

Dimension Reduction in Structured Dynamical Systems: Optimal- \mathcal{H}_2 Approximation, Data-Driven Balancing, and Real-Time Monitoring

Sean J. Reiter

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

Mark Embree, Chair
Serkan Gugercin, Co-chair
Christopher Beattie
Steffen W. R. Werner

May 21, 2025
Blacksburg, Virginia

Keywords: Model-order reduction, data-driven reduced modeling, \mathcal{H}_2 optimality, balanced truncation, transfer function data, multivariate rational interpolation, quadratic outputs, low-rank, interpolatory decompositions, phasor measurement unit data

Copyright 2025, Sean J. Reiter

Dimension Reduction in Structured Dynamical Systems: Optimal- \mathcal{H}_2 Approximation, Data-Driven Balancing, and Real-Time Monitoring

Sean J. Reiter

(ABSTRACT)

This dissertation considers a variety of problems pertaining to the model-order reduction, data-driven reduced-order modeling, and real-time monitoring of large-scale and structured dynamical systems. In the first part, balancing-based methods for system-theoretic model reduction of linear time-invariant systems are considered. We generalize conditions for which the balanced truncation \mathcal{H}_∞ error bound is known to hold with equality. Specifically, we show that the bound is tight for single-input, single-output systems for which the truncated part of the model is a mild generalization of state-space symmetric. After this, we develop data-driven reformulations of various kinds of balancing-based model reduction for linear first-order and second-order systems. The variants considered are balanced stochastic truncation, positive-real, bounded-real, bounded-real balanced truncation, frequency-weighted balanced truncation, and position-velocity balanced truncation. For each variant, we show how to approximately construct the balanced truncation reduced model from various kinds of input-output invariant frequency response data. In the second part of this dissertation, we consider the \mathcal{H}_2 -optimal model reduction problem for linear dynamical systems with quadratic-output functions. As the significant theoretical contributions of this portion, we establish the Sylvester equation-based (Wilson) and interpolation-based (Meier-Luenberger) \mathcal{H}_2 -optimality frameworks for this class of systems. These frameworks are based on two independent sets of first-order necessary conditions for optimality. We additionally show how to enforce the established necessary optimality conditions using a Petrov-Galerkin projection, and prove that the Wilson optimality conditions imply the interpolation-based optimality conditions under some mild assumptions. Based on the theoretical optimality frameworks, two iterative algorithms for the optimal- \mathcal{H}_2 approximation of linear quadratic-output systems are proposed. In the final portion, we investigate low-rank interpolatory matrix decompositions for reducing the dimensionality of large matrices of Phasor Measurement Unit (PMU) data in the real-time monitoring of electrical power networks. We propose a theoretical framework for analyzing sparse reconstructions of PMU data during online operations. Drawing upon the numerical linear algebra literature, this framework allows us to state a rigorous, computable upper bound for the interpolatory reconstruction error. This bound can be used to certify whether a collection of PMUs or time instances truly captures the low-rank character of the data, and can be leveraged toward various operational functions. Specifically, we propose a data-driven algorithm for real-time disturbance monitoring that is based upon interpolatory approximations and the discrete empirical interpolation method.

Numerical results are included in each portion to validate the theoretical results of the dissertation.

Dimension Reduction in Structured Dynamical Systems: Optimal- \mathcal{H}_2 Approximation, Data-Driven Balancing, and Real-Time Monitoring

Sean J. Reiter

(GENERAL AUDIENCE ABSTRACT)

Dynamical systems are mathematical models of physical phenomena that evolve and change their behavior over time. They are widely used as computational tools for understanding and making reliable predictions about physical systems. These models may take various forms in order to accurately reflect the underlying problem. Often, these models are large in some sense, requiring many degrees of freedom to make accurate predictions and thus are computationally expensive to evaluate. In these instances, it becomes desirable to replace the large-scale, expensive-to-evaluate system with a smaller-scale, cheaper-to-evaluate system that accurately reflects the behavior of the original. This dissertation deals with a variety of problems relating to this approximation procedure. First, we show how to compute approximations that are guaranteed to have certain theoretical properties using data when the underlying model is unavailable. These data can be obtained from computer simulations or real-world measurements of some physical process. Second, we consider the problem of computing the best approximation among smaller, more tractable models that is the best out of all possible approximations, for a certain kind of nonlinear dynamical system. Finally, we develop a method for monitoring the behavior of a system in real time using data. The specific application we consider involves detecting and locating faults, e.g., a fallen power line, in electrical power networks.

Dedication

*For my mom, who gave me everything.
“Everything good in me is due to you, the rest is all my fault.”*

Acknowledgments

I would first like to acknowledge the support of this research by the U. S. National Sciences Foundation (NSF) grants: NSF DMS-2318880 and NSF DMS-1923221. I am overwhelmingly thankful to many, many individuals who have supported me throughout my doctoral work and my entire educational journey. I am fortunate to have been taught and led in the right direction by many great teachers and mentors. I don't think I would have pursued an undergraduate degree in mathematics, let alone a doctorate, had it not been for my high school math teacher, Mr. Berto, showing me the beauty of calculus. I still hear his voice shouting between my ears whenever I use the product rule. I am also thankful to Mr. Pushee, for being a mentor into my adult years, and to Mr. Harrington, for putting Blacksburg on my radar and teaching me how to face hard problems. I am incredibly grateful to my academic advisors, Dr. Mark Embree and Dr. Serkan Gugercin, both of whom have contributed a great deal to my growth as a researcher and as an individual. From my time as an undergraduate, they have supported me professionally, academically, and personally, as well as introduced me to many wonderful and beautiful areas of mathematics. Both have been wonderful mentors and teachers to me, and have shaped me into the mathematician I am today. It has been an absolute pleasure to be their Ph.D. student. I would also like to thank Dr. Chris Beattie and Dr. Steffen Werner for serving on my committee. In particular, I would like to thank Steffen for being a generous collaborator and mentor during the past year or so. Many future graduate students will be lucky to have him as an advisor. I am grateful for the camaraderie and friendships I have developed with the many other graduate students I have met during my studies, especially Mike Ackermann, Art Pelling, Jonas Nicodemus, Linus Balicki, and Yichen Guo. Mike, in particular, has been a consistent force and great friend during my graduate school experience; it is hard to imagine going through it without him. I would also like to thank Petar Mlinarić for being a kind, patient, and willing postdoctoral mentor during his time at Virginia Tech. Thank you to my dear friends Joseph Mills and Colby Weit, who have been foundational pillars of support for me for many years, and these past six in particular. They have helped me keep my sanity on more than one occasion. Finally, I am most thankful for my partner and soon-to-be wife, Maggie, as well as my mother, Kathleen. Maggie and I met before I had even finished my master's degree, and she has been my greatest pillar of support in life, and in the completion of this work, ever since. In her own words: "Optimization matters". My mother Kathleen instilled in me that I could always do what I put my mind to, and that I *should* put my mind to what I love doing the most. It turns out, I love doing math more than I do most other things.

