $m_{j,k}$ from (5.3.12), we can approximate the action of $\mathcal{M}$ on discrete values of $\phi(t)$ and $x(t)$ from (5.3.9) as a matrix multiplication:

$$\begin{bmatrix} x(t_0) \\ \vdots \\ x(t_N) \end{bmatrix} \approx \mathbf{M}_N \begin{bmatrix} \phi(t_0) \\ \vdots \\ \phi(t_N) \end{bmatrix} = \mathbf{M}_N \begin{bmatrix} u_0^{(0)} \\ \vdots \\ u_N^{(0)} \end{bmatrix}, \tag{5.3.13}$$

where the entries of $\mathbf{M}_N \in \mathbb{C}^{(N+1)\times(N+1)}$ are given by

$$[\mathbf{M}_N]_{j,k} = \delta_{k,0}\,\mathrm{e}^{t_j a} + b\underbrace{\int_0^{t_j} \mathrm{e}^{a(t_j-s)}\ell_k(s)\,\mathrm{d}s}_{=\ m_{j,k}\ \text{from (5.3.12)}} = \delta_{k,0}\,\mathrm{e}^{t_j a} + bm_{j,k}, \qquad j,k = 0,\ldots,N. \tag{5.3.14}$$

The quantities $m_{j,k}$ in (5.3.12) or (5.3.14) can be precomputed and stored. As for $\mathcal{M}$ from (5.3.2), iterated applications of $\mathbf{M}_N$ yield the approximate solution on intervals $[(k-1)\tau, k\tau]$

at the discretization points:

$$\begin{bmatrix} x^{(k+1)}(t_0) \\ \vdots \\ x^{(k+1)}(t_N) \end{bmatrix} \approx \mathbf{M}_N \begin{bmatrix} x^{(k)}(t_0) \\ \vdots \\ x^{(k)}(t_N) \end{bmatrix} \approx \mathbf{M}_N^{k+1} \begin{bmatrix} u_0^{(0)} \\ \vdots \\ u_N^{(0)} \end{bmatrix}. \tag{5.3.15}$$

We emphasize that the discrete solution matrix $\mathbf{M}_N$ advances the solution at discrete points t_j, $j = 0, \ldots, N$ by one τ interval. The following decomposition of $\mathbf{M}_N$ in (5.3.14) is useful in our following investigations:

$$\mathbf{M}_N = \mathbf{M}_{N,E} + b\mathbf{M}_{N,0}, \qquad [\mathbf{M}_{N,E}]_{j,k} = \delta_{k,0}\, \mathrm{e}^{t_j a}, \qquad [\mathbf{M}_{N,0}]_{j,k} = m_{j,k}. \tag{5.3.16}$$

The coefficient a from (5.3.4) enters both in $\mathbf{M}_{N,E}$ and $\mathbf{M}_{N,0}$, without explicit notation.

The authors in [26, 110] show the convergence of $\mathbf{M}_N$ to $\mathcal{M}$.

It is important to mention that our solution approximation using powers of $\mathbf{M}_N$ is not exact. On $[-\tau, 0]$, the coefficients of the interpolant $u_j^{(0)}$ match $\phi(t_j)$ exactly ($j = 0, \ldots, N$) by construction using Lagrange interpolation in (5.3.9). The accuracy of the approximation is lost, however, by evolving the discrete system in time using the solution matrix $\mathbf{M}_N$ However, for an accurate enough discretization, we retain $x^{(k)}(t_j) \approx u_j^{(k)}$, $t \in [k\tau, (k+1)\tau]$, $k \in \mathbb{N}$, $j = 0, \ldots, N$. By refining the discretization, we get

$$\left(\mathbf{M}_N^k \phi\right)(t_j) \to \left(\mathcal{M}^k \phi\right)(t_j), \qquad N \to \infty, \quad k \in \mathbb{N}, \quad j = 0, \ldots, N. \tag{5.3.17}$$

Remark 5.3.3 In our experience, we observe spectral convergence of $\mathbf{M}_N$ to $\mathcal{M}$. This justifies the relatively low order approximations in our experiments; usually $N = 16$ or $N = 32$ is sufficient.

To study transient behavior, we connect the NLEVP (5.2.13) with the solution operator $\mathcal{M}$ from (5.3.2) in the following lemma; for more details, see [37, 65].

Lemma 5.3.4 *Let $\tau = 1$. If $\lambda \in \mathbb{C}$ solves the NLEVP, i.e., $F(\lambda) = 0$ from (5.3.4) , then e^{λ} is an eigenvalue of $\mathcal{M}$ corresponding to (5.3.4).*

Proof Rearranging $F(\lambda) = 0$ gives

$$\lambda - a = \mathrm{e}^{-\lambda} b. \tag{5.3.18}$$

Apply $\mathcal{M}$ to the ansatz $x(t) = \mathrm{e}^{\lambda t}$ to get

$$\begin{aligned} \mathcal{M}\,\mathrm{e}^{\lambda t} &= \mathrm{e}^{ta}\,\mathrm{e}^{\lambda\tau} + b\int_0^t \mathrm{e}^{(t-s)a}\,\mathrm{e}^{\lambda(s-1)}\,\mathrm{d}s \\ &= \mathrm{e}^{ta+\lambda} + \mathrm{e}^{ta}\,\mathrm{e}^{-\lambda} b\left[\frac{1}{\lambda - a}\,\mathrm{e}^{(\lambda-a)s}\right]_{s=0}^{s=t} \\ &= \mathrm{e}^{ta+\lambda} + \mathrm{e}^{ta}\big(\mathrm{e}^{(\lambda-a)t} - 1\big) \\ &= \mathrm{e}^{\lambda}\,\mathrm{e}^{\lambda t}. \end{aligned} \tag{5.3.19}$$

So we have shown that e^{λ} is an eigenvalue of $\mathcal{M}$ with corresponding eigenfunction $\mathrm{e}^{\lambda t}$. ■

Hence the map $z \mapsto \mathrm{e}^{z}$ maps solutions of the nonlinear eigenvalue problem (5.2.14) to eigenvalues of the solution operator $\mathcal{M}$.

We refer to the solutions of the NLEVP, $F(\lambda) = 0$ with

$$F(\lambda) = \lambda - a - b\,\mathrm{e}^{-\lambda\tau} \tag{5.3.20}$$

as λ. Stability requires $\Re(\lambda) < 0$ and for transient analysis. The eigenvalue $\lambda = 0$ on the boundary of the stability region.

The operator $\mathcal{M}$ and its discretization $\mathbf{M}_N$ are discrete-time systems. We refer to the discrete-time eigenvalues as $\mu = \mathrm{e}^{\lambda}$ by Lemma 5.3.4. The eigenvalue on the stability boundary corresponding to $\lambda = 0$ then is $\mu = 1$.

We illustrate the convergence of $\mathbf{M}_N$ to $\mathcal{M}$ on the simple example $\dot{x}(t) = -x(t) + x(t-1)$ in Figure 5.5.

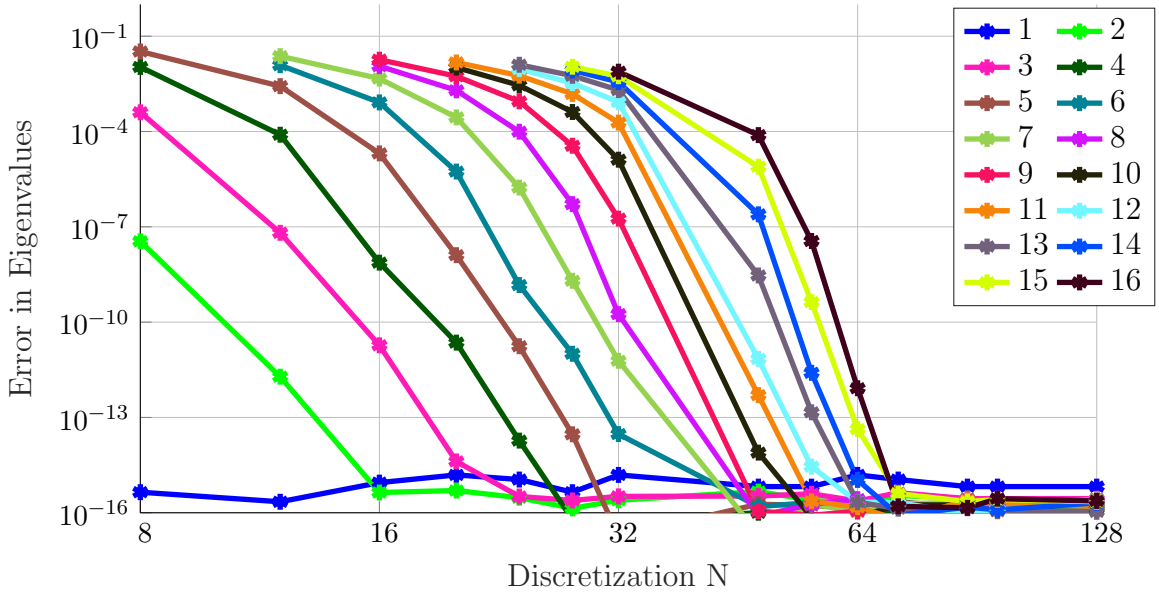


Figure 5.5: Error between eigenvalues of $\mathbf{M}_N$ and those of $\mathcal{M}$ for various discretization orders N. Different colors represent different eigenvalues. Note that we only plot eigenvalues on or above the real axis since they appear in complex conjugate pairs.

We label the eigenvalues by absolute value in descending order, i.e., $\lambda_1(\mathbf{M}_N)$ denotes the largest magnitude eigenvalue of $\mathbf{M}_N$. We observe the spectral convergence of the eigenvalues of the discrete solution matrix $\mathbf{M}_N$ to the eigenvalues of the continuous solution operator $\mathcal{M}$, where the eigenvalues of $\mathcal{M}$ are computed using the Lambert W function from (5.2.12). We remark that the largest magnitude eigenvalues of $\mathcal{M}$ are approximated best by the eigenvalues of $\mathbf{M}_N$. This is important since those are precisely the eigenvalues we expect to be most closely associated with transient growth of the solution.

5.3.4 A Simple Example

We illustrate the insight from the discrete solution matrix $\mathbf{M}_N$ on transient behavior with the following simple example. Let $\tau = 1$ and consider

$$\dot{x}(t) = ax(t) + bx(t-1), \qquad a, b, \in \mathbb{R}. \tag{5.3.21}$$

For $\mathbf{M}_N$ corresponding to (5.3.21), we compare the following two quantities:

$$\rho(\mathbf{M}_N) = \max_{i=1,\dots,N+1} |\lambda_i(\mathbf{M}_N)|, \qquad \text{and} \qquad \|\mathbf{M}_N\|_{L_\infty} = \max_{\|\mathbf{v}\|_{L_\infty}=1} \|\mathbf{M}_N\mathbf{v}\|_{L_\infty}. \tag{5.3.22}$$

Stability of the solution is determined by $\rho(\mathcal{M})$, which we approximate by $\rho(\mathbf{M}_N)$, so we require $\rho(\mathbf{M}_N) < 1$ for asymptotic stability of the solution $x(t)$. On the other hand $\|\mathbf{M}_N\|_{L_\infty}$ suggests maximum growth of $x(t)$. To analyze transient behavior of solutions to (5.3.21), we aim to maximize $\|\mathbf{M}_N\|_{L_\infty}$ under the constraint that $\rho(\mathbf{M}_N) < 1$. To find such a parameter configuration (a, b), we examine level sets of $\rho(\mathbf{M}_N)$ and $\|\mathbf{M}_N\|_{L_\infty}$ for varying a and b in Figure 5.6.

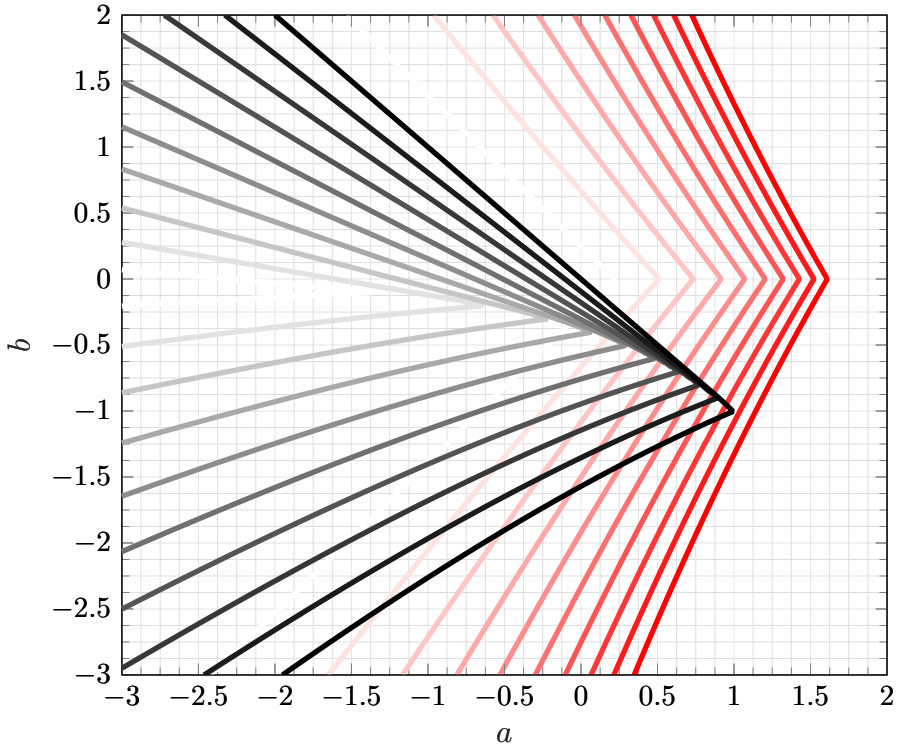
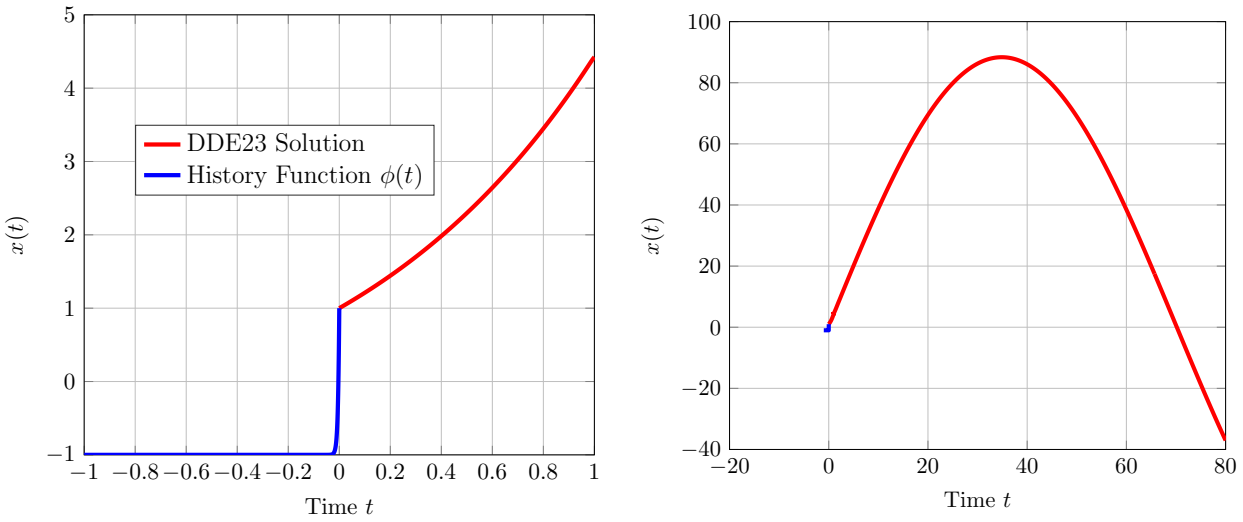


Figure 5.6: Level set plot of the spectral radius $\rho(\mathbf{M}_{32})$ and $\|\mathbf{M}_{32}\|_{L_\infty}$ for varying a and b values. Here $\rho(\mathbf{M}_{32})$ is shown in black ranging from 0.1 to 1, $\|\mathbf{M}_{32}\|_{L_\infty}$ in red from 1.25 to 5, each light to dark.