Contents

List of Figures	xii
List of Tables	xiv
1 Introduction	1
1.1 Motivation and problem setting	1
1.2 Outline and contributions of the dissertation	3
2 Mathematical Preliminaries	8
2.1 Linear algebra concepts and notation	8
2.1.1 The Kronecker product and vectorization operator	10
2.2 Analysis preliminaries	12
2.2.1 Functional analysis preliminaries	12
2.2.2 Complex analysis preliminaries	14
2.2.3 \mathcal{L}_p and \mathcal{H}_p spaces	15
2.3 Systems theory for linear dynamical systems	17
2.3.1 Basic notation and definitions	17
2.3.2 Solutions, input-to-output representations, and stability	18
2.3.3 Algebraic operations on linear time-invariant systems	20
2.3.4 Reachability, observability, and infinite Gramians	21
2.3.5 Linear system norms	24
2.4 Model reduction of linear time-invariant systems	26
2.4.1 Interpolatory model reduction of linear systems	28
2.4.2 \mathcal{H}_2 -optimal model reduction of linear systems	29
2.4.3 Balanced truncation model reduction	34

3	On the balanced truncation error bound	41
3.1	A motivating example from power systems modeling	41
3.2	The sign symmetry of balanced realizations	42
3.3	Generalized conditions for the balanced truncation error bound to hold with equality	44
3.4	On the sign parameters of a linear system	49
3.5	Systems with arrowhead realizations	51
3.6	A special case of arrowhead systems	54
3.7	Conclusions	56
4	Data-driven balancing	58
4.1	Introduction	58
4.1.1	Literature review	59
4.1.2	Chapter contents	60
4.2	Generalized quadrature-based balancing	61
4.2.1	An abstract framework for Lyapunov balanced truncation	61
4.2.2	Theoretical formulation for data-driven balancing	62
4.3	What to sample for balanced truncation variants	67
4.3.1	Data-driven balanced stochastic truncation	69
4.3.2	Data-driven positive-real balanced truncation	72
4.3.3	Data-driven bounded-real balanced truncation	76
4.3.4	Common setup for numerical experiments	81
4.3.5	Numerical results	82
4.4	Frequency-weighted balanced truncation	86
4.4.1	Intrusive frequency-weighted balanced truncation	86
4.4.2	Quadrature-based frequency-weighted balanced truncation	89
4.4.3	Numerical results	92
4.5	Data-driven balancing for second-order systems	94
4.5.1	Second-order linear systems theory	96

4.5.2	Second-order balanced truncation	98
4.5.3	Quadrature-based position-velocity balanced truncation	101
4.5.4	Numerical results	105
4.5.5	Butterfly gyroscope	107
4.5.6	Plate with tuned vibration absorbers	109
4.5.7	Mass-spring-damper network with velocity outputs	110
4.6	Conclusions	112
5	System-theoretic concepts for linear systems with quadratic outputs	113
5.1	Introduction	113
5.1.1	Chapter contents	115
5.2	Motivating examples	115
5.2.1	Vibration of a plate with tuned vibration absorbers	116
5.2.2	The Hamiltonian energy functional	117
5.2.3	1D advection-diffusion equation with a quadratic cost	118
5.3	System-theoretic concepts	119
5.3.1	Basic concepts and definitions	119
5.3.2	Subsystem Volterra kernels and transfer functions	120
5.3.3	Gramians and the observability energy functional	123
5.3.4	Generic model reduction of linear quadratic-output systems	126
5.4	Two computationally tractable expressions for the system \mathcal{H}_2 norm and inner product	130
5.4.1	Sylvester-equation based formulation	131
5.4.2	Pole residue-based formulation	134
6	Optimal-\mathcal{H}_2 approximation of linear systems with quadratic outputs	141
6.1	Introduction	141
6.1.1	Chapter contents	141
6.2	Motivating the optimal- \mathcal{H}_2 approximation problem	143

6.3	Sylvester equation-based \mathcal{H}_2 optimality framework	144
6.3.1	The theoretical optimality framework	144
6.3.2	A two-sided iterative algorithm for optimal- \mathcal{H}_2 approximation of linear quadratic-output systems	157
6.4	Interpolation-based \mathcal{H}_2 optimality framework	161
6.4.1	Theoretical optimality framework	162
6.4.2	An iterative rational Krylov algorithm for optimal- \mathcal{H}_2 approximation of linear quadratic-output systems	178
6.5	Comparing the two optimality frameworks	182
6.5.1	Comparing the two iterative algorithms	185
6.6	Numerical examples	185
6.6.1	1D advection-diffusion equation	186
6.6.2	Vibration of a plate with tuned vibration absorbers	192
6.7	Conclusions	195
7	Interpolatory Matrix Factorizations for Phasor Measurement Unit Data	197
7.1	Introduction	197
7.1.1	Background and motivation	197
7.1.2	Chapter contents	198
7.2	Low-rank matrix factorizations of Phasor Measurement Unit data	199
7.2.1	Basic setup and the Singular Value Decomposition	200
7.2.2	Interpolatory approximations of PMU data	201
7.2.3	Analyzing the Interpolatory Approximation Error	206
7.3	Strategies for the pilot bus and time snapshot selection problems	209
7.3.1	The discrete empirical interpolation method	209
7.3.2	Alternative strategies from the power systems literature	213
7.3.3	Numerical experiments	214
7.4	Data-driven monitoring with ID-DEIM	217
7.4.1	Adaptive DEIM-based training of pilot bus configurations	218

7.4.2	Event detection using the error bound (7.15)	220
7.4.3	Event localization using DEIM	221
7.5	Conclusions	226
8	Conclusions and outlook	227
8.1	Summary of contributions	227
8.2	Opportunities for future research	228
	Bibliography	231
	Appendices	252

List of Figures

3.1	An arrowhead network with $n = 6$, with input \mathbf{u} and output \mathbf{y}_o restricted to state x_1	51
4.1	Results for the quadrature-based and intrusive BST reduced models of the RLC circuit benchmark.	83
4.2	Results for the quadrature-based and intrusive PRBT reduced models of the RLC circuit benchmark.	84
4.3	Results for the quadrature-based and intrusive BRBT reduced models of the RLC circuit benchmark.	85
4.4	Frequency response results for order $r = 10$ and $r = 20$ reduced-order models of the RLC circuit benchmark.	92
4.5	Visual representation of the butterfly gyroscope [40, 158].	107
4.6	Frequency response results for reduced-order models of the butterfly gyroscope benchmark for order $r = 10$	108
4.7	Frequency response results for reduced-order models of the plate with TVAs for order $r = 50$	110
4.8	Frequency response results for reduced-order models of the coupled mass-spring-damper network with velocity outputs for order $r = 10$	111
5.1	Plate equipped with TVAs from [9].	116
6.1	Output magnitudes and pointwise relative errors (6.52) of the order $r = 30$ LQO-IRKA and benchmark reduced models in response to the input signals $u_0(t) = 0$, and $u_{\text{sync}}, u_{\text{exp}}$	189
6.2	Output magnitudes and pointwise relative errors (6.52) of the order $r = 30$ LQO-TSIA and benchmark reduced models in response to the input signals $u_0(t) = 0$, and $u_{\text{sync}}, u_{\text{exp}}$	190
6.3	Relative \mathcal{H}_2 errors of the intermediate reduced models produced during the first 50 iterations of LQO-IRKA and LQO-TSIA.	191
6.4	Relative \mathcal{H}_2 errors (6.55) due to the hierarchy of reduced models for orders $r = 2, 4, \dots, 30$	192

6.5	Frequency response magnitude and pointwise relative errors (6.56) for the order $r = 20$ reduced models of the plate with TVAs.	194
6.6	Frequency response magnitude and pointwise relative errors (6.56) for the order $r = 50$ reduced models of the plate with TVAs.	195
7.1	PMU data organized in an $N \times T$ matrix; each of the N rows corresponds to a bus; each of the T columns is a snapshot of the system in time.	200
7.2	Visual illustration of low-rank approximations \mathbf{Y}_K , \mathbf{Y}_s , \mathbf{Y}^t , and \mathbf{Y}_s^t to the PMU data matrix \mathbf{Y}	203
7.3	Visual illustration of the online pilot-based reconstruction of non-pilots described in Algorithm 7.2.1.	204
7.4	Relative errors (7.21) and associated Lebesgue constants for rank $k = 2, 3, \dots, 20$ interpolatory matrix approximations \mathbf{Y}_s and \mathbf{Y}^t of the data generated using the 68-bus 16-machine NETS-NYPS test system.	216
7.5	Relative errors (7.21) and associated Lebesgue constants for rank $k = 1, 2, \dots, 20$ interpolatory matrix approximations \mathbf{Y}_s and \mathbf{Y}^t of the data generated from the far-west region of the ACTIVsg-2000 test case.	217
7.6	Evolution of the Lebesgue constant η_{s_k} and the interpolatory error bound (7.15) throughout the adaptive training as more buses are added.	219
7.7	Interpolatory reconstructions of various (pilot and non-pilot) PMU datastreams using pilots chosen by DEIM during a three-phase fault of the line between buses 28 and 29.	222
7.8	Interpolatory reconstructions of various (pilot and non-pilot) PMU datastreams using pilots chosen by Q-DEIM during a three-phase fault of the line between buses 28 and 29.	223
7.9	Interpolatory reconstructions of various (pilot and non-pilot) PMU datastreams using pilots chosen by MILP during a three-phase fault of the line between buses 28 and 29.	224
7.10	Percentage of event simulations in which the indicated method correctly identified both buses associated with the faulted line in the first k indices for $k = 2, \dots, 8$	225

List of Tables

3.1	\mathcal{H}_∞ norm of the error system, compared to the balanced truncation upper bound (2.77) for a system where the hypothesis (3.8) of Theorem 3.4 holds for $r = 2$ and $r = 3$, but <i>not</i> for $r = 1$	48
3.2	\mathcal{H}_∞ norm of the error system, compared to the balanced truncation upper bound (2.77) for a power system, $n = 5$	56
4.1	Relative Frobenius errors (4.53) in the first 20 (stochastic (4.18), positive-real (4.27), or bounded-real (4.38)) singular values of the RLC circuit benchmark. The smallest error for each set of quadrature-based reduced models is highlighted in boldface	82
4.2	Relative \mathcal{H}_∞ errors according to (4.71) for the BT and frequency-weighted reduced-order models of the RLC circuit benchmark for $r = 10$ and 20. The smallest error for each order is highlighted in boldface	93
4.3	Relative \mathcal{H}_∞ errors (4.95) and \mathcal{H}_2 errors (4.96) for the order $r = 10$ reduced models of the butterfly gyroscope benchmark. The smallest error is highlighted in boldface	108
4.4	Relative \mathcal{H}_∞ errors (4.95) and \mathcal{H}_2 errors (4.96) for the order $r = 50$ reduced models of the plate with TVAs. The smallest error is highlighted in boldface	109
4.5	Relative \mathcal{H}_∞ errors (4.95) and \mathcal{H}_2 errors (4.96) for the order $r = 10$ reduced models of the MSD chain. The smallest error is highlighted in boldface	111
6.1	Relative errors (6.53)–(6.55) for the order $r = 30$ reduced models. The smallest error for each metric is highlighted in boldface	188
6.2	Relative \mathcal{H}_∞ errors (6.57) and \mathcal{H}_2 errors (6.58) for the order $r = 20$ and $r = 50$ reduced models of the plate with TVAs. The smallest error is highlighted in boldface	193
7.1	Pilot buses as chosen by DEIM, Q-DEIM and MILP and corresponding error bounds (7.15).	221

List of Symbols

\mathbb{R}, \mathbb{C}	Fields of real, complex numbers
$\mathbb{R}_{\geq 0}, \mathbb{R}_{> 0}$	Nonnegative, Strictly positive real numbers
$\mathbb{C}_{\geq 0}, \mathbb{C}_{> 0}$	Closed, open right complex half plane
$\mathbb{C}_{\leq 0}, \mathbb{C}_{< 0}$	Closed, open left complex half plane
$\mathbb{Z}, \mathbb{Z}_{\geq 0}$	Integers, nonnegative integers
$\mathbb{R}^{n_1 \times n_2}, \mathbb{C}^{n_1 \times n_2}$	Sets of real-, complex-valued $n_1 \times n_2$ matrices
$\mathbb{R}^n, \mathbb{C}^n$	Sets of real-, complex-valued n -tuples
$ \alpha $	Absolute value or modulus of a real or complex scalar
$\arg(z)$	Argument of a complex scalar
i	Imaginary unit ($i^2 = -1$)
$\operatorname{Re}(z), \operatorname{Im}(z)$	Real, imaginary parts of a complex number $z = \operatorname{Re}(z) + i \operatorname{Im}(z)$
\bar{z}	Conjugate of a complex number $\bar{z} = \operatorname{Re}(z) - i \operatorname{Im}(z)$
$\mathbf{X}_{i,j}, x_{i,j}$	(i, j) -th entry of a matrix \mathbf{X}
$\mathbf{X}_{:,i}$	i -th column of a matrix \mathbf{X}
$\mathbf{X}_{i,:}$	i -th row of a matrix \mathbf{X}
$\mathbf{X}_{i:j,k:\ell}$	Rows i, \dots, j and columns k, \dots, ℓ of a matrix \mathbf{X}
\mathbf{x}_i	i -th entry of a column or row vector \mathbf{x}
$\mathbf{X}_{:,j}$	Columns of a matrix \mathbf{X} given by the indices $\mathbf{j} = [j_1 \ \cdots \ j_k]^\top$
$\mathbf{X}_{i,:}$	Rows of a matrix \mathbf{X} given by the indices $\mathbf{i} = [i_1 \ \cdots \ i_k]$
\mathbf{x}_i	Entries of a vector \mathbf{x} given by the indices $\mathbf{i} = [i_1 \ \cdots \ i_k]$
\mathbf{I}_n	$n \times n$ identity matrix
$\mathbf{e}_i^n, \mathbf{e}_i$	i -th canonical basis vector of dimension n
$\mathbf{0}_{n_1 \times n_2}$	$n_1 \times n_2$ matrix of all zeros

$\mathbf{0}_n$	n -dimensional vector of all zeros
$\mathbf{1}_n$	n -dimensional vector of all ones
$\lambda(\mathbf{A})$	Spectrum of a matrix \mathbf{A}
$\lambda(\mathbf{A}, \mathbf{E})$	Spectrum of a matrix pencil $\lambda\mathbf{E} - \mathbf{A}$
$\sigma(\mathbf{A})$	Singular values of a matrix \mathbf{A}
$\text{Range}(\mathbf{X})$	Range of a matrix \mathbf{X}
$\text{span}(\mathbf{X})$	Span of a matrix \mathbf{X}
$\text{Ker}(\mathbf{X})$	Kernel of a matrix \mathbf{X}

Chapter 1

Introduction

1.1 Motivation and problem setting

Mathematical models in the form of dynamical systems are essential tools for forecasting and deciphering the behavior of many complex physical phenomena. These dynamical systems typically appear as collections of ordinary differential equations (ODEs) resulting from, e.g., detailed spatial discretizations of a partial differential equation (PDE), or interconnected subsystems of large-scale networks like the electrical power grid and coupled mass-spring-damper systems. In the presence of external control variables (inputs) and selected quantities of interest (outputs) the time evolution of a real-valued, finite-dimensional dynamical system can be formally expressed as

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), & \mathbf{x}_0 &= \mathbf{x}(0), \\ \mathbf{y}(t) &= \mathbf{g}(t, \mathbf{x}(t), \mathbf{u}(t)), \end{aligned} \tag{1.1}$$

where $\mathbf{x}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^n$ contains the internal state variables, $\mathbf{u}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m$ the external inputs used to control or influence the system, and $\mathbf{y}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ the outputs or quantities of interest; $\mathbf{E} \in \mathbb{R}^{n \times n}$ is called the descriptor matrix. The time-dependent functions $\mathbf{f}: \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $\mathbf{g}: \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^p$ model the forward evolution of the state and quantities of interest, respectively. This dissertation considers three connected problems relating to dimensionality reduction and reduced-order modeling in the context of large-scale dynamical systems of the general form (1.1).