Visually, we detect the largest transient growth by finding the rightmost point of the $\|\mathbf{M}_N\|_{L_\infty}$ level sets under the constraint $\rho(\mathbf{M}_N) < 1$, which yields $\|\mathbf{M}_N\|_{L_\infty} \approx 4.43$, corresponding to $(a, b) = (1, -1)$. However, this parameter configuration is on the boundary of the stability region. To find a stable, nearby system with similar transient behavior, we choose values (a, b) that lie within $\rho(\mathbf{M}_N) < 1$ but close to $(a, b) = (1, -1)$. Specifically, we fix $b = -1$ and $a < 1$ but a close to 1.

In Figure 5.7, we plot the solution corresponding to $a = 0.999$, $b = -1$ and the initial history

function $\phi(t) = 2\,\mathrm{e}^{2000(t-1)} - 1$.



(a) Solution on $t \in [0, 1]$

(b) Solution on $t \in [0, 80]$

Figure 5.7: Solution $x(t)$ on $[0, 1]$, $a = 0.999$, $b = -1$, $\phi(t) = 2\,\mathrm{e}^{2000t} - 1$

The maximum value of $\max_{a,b} |\mathbf{M}_N\|_{L_\infty} \approx 4.43$; note that $x(1)$ comes close to attaining this value in Figure 5.7a. However, we observe in Figure 5.7b that the maximum transient growth is not attained in $[0, \tau]$, but at some time around $t \approx 35$. In Figure 5.8, we fix $b = -1$ and show $\|\mathbf{M}_N^k\|_{L_\infty}$ over powers $k = 0, 1, \ldots$ for various values of $a \to 1$, approaching the boundary of the stability region in Figure 5.6.

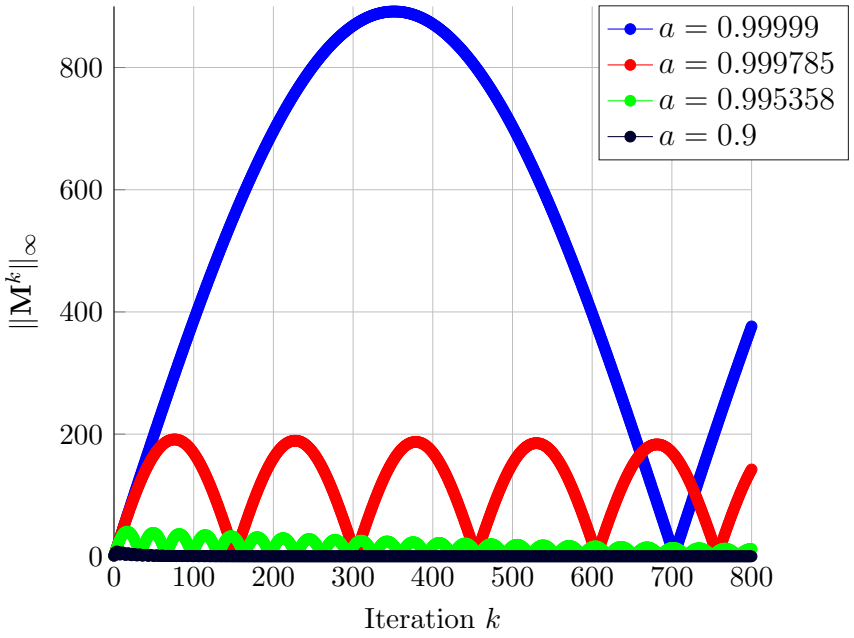


Figure 5.8: $\|\mathbf{M}_{32}^k\|_{L_\infty}$ for different values of a, $b = -1$ fixed.

In Figure 5.8, we observe that we can force arbitrary transient growth by choosing a close to 1. More generally, choosing a parameter configuration (a, b) that yields an $\mathbf{M}_N$ that is stable but close to the maximum transient one yields such behavior.

Remark 5.3.5 We construct the initial history function $\phi(t)$ that realizes the maximum transient growth as follows. For (a, b) with maxium transient growth, we find $\mathbf{y}^* \in \mathbb{C}^{N+1}$ so that

$$\mathbf{y}^* = \arg \max_{\mathbf{y} \in \mathbb{C}^{N+1}, \|\mathbf{y}\|_{L_\infty} = 1} \|\mathbf{M}_N \mathbf{y}\|_{L_\infty} \tag{5.3.23}$$

Such a $\mathbf{y}^*$ necessarily has entries $\mathbf{y}_i = \pm 1$ to maximize $\|\cdot\|_{L_\infty}$. Then ϕ is chosen as a smooth approximation of $\mathbf{y}^*$.

Motivated by the findings in this section, we want to determine parameter configurations

(a, b) that yield maximum transient growth in more general settings. A natural next step is the extension of the discrete solution matrix $\mathbf{M}_N$ to several commensurate delays.

5.3.5 Several Commensurate Delays

Consider a scalar delay differential equation with d commensurate delays:

$$\dot{x}(t) = a_0 x(t) + a_1 x(t-\tau) + \cdots + a_d x(t-d\tau) = \sum_{k=0}^{d} a_k x(t-k\tau), \qquad d \in \mathbb{N}. \tag{5.3.24}$$

Investigating the transient behavior of (5.3.24) via the discrete solution matrix (5.3.14) requires a generalization of the discretization $\mathbf{M}_N$ to several commensurate delays. We present a detailed derivation for $d = 2$ for illustration; a similar construction holds for the general case of d commensurate delays. Consider the delay equation

$$\begin{aligned} \dot{x}(t) &= ax(t) + bx(t-\tau) + cx(t-2\tau), \qquad a, b, c \in \mathbb{R}, \\ x(t) &= \phi(t+2\tau), \qquad t \in [-2\tau, 0]. \end{aligned} \tag{5.3.25}$$

To advance the solution by one τ interval, we split the initial history $\phi(t)$, defined on $[0, 2\tau]$ into

$$\begin{aligned} \phi_1(t) &= \phi(t+\tau) \quad t \in [0, \tau], \quad \text{and} \\ \phi_2(t) &= \phi(t), \qquad\qquad\qquad\qquad t \in [0, \tau]. \end{aligned} \tag{5.3.26}$$

So then we get

$$\begin{aligned} x(0) &= \phi(2\tau) = \phi_1(\tau), \\ x(-\tau) &= \phi(\tau) = \phi_1(0) = \phi_2(\tau), \quad \text{and} \\ x(-2\tau) &= \phi(0) = \phi_2(0). \end{aligned} \tag{5.3.27}$$

We approximate the initial history functions in (5.3.26) on $[0,\tau]$ with the Lagrange basis functions $\ell_k(t)$ from (5.3.8) by

$$\phi_1(t) \approx \sum_{k=0}^{N} u_k^{(0)} \ell_k(t), \quad \text{and} \quad \phi_2(t) \approx \sum_{k=0}^{N} u_k^{(-1)} \ell_k(t). \tag{5.3.28}$$

Then we can write the solution $x(t)$ on $[0,\tau]$ based on the (given) input functions from (5.3.26) as

$$x(t) = \mathrm{e}^{ta} x_0 + \int_0^t \mathrm{e}^{(t-s)a} c\phi_2(s)\,\mathrm{d}s + \int_0^t \mathrm{e}^{(t-s)a} b\phi_1(s)\,\mathrm{d}s. \tag{5.3.29}$$

Substituting (5.3.28) into (5.3.29), $x(t)$ can be approximated at the Chebyshev points $\{t_0,\ldots,t_N\}$ as

$$\begin{aligned} x(t_j) &\approx \mathrm{e}^{t_j a}\phi_1(\tau) + \int_0^{t_j} \mathrm{e}^{(t_j-s)a} c \sum_{k=0}^{N} u_k^{(-1)} \ell_k(s)\,\mathrm{d}s + \int_0^{t_j} \mathrm{e}^{(t_j-s)a} b \sum_{k=0}^{N} u_k^{(0)} \ell_k(s)\,\mathrm{d}s \\ &= \mathrm{e}^{t_j a}\phi_1(\tau) + c\sum_{k=0}^{N} u_k^{(-1)} \underbrace{\mathrm{e}^{t_j a}\int_0^{t_j} \mathrm{e}^{-sa}\ell_k(s)\,\mathrm{d}s}_{=m_{j,k}} + b\sum_{k=0}^{N} u_k^{(0)} \underbrace{\mathrm{e}^{t_j a}\int_0^{t_j} \mathrm{e}^{-sa}\ell_k(s)\,\mathrm{d}s}_{=m_{j,k}} \\ &= \mathrm{e}^{t_j a} u_0^{(0)} + c\sum_{k=0}^{N} u_k^{(-1)} m_{j,k} + b\sum_{k=0}^{N} u_k^{(0)} m_{j,k}. \end{aligned} \tag{5.3.30}$$

Observe how the precomputed quantities $m_{j,k}$ from (5.3.12) reappear in (5.3.30) for several delays. The (discrete) solution operator that advances the solution of (5.3.25) by one τ interval can be expressed in block matrix form as

$$\mathbf{M}_N = \begin{bmatrix} b\mathbf{M}_{N,0} + \mathbf{M}_{N,E} & c\mathbf{M}_{N,0} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \text{with} \quad [\mathbf{M}_{N,0}]_{j,k} = m_{j,k}, \quad \text{and} \quad [\mathbf{M}_{N,E}]_{j,k} = \mathrm{e}^{t_j a}\delta_{k,0}. \tag{5.3.31}$$

The definition of $\mathbf{M}_{N,0}$ and $\mathbf{M}_{N,E}$ matches (5.3.14). Let $x(t)$ on $[0,\tau]$ be approximated as

$$x(t) \approx \sum_{k=0}^{N} u^{(0)} \ell_k(t), \qquad t \in [0,\tau], \tag{5.3.32}$$

with coefficient vector $\mathbf{u}^{(0)} = [u_0^{(0)}, \ldots, u_N^{(0)}]^\top \in \mathbb{C}^{n+1}$, (5.3.30) can be represented by block matrix multiplication:

$$\begin{bmatrix} \mathbf{u}^{(1)} \\ \\ \mathbf{u}^{(0)} \end{bmatrix} = \begin{bmatrix} b\mathbf{M}_{N,0} + \mathbf{M}_{N,E} & c\mathbf{M}_{N,0} \\ \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(0)} \\ \\ \mathbf{u}^{(-1)} \end{bmatrix}. \tag{5.3.33}$$

Now consider the case of d commensurate delays, as in (5.3.24). Then $\phi(t)$ is split up into d functions $\phi_k(t)$, $k = 1, \ldots, d$ as in (5.3.26). Let $\mathbf{u}_N^{(k)} = [u_0^{(k)}, \ldots, u_N^{(k)}]^\top \in \mathbb{C}^{N+1}$ be the coefficient vector for $\phi_k(t)$, $j = 0, \ldots, d$. The action of $\mathcal{M}$ is approximated by $\mathbf{M}_N$ as follows:

$$\begin{bmatrix} \mathbf{u}^{(1)} \\ \mathbf{u}^{(0)} \\ \vdots \\ \mathbf{u}^{(-d+1)} \end{bmatrix} = \begin{bmatrix} a_1\mathbf{M}_{N,0} + \mathbf{M}_{N,E} & a_2\mathbf{M}_{N,0} & \cdots & a_d\mathbf{M}_{N,0} \\ \mathbf{I} & 0 & & \\ & & \ddots & \\ & & \mathbf{I} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u}^{(0)} \\ \mathbf{u}^{(-1)} \\ \vdots \\ \mathbf{u}^{(-d)} \end{bmatrix} = \mathbf{M}_N \begin{bmatrix} \mathbf{u}^{(0)} \\ \vdots \\ \mathbf{u}^{(-d)} \end{bmatrix}. \tag{5.3.34}$$

Note that this structure of $\mathbf{M}_N$ resembles the companion linearization of a degree d polynomial eigenvalue problem [64].

5.3.6 Connecting the NLEVP and the Solution Operator $\mathcal{M}$

To compare the eigenvalues of the solution operator $\mathcal{M}$ to the solutions of the nonlinear eigenvalue problem $F(\lambda) = 0$ from (5.2.14), we lay out the details of the corresponding map $\Phi : \mathbb{C} \to \mathbb{C}$ that maps the solutions of $F(\lambda) = 0$ to the eigenvalues of $\mathcal{M}$. This directly extends the case $d = 1$ in Lemma 5.3.4 to several delays.