1. System-theoretic model-order reduction. An important measure of the complexity of a dynamical system (1.1) is its *dimension* or *order*: the number of differential equations in (1.1). This metric is particularly useful for assessing complexity because it is agnostic to any specific formulation of (1.1); e.g., whether the system is linear or nonlinear. In real-world applications, the dimension n is often very large, e.g., $n \gtrsim 10^6$, due to the need for highly accurate numerical predictions and fine spatial or temporal resolutions. This, in turn, places significant demands on computational resources such as time and memory when the large-scale system (1.1) is simulated in computations. A remedy to this problem is *model-order reduction* (MOR): the construction of cheap-to-evaluate surrogate models that replicate the input-to-output response of (1.1) but are described by far fewer differential equations. The computed reduced-order model (ROM) can thereby be used as a high-fidelity approximation in place of the original large-scale system (1.1) in downstream computational tasks such as

solving for the state $\mathbf{x}(t)$, controller design, or ODE-constrained optimization. Precisely, the (generic) model reduction problem is stated as follows: Given an order- n system of the form (1.1), we seek a comparatively low-order reduced model of the form

$$\begin{aligned}\tilde{\mathbf{E}}\dot{\tilde{\mathbf{x}}}(t) &= \tilde{\mathbf{f}}(t, \tilde{\mathbf{x}}(t), \mathbf{u}(t)), & \tilde{\mathbf{x}}_0 &= \tilde{\mathbf{x}}(0) \\ \tilde{\mathbf{y}}(t) &= \tilde{\mathbf{g}}(t, \tilde{\mathbf{x}}(t), \mathbf{u}(t)),\end{aligned}\tag{1.2}$$

where $\tilde{\mathbf{x}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^r$ contains the r reduced states with $r \ll n$, $\tilde{\mathbf{E}} \in \mathbb{R}^{r \times r}$, $\tilde{\mathbf{f}}: \mathbb{R}_{\geq 0} \times \mathbb{R}^r \times \mathbb{R}^m \rightarrow \mathbb{R}^r$, $\tilde{\mathbf{g}}: \mathbb{R}_{\geq 0} \times \mathbb{R}^r \times \mathbb{R}^m \rightarrow \mathbb{R}^p$, and $\tilde{\mathbf{y}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ are the approximated outputs. To serve as an effective surrogate, (1.2) should replicate the input-to-output response of the original large-scale system; i.e., for a tolerance $\tau > 0$ the approximate output $\tilde{\mathbf{y}}$ should be close to the full-order output \mathbf{y} in the sense that

$$\|\mathbf{y} - \tilde{\mathbf{y}}\| \leq \tau \cdot \|\mathbf{u}\|,$$

in suitable norms for all admissible inputs \mathbf{u} . Other considerations are the preservation of internal structural or qualitative features, and the complexity or offline cost of the computational procedure used to produce (1.2). Because of its broad applicability, MOR is an enabling technology in numerous areas of engineering and the physical sciences; it plays a critical role in outer-loop applications that require many queries of the original system (1.1) for multiple parameter values, initial conditions, or control inputs. We refer to the works [4, 5, 19, 31, 32] and the references therein for a comprehensive overview of the state of the art.

2. Data-driven reduced-order modeling. Most traditional system approximation techniques are intrusive, insofar as they require an explicit mathematical formulation of the system (1.1) and the associated operators \mathbf{E} , \mathbf{f} , and \mathbf{g} to compute (1.2) by projection. In complex applications, however, an explicit computational model of the form (1.1) may be difficult to obtain or wholly unavailable; rather, the underlying system is only accessible in the form of *data*. These data may appear in the form of, e.g., input-to-output invariants like frequency-response measurements, or state trajectories obtained via numerical evaluation of black-box or legacy codes. This motivates the discipline of *data-driven reduced-order modeling*: the construction of low-dimensional surrogate models (1.2) solely from system data. In these instances, the goal is to learn a compact representation of the underlying dynamics in (1.1), which can then be used in arbitrary follow-up tasks that require a computational model.

3. Real-time monitoring of physical systems. While data can be used to construct reduced-order models of dynamical systems, another important area of research is the use of streaming data, e.g., data collected dynamically from *in situ* sensors, for the *real-time monitoring* of some physical asset or infrastructure. When paired with a computational model such as (1.1), these data may also be used to infer the internal state of the system or derive control actions used to drive the asset to a favorable operating condition. While actual sensor measurements reveal a wealth of information about the system, data accumulation often poses a significant roadblock to real-time operational benefits. Depending

on the application, several gigabytes of data can be generated or collected each day. An application of interest for this dissertation is the wide-area monitoring of electrical power grids, which are particularly susceptible to low-probability, high-impact events. In settings such as these, dimensionality reduction is employed to reduce the scale of streaming data or impute the full dataset from fewer sparse measurements, thereby enabling more efficient storage, transmission, and subsequent analysis.

This dissertation will make contributions to different aspects of each of these three topics. A theme of this dissertation is the consideration of dynamical systems with *structure*. Depending on the modeling application or underlying physical phenomena, generic dynamical systems of the form (1.1) often inherit additional features. These commonly appear in the time-evolution of the model described by \mathbf{f} ; relevant examples include systems with second-order differential structure, such as the dynamics of an electrical power network or mechanical structures [226], both of which are considered in this dissertation. On the other hand, if the quantities of interest in \mathbf{y} are governed by a *nonlinear* function \mathbf{g} of the state \mathbf{x} , such structure may appear in the output equation of (1.1). For instance, this dissertation considers the case where \mathbf{g} is a *quadratic* function of the state \mathbf{x} . Quadratic outputs arise whenever one is interested in observing quantities computed as the product of time- or frequency-domain components of the state [215, Section 2]. In addition to structured formulations of \mathbf{f} and \mathbf{g} , one might also consider *qualitative* system-theoretic features, such as stability or conservation laws, as another kind of structure. In the context of model reduction or reduced-order modeling, it is desirable that the computed reduced model (1.2) preserve any structural features of the underlying system (1.1) being approximated. Structured surrogates tend to produce more accurate approximations compared to generic (unstructured) reduced models of the same order; they also allow for the re-interpretation of the reduced-order quantities in the original modeling context.

1.2 Outline and contributions of the dissertation

This dissertation makes contributions to each of the three previously introduced research areas in the context of large-scale and structured dynamical systems. Chapter 2 lays out the mathematical background and preliminaries required for the remainder of the dissertation. The chapter begins with the relevant ideas from linear algebra, functional and complex analysis, and linear systems theory, and then reviews state-of-the-art methods for system-theoretic model reduction of linear dynamical systems. This material grounds the results of later chapters.

In Chapter 3, we present new results pertaining to (intrusive) balanced truncation (BT) model reduction for linear time-invariant dynamical systems. First, we show that the BT \mathcal{H}_∞ error bound holds with equality for single-input, single-output (SISO) systems for which the truncated part of the model satisfies a particular state-space symmetry in its canonical balanced form. The *sign parameters* associated with the Hankel singular values provide

the main tool for determining this generalized state-space symmetry of the balanced model. Secondly, we show that these sign parameters are determined by *any* realization of the full-order model that satisfies this generalized state-space symmetry condition. These results are illustrated on a model of the aggregate dynamics of a network of coherent generators that motivated the study. The content of Chapter 3 is published in [182] and also (in a preliminary form) in the author’s M. S. dissertation [189].

1. Reiter, S., Damm, T., Embree, M., and Gugercin, S. (2024). [On the balanced truncation error bound and sign parameters from arrowhead realizations.](#) *Advances in Computational Mathematics*, 50(1):10.
2. Reiter, S. J. (2022). [On the Tightness of the Balanced Truncation Error Bound with an Application to Arrowhead Systems.](#) M. S. dissertation, Virginia Tech.

Balancing-based methods are some of the gold standards for linear model reduction because they (i) preserve desirable qualitative features of the full-order model, e.g., asymptotic stability or passivity, and (ii) often provide error bounds. However, BT and its variants are intrusive projection-based algorithms, and thus cannot be implemented without access to an explicit computational model. Towards developing non-intrusive implementations of balancing-based approximation, Chapter 4 presents novel data-driven reformulations for various types of balancing-related model reduction that build upon the quadrature-based balancing framework of [89]. Specifically, we develop data-driven formulations of balanced stochastic truncation [61, 92, 93], positive-real or passivity-preserving balanced truncation [61], bounded-real balanced truncation [159], the frequency-weighted balanced truncation of Enns [69, 116, 244], and position-velocity balanced truncation [181] for second-order systems. In each case, it is shown how to (approximately) construct the reduced-order quantities underlying the BT-ROM from different kinds of state-space invariant frequency-response data. For the linear BT-variants, these data are either evaluations of particular *spectral factors* [245] associated with the full-order model transfer function, or input- and output-weight functions that specify a frequency range of interest. For the second-order position-velocity BT, the requisite data are evaluations of the system’s position- and velocity-output transfer functions. In effect, these results provide the theoretical foundation for the (approximate) construction of different BT reduced models *directly from transfer function data*. Numerical examples are provided to validate the data-based reduced models. The results for balanced stochastic truncation, positive-real BT, and bounded-real BRBT are contained in the preprint [184].