Lemma 5.3.6 *Consider the delay equation (5.3.24) with corresponding eigenvalue problem* $F(\lambda) = 0$ *from (5.2.14):*

$$F(\lambda) = \lambda - a_0 - \sum_{k=1}^{d} a_k \, \mathrm{e}^{-\lambda k \tau}. \tag{5.3.35}$$

If $F(\lambda) = 0$*, then* $\mathrm{e}^{\lambda\tau}$ *is an eigenvalue of* $\mathcal{M}$ *corresponding to (5.3.24).*

Proof Let $\lambda \in \mathbb{C}$ be such that $F(\lambda) = 0$. We rearrange (5.3.35) so that

$$\lambda - a_0 = \sum_{k=1}^{d} a_k \, \mathrm{e}^{-\lambda k \tau}. \tag{5.3.36}$$

Applying the solution operator $\mathcal{M}$ to the ansatz $x(t) = \mathrm{e}^{\lambda t}$ yields

$$\begin{aligned}
\mathcal{M}\,\mathrm{e}^{\lambda t} &= \mathrm{e}^{t a_0}\,\mathrm{e}^{\lambda\tau} + \sum_{k=1}^{d} \int_0^t \mathrm{e}^{a_0(t-s)} a_k \, \mathrm{e}^{\lambda(s-(k-1)\tau)} \, \mathrm{d}s \\
&= \mathrm{e}^{t a_0}\,\mathrm{e}^{\lambda\tau} + \mathrm{e}^{a_0 t + \lambda\tau} \sum_{k=1}^{d} a_k \, \mathrm{e}^{-k\lambda\tau} \int_0^t \mathrm{e}^{(\lambda - a_0)s} \, \mathrm{d}s \\
&= \mathrm{e}^{t a_0 + \lambda\tau} + \mathrm{e}^{a_0 t + \lambda\tau} \left(\mathrm{e}^{(\lambda - a_0)t} - 1 \right) \frac{1}{\lambda - a_0} \sum_{k=1}^{d} a_k \, \mathrm{e}^{-k\lambda\tau} \\
&= \mathrm{e}^{t a_0 + \lambda\tau} + \mathrm{e}^{a_0 t + \lambda\tau}\,\mathrm{e}^{(\lambda - a_0)t} - \mathrm{e}^{a_0 t + \lambda\tau} \\
&= \mathrm{e}^{\lambda\tau}\,\mathrm{e}^{\lambda t}.
\end{aligned} \tag{5.3.37}$$

Thus $x(t) = \mathrm{e}^{\lambda t}$ is an eigenfunction of $\mathcal{M}$ corresponding to the eigenvalue $\mathrm{e}^{\lambda\tau}$. ■

In this case $z \mapsto \mathrm{e}^{\tau z}$ maps solutions to the nonlinear eigenvalue problem $F(\lambda) = 0$ to eigenvalues of the solution operator $\mathcal{M}$.

To compute solutions of the NLEVP for several commensuare delays in (5.3.24), we use a result from [29], summarized in the following lemma. We omit the proof since the methods are not relevant to our investigations.

Lemma 5.3.7 *Consider the delay differential equation (5.3.24) with corresponding nonlinear eigenvalue problem $F(\lambda) = 0$ from (5.3.35). If $F(\lambda_{s,k}) = 0$, then*

$$\lambda_{s,k} = \frac{1}{\tau} W_k \left(\sum_{j=1}^{d} a_j \tau \, \mathrm{e}^{-j a_0 \tau} \, \mathrm{e}^{(1-j)(\lambda_{s,k} - a_0)\tau} \right) + a_0, \tag{5.3.38}$$

where W_k denotes branch k of the Lambert W function and $s = 1, \ldots, d$.

We illustrate the application of Lemma 5.3.7 to a delay equation with the following example.

Example 5.3.8 Consider the delay equation

$$\dot{x}(t) = -x(t) + x(t-1) + \tfrac{1}{2} x(t-2), \qquad t \geq 0, \tag{5.3.39}$$

with corresponding nonlinear eigenvalue problem

$$F(\lambda) = \lambda + 1 - \mathrm{e}^{-\lambda} - \tfrac{1}{2} \mathrm{e}^{-2\lambda} = 0, \qquad z \in \mathbb{C}. \tag{5.3.40}$$

In Figure 5.10, we compare the level sets of $F(z)$, $z \in \mathbb{C}$ to the solutions of the fixed point problem (5.3.38). As starting values, we follow the suggestion from [29] and choose

$$\lambda_{s,k}^{(0)} := W_k(\tfrac{\mathrm{e}}{2}) - 1 - \mathrm{i}(s-2) \cdot \begin{cases} \frac{\pi}{2}, & k = 0; \\ 3\pi, & k > 0. \end{cases} \tag{5.3.41}$$

We use a MATLAB®'s `fsolve` routine to solve the fixed point iteration (5.3.38) for the delay equation (5.3.39).

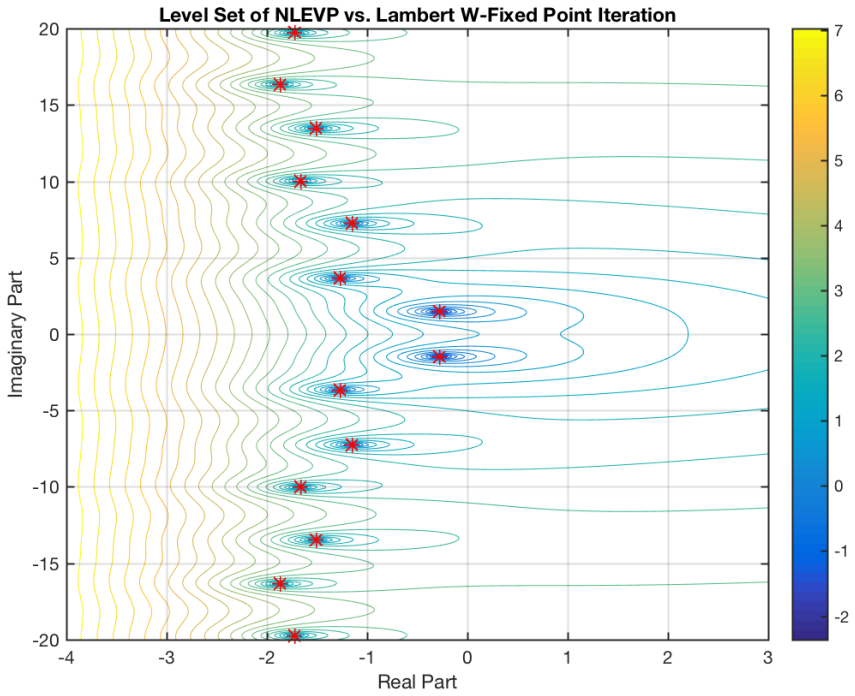
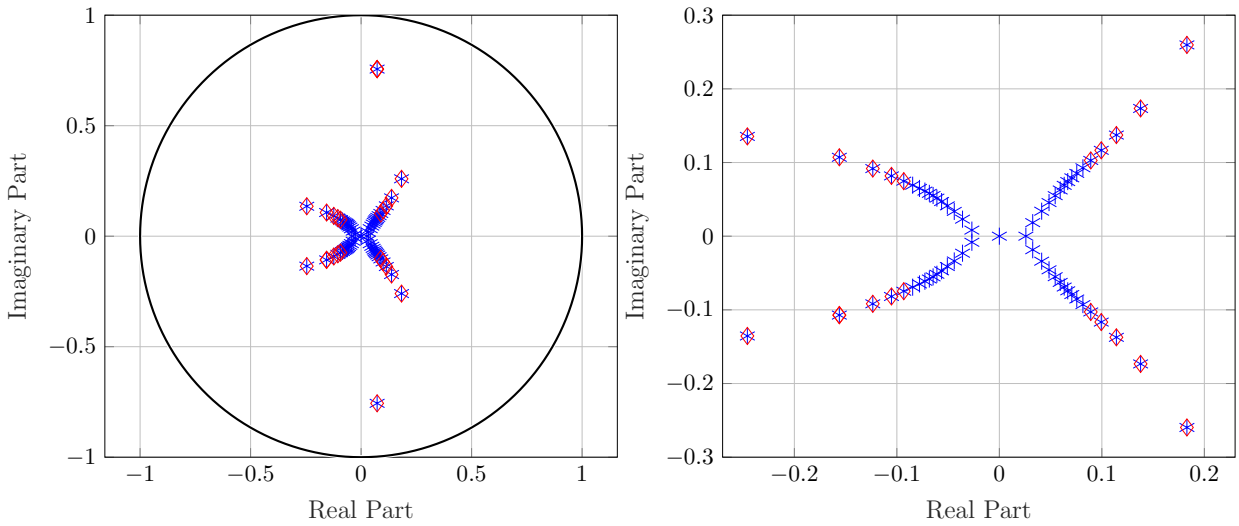


Figure 5.9: Level set plot of $\log(|F(\lambda)|)$ together with the computed eigenvalues from Lemma 5.3.7 (red stars).

In Figure 5.9, observe how the fixed points of (5.3.38) correspond to the solutions of the nonlinear eigenvalue problem $F(\lambda) = 0$.

With the fixed point characterization of the eigenvalues for (5.3.39), we can we visualize the approximation of eigenvalues for $N = 32$ in Figure 5.10.



(a) Plot on the unit disc

(b) Zoom

Figure 5.10: Fixed points from Lemma 5.3.7 (blue), mapped to the unit disc via $\mathbf{\Phi}$, compared to the eigenvalues of $\mathbf{M}_{32}$ (red).

We observe in Figure 5.10 the largest magnitude eigenvalues are approximated.

With the approximation properties shown in this section, we can use $\mathbf{M}_N$ to gain insight into the eigenvalues of the nonlinear eigenvalue problem (5.2.13). We continue with our investigation of parameter configuration for maximum transient growth in the next section.

5.4 Parameter Configuration for Maximum Transient Growth

With our intuition from Section 5.3.4, we aim to find an explicit representation of the parameter configuration for (5.3.24) that yields maximum transient growth for stable solutions $x(t)$. This far, we have considered the ansatz $x(t) = \mathrm{e}^{\lambda t}$ which yields the nonlinear eigenvalue problem. In an effort to find parameter configurations that support significant transient growth, we shall seek solutions of the form $x(t) = t^d \mathrm{e}^{\lambda t}$, which we expect to exists only in cases where $\mathrm{e}^{\lambda \tau}$ is an eigenvalue of $\mathcal{M}$, corresponding to a $(d+1) \times (d+1)$ Jordan block. We are especially interested in cases where $\Re(\lambda) < 0$ is close enough to zero that the t^d term in $x(t)$ initially grows fast enough to give transient growth, before the $\mathrm{e}^{\lambda t}$ term eventually gives decay in $x(t) = t^d \mathrm{e}^{\lambda t}$.

We first consider eigenvalues $\lambda = 0$ on the boundary of the stability region, corresponding to unstable solutions $x(t) = t^d$, then consider stable solutions with $\Re(\lambda) < 0$ in Section 5.4.2.

5.4.1 Maximum Size Jordan Block

The problem in this section is to find a parameter configuration $\mathbf{a}^* = \begin{bmatrix} a_0 & \dots & a_d \end{bmatrix}$ that solves

$$\mathbf{a}^* = \arg \sup_{a_0,\dots,a_d} \|\mathbf{M}_N\|_{L_\infty}, \quad \text{with} \quad \rho(\mathbf{M}_N) \leq 1. \tag{5.4.1}$$

For maximum transient growth, we construct a maximum $(d+1)\times(d+1)$ size Jordan block of $\mathcal{M}$ corresponding to $\mu = 1$. We chose the point of view of the nonlinear eigenvalue problem $F(\lambda) = 0$ and keep in mind that $\mu = \mathrm{e}^{\lambda\tau}$ yields corresponding eigenvalues for $\mathcal{M}$.

Consider $\mu = 1$, or, equivalently, $\lambda = 0$. Substituting the solution ansatz $x_0(t) = \mathrm{e}^{\lambda t} = 1$ into (5.3.24) yields

$$0 = a_0 + a_1 + \cdots + a_d. \tag{5.4.2}$$

Further consider $x_1(t) = t$ as a second, linearly independent solution corresponding the same putative eigenvalue $\lambda = 0$ associated with the same Jordan block. Substituting $x_1(t)$ into the DDE (5.3.24) yields

$$\begin{aligned} 1 &= a_0 t + a_1(t-\tau) + \cdots + a_d(t-d\tau) \\ &= t\,(a_0 + a_1 + \cdots + a_d) - (\tau a_1 + 2\tau a_2 + \cdots + d\tau a_d) \\ \Rightarrow \qquad -1 &= \tau a_1 + 2\tau a_2 + \cdots + d\tau a_d, \end{aligned} \tag{5.4.3}$$

where the coefficients $O(t)$ sum to 0 by (5.4.2). If $d > 2$, we continue by substituting $x_2(t) = t^2$ into (5.3.24) to get

$$\begin{aligned} 2t &= a_0 t^2 + a_1(t-\tau)^2 + a_2(t-2\tau)^2 + \cdots + a_d(t-d\tau)^2 \\ &= t^2\,(a_0 + a_1 + \cdots + a_d) - t\tau\,(a_1 + 2a_2 + \cdots + 2da_d) \\ &\qquad + \tau^2\left(a_1 + 4a_2 + 9a_3 + \cdots + d^2 a_d\right) \\ \Rightarrow \qquad 0 &= a_1 + 4a_2 + 9a_3 + \cdots + d^2 a_d. \end{aligned} \tag{5.4.4}$$

Here the coefficients $O(t^2)$ cancel, again, by (5.4.2) and those of $O(t)$ by (5.4.3). If necessary, we iteratively substitute $x_k(t) = t^k$ into (5.3.24) and find that the coefficients for powers t^k

($k > 1$) cancel by previous substitutions of $x_{k-1}(t), \ldots, x_0(t)$ and only the scalar terms remain. The resulting set of equations can be formulated as a linear system $\mathcal{Q}_{d,0}\mathbf{a} = \mathbf{b}$ with $\mathbf{b} := \begin{bmatrix} 0 & -1 & 0 & \cdots & 0 \end{bmatrix}^\top \in \mathbb{C}^{d+1}$ and $\mathcal{Q}_{d,0} \in \mathbb{C}^{(d+1)\times(d+1)}$ in the following way:

$$\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & \tau & \cdots & d\tau \\ \vdots & & & \vdots \\ 0 & \tau^d & \cdots & d^d\tau^d \end{bmatrix}}_{=:\ \mathcal{Q}_{d,0}} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ \vdots \\ 0 \end{bmatrix}. \tag{5.4.5}$$

We show the solution vector $\mathbf{a}_{\max}$ to (5.4.5) for several numbers of delays in Table 5.1.

d	a_0	a_1	a_2	a_3	a_4	a_5	$\|\mathbf{M}_{64}\|_\infty$
1	1.00	−1.00					4.44
2	1.50	−2.00	0.50				10.28
3	1.83	−3.00	1.50	−0.33			20.11
4	2.08	−4.00	3.00	−1.33	0.25		36.99
5	2.28	−5.00	5.00	−3.33	1.25	−0.20	66.84

Table 5.1: Parameter configurations for maximum size Jordan block in $\mathcal{M}$ for eigenvalue $\mu = 1$ and numbers of delays $d = 1, \ldots, 5$, $\tau = 1$.