3. Reiter, S., Gosea, I. V., and Gugercin, S. (2023). [Generalizations of data-driven balancing: what to sample for different balancing-based reduced models.](#) arXiv 2312.12561. (Provisionally accepted for publication in *Automatica*.)

The results for frequency-weighted BT are unpublished, and were developed in collaboration with Serkan Gugercin and Steffen W. R. Werner from Virginia Tech, and Ion Victor Gosea

from the Max Planck Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany. The results for second-order position-velocity BT were developed in collaboration with Steffen W. R. Werner, and are in preparation [187].

4. Reiter, S., and Werner, S. W. R. (2025). [Data-driven balanced truncation for second-order systems with generalized proportional damping](#). In preparation.

Chapters 5 and 6 consider the interpolation-based and optimal- \mathcal{H}_2 approximation of *linear* dynamical systems with *quadratic-output* functions, or *linear quadratic-output* (LQO) systems. Quadratic-output systems have received increased attention in the recent model reduction literature due to the abundance of quadratic quantities of interest in applications; see the motivating examples provided in Section 5.2. Traditional approaches for approximating quadratic-output systems involve first (equivalently) rewriting the quadratic outputs as multiple linear outputs. Then, any well-established technique from linear model reduction can be applied to determine suitable subspaces for approximation. However, this approach usually results in a large number of outputs, $p \sim n$, which can be disadvantageous in the design of reduced models. Instead, Chapters 5 and 6 consider methods that leverage the quadratic state-to-output structure *directly* to compute a reduced-order model. Outside of the author’s previous work [186], this dissertation is the first to consider the optimal- \mathcal{H}_2 approximation of linear quadratic-output systems. In Chapter 5, we begin by introducing the relevant systems theory for linear quadratic-output systems. As our first contribution in this realm, we present two new expressions for calculating the system \mathcal{H}_2 norm and inner product. Chapter 6 then formally considers the \mathcal{H}_2 -optimal model reduction of LQO systems. Therein, we present a pair of solutions to the \mathcal{H}_2 -optimal approximation problem in the form of two independent \mathcal{H}_2 -optimality frameworks; each is based on a different set of (structured) first-order necessary conditions for \mathcal{H}_2 optimality. In effect, these results establish the Sylvester equation-based (or Wilson) [217, 233] and the interpolation-based (or Meier-Luenberger) [97, 142] optimality frameworks for the optimal- \mathcal{H}_2 approximation of LQO systems. Based on these theoretical optimality frameworks, two iterative algorithms for \mathcal{H}_2 -optimal model reduction of LQO systems are developed. The results for the Wilson and interpolation-based \mathcal{H}_2 -optimality frameworks were developed in collaboration with Serkan Gugercin and Ion Victor Gosea, as well as Igor Pontes Duff from the Max Planck Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany. The Wilson and interpolation-based \mathcal{H}_2 -optimality frameworks are respectively contained in the preprints [186] and [185].

5. Reiter, S., Pontes Duff, I., Gosea, I. V., and Gugercin, S. (2024a). [\$\mathcal{H}_2\$ -optimal model reduction of linear systems with multiple quadratic outputs](#). arXiv 2405.05951. (Under review.)
6. Reiter, S., Gosea, I. V., Pontes Duff, I., and Gugercin, S. (2025). [\$\mathcal{H}_2\$ -optimal model reduction of linear quadratic-output systems by multivariate rational interpolation](#). arXiv, 2505.03057.

Portions of the numerical results in Chapter 6 are published in [188].

7. Reiter, S. and Werner, S. W. R. (2025b). [Interpolatory model reduction of dynamical systems with root mean squared error](#). *IFAC-PapersOnLine*, 59(1):385–390.

Chapter 7 considers the low-rank matrix approximation of streaming Phasor Measurement Unit (PMU) data for the real-time monitoring of electrical power networks. PMUs are sensor devices that provide global positioning system (GPS)-synchronized phasor readings of various grid quantities at a rate of 30–200 samples per second. These synchrophasor data reflect the current operating condition of the physical network, and can be used to alert system operators of irregularities and disturbances in a timely manner [125, 235]. However, data accumulation presents a significant roadblock to real-time operational benefits; e.g., a network consisting of 100 PMUs, each with a sampling rate of 120 Hz, generates 200 gigabytes of data per day [79]. We propose using the technology of interpolatory matrix decompositions (IMD) [138, 206] and the (discrete) empirical interpolation method (DEIM) [17, 52] for the sparse reconstruction of PMU data, and related problems in power systems monitoring. From the interpolatory approximation, we can state a rigorous, computable estimate of the reconstruction error during online operations; violation of the estimate is used as a mechanism for detecting changes in the network’s operating condition. In contrast to other, more commonly used low-rank data reduction techniques, interpolatory approximations offer several benefits for this particular application. We describe how this joint IMD-DEIM framework for reducing the dimensionality of PMU data can be leveraged towards operational benefits in real-time disturbance event monitoring, such as event detection and pilot PMU selection [235]. These results are currently in preparation with Serkan Gugercin, Mark Embree, from Virginia Tech, and Vassilis Kekatos from Purdue University. [183]

7. Reiter, S., Embree, M., Gugercin, S., and Kekatos, V. (2025). [Interpolatory matrix approximations for PMU data: Dimension reduction, pilot bus selection, and voltage event monitoring](#). In preparation

Finally, Chapter 8 summarizes the contributions and provides an outlook toward future research problems relating to those considered in this dissertation.

In summary, the major contributions of the dissertation are as follows.

1. In Chapter 3, we generalize conditions for which the balanced truncation \mathcal{H}_∞ error bound is known to hold with equality. Specifically, we show that the bound is tight for single-input, single-output systems for which the truncated part of the model is a mild generalization of state-space symmetric.
2. In Chapter 4, we develop a theoretical framework for data-driven balancing that shows how to construct various balancing-based reduced models directly from different input-to-output invariant transfer function data. The variants considered are

balanced stochastic truncation, positive-real or passivity-preserving balanced truncation, bounded-real balanced truncation, the frequency-weighted balanced truncation of Enns, and position-velocity balanced truncation.

3. In Chapters 5 and 6, we establish the Sylvester equation-based (Wilson) and rational interpolation-based (Meier-Luenberger) optimality frameworks for the optimal- \mathcal{H}_2 approximation of linear quadratic-output systems. We propose a pair of iterative algorithms based on the pair of optimality frameworks that generalize the two-sided iterative algorithm and the iterative rational Krylov algorithm from linear model reduction.
4. In Chapter 7, we propose a theoretical framework for the dimensionality reduction (compression) of large matrices of Phasor Measurement Unit data based on interpolatory matrix decompositions and the discrete empirical interpolation method. This leads to a joint IMD-DEIM framework for wide-area event monitoring using sparse reconstructions of Phasor Measurement Unit data.

Chapter 2

Mathematical Preliminaries

In this chapter, we establish the mathematical preliminaries and basic ideas that are assumed for the remainder of the dissertation, as well as the general problem setting. Certain material, specifically system-theoretic concepts for second-order linear time-invariant systems and linear quadratic-output systems relevant to the results of Chapter 4 and Chapters 5 and 6, is sequestered to those chapters in an effort to make them as self-contained as possible.