Observe how the $d = 1$ line in Table 5.1 corresponds to Section 5.3.4, where $(a, b) = (1, -1)$ is the rightmost point on the $\rho(\mathbf{M}_N) = 1$ line.

We have designed these examples to give large Jordan blocks at $\mu = 1$ on the boundary of the stability region. This procedure gives no guarantee that all the other eigenvalues of $\mathcal{M}$ will be stable. We observe such stability for $d = 1, \ldots, 4$, with eigenvalues of $\mathbf{M}_N$ getting increasingly close to the unit circle as d increases. When $d = 5$ (and for all larger d we have investigated), some of the other eigenvalues fall outside the unit disc, causing the resulting equations to be unstable. We show the example for $d = 5$ in Figure 5.11.

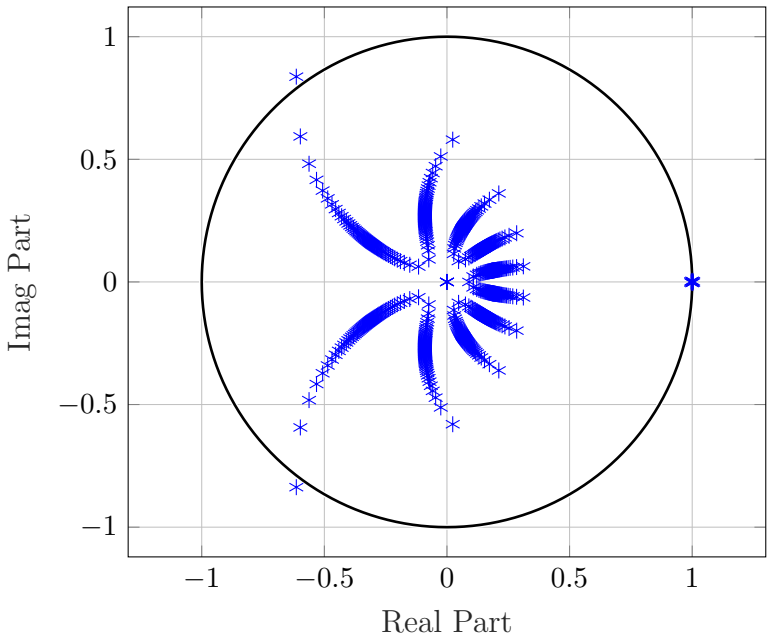


Figure 5.11: Eigenvalues of the discrete solution operator $\mathbf{M}_{128}$ for the maximum transient parameter selection $d = 5$ from Table 5.1

To avoid such instability, we propose two approaches:

- Seek lower order Jordan blocks instead of the maximum possible one, for a given number of delays $d \geq 5$.

- Force a full order Jordan block for $\lambda < 0$ small, instead of $\lambda = 0$.

The following subsections pursue these two approaches.

5.4.2 Lower Order Jordan Blocks

Figure 5.11 indicates that it may be too demanding to force a maximum $(d+1)\times(d+1)$ size Jorden block in $\mathcal{M}$ at $\mu = 1$ ($\lambda = 0$), causing instability in the resulting equation. Instead, we investigate lower order $d \times d$ Jordan blocks that yield a stable solution. To find those, we truncate the system (5.4.5) and only consider the solution ansatz $x_0(t) = 1$, $x_1(t) = t$, up to $x_{d-1}(t) = t^{d-1}$. This leads to the following linear system:

$$\underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 0 & \tau & \cdots & d\tau \\ \vdots & & & \vdots \\ 0 & \tau^{d-1} & \cdots & \tau d^{d-1} \end{bmatrix}}_{=:\mathcal{Q}_{d-1,0}} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ \vdots \\ 0 \end{bmatrix}. \tag{5.4.6}$$

Here $\mathcal{Q}_{d-1,0}$ is full rank of size $d \times (d+1)$. The solution to (5.4.6) is no longer unique. To parameterize the solutions of (5.4.6), let $\mathcal{B}_{d-1,\lambda} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be the singular value decomposition of $\mathcal{Q}_{d-1,0}$. The null space of $\mathcal{Q}_{d-1,0}$ is spanned by the right singular vector corresponding to the zero singular value, $\mathbf{V}\mathbf{e}_{d+1}$, where $\mathbf{e}_{d+1} = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^\top$. Let $\mathbf{a}_0 := \mathcal{Q}_{d-1,0}^{+}\mathbf{b}$, where $\mathcal{Q}_{d-1,0}^{+}$ denotes the pseudo-inverse of $\mathcal{Q}_{d-1,0}$. We parameterize the solutions as follows:

$$\mathbf{a}_{\max}(\gamma) := \mathbf{a}_0 + \gamma \mathbf{V}\mathbf{e}_d, \qquad \gamma \in \mathbb{R}. \tag{5.4.7}$$

With the parametrization of the solutions in (5.4.6), we can find a $\gamma \in \mathbb{R}$ that results in a stable system. By construction, any $\mathbf{a}_{\max}(\gamma)$ yields a $d \times d$ size Jordan block corresponding to $\mu = 1$, the maximum size Jordan block we can expect to be stable. Among all such γ, we aim to find the one that maximizes $\|\mathbf{M}_N\|_{L_\infty}$. As Figure 5.12 shows, many values of $\gamma \in \mathbb{R}$ yield a stable system, except for the eigenvalue $\mu = 1$.

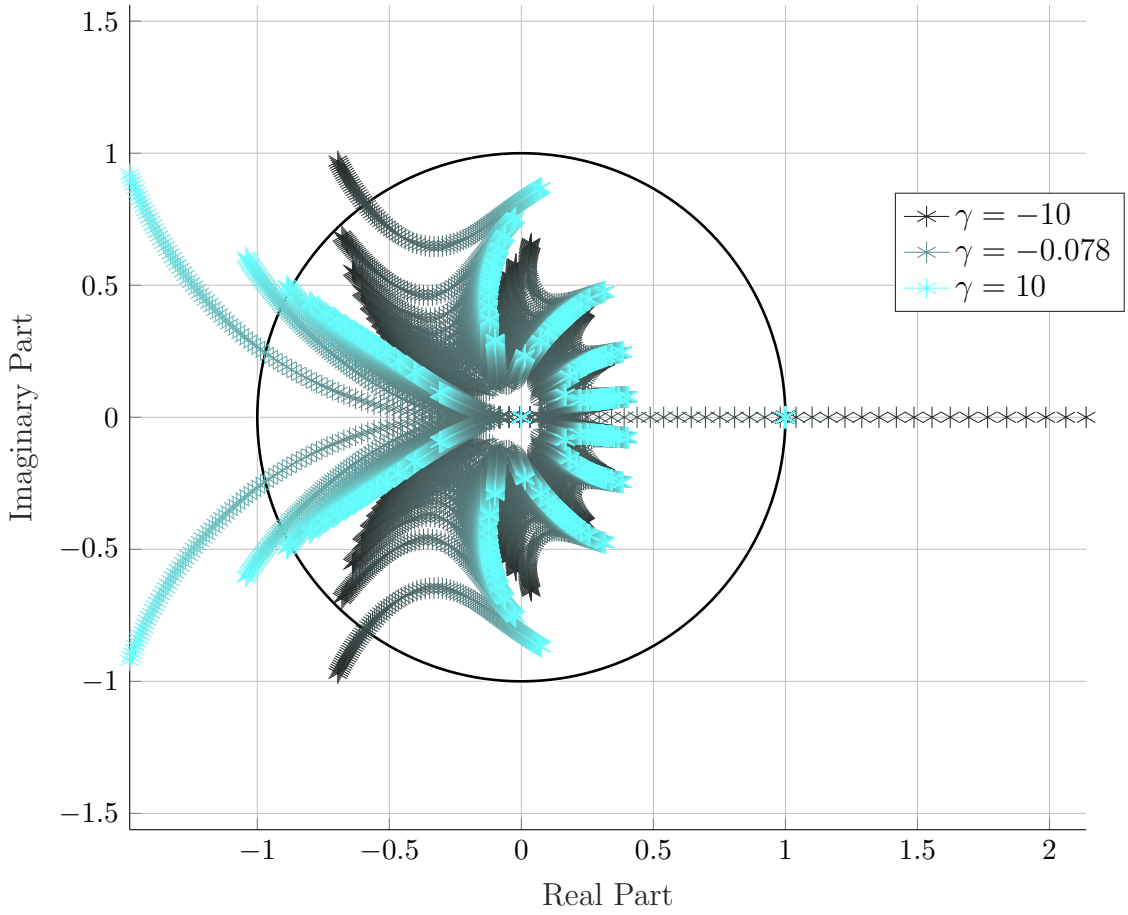


Figure 5.12: Movement of the eigenvalues corresponding to a $d \times d$ Jordan block for $\mu = 1$. Parametrization by γ as in (5.4.7).

To find γ^* so that $\mathbf{M}_N$ with $a(\gamma^*)$ has maximum transient growth, we compare the spectral

radius $\rho(\mathbf{M}_N)$ and $\|\mathbf{M}_N\|_{L_\infty}$, similar to Figure 5.6 for different values of γ in Figure 5.13.

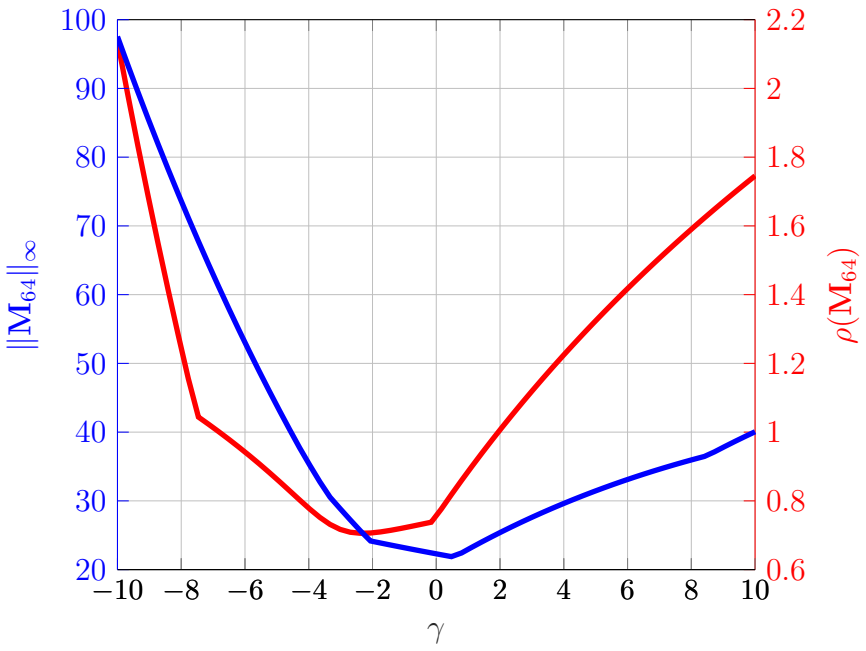


Figure 5.13: Spectral radius $\rho(\mathbf{M}_{128})$ and $\|\mathbf{M}_{128}\|_{L_\infty}$ for varying γ, $d = 5$ delays, Jordan block of size 5×5 corresponding to Figure 5.12 and parametrization (5.4.7).

Since only stable solutions are of interest, the maximum transient behavior appears for $\gamma \approx -6.1$, leading to $\|\mathbf{M}_N\|_{L_\infty} \approx 58.1$. For an (unstable) $(d+1) \times (d+1)$ Jordan block, we find $\|\mathbf{M}_N\|_{L_\infty} \approx 58.2$, suggesting that a lower order Jordan block can yield a similar transient growth as the maximum order Jordan block while maintaining stability. In the next subsection, we investigate maximal $(d+1) \times (d+1)$ Jordan blocks in $\mathbf{M}_N$ for other, stable, choices of λ.

5.4.3 Maximal Order Jordan Blocks for Asymptotically Stable Solutions

In order to construct maximum order Jorden blocks for $\lambda < 0$, we require an extension of the procedure in Section 5.4.1. Let $\lambda < 0$. Substitute the ansatz $\mathrm{e}^{\lambda t}$ into the DDE (5.3.24) and factor out $\mathrm{e}^{\lambda t}$ to obtain

$$\lambda = a_0 + a_1 \mathrm{e}^{-\tau\lambda} + \cdots + a_d \mathrm{e}^{-d\tau\lambda}. \tag{5.4.8}$$

Note that for $\lambda = 0$, we recover (5.4.2). In the case $d > 1$, we further consider $x_1(t) = t\,\mathrm{e}^{\lambda t}$ as a second, linearly independent solution corresponding the same putative eigenvalue $\lambda < 0$ associated with the same Jordan block at $\mu = \mathrm{e}^{\lambda\tau}$ of $\mathcal{M}$. Substituting $x_1(t)$ into (5.3.24) and dividing by the common factor $\mathrm{e}^{\lambda t}$ leads to

$$1 + t\lambda = a_0 t + a_1(t-\tau)\,\mathrm{e}^{-\tau\lambda} + \cdots + a_d(t-d\tau)\,\mathrm{e}^{-d\tau\lambda}. \tag{5.4.9}$$

Equating coefficients for the powers of t, we note that the coefficients for t cancel by (5.4.8). Additional information is contained in the scalars that need to satisfy

$$0 = 1 + \tau a_1 \mathrm{e}^{-\tau\lambda} + \cdots + d\tau a_d \mathrm{e}^{-d\tau\lambda}. \tag{5.4.10}$$

Iteratively substituting $x_k(t) := t^k \, \mathrm{e}^{\lambda t}$ into (5.3.24) ($k = 0, 1, \ldots, d$), and taking cancellations from previous substitutions of $x_{k-1}(t)$ into account, we arrive at the linear system

$$\begin{bmatrix} 1 & \mathrm{e}^{-\tau\lambda} & \cdots & \mathrm{e}^{-d\tau\lambda} \\ 0 & \tau \, \mathrm{e}^{-\tau\lambda} & \cdots & d\tau \, \mathrm{e}^{-d\tau\lambda} \\ \vdots & & & \vdots \\ 0 & \tau^d \, \mathrm{e}^{-\tau\lambda} & \cdots & d^d \tau^d \, \mathrm{e}^{-d\tau\lambda} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} \lambda \\ -1 \\ \vdots \\ 0 \end{bmatrix}. \tag{5.4.11}$$

Note that for $\lambda = 0$, (5.4.11) reduces to (5.4.6).