2.1 Linear algebra concepts and notation

Throughout this dissertation, we use the following notation to denote the rows, columns, and entries of a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$: The (i, j) -th entry of \mathbf{X} is denoted $\mathbf{X}_{i,j} \in \mathbb{C}$ and occasionally $x_{i,j} \in \mathbb{C}$; the i -th row of \mathbf{X} is denoted $\mathbf{X}_{i,:} \in \mathbb{C}^{1 \times n_2}$; the j -th column of \mathbf{X} is denoted $\mathbf{X}_{:,j} \in \mathbb{C}^{n_1}$ and occasionally $\mathbf{x}_j \in \mathbb{C}^{n_1}$. For a block matrix $\mathbf{X} \in \mathbb{C}^{p \times m}$ consisting of $p \times m$ submatrices, we introduce the notation

$$\mathbf{X}_{\mathbf{k},\mathbf{j}} \stackrel{\text{def}}{=} \mathbf{X}_{(k-1)p+1:kp, (j-1)m+1:jm} \in \mathbb{C}^{p \times m} \quad (2.1)$$

to denote the (k, j) -th block-wise entry of \mathbf{X} of size $p \times m$ over the indicated indices. If $p = 1$ (or $m = 1$) we instead write $\mathbf{X}_{:,j}$ ($\mathbf{X}_{\mathbf{k},:}$) to denote the j -th (k -th) block column (row) of \mathbf{X} . We take $\mathbf{X}^H \stackrel{\text{def}}{=} \overline{\mathbf{X}}^T$ to denote the Hermitian transpose of the matrix, where $\overline{\mathbf{X}}$ is taken to mean entrywise complex conjugation. For a complex number $z \in \mathbb{C}$, the *modulus* of z , $|z| \geq 0$, is defined via $|z|^2 = z\bar{z} \in \mathbb{R}$. We take $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ to denote the $n \times n$ identity matrix, $\mathbf{0}_{n_1 \times n_2} \in \mathbb{R}^{n_1 \times n_2}$ and $\mathbf{0}_{n_1} \in \mathbb{R}^{n_1}$ to respectively denote the $n_1 \times n_2$ matrix and n_1 -dimensional vector of all zeros, and $\mathbf{e}_i^n \in \mathbb{R}^n$ to denote the i -th canonical basis vector in \mathbb{R}^n . When the dimension of \mathbf{e}_i^n is obvious from the discussion, we drop the superscript. Similarly, we drop the subscripts and write \mathbf{I} or $\mathbf{0}$ when the dimensions of the identity and zero matrix (vector) are obvious from context. We denote the range and kernel of a matrix \mathbf{X} by $\text{Range}(\mathbf{X})$ and $\text{Ker}(\mathbf{X})$. Most of the tools and definitions from matrix theory and linear algebra presented in this dissertation are taken from [86], but can be found in other standard texts.

Definition 2.1 (Frobenius and p -norms [86]). Consider a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$. The *Frobenius norm* of \mathbf{X} is defined as

$$\|\mathbf{X}\|_F \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} |x_{ij}|^2} = \sqrt{\text{tr}(\mathbf{X}^H \mathbf{X})}, \quad (2.2)$$

where $\text{tr}: \mathbb{C}^{n \times n} \rightarrow \mathbb{C}$ returns the *trace* of a square matrix defined as the sum of its diagonal entries. For $1 \leq p \leq \infty$, the p -norm of \mathbf{X} is defined as

$$\|\mathbf{X}\|_p = \sup_{\mathbf{z} \neq \mathbf{0}} \frac{\|\mathbf{X}\mathbf{z}\|_p}{\|\mathbf{z}\|_p}. \quad (2.3)$$

◇

The singular value decomposition (SVD) is one of the most fundamental matrix decompositions. It is an essential tool in the analysis of high-dimensional matrices.

Theorem 2.2 (Singular value decomposition [86, Section 2.4]). For a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ with $\text{rank}(\mathbf{X}) = r \leq \min\{n_1, n_2\}$, there exist matrices with orthonormal columns

$$\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_r] \in \mathbb{C}^{n_1 \times r} \text{ and } \mathbf{Y} = [\mathbf{y}_1 \ \mathbf{y}_2 \ \cdots \ \mathbf{y}_r] \in \mathbb{C}^{n_2 \times r},$$

such that

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}^H, \text{ where } \mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_r),$$

and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the *singular values* of \mathbf{X} . The columns of \mathbf{U} and \mathbf{Y} are called the left and right *singular vectors* of \mathbf{X} . ◇

Using the SVD, an arbitrary matrix \mathbf{X} can be decomposed into a sum of r rank-1 matrices:

$$\mathbf{X} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{y}_i^H. \quad (2.4)$$

We call (2.4) the *dyadic form* of \mathbf{X} . Significantly, the SVD produces best low-rank approximations to a matrix in the 2-norm (the spectral norm), and *the* best low-rank approximation to a matrix in the Frobenius norm.

Theorem 2.3 (Eckart-Young-Mirsky Theorem [86, Theorem 2.4.8]). Consider a matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$, and consider $k \in \mathbb{Z}_{>0}$ such that $1 \leq k < r$, where $\text{rank}(\mathbf{X}) = r \leq \min\{n_1, n_2\}$. Let

$$\mathbf{U}_k = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_k] \in \mathbb{C}^{n_1 \times k}, \ \mathbf{Y}_k = [\mathbf{y}_1 \ \cdots \ \mathbf{y}_k] \in \mathbb{C}^{n_2 \times k}, \text{ and } \mathbf{\Sigma}_k = \text{diag}(\sigma_1, \dots, \sigma_k)$$

denote the leading k left singular vectors, right singular vectors, and singular values of \mathbf{X} . Define the matrix

$$\mathbf{X}_k \stackrel{\text{def}}{=} \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{Y}_k^H = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{y}_i^H,$$

which is obtained by truncating the trailing $r - k$ components of the SVD of \mathbf{X} . Then

$$\mathbf{X}_k = \arg \min_{\mathbf{Z} \in \mathbb{C}^{n_1 \times n_2}} \|\mathbf{X} - \mathbf{Z}\|_F \text{ subj. to } \text{rank}(\mathbf{Z}) = k,$$

and the minimizer \mathbf{X}_k attains the approximation errors

$$\sigma_{k+1} = \|\mathbf{X} - \mathbf{X}_k\|_2 \quad \text{and} \quad \sqrt{\sum_{i=k+1}^r \sigma_i^2} = \|\mathbf{X} - \mathbf{X}_k\|_F.$$

◇

Evidently, \mathbf{X}_k provides a satisfactory low-rank approximation to \mathbf{X} if its trailing singular values are negligibly small.

2.1.1 The Kronecker product and vectorization operator

Here, we introduce some algebraic properties of the Kronecker product and vectorization operator. These will primarily be used in Chapters 5 and 6 for simplifying certain mathematical expressions therein. The content of this section is collected from [86, Chapter 12.3] as well as [47, 137].

Definition 2.4 (Kronecker product and vectorization [47]). Given $\mathbf{X} \in \mathbb{C}^{n_1 \times m_1}$ and $\mathbf{Y} \in \mathbb{C}^{n_2 \times m_2}$, the *Kronecker product* of \mathbf{X} and \mathbf{Y} is the matrix $\mathbf{X} \otimes \mathbf{Y} \in \mathbb{C}^{n_1 n_2 \times m_1 m_2}$ defined by

$$\mathbf{X} \otimes \mathbf{Y} \stackrel{\text{def}}{=} \begin{bmatrix} x_{11}\mathbf{Y} & \cdots & x_{1m_2}\mathbf{Y} \\ \vdots & \ddots & \vdots \\ x_{n_1 1}\mathbf{Y} & \cdots & x_{n_1 m_2}\mathbf{Y} \end{bmatrix}. \quad (2.5)$$

The vectorization operator $\text{vec}: \mathbb{C}^{n_1 \times n_2} \rightarrow \mathbb{C}^{n_1 n_2}$ reshapes a matrix into a column vector by column concatenation. The *vectorization* of \mathbf{X} is defined as

$$\text{vec}(\mathbf{X}) \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{X}_{:,1} \\ \vdots \\ \mathbf{X}_{:,n_1} \end{bmatrix} \in \mathbb{C}^{n_1 n_2}. \quad (2.6)$$

◇

As a direct consequence of Definition 2.4, one can deduce the following properties of the Kronecker product and vectorization operator.

Proposition 2.5 (Properties of the Kronecker product and vectorization operator [86, Chapter 12.3]). Given the matrices $\mathbf{A} \in \mathbb{C}^{n_1 \times m_1}$, $\mathbf{B} \in \mathbb{C}^{n_2 \times m_2}$, $\mathbf{C} \in \mathbb{C}^{m_1 \times k_1}$, $\mathbf{D} \in \mathbb{C}^{m_2 \times k_2}$, and $\mathbf{E} \in \mathbb{C}^{k_1 \times q_1}$, the following properties hold:

$$\text{vec}(\mathbf{ACE}) = (\mathbf{E}^\top \otimes \mathbf{A}) \text{vec}(\mathbf{C}), \quad (2.7)$$

$$\text{tr}(\mathbf{AC}) = \text{vec}(\mathbf{A}^\top)^\top \text{vec}(\mathbf{C}), \quad (2.8)$$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD}). \quad (2.9)$$

◇

We refer to (2.9) as the *mixed product property* of the Kronecker product. In general, the Kronecker product of two matrices $\mathbf{X} \in \mathbb{C}^{n_1 \times m_1}$ and $\mathbf{Y} \in \mathbb{C}^{n_2 \times m_2}$ is not commutative in the sense that $(\mathbf{X} \otimes \mathbf{Y}) \neq (\mathbf{Y} \otimes \mathbf{X})$. However, these matrices are *permutation* equivalent, i.e., there exist $\mathbf{K}_{n_1 n_2} \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$ and $\mathbf{K}_{m_1 m_2} \in \mathbb{R}^{m_1 m_2 \times m_1 m_2}$ so that

$$\mathbf{K}_{n_1 n_2}(\mathbf{X} \otimes \mathbf{Y})\mathbf{K}_{m_1 m_2} = (\mathbf{Y} \otimes \mathbf{X}), \quad (2.10)$$

where the so-called *perfect shuffle* (or *commutation*) matrices $\mathbf{K}_{n_1 n_2}$ and $\mathbf{K}_{m_1 m_2}$ are defined according to

$$\mathbf{K}_{n_1 n_2} \stackrel{\text{def}}{=} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(\mathbf{e}_i^{n_1} \mathbf{e}_j^{n_2 \top} \otimes \mathbf{e}_j^{n_1} \mathbf{e}_i^{n_2 \top} \right), \quad (2.11)$$

where $\mathbf{e}_i^{n_1} \in \mathbb{R}^{n_1}$ is the i -th canonical basis vector. See, for instance [137], [86, Chapter 1.2.11]. By definition of the commutation matrices as well as (2.10), we have the following identities from [137, Theorem 3.1].

Proposition 2.6 (Properties of the Kronecker product and perfect shuffle matrices [137, Theorem 3.1]). Let $\mathbf{K}_{n_1 n_2} \in \mathbb{R}^{n_1 n_2 \times n_1 n_2}$ denote the perfect shuffle matrix defined according to (2.11). For any matrix $\mathbf{X} \in \mathbb{C}^{n_1 \times n_2}$ and vector $\mathbf{v} \in \mathbb{C}^{n_2}$, the following identities hold:

$$\mathbf{K}_{n_1 n_2}^\top = \mathbf{K}_{n_2 n_1}, \quad (2.12)$$

$$\mathbf{K}_{n_1 n_2}^\top \mathbf{K}_{n_1 n_2} = \mathbf{K}_{n_1 n_2} \mathbf{K}_{n_1 n_2}^\top = \mathbf{I}_{n_1 n_2}, \quad \text{i.e., } \mathbf{K}_{n_1 n_2}^\top = \mathbf{K}_{n_1 n_2}^{-1}, \quad (2.13)$$

$$\mathbf{K}_{n_1 n_2}(\mathbf{X} \otimes \mathbf{v}) = (\mathbf{v} \otimes \mathbf{X}), \quad (2.14)$$

$$\mathbf{K}_{n_1 n_2} \text{vec}(\mathbf{X}) = \text{vec}(\mathbf{X}^\top) \quad \text{and} \quad \text{vec}(\mathbf{X})^\top \mathbf{K}_{n_1 n_2}^\top = \text{vec}(\mathbf{X}^\top)^\top. \quad (2.15)$$

◇

For the results in Chapter 4, we will rely on the following identities.

Lemma 2.7 (Resolvent identities [68]). For matrices $\mathbf{E}, \mathbf{A} \in \mathbb{C}^{n \times n}$, the identities

$$(s\mathbf{E} - \mathbf{A})^{-1} \mathbf{E} (z\mathbf{E} - \mathbf{A})^{-1} = \frac{(z\mathbf{E} - \mathbf{A})^{-1} - (s\mathbf{E} - \mathbf{A})^{-1}}{s - z}, \quad (2.16)$$

$$(s\mathbf{E} - \mathbf{A})^{-1} \mathbf{A} (z\mathbf{E} - \mathbf{A})^{-1} = \frac{z(z\mathbf{E} - \mathbf{A})^{-1} - s(s\mathbf{E} - \mathbf{A})^{-1}}{s - z}, \quad (2.17)$$

hold for any $s, z \in \mathbb{C}$ such that $s \neq z$ and $s\mathbf{E} - \mathbf{A}$ and $z\mathbf{E} - \mathbf{A}$ are nonsingular. ◇

Proof of Lemma 2.7. For any $s, z \in \mathbb{C}$ such that $s\mathbf{E} - \mathbf{A}$ and $z\mathbf{E} - \mathbf{A}$ are nonsingular, observe:

$$\begin{aligned} (s - z)(s\mathbf{E} - \mathbf{A})^{-1} \mathbf{E} (z\mathbf{E} - \mathbf{A})^{-1} &= (s\mathbf{E} - \mathbf{A})^{-1} (s\mathbf{E} - \mathbf{A} - (z\mathbf{E} - \mathbf{A})) (z\mathbf{E} - \mathbf{A})^{-1} \\ &= (z\mathbf{E} - \mathbf{A})^{-1} - (s\mathbf{E} - \mathbf{A})^{-1}, \end{aligned}$$

thus proving (2.16). Similarly, observe:

$$\begin{aligned} (s-z)(s\mathbf{E}-\mathbf{A})^{-1}\mathbf{A}(z\mathbf{E}-\mathbf{A})^{-1} &= (s\mathbf{E}-\mathbf{A})^{-1}(sz\mathbf{E}-z\mathbf{A}-(sz\mathbf{E}-s\mathbf{A}))(z\mathbf{E}-\mathbf{A})^{-1} \\ &= z(z\mathbf{E}-\mathbf{A})^{-1}-s(s\mathbf{E}-\mathbf{A})^{-1}, \end{aligned}$$

thus proving (2.17). □

2.2 Analysis preliminaries

This section recalls the relevant definitions and results from functional and complex analysis. Throughout, we use $\mathbb{R}_{\geq 0}$, $\mathbb{R}_{> 0}$ to denote the nonnegative and strictly positive real line, $\mathbb{C}_{\geq 0}$, $\mathbb{C}_{> 0}$ to denote the closed and open right complex half plane, and $\mathbb{C}_{\leq 0}$, $\mathbb{C}_{< 0}$ to denote the closed and open left complex half plane.

2.2.1 Functional analysis preliminaries

Because a portion of this dissertation deals with minimization over Hilbert spaces, we review here the relevant definitions and results for performing calculus over normed vector spaces; our presentation follows that of [55]. Throughout, we consider arbitrary normed vector spaces X and Y over the field \mathbb{R} , and take $L(X, Y)$ to denote the space of bounded linear operators $A: X \rightarrow Y$. A vector space X is said to be *complete* if every Cauchy sequence in X converges. A Banach space is a complete vector space X equipped with the norm $\|\cdot\|: X \rightarrow \mathbb{R}_{\geq 0}$; a Hilbert space is a complete vector space X equipped with the inner product $\langle \cdot, \cdot \rangle_X: X \times X \rightarrow \mathbb{R}$.