By solving (5.4.11), we now are able to find parameter configurations $\mathbf{a}_{\max}$ that yield maximum transient growth in $\mathbf{M}_N$, corresponding to eigenvalues $\mu = \mathrm{e}^{\lambda\tau}$. Note that $|\mu| < 1$ if $\Re\lambda < 0$, so μ is within the stability region of $\mathcal{M}$ and $\mathbf{M}_N$, respectively. We expect the resulting parameter configuration to result in a stable system if λ is not chosen too close to 0.

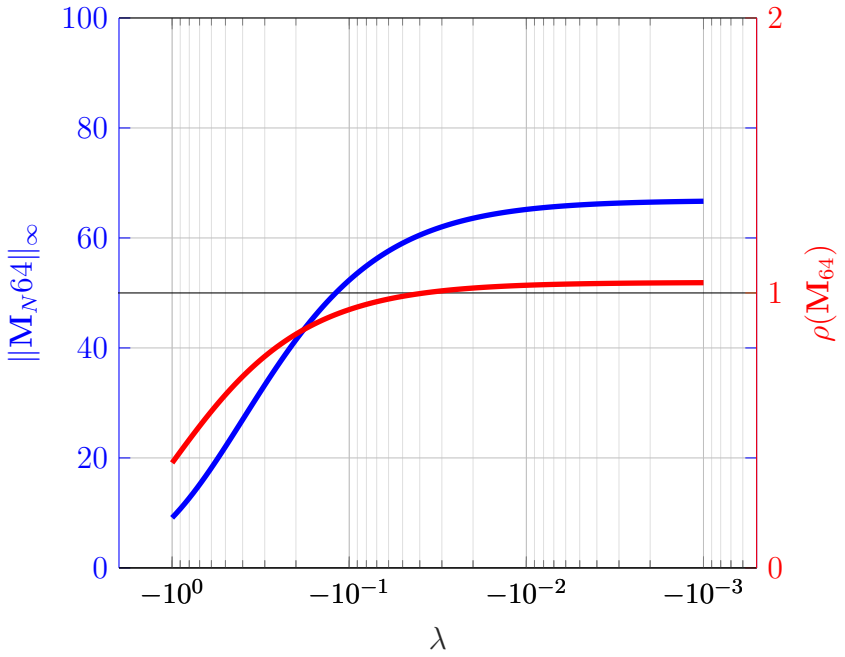
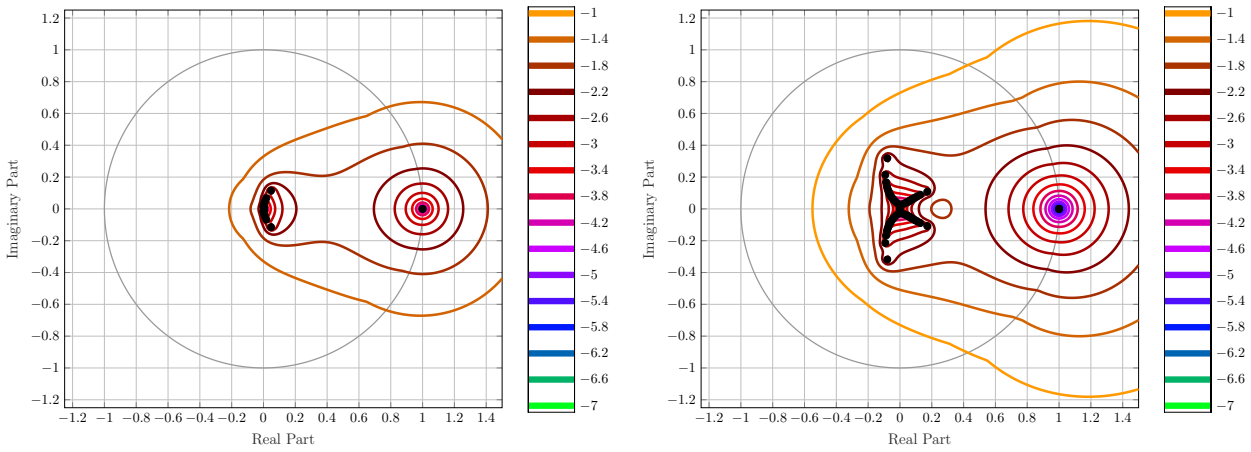


Figure 5.14: Comparison $\|\mathbf{M}_N^k\|_{L_\infty}$ and $\rho(\mathbf{M}_N)$ for various $\lambda < 0$.

Similar to Figure 5.13, in Figure 5.14, we compare $\|\mathbf{M}_N\|_{L_\infty}$ and $\rho(\mathbf{M}_N)$ for values $\lambda < 0$. The maximum value of $\|\mathbf{M}_N\|_{L_\infty}$ is ≈ 60.34 with $\rho(\mathbf{M}_N) \leq 1$ for $\lambda_* \approx -0.0411$.

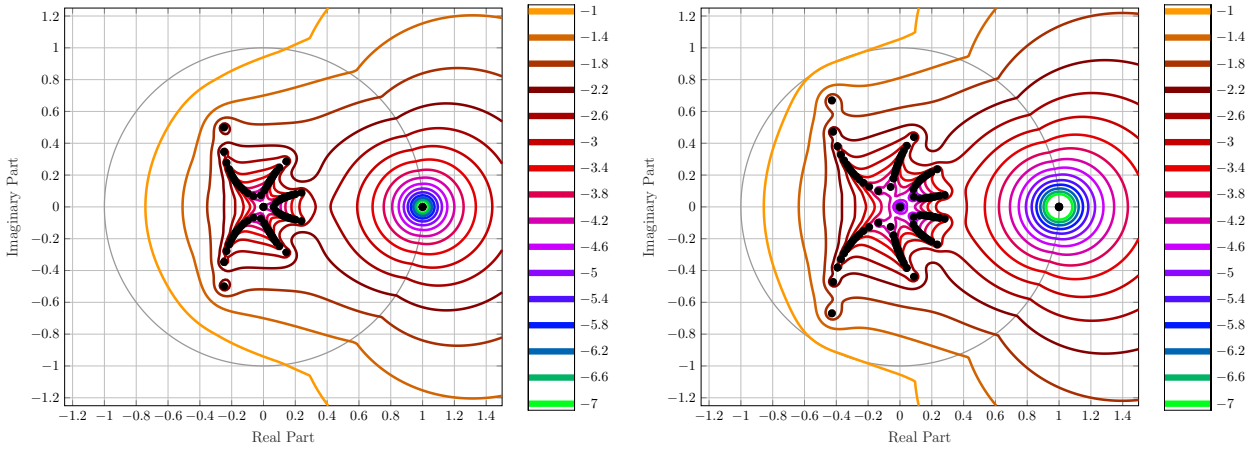
In Figure 5.15, we show the level set lines of the pseudospectra for the cases of $d = 1$ and $d = 2$ delays.



(a) $d = 1$ (b) $d = 2$

Figure 5.15: Pseudospectral plots for the $\|\cdot\|_{L_\infty}$ norm pseudospectra of the maximum Jordan block at $\lambda = -0.0001$.

One can see how the pseudospectral level sets reach outside the unit disc for even large values of $\epsilon > 0$, indicating a large Kreiss constant $\mathcal{K}(\mathbf{M}_N)$. With more delay terms available, Figure 5.16 shows that the pseudospectra get even larger with the number of delay terms d.



(a) $d = 3$ (b) $d = 4$

Figure 5.16: Pseudospectral plots for the $\|\cdot\|_{L_\infty}$ norm pseudospectra of the maximum Jordan block at $\lambda = -0.0001$.

Also note that all the systems presented in these figures are stable.

The Kreiss constant is a more precise estimator of the maximum transient growth. In Figure 5.17, we compare $\mathcal{K}(\mathbf{M}_N)$ for various $\lambda < 0$ and $\mathbf{a}_{\max}$ solving (5.4.11).

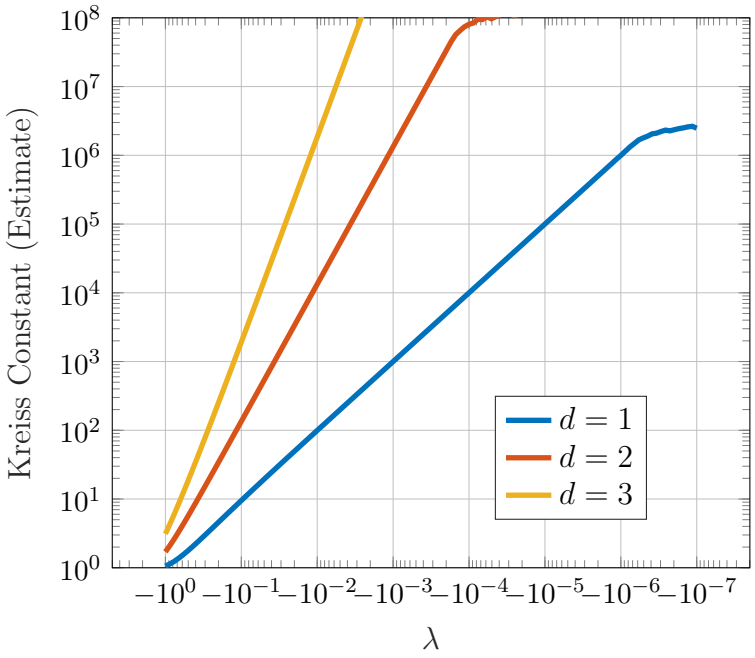
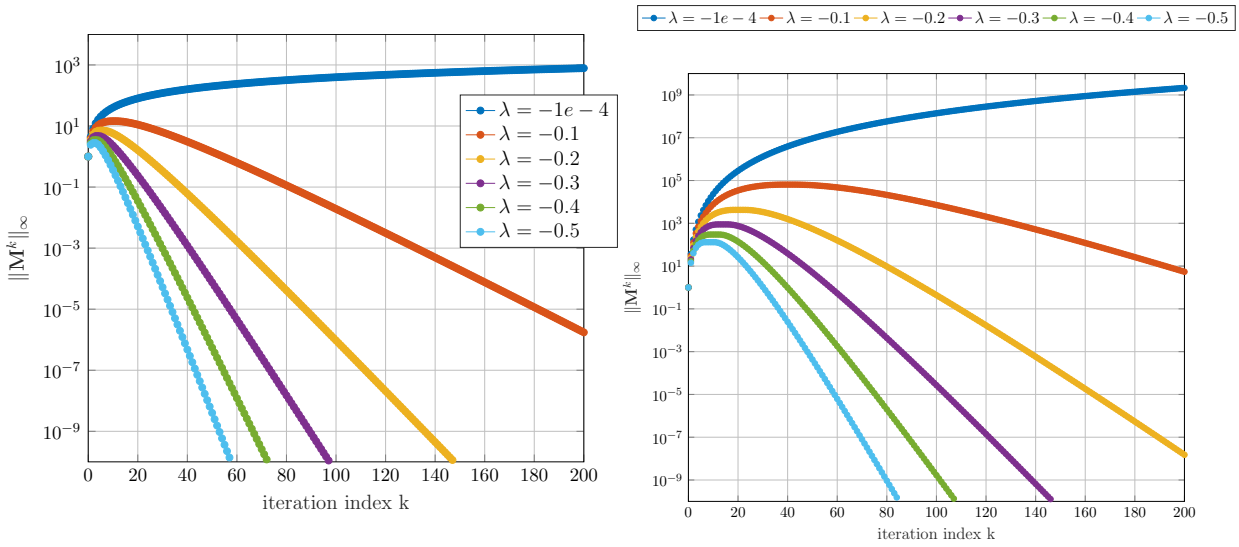


Figure 5.17: Numerical estimate of the Kreiss constant $\mathcal{K}(\mathbf{M}_N)$ for various numbers of delays. We consider parameter configurations $\mathbf{a}_{\max}$ for maximal Jordan blocks at $\lambda < 0$.

In Figure 5.17, we show conservative estimates of the Kreiss constant, computed by a sampling approach of $\mathbb{C}\backslash\overline{\mathbb{D}}$ to estimate $\mathcal{K}(\mathbf{M}_N)$ instead of a line search algorithm [60]. We see how the Kreiss constant increases at larger rates for higher number of delays.

The $\max_{t\geq 0}|x(t)|$ need not be attained on $[0,\tau]$, so we compare $\|\mathbf{M}_N^k\|_{L_\infty}$, $k = 0, 1, \ldots$. Figure 5.18 shows $\|\mathbf{M}_{32}^k\|_{L_\infty}$ for the discretized solution operator and coefficients $\mathbf{a}_{\max}$ solving (5.4.11).



(a) $d = 1$

(b) $d = 4$

Figure 5.18: Norm $\|\mathbf{M}_{32}^k\|_{L_\infty}$ for different numbers of delays d and selections of eigenvalues $\lambda < 0$.

We observe that the $d = 4$ delay case in Figure 5.18 yields more transient growth with increasing matrix powers then the $d = 1$ case.

5.5 Higher Dimensional State Space

We have focused on scalar systems ($n = 1$), since transient growth for $n = 1$ contrasts the standard ODE case, where no transient growth is possible. Unsurprisingly, delay equations with $n > 1$ exhibit transient behavior as well. In this section, we address the generalization of $\mathbf{M}_N$ to $n > 1$. We start by introducing the corresponding nonlinear eigenvalue problem.

Consider the DDE

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{x}(t-\tau), \qquad \mathbf{A}, \mathbf{B} \in \mathbb{C}^{n\times n}, \\ \mathbf{x}(t) &= \boldsymbol{\phi}(t+\tau), \qquad t \in [-\tau, 0], \end{aligned} \tag{5.5.1}$$

with history function $\boldsymbol{\phi} : [0, \tau] \to \mathbb{C}^n$, $n \geq 1$. In this case, the NLEVP (5.0.1) becomes

$$\mathbf{F}(\lambda) := \det\left(\lambda\mathbf{I} - \mathbf{A} - \mathbf{B}\,\mathrm{e}^{-\lambda\tau}\right) = 0. \tag{5.5.2}$$

If the matrices $\mathbf{A}$ and $\mathbf{B}$ are simultaneously diagonalizable, i.e., there exists $\mathbf{S} \in \mathbb{C}^{n\times n}$ of full rank and diagonal matrices $\boldsymbol{\Sigma}, \boldsymbol{\Gamma} \in \mathbb{C}^{n\times n}$ so that

$$\mathbf{A} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^{-1} \qquad \text{and} \qquad \mathbf{B} = \mathbf{S}\boldsymbol{\Gamma}\mathbf{S}^{-1}, \tag{5.5.3}$$

then premultiplying (5.5.1) by $\mathbf{S}^{-1}$ yields

$$\mathbf{S}^{-1}\dot{\mathbf{x}}(t) = \boldsymbol{\Sigma}\mathbf{S}^{-1}\mathbf{x}(t) + \boldsymbol{\Gamma}\mathbf{S}^{-1}\mathbf{x}(t-\tau). \tag{5.5.4}$$

Changing variables to $\mathbf{y}(t) := \mathbf{S}^{-1}\mathbf{x}(t)$ then gives

$$\dot{\mathbf{y}}(t) = \boldsymbol{\Sigma}\mathbf{y}(t) + \boldsymbol{\Gamma}\mathbf{y}(t-\tau). \tag{5.5.5}$$

The corresponding characteristic equations $\mathbf{F}(\lambda) := \det\left(\lambda\mathbf{I} - \boldsymbol{\Sigma} - \boldsymbol{\Gamma}\,\mathrm{e}^{-\lambda\tau}\right) = 0$ has the same solutions as (5.5.2). Since the delay system in $\mathbf{y}(t)$ has diagonal state space matrices, the equations fully decouple and can be solved using the Lambert W function for each diagonal entry, as in the scalar case:

$$\mathbf{F}(\lambda) = 0 \qquad \Leftrightarrow \qquad \lambda \in \bigcup_{j=1,\ldots,n} \left\{ \frac{1}{\tau} W_k(\tau\boldsymbol{\Gamma}_j\,\mathrm{e}^{-\boldsymbol{\Sigma}_j\tau}) + \boldsymbol{\Sigma}_j : \ k \in \mathbb{Z} \right\}, \tag{5.5.6}$$

where $\mathbf{\Gamma}_j$ and $\mathbf{\Sigma}_j$ referr to the jth diagonal entries of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$, respectively. Alternatively, the *matrix valued* Lambert W function directly provides the solution of (5.5.2):

$$\mathbf{F}(\lambda) = 0 \qquad \Leftrightarrow \qquad \lambda \in \bigcup_{k\in\mathbb{Z}} \left\{ \frac{1}{\tau} W_k \left(\tau \mathbf{B}\, \mathrm{e}^{-\mathbf{A}\tau} \right) + \mathbf{A} \right\}. \tag{5.5.7}$$

without requiring the simultaneous diagonizability of $\mathbf{A}$ and $\mathbf{B}$. We extend Section 5.3.1 to the $n > 1$ case. The solution of (5.5.1) can be expressed using the variation of parameters formula for $t \in [0, \tau]$:

$$\mathbf{x}(t) = \mathrm{e}^{t\mathbf{A}}\mathbf{x}(0) + \int_0^t \mathrm{e}^{\mathbf{A}(t-s)}\mathbf{B}\boldsymbol{\phi}(s)\,\mathrm{d}s. \tag{5.5.8}$$

Note that for $n > 1$, we have to be mindful of the order of multiplication since $\mathbf{AB} \neq \mathbf{BA}$ in general. We define the operator $\boldsymbol{\mathcal{M}} : C\left([0,\tau], \mathbb{C}^n\right) \to C\left([0,\tau], \mathbb{C}^n\right)$ as

$$(\boldsymbol{\mathcal{M}}\mathbf{x})(t) := \mathrm{e}^{t\mathbf{A}}\mathbf{x}(\tau) + \int_0^t \mathrm{e}^{\mathbf{A}(t-s)}\mathbf{B}\mathbf{x}(s)\,\mathrm{d}s. \tag{5.5.9}$$

As for the scalar case, we represent the initial history function $\boldsymbol{\phi}$ and the solution $\mathbf{x}(t)$ by Lagrange interpolation at Chebyshev points in every dimension:

$$[\boldsymbol{\phi}(t)]_j \approx \sum_{k=0}^{N} \mathbf{u}^{(0)}_{j,k} \ell_k(t), \qquad [\mathbf{x}(t)]_j \approx \sum_{k=0}^{N} \mathbf{u}^{(1)}_{j,k} \ell_k(t), \quad j = 1, \ldots, n. \tag{5.5.10}$$

Note that $\mathbf{u}^{(\nu)} \in \mathbb{C}^{n\times(N+1)}$, $\nu = 0, 1$. The solution formula (5.5.8) leads to

$$\begin{aligned}
\mathbf{x}(t_j) &\approx \mathrm{e}^{t_j\mathbf{A}}\mathbf{x}(\tau) + \int_0^{t_j} \mathrm{e}^{(t_j-s)\mathbf{A}}\mathbf{B}\boldsymbol{\phi}(s)\,\mathrm{d}s \\
&\approx \mathrm{e}^{t_j\mathbf{A}}\mathbf{u}^{(0)}_{:,0} + \mathrm{e}^{t_j\mathbf{A}} \int_0^{t_j} \mathrm{e}^{-s\mathbf{A}}\mathbf{B} \sum_{k=0}^{N} \mathbf{u}^{(0)}_{:,k} \ell_k(s)\,\mathrm{d}s \\
&= \mathrm{e}^{t_j\mathbf{A}}\mathbf{u}^{(0)}_{:,0} + \sum_{k=0}^{N} \mathrm{e}^{t_j\mathbf{A}} \int_0^{t_j} \mathrm{e}^{-s\mathbf{A}}\mathbf{B}\ell_k(s)\,\mathrm{d}s\ \mathbf{u}^{(0)}_{:,k} \\
&= \mathrm{e}^{t_j\mathbf{A}}\mathbf{u}^{(0)}_{:,0} + \sum_{k=0}^{N} \mathbf{w}_{j,k}\mathbf{u}^{(0)}_{:,k},
\end{aligned} \tag{5.5.11}$$

where we define

$$\mathbf{w}_{j,k} := \delta_{k,0}\, \mathrm{e}^{t_j \mathbf{A}} + \mathrm{e}^{t_j \mathbf{A}} \int_0^{t_j} \mathrm{e}^{-s\mathbf{A}} \mathbf{B} \ell_k(s)\, \mathrm{d}s \in \mathbb{C}^{n \times n}. \tag{5.5.12}$$

For practical purposes and implementation, we vectorize $\mathbf{w}_{j,k}$. More precisely, we vectorize the coefficients $\mathbf{u}_k^{(\nu)}$ $(\nu = 0, 1)$ from (5.5.10) in the following way

$$\mathbf{U}_{\phi,N} := \begin{bmatrix} \begin{bmatrix} \mathbf{u}_{:,0}^{(0)} \end{bmatrix}^\top & \begin{bmatrix} \mathbf{u}_{:,1}^{(0)} \end{bmatrix}^\top & \cdots & \begin{bmatrix} \mathbf{u}_{:,N}^{(0)} \end{bmatrix}^\top \end{bmatrix}^\top \in \mathbb{C}^{n(N+1)} \tag{5.5.13}$$

Then $\mathbf{U}_{\mathbf{x},N}$ corresponds to the same vectorization of the entries in $\mathbf{x}$, where we add the N to emphasize that $\mathbf{U}_{\mathbf{x},N}$ depends on the discretization of $[0, \tau]$ in (5.5.10).

We then can rewrite (5.5.11) as a matrix-vector product.

$$\mathbf{U}_{\mathbf{x}} = \begin{bmatrix} \mathbf{x}_1(t_0) \\ \mathbf{x}_2(t_0) \\ \vdots \\ \mathbf{x}_n(t_N) \end{bmatrix} \approx \underbrace{\begin{bmatrix} \mathbf{w}_{0,1} & \mathbf{w}_{0,2} & \cdots & \mathbf{w}_{0,n} \\ \mathbf{w}_{1,1} & & & \\ \vdots & & \vdots & \\ \mathbf{w}_{N,1} & \cdots & & \mathbf{w}_{N,n} \end{bmatrix}}_{=:\, \mathfrak{W}} \begin{bmatrix} \mathbf{u}_{:,0}^{(0)} \\ \mathbf{u}_{:,1}^{(0)} \\ \vdots \\ \mathbf{u}_{:,N}^{(0)} \end{bmatrix} =: \mathfrak{W} \mathbf{U}_\phi \tag{5.5.14}$$

Note that all the entries of $\mathfrak{W} \in \mathbb{C}^{n(N+1) \times n(N+1)}$ are precomputed and stored for fast simulation of the DDE for any given input. We remark that when using (5.5.14) to simulate a system, we have to be careful about the ordering of the variables.

For $n > 1$ the choice of norms becomes more delicate. Both the entries of the vector $\mathbf{x}(t)$ as well as the discretization order enters into the picture. The notion of combined norms

applies here as follows. Given p and q, we define $|||\cdot|||_{p,q}$ by:

$$|||\mathbf{U}_{\mathbf{x}}|||_{p,q} := \left\| \begin{bmatrix} \|\mathbf{x}_1(t)\|_p \\ \vdots \\ \|\mathbf{x}_n(t)\|_p \end{bmatrix} \right\|_q \tag{5.5.15}$$

For transient behavior, the $(\infty - \infty)$ norm appears to be both insightful and computable, since $|||\mathbf{U}_{\mathbf{x},N}|||_{\infty,\infty} = \|\mathbf{U}_{\mathbf{x},N}\|_{L_\infty}$.

5.6 Summary of Contributions and Future Work

In this section, we presented a method to construct coefficients for delay differential equations, in particular scalar ones, that yield arbitrarily large transient growth.

We focus on the scalar case, since the distinction to the non-delay case is most prominent.

Unlike other approaches that are based on the nonlinear eigenvalue problem, we chose the discrete-time solution operator as a basis for our investigation. For our analysis, we use a particular discretization at Chebyshev points that yields fast convergence of the eigenvalues. With moderate discretization orders, we can assess stability and transient behavior of the system.

We derive a system of equations that characterizes coefficients $\mathbf{a}_{\max}$ that yield strong transient growth. To ensure stability of the solution corresponding to the constructed coefficients $\mathbf{a}_{\max}$, we adjust the size of Jordan blocks at the maximum eigenvalue μ or, if necessary, consider eigenvalues $|\mu| < 1$ inside the stability region.

A possible next step is the application to nonlinear delay equations, delay algebraic equations and neutral delay equations.

Chapter 6

Conclusions

We investigated parametric modeling methods for data driven and measurement driven settings. Model reduction via optimal rational interpolation was extended to the parametric setting. For parametric interpolatory model reduction, the choice of function spaces and norms was analyzed. In particular an extension of classic Hardy spaces to the two variable case was used as a measure for approximating parametric dynamical systems. We chose a particular form of the reduced model with separable poles in frequency and parameter, which span a dense subspace in the two-variable Hardy space under consideration. The main results in Chapter 3 established optimality conditions that characterize (locally) optimal reduced models in a tensor norm in frequency and parameter. We showed the approximation quality on synthetic and realistic models, including a model that does not admit a separable pole structure in frequency and parameter. Our gradient descent algorithm converged to a joint set of interpolation points.

While our implementations perform well on the examples presented here, there are several possible extensions of our work and new research directions. Since many physical applications require a real parameter, for example $\mathcal{P} = [-1, 1]$, we want to focus the approximation quality of our algorithm on such a region of interest. This leads to a weighted optimality measure in frequency and parameter, which is subject to future work.

We used a separable parameterization of the poles in p to establish the theory for optimality conditions in frequency and parameter. Physical models may not admit such a separable parametrization. Hence we plan to investigate other parametrizations of the poles in p that are non-separable, allowing the poles of the optimal reduced model to move with the frequency variable.

Measurement based model reduction has been extended in Chapter 4 using a least-squares approach in frequency and parameter, based on the Vector Fitting algorithm. To that end, we solve a two-variable least-squares problem, directly extending the single variable case. Our approach is based on local approximations, computed at fixed parameter values with classic Vector Fitting. The final model is constructed using basis functions in p with coefficients solving the two-variable least-squares problem. We investigated polynomial and rational basis functions to combine local models. For rational functions in particular, our implementation uses the variable projection method, which takes advantage of the problem structure. This allowed us to adaptively choose poles of the rational functions, based on the given measurements. In comparison to other interpolatory approaches, our method exhibits resistance to measurement outliers while remaining feasible from a computational

perspective.

Depending on the number of measurements and orders of the local models, parametric Vector Fitting may result in a moderate size parametric approximation. To further reduce the complexity of the parametric Vector Fitting result, we make use of optimal rational interpolation, implemented by IRKA for a special parametric dependence.

In Chapter 5, we presented a particular class of parametric dynamical systems where the parameter enters as a delay. We sought parameter configurations that yield stable systems with significant transient growth, focusing on the scalar case, since for higher dimensional state spaces, the non-delay case already exhibits transient behavior. To analyze stability and transient behavior, we implemented a spectral discretization of the solution operator that enables the use of classical tools for spectral and pseudospectral analysis. This allows us insight into possible transient behavior and the parameter configurations associated with it. Moreover, we have presented a method to construct parameter configurations for delay equations with several delays that correspond to maximum transient growth in the solution.

A further extension to the nonlinear case is of interest, particularly for control theory. This would allow us to access how robust a controller needs to be designed to overcome possible transient behavior of the solution to stabilize a delay dynamical system around an equilibrium point.

Bibliography

[1] B. Anic, C. A. Beattie, S. Gugercin, and A. C. Antoulas. "Interpolatory Weighted-H2 Model Reduction". In: *Automatica J. IFAC* 49.2 (2012), pp. 1–7. DOI: 10.1016/j.automatica.2013.01.040 (cited on page 99).

[2] G Antonini, D. Deschrijver and T. Dhaene. "A Comparative Study of Vector Fitting and Orthonormal Vector Fitting Techniques for EMC Applications". In: *International Symposium on Electromagnetic Compatibility*. 2006, pp. 6–11. ISBN: 142440293X (cited on page 39).

[3] A. C. Antoulas. *Approximation of large-scale dynamical systems*. SIAM, 2005 (cited on pages 2, 3, 21, 23, 26, 27, 65).

[4] A. C. Antoulas and D. C. Sorensen. "Approximation of large-scale dynamical systems: An overview". In: *Int. J. Appl. Math. Comput. Sci.* 11.5 (2001), pp. 1093–1121 (cited on page 17).

[5] C. Balazs and J. Szabados. "Approximation by Bernstein Type Rational Functions. II". In: *Acta Math. Acad. Sci. hungar.* 40 (1982), pp. 331–337 (cited on page 109).

[6] M. Bando, K. Hasebe, K. Nakanishi, and A. Nakayama. "Analysis of optimal velocity model with explicit delay". In: *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 58.5 (1998), pp. 5429–5435. ISSN: 1063651X. DOI: 10.1103/PhysRevE.58.5429. arXiv: 9805002 [patt-sol] (cited on page 142).

[7] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. "An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations". In: *Comptes Rendus Mathematique* 339.9 (2004), pp. 667–672 (cited on page 42).

[8] R. H. Bartels and G. W. Stewart. "Algorithm 432 Solution of the Matrix Equation AX + XB = C [F4]". In: *Communications of the ACM* 15.9 (1972), pp. 820–826 (cited on page 67).

[9] A. Bátkai and S. Piazzera. "Semigroups and linear partial differential equations with delay". In: *Journal of Mathematical Analysis and Applications* 264 (2001), pp. 1–20. ISSN: 0022247X. DOI: 10.1006/jmaa.2001.6705. arXiv: 1211.7197 (cited on page 152).

[10] U Baur and P Benner. "Modellreduktion für parametrisierte Systeme durch balanciertes Abschneiden und Interpolation (Model Reduction for Parametric Systems Using Balanced Truncation and Interpolation)". In: *at-Automatisierungstechnik* 57.8 (2009), pp. 411–420 (cited on page 104).

[11] U. Baur, C. A. Beattie, P. Benner, and S. Gugercin. "Interpolatory Projection Methods for Parameterized Model Reduction". In: *SIAM J. Sci. Comput.* 33.5 (2011),

pp. 2489–2518. ISSN: 1064-8275. DOI: 10.1137/090776925 (cited on pages 10, 44, 131, 133).

[12] U. Baur and P. Benner. "Modellreduktion für parametrisierte Systeme durch balanciertes Abschneiden und Interpolation". In: *Automatisierungstechnik* 57.8 (2009), pp. 411–419. ISSN: 01782312. DOI: 10.1524/auto.2009.0787 (cited on page 50).

[13] C. Beattie and S. Gugercin. "A trust region method for optimal H_2 model reduction". In: *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference* (2009), pp. 5370–5375. ISSN: 0191-2216. DOI: 10.1109/CDC.2009.5400605 (cited on page 31).

[14] C. Beattie and S. Gugercin. "Realization-independent H2-approximation". In: *Proceedings of the 51st IEEE Conference on Decision and Control*. Vol. 1. Citeseer, 2012 (cited on pages 4, 33, 82).

[15] C. Beattie, S. Gugercin, and V. Mehrmann. "Model reduction for systems with inhomogeneous initial conditions". In: *Systems and Control Letters* 99 (2017), pp. 99–106. ISSN: 0167-6911. DOI: 10.1016/j.sysconle.2016.11.007 (cited on page 17).

[16] A. Bellen and M. Zennaro. *Numerical methods for delay differential equations*. Oxford university press, 2013 (cited on page 140).

[17] P. Benner and L. Feng. "A Robust Algorithm for Parametric Model Order Reduction Based on Implicit Moment Matching". In: *Reduced Order Methods for Mod-*

eling and Computaional Reduction. July 2016. Springer Switzerland, 2014. ISBN: 9783319020907. DOI: 10.1007/978-3-319-02090-7 (cited on page 44).

[18] P. Benner, S. Gugercin, and K. Willcox. "A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems". In: *SIAM Review* 57.4 (2015), pp. 483–531. ISSN: 0036-1445. DOI: 10.1137/130932715 (cited on pages 6, 41, 43).

[19] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox. *Model Reduction and Approximation: Theory and Algorithms*. 2017 (cited on page 2).

[20] P. Benner, E. W. Sachs, and S. Volkwein. "Model Order Reduction for PDE Constrained Optimization". In: *Trends in PDE Constrained Optimization* (2014), pp. 303–326 (cited on page 6).

[21] G. Berkooz, P. Holmes, and J. L. Lumley. "The proper orthogonal decomposition in the analysis of turbulent flows". In: *Annual review of fluid mechanics* 25.1 (1993), pp. 539–575 (cited on page 3).

[22] J.-P. Berrut and L. N. Trefethen. "Barycentric Lagrange Interpolation". In: *SIAM Review* 46.3 (2004), pp. 501–517. ISSN: 0036-1445. DOI: 10.1137/S0036144502417715 (cited on page 36).

[23] A. Beygi and A. Dounavis. "An instrumental variable vector-fitting approach for noisy frequency responses". In: *IEEE Transactions on Microwave Theory and Techniques* 60.9 (2012), pp. 2702–2712. ISSN: 00189480. DOI: 10.1109/TMTT.2012.2206399 (cited on page 39).

[24] B. Bond and L. Daniel. "Parameterized Model Order Reduction of Nonlinear Dynamical Systems". In: *Proc. IEEE/ACM Conference on Computer-Aided Design, ICCAD-2005*. 2005, pp. 487–494. DOI: 10.1109/ICCAD.2005.1560117 (cited on page 44).

[25] J. P. Boyd. "Exponentially convergent Fourier-Chebshev quadrature schemes on bounded and infinite intervals". In: *Journal of Scientific Computing* 2.2 (1987), pp. 99–109. ISSN: 08857474. DOI: 10.1007/BF01061480 (cited on page 40).

[26] E. Bueler. "Chebyshev collocation for linear, periodic ordinary and delay differential equations: a posteriori estimates". In: *arXiv preprint* 0114500 (2004). arXiv: 0409464 [math] (cited on pages 155, 157).

[27] Y. Chahlaoui and P. V. Dooren. "A collection of Benchmark examples for model reduction of linear time invariant dynamical systems". In: *SLICOT Working Notes* (2002), pp. 1–28. DOI: 10.1007/3-540-27909-1_24 (cited on page 8).

[28] S. Chaturantabut and D. C. Sorensen. "Nonlinear model reduction via discrete empirical interpolation". In: *SIAM Journal on Scientific Computing* 32.5 (2010), pp. 2737–2764. ISSN: 1064-8275. DOI: 10.1137/090766498 (cited on page 42).

[29] S.-T. Chen, S.-P. Hsu, H.-N. Huang, and B.-Y. Yang. "Time response of a scalar dynamical system with multiple delays via Lambert W functions". In: *arXiv preprint* (2016). arXiv: 1609.02034 (cited on pages 167, 168).

[30] A. Chinea and S. Grivet-Talocia. "On the parallelization of vector fitting algorithms". In: *IEEE Transactions on Components, Packaging and Manufacturing Technology*

1.11 (2011), pp. 1761–1773. ISSN: 21563950. DOI: 10.1109/TCPMT.2011.2167973 (cited on page 39).

[31] J. Chung and J. G. Nagy. "An Efficient Iterative Approach for Large-Scale Separable Nonlinear Inverse Problems". In: *SIAM Journal on Scientific Computing* 31.6 (2010), pp. 4654–4674. ISSN: 1064-8275. DOI: 10.1137/080732213 (cited on page 115).

[32] R. R. Coifman, R. Rochberg, and G. Weiss. "Factorization theorems for Hardy spaces in several variables". In: *Annals of Mathematics* 103.3 (1976), pp. 611–635 (cited on page 51).

[33] R. F. Curtain and H. J. Zwart. *Introductino to Infinite-Dimensional Linear System Theory*. Springer-Verlag New York, Inc., 1995, p. 698. ISBN: 0-387-94475-3 (cited on pages 144, 153).

[34] D. Deschrijver, L. Knockaert, and T. Dhaene. "Improving the Robustness of Vector Fitting to Outliers in the Data". In: *Electronics letters* 46.17 (2010), pp. 1200–1201 (cited on page 39).

[35] D. Deschrijver and T. Dhaene. "A note on the multiplicity of poles in the vector fitting macromodeling method". In: *IEEE Transactions on Microwave Theory and Techniques* 55.4 (2007), pp. 736–741. ISSN: 00189480. DOI: 10.1109/TMTT.2007.893651 (cited on page 39).

[36] D. Deschrijver, B. Haegeman, and T. Dhaene. "Orthonormal Vector Fitting : A Robust Macromodeling Tool for Rational Approximation of Frequency Domain Re-

sponses". In: *IEEE Transactions on Advanced Packaging* 30.2 (2007), pp. 216–225 (cited on page 39).

[37] O. Diekmann, S. A. van Gils, S. M. V. Lunel, and H.-O. Walther. *Delay Equations: Functional- Complex-, and Nonlinear Analysis*. Vol. 33. 2012, pp. 3–8. ISBN: 9781604138795. DOI: 10.1073/pnas.0703993104 (cited on page 158).

[38] R. Douglas, K. Davidson, M. Putinar, and J. Eschmeier. "Multivariate Operator Theory". In: *Citeseer* (2010), pp. 1–13 (cited on page 51).

[39] Z. Drmač and S. Gugercin. "A New Selection Operator for the Discrete Empirical Interpolation Method – improved a priori error bound and extensions". In: *SIAM Journal on Scientific Computing* 38.2 (2015), A631–A648. ISSN: 10957200. DOI: 10.1137/15M1019271 (cited on page 42).

[40] Z. Drmač, S. Gugercin, and C. Beattie. "Quadrature-based vector fitting for discretized H2 approximation". In: *SIAM J. Sci. Comput.* 37.2 (2015), pp. 625–652 (cited on pages 40, 139).

[41] Elmar Plischke. "Transient Effects of Linear Dynamical Systems". PhD thesis. 2005 (cited on page 151).

[42] M. Embree and L. N. Trefethen. "Generalizing Eigenvalue Theorems to Pseudospectra Theorems". In: 23.2 (2000), pp. 583–590. ISSN: 1064-8275. DOI: 10.1137/S1064827500373012 (cited on pages 141, 142).

[43] K.-J. Engel and R. Nagel. *One-parameter semigroups for linear evolution equations*. Vol. 63. 2. Springer, 2001, pp. 278–280. ISBN: 0-387-98463-1. DOI: 10.1007/s002330010042 (cited on pages 144, 152).

[44] R. T. Farouki, T. Goodman, and T. Sauer. "Construction of orthogonal bases for polynomials in Bernstein form on triangular and simplex domains". In: *Computer Aided Geometric Design* 20.4 (2003), pp. 209–230. ISSN: 01678396. DOI: 10.1016/S0167-8396(03)00025-6 (cited on page 109).

[45] F. Ferranti, L. Knockaert, and T. Dhaene. "Guaranteed Passive Parameterized Admittance-Based Macromodeling". In: *IEEE Transactions on Advanced Packaging* 33.3 (2010), pp. 623–629 (cited on page 128).

[46] G. Flagg, C. Beattie, and S. Gugercin. "Convergence of the iterative rational Krylov algorithm". In: *Systems and Control Letters* 61.6 (2012), pp. 688–691. ISSN: 01676911. DOI: 10.1016/j.sysconle.2012.03.005. arXiv: arXiv:1107.5363v1 (cited on page 28).

[47] P. Fuhrmann. *A polynomial approach to linear algebra*. 2011. ISBN: 9781461403371 (cited on page 20).

[48] T. W. Gamelin. *Complex Analysis*. Spinger, 2001. ISBN: 6811008466 (cited on page 54).

[49] J. B. Garnett and R. H. Latter. "The atomic decomposition for Hardy spaces in several complex variables". In: *Duke Math. J* 45 (1978), pp. 815–845 (cited on page 51).

[50] M. Geuss, H. Panzer, and B. Lohmann. "On Parametric Model Order Reduction by Matrix Interpolation". In: *European Control Conference (ECC), July 17-19* 58 (2013), pp. 3433–3438. ISSN: 0178-2312. DOI: 10.1524/auto.2010.0863 (cited on page 50).

[51] G. Golub. "Some Modified Matrix Eigenvalue Problems". In: *SIAM Review* 15.2 (1973), pp. 318–334. ISSN: 00361445 (cited on page 38).

[52] G. Golub and V. Pereyra. "Separable Nonlinear Least Squares : the Variable Projection Method and its Applications". In: *Inverse problems* 19.2 (2003), R1 (cited on page 114).

[53] A. Gombani and G. Michaletzky. "On the Nevanlinna-Pick interpolation problem: Analysis of the McMillan degree of the solutions". In: *Linear Algebra and Its Applications* 425.2-3 (2007), pp. 486–517. ISSN: 00243795. DOI: 10.1016/j.laa.2006.02.045 (cited on page 99).

[54] S. Grivet-Talocia. "A Perturbation Scheme for Passivity Verification and Enforcement of Parameterized Macromodels". In: *IEEE Transactions on Components, Packaging and Manufacturing Technology* (2017), pp. 1–12 (cited on page 104).

[55] S. Grivet-Talocia and M. Bandinu. "Improving the convergence of vector fitting for equivalent circuit extraction from noisy frequency responses". In: *IEEE Transactions on Electromagnetic Compatibility* 48.1 (2006), pp. 104–120. ISSN: 00189375. DOI: 10.1109/TEMC.2006.870814 (cited on page 39).

[56] S. Grivet-Talocia and E. Fevola. "Compact Parameterized Black-Box Modeling via Fourier-Rational Approximations". In: *IEEE Transactions on Electromagnetic Compatibility* 59.4 (2017), pp. 1133–1142 (cited on page 104).

[57] K. Gu, V. L. Kharitonov, and J. Chen. *Stability of Time-Delay Systems*. Springer Science + Business Media, LLC, 2003, p. 367. ISBN: 9781461265849 (cited on page 152).

[58] S. Gugercin and A. C. Antoulas. "A survey of model reduction by balanced truncation and some new results". In: *Internat. J. Control* 77.8 (2004), pp. 748–766. ISSN: 0020-7179. DOI: 10.1080/00207170410001713448 (cited on page 26).

[59] S. Gugercin, A. C. Antoulas, and C. Beattie. "H2 Model Reduction for Large-Scale Linear Dynamical Systems". In: *SIAM J. Matrix Anal. Appl.* 30.2 (2008), pp. 609–638. ISSN: 0895-4798. DOI: 10.1137/060666123 (cited on pages 4, 28, 99).

[60] N. Guglielmi and M. L. Overton. "Fast Algorithms for the Approximation of the Pseudospectral Abscissa and Pseudospectral Radius of a Matrix". In: *SIAM J. Matrix Anal. Appl.* 32.4 (2011), pp. 1166–1192 (cited on page 183).

[61] B. Gustavsen. "Comments on "a comparative study of vector fitting and orthonormal vector fitting techniques for EMC applications"". In: *Proceedings of the 18th International Zurich Symposium on Electromagnetic Compatibility, EMC* 2.6 (2007), pp. 131–134. ISSN: 10774076. DOI: 10.1109/EMCZUR.2007.4388213 (cited on page 39).

[62] B. Gustavsen. "Improving the pole relocating properties of vector fitting". In: *IEEE Transactions on Power Delivery* 21.3 (2006), pp. 1587–1592. ISSN: 08858977. DOI: 10.1109/TPWRD.2005.860281 (cited on pages 37, 39).

[63] B. Gustavsen and A. Semlyen. "Rational approximation of frequency domain responses by vector fitting". In: *IEEE Transactions on Power Delivery* 14.3 (1999), pp. 1052–1061. ISSN: 0001-8708 (cited on pages 31, 38, 100).

[64] S. Guttel and F. Tisseur. "The Nonlinear Eigenvalue Problem". In: *MIMS EPrint* 7 (2017), pp. 1–96 (cited on pages 7, 144, 166).

[65] J. K. Hale and S. M. Verduyn Lunel. *Introduction to Functional Differential Equations*. 1993, p. 458. ISBN: 9780387877136 (cited on page 158).

[66] G. H. Hardy. "The mean value of the modulus of an analytic function". In: *Proceedings of the London Mathematical Society* s2-14.1 (1915), pp. 269–277. ISSN: 1460244X. DOI: 10.1112/plms/s2_14.1.269 (cited on page 19).

[67] M. Heinkenschloss, T. Reis, and A. C. Antoulas. "Balanced truncation model reduction for systems with inhomogeneous". In: *Automatica* 47.3 (2011), pp. 559–564. ISSN: 0005-1098. DOI: 10.1016/j.automatica.2010.12.002 (cited on page 17).

[68] W. Hendrickx, D. Deschrijver, and T. Dhaene. "Some remarks on the Vector Fitting iteration". In: *Progress in Industrial Mathematics at ECMI 2004* (2004), pp. 134–138. DOI: 10.1007/3-540-28073-1_15 (cited on page 39).

[69] J. S. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*. Springer, 2016. ISBN: 9783319224701. DOI: 10.1007/978-3-319-22470-1 (cited on page 43).

[70] N. J. Higham. "The numerical stability of barycentric Lagrange interpolation". In: *IMA Journal of Numerical Analysis* 24.4 (2004), pp. 547–556. ISSN: 02724979. DOI: 10.1093/imanum/24.4.547 (cited on page 37).

[71] M. Hinze and S. Volkwein. "Proper Orthogonal Decomposition Surrogate Models for Nonlinear Dynamical Systems: Error Estimates and Suboptimal Control". In: *Dimension Reduction of Large-Scale Systems*. Ed. by P Benner, V Mehrmann, and D. C. Sorensen. Vol. 45. Lect. Notes Comput. Sci. Eng. Springer-Verlag, Berlin/Heidelberg, Germany, 2005, pp. 261–306 (cited on page 3).

[72] J. M. Hokanson and P. G. Constantine. "Data-driven polynomial ridge approximation using variable projection". In: February (2017), pp. 1–16. arXiv: 1702.05859 (cited on page 114).

[73] T. Insperger and G. Stepan. *Semi-Discretization for Time-Delay Systems*. Ed. by S. S. Antman, P Holmes, L Sirovich, and K Sreenivasan. Vol. 178. Springer, 2011, p. 181. ISBN: 9781461403357. DOI: 10.1007/978-1-4614-0335-7 (cited on page 142).

[74] A. Ionita and A. C. Antoulas. "Data Driven Parameterized Model Reduction in the Loewner Framework". In: *SIAM J. Sci. Comput.* 36.3 (2014), pp. 984–1007 (cited on pages 5, 31, 45).

[75] A. Ionita. "Lagrange rational interpolation and its applications to approximation of large-scale dynamical systems". PhD thesis. Rice University, 2013 (cited on page 45).

[76] A. C. Ionita and A. C. Antoulas. "Data-Driven Parametrized Model Reduction in the Loewner Framework". In: *SIAM J. Sci. Comput.* 36.3 (2014), A984–A1007. ISSN: 0022-3999. DOI: 10.1137/090750688 (cited on page 106).

[77] S. Ito and Y. Nakatsukasa. "Stable polefinding and rational least-squares fitting via eigenvalues". 2016 (cited on page 39).

[78] E. Jarlebring. "Some numerical methods to compute the eigenvalues of a time-delay system using MATLAB®". In: *The delay e-letter* 2 (2008) (cited on page 155).

[79] M. Köhler. "On the closest stable descriptor system in the respective spaces RH_2 and RH_∞". In: *Linear Algebra and Its Applications* 443 (2014), pp. 34–49. ISSN: 00243795. DOI: 10.1016/j.laa.2013.11.012 (cited on page 20).

[80] P. Koosis. *Introduction to Hp spaces*. Vol. 115. Cambridge University Press, 1998 (cited on pages 16, 52).

[81] J. G. Korvink and E. B. Rudnyi. "Oberwolfach Benchmark Collection". In: *Dimension Reduction of Large-Scale Systems*. Ed. by P Benner, D. Sorensen, and V Mehrmann. Vol. 45. Lecture Notes in Computational Science and Engineering. Springer Berlin Heidelberg, 2005, pp. 311–315. DOI: 10.1007/3-540-27909-1_11 (cited on page 8).

[82] Y. Kuang. *Delay differential equations: with applications in population dynamics*. Vol. 191. Academic Press, 1993 (cited on page 142).

[83] K. Kunisch and S. Volkwein. "Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics". In: *SIAM J. Numer. Anal.* 40.2 (2002), pp. 492–515. ISSN: 0036-1429. DOI: 10.1137/S0036142900382612 (cited on page 3).

[84] J. Lambert. "Observations Analytiques". In: *Nouveaux mémoires de l'Académie royale des sciences et belles-lettres* 1 (1772) (cited on page 144).

[85] S. Lefteriu and A. C. Antoulas. "On the convergence of the vector-fitting algorithm". In: *IEEE Transactions on Microwave Theory and Techniques* 61.4 (2013), pp. 1435–1443. ISSN: 00189480. DOI: 10.1109/TMTT.2013.2246526 (cited on page 39).

[86] S. Lefteriu, A. C. Antoulas, and A. C. Ionita. "Parametric model reduction in the Loewner framework". In: *IFAC Proceedings Volumes (IFAC-PapersOnline)* 18.PART 1 (2011), pp. 12751–12756. ISSN: 14746670. DOI: 10.3182/20110828-6-IT-1002.02651 (cited on pages 82, 130).

[87] J. L. Lumley. "The structure of inhomogeneous turbulent flows". In: *Atmospheric turbulence and radio wave propagation* (1967) (cited on page 3).

[88] Y. Maday, O. Mula, A. T. Patera, and M. Yano. "The Generalized Empirical Interpolation Method: Stability theory on Hilbert spaces with an application to the Stokes equation". In: *Computer Methods in Applied Mechanics and Engineering* 287 (2015), pp. 310–334. ISSN: 00457825. DOI: 10.1016/j.cma.2015.01.018 (cited on page 42).

[89] W. Mai and T. Qian. "Rational Approximation of Functions in the Hardy Spaces on Tubes". In: *arXiv preprint* 2 (2016), pp. 1–28. arXiv: 1604.07597 (cited on page 51).

[90] A. J. Mayo and A. C. Antoulas. "A framework for the solution of the generalized realization problem". In: *Linear Algebra and Its Applications* 425.2-3 (2007), pp. 634–662. ISSN: 00243795. DOI: 10.1016/j.laa.2007.03.008 (cited on pages 5, 31, 32).

[91] L. Meier III and D. G. Luenberger. "Approximation of linear constant systems". In: *Automatic Control, IEEE Transactions on* 12.5 (1967), pp. 585–588. ISSN: 0018-9286. DOI: 10.1109/TAC.1967.1098680 (cited on pages 27, 47).

[92] W. Michiels, K. Green, T. Wagenknecht, and S. I. Niculescu. "Pseudospectra and stability radii for analytic matrix functions with application to time-delay systems". In: *Linear Algebra and Its Applications* 418.1 (2006), pp. 315–335. ISSN: 00243795. DOI: 10.1016/j.laa.2006.02.036 (cited on page 145).

[93] W. Michiels and S.-I. Niculescu. *Stability, Control, and Computation for Time-Delay Systems*. SIAM, 2014, p. 435. ISBN: 978-1-61197-362-4. DOI: 10.1137/1.9781611973631 (cited on pages 145, 146).

[94] B. Moore. "Principal component analysis in linear systems: Controllability, observability, and model reduction". In: *IEEE Transactions on Automatic Control* 26.1 (1981), pp. 17–32. ISSN: 0018-9286. DOI: 10.1109/TAC.1981.1102568 (cited on pages 4, 26).

[95] C. T. Mullis and R. A. Roberts. "Synthesis of Minimum Roundoff Noise Fixed Point Digital Filters". In: *IEEE Transactions on Circuits and Systems* 23.9 (1976), pp. 551–562. ISSN: 00984094. DOI: 10.1109/TCS.1976.1084254 (cited on pages 4, 26).

[96] S. Neild, P. D. McFadden, and M. S. Williams. "A Discrete Model of a Vibrating Beam Using a Time-Stepping Approach". In: *Journal of Sound and Vibration* 239.1 (2001), pp. 99–121. ISSN: 0022460X. DOI: 10.1006/jsvi.2000.3158 (cited on page 8).

[97] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer S. Springer DE, 1999 (cited on page 117).

[98] G. Orosz, B. Krauskopf, and R. E. Wilson. "Bifurcations and multiple traffic jams in a car-following model with reaction-time delay". In: *Physica D: Nonlinear Phenomena* 211.3-4 (2005), pp. 277–293. ISSN: 01672789. DOI: 10.1016/j.physd.2005.09.004 (cited on page 142).

[99] H. Panzer, R. Eid, and B. Lohmann. "Generating a Parametric Finite Element Model of a 3D Cantilever Timoshenko Beam Using Matlab". In: (2009), pp. 1–8 (cited on page 109).

[100] A. Patera and G Rozza. *Reduced Basis Approximation and a Posteriori Error Estimation for Parametrized Partial Differential Equations*. MIT, 2007 (cited on page 43).

[101] T. Penzl. "Algorithms for model reduction of large dynamical systems". In: *Linear Algebra Appl.* 415.2–3 (2006), pp. 322–343. ISSN: 00243795. DOI: 10.1016/j.laa.2006.01.007 (cited on pages 83, 118).

[102] M. Reed and B. Simon. *Methods of mathematical physics I: Functional analysis*. 1972 (cited on page 52).

[103] F. Riesz. "Über die Randwerte einer analytischen Funktion". In: *Mathematische Zeitschrift* 18.1 (1923), pp. 87–95 (cited on page 19).

[104] M. Roussel. "The use of delay differential equations in chemical kinetics". In: *The journal of physical chemistry* 100.96 (1996), pp. 8323–8330. ISSN: 0022-3654. DOI: 10.1021/jp9600672 (cited on page 142).

[105] C. Sanathanan and J. Koerner. "Transfer function synthesis as a ratio of two complex polynomials". In: *IEEE Transactions on Automatic Control* 8.1 (1963), pp. 56–58. ISSN: 0018-9286. DOI: 10.1109/TAC.1963.1105517 (cited on page 34).

[106] W. H. Schilders, H. A. van der Vorst, and J. Rommes. *Model Order Reduction - Theory, Research Aspects and Applications*. Springer, 2000, p. 466. ISBN: 9783540788409 (cited on page 3).

[107] A. Semlyen and B. Gustavsen. "Vector fitting by pole relocation for the state equation approximation of nonrational transfer matrices". In: *Circuits, Systems, and Signal Processing* 19.6 (2000), pp. 549–566. ISSN: 0278081X. DOI: 10.1007/BF01271288 (cited on pages 5, 39).

[108] G. Shi. "On the Nonconvergence of the Vector Fitting Algorithm". In: *IEEE Trans. Circuits Syst.* 63.8 (2016), pp. 718–722 (cited on page 39).

[109] H. Smith. *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Vol. 57. 2011, p. 178. ISBN: 978-1-4419-7645-1. DOI: 10.1007/978-1-4419-7646-8. arXiv: 1011.1669 (cited on page 142).

[110] J. N. Stroh. "Non-normality in scalar delay differential equations". PhD thesis. University of Alaska Fairbanks, 2006 (cited on pages 7, 145, 151, 155, 157).

[111] S. R. Taylor and S. A. Campbell. "Approximating chaotic saddles for delay differential equations". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 75.4 (2007), pp. 1–8. ISSN: 15393755. DOI: 10.1103/PhysRevE.75.046215 (cited on pages 142, 152).

[112] L. N. Trefethen. "Pseudospectra of Linear Operators". In: *SIAM Review* 39.3 (1997), pp. 383–406. ISSN: 0036-1445. DOI: 10.1137/S0036144595295284 (cited on page 142).

[113] L. N. Trefethen. "Spectral Methods in Matlab". In: *Lloydia Cincinnati* 10 (2000), p. 184. ISSN: 0586-7614. DOI: 10.1137/1.9780898719598. arXiv: arXiv:1011.1669v3 (cited on page 155).

[114] L. N. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005 (cited on pages 7, 141, 142, 143).

[115] P. M. Van Dooren, K. A. Gallivan, and P. A. Absil. "H2-optimal model reduction of MIMO systems". In: *Applied Mathematics Letters* 21.12 (2008), pp. 1267–1273. ISSN: 08939659. DOI: 10.1016/j.aml.2007.09.015 (cited on page 99).

[116] P. Vuillemin, C. Poussot-Vassal, and D. Alazard. "H 2 optimal and frequency limited approximation methods for large-scale LTI dynamical systems". In: *IFAC Proceedings*

on System Structure and Control. Vol. 5. IFAC, 2013. ISBN: 9783902823250. DOI: 10.3182/20130204-3-FR-2033.00061 (cited on page 99).

[117] T. Vyhlídal, J.-F. Lafay, and R. Sipahi. *Delay systems : from theory to numerics and applications*. Vol. 1. 2013, p. 404. ISBN: ISBN-13: 978-3-319-01694-8 e-ISBN-13: 978-3-319-01695-5. DOI: 10.1007/978-3-319-01695-5 (cited on page 153).

[118] M. Webb, L. N. Trefethen, and P. Gonnet. "Stability of Barycentric Interpolation Formulas for Extrapolation". In: *SIAM J. Sci. Comput.* 34.6 (2012), A3009–A3015 (cited on page 36).

[119] E. M. Wright. "XII.Solution of the Equation $ze^z = a$". In: *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* 65.2 (1959), pp. 193–203 (cited on page 144).