Definition 2.8 (Fréchet derivative [55, Section 2.2]). Let X and Y be normed vector spaces with the norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively, let $U \subset X$ be open, and $f: U \rightarrow Y$ a function. Then f is said to be *Fréchet differentiable* at $x_0 \in U$ if there exists a bounded linear operator $Df(x_0) \in L(X, Y)$ such that

$$\lim_{h \rightarrow 0} \frac{\|f(x_0 + h) - f(x_0) - Df(x_0)h\|_Y}{\|h\|_X} = 0.$$

We call $Df(x_0)$ the *Fréchet derivative* of f at $x_0 \in U$. ◇

For a function $f: X \rightarrow Y$ that is Fréchet differentiable at $x_0 \in X$, we can write

$$f(x_0 + h) = f(x_0) + Df(x_0)h + O(\|h\|_X^2),$$

for all $h \in X$ in a neighborhood of zero. For functions $f: X \rightarrow Y$ and $g: X \rightarrow Y$ and a point $x_0 \in X$, we write $f(x_0 + h) = g(x_0) + O(\|h\|_X^2)$ if $\|f(x_0 + h) - g(x_0)\|_Y \leq C\|h\|_X^2$

for all $h \in X$ in a sufficiently small neighborhood of x_0 and a real constant $C < \infty$. If it exists, the Fréchet derivative is unique [55, Proposition 2.2], thereby justifying the notation $Df(x_0)$. Moreover, f is continuous at x_0 if it is Fréchet differentiable at x_0 .

Consider the special case of real-valued functions f (so that $Y = \mathbb{R}$) over a Hilbert space X equipped with the inner product $\langle \cdot, \cdot \rangle_X: X \times X \rightarrow \mathbb{R}$. Then $L(X, \mathbb{R}) = X^*$, i.e., the Fréchet derivative in Definition 2.8 is a linear functional and an element of the *dual* of X , denoted X^* . $Df(x_0)$ can thus be identified with a unique element of X via the Riesz Representation Theorem [55, Theorem 6.4].

Definition 2.9 (Gradient of a Fréchet differentiable function [55, Section 6.4]). Let X be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_X$, let $U \subset X$ be open, and $f: U \rightarrow \mathbb{R}$ be a function. Suppose that f is Fréchet differentiable at $x_0 \in U$. The *Riesz representative* of $Df(x_0)$, i.e., the unique element $\nabla f(x_0) \in X$ such that

$$Df(x_0)h = \langle \nabla f(x_0), h \rangle_X, \quad (2.18)$$

for all $h \in X$, is called the *gradient* of f at x_0 . ◇

For a multivariate function $f: X_1 \times \cdots \times X_\ell \rightarrow \mathbb{R}$, partial gradients $\nabla_{x_i} f(x_1, \dots, x_\ell)$ are defined analogously. Consider $f: U \rightarrow \mathbb{R}$ in the context of Definition 2.9. Combining Definition 2.8 and Definition 2.9, the condition for the existence of a Fréchet derivative of f at $x_0 \in U$ becomes

$$f(x_0 + h) = f(x_0) + \langle \nabla f(x_0), h \rangle_X + O(\|h\|_X^2), \quad (2.19)$$

for all sufficiently small perturbations $h \in X$. Thus, the problem of computing the gradient of a real-valued function f becomes, for an arbitrary perturbation h , finding the unique element $\nabla f(x_0) \in X$ so that (2.19) holds.

If $Df(x_0) = 0$, we call $x_0 \in X$ a *critical point* of f ; if f has a local extremum at a point x_0 , then necessarily x_0 is a critical point [55, Corollary 2.5]. Now, let X be a Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle_X$ and $U \subset X$ be open. For a real-valued function $f: U \rightarrow \mathbb{R}$ to have a local minimum (or maximum) at a point $x_0 \in U$, then by (2.18) for any point $h \in X$ it holds that $0 = Df(x_0)h = \langle \nabla f(x_0), h \rangle_X$, meaning $\nabla f(x_0)$ is identically 0 by properties of the inner product.

Next, we introduce the relevant ideas relating to optimization on normed vector spaces. The following definition says nothing about the existence of extrema. Existence of such points can be guaranteed if, e.g., f is continuous on a compact set.

Definition 2.10 (Local extrema [55, Section 2.5]). Let X be a normed vector space with norm $\|\cdot\|_X$, $U \subset X$ be open, and $f: U \rightarrow \mathbb{R}$ be a function. We say that f has a *local minimum* (*local maximum*) at $x_0 \in U$ if $f(x_0) \leq f(x)$ ($f(x) \leq f(x_0)$) for all x in a neighborhood of x_0 . ◇

Proposition 2.11 (Critical points [55, Corollary 2.5]). Let X be a normed vector space with norm $\|\cdot\|_X$ and $U \subset X$ be open. If $f: U \rightarrow \mathbb{R}$ is Fréchet differentiable at $x_0 \in U$ and f has a local extremum at x_0 , then $Df(x_0) = 0$. \diamond

Now, let X be a Hilbert space equipped with the inner product $\langle \cdot, \cdot \rangle_X$ and $U \subset X$ be open. Proposition 2.11 provides a necessary condition for a real-valued function $f: U \rightarrow \mathbb{R}$ to have a local minimum (or maximum) at a point $x_0 \in U$. Namely, by (2.18), for any point $h \in X$ we have

$$0 = Df(x_0)h = \langle \nabla f(x_0), h \rangle_X,$$

meaning $\nabla f(x_0)$ is identically 0.

2.2.2 Complex analysis preliminaries

We review the relevant definitions and results for functions of a complex variable. Our presentation pulls from [13, 81]. Throughout, we take $g: D \rightarrow \mathbb{C}$ to denote a function of the complex variable s on an open domain $D \subseteq \mathbb{C}$. The definitions and results here generalize to complex-matrix-valued functions of the form $\mathbf{G}: D \rightarrow \mathbb{C}^{n_1 \times n_2}$ straightforwardly by applying them entrywise to $\mathbf{G}_{i,j} = g_{i,j}$, which are scalar-complex-valued functions. Throughout, we take i to be the imaginary unit such that $i^2 = -1$.

Definition 2.12 (Definition of an analytic function [81]). A function $g: D \rightarrow \mathbb{C}$ is *analytic* on the domain $D \subseteq \mathbb{C}$ if $g(z)$ is complex differentiable at each point $z \in D$. \diamond

There are a few equivalent characterizations of analyticity. One characterization is that a function $g: D \rightarrow \mathbb{C}$ is *analytic* on the domain D if and only if $g(z)$ can be expanded as a power series on a disc about any point $z \in D$. From this characterization, it follows that an analytic function g is in fact infinitely complex differentiable, i.e., $g \in C^\infty(D)$.

Definition 2.13 (Isolated singularities, poles, and residues [81, Section VI.2]). A point z_0 is an *isolated singularity* of $g: D \rightarrow \mathbb{C}$ if g is analytic in some punctured disk $\{z \in \mathbb{C}: 0 < |z - z_0| < R\}$. The *residue* of g corresponding to z_0 is defined as

$$\text{res}(g(z), z = z_0) = \frac{1}{2\pi i} \int_{|z-z_0|=R} g(z) dz.$$

We further call z_0 is a *pole* of order N if and only if $g(z) = \frac{\check{g}(z)}{(z-z_0)^N}$ for some $\check{g}: \mathbb{C} \rightarrow \mathbb{C}$ that is analytic and nonzero at z_0 . \diamond

The residue of an isolated singularity is the first coefficient corresponding to a negative power in the function's Laurent series expansion. A very powerful technique for evaluating contour integrals is provided by the Residue Theorem.

Theorem 2.14 (The Residue Theorem [81, Chapter VII]). Let $g: D \rightarrow \mathbb{C}$ be an analytic function on the bounded domain $D \cup \partial D$ except at a finite number of isolated singularities $z_1, \dots, z_k \in D$, where the boundary ∂D of D is piecewise smooth. Then

$$\int_{\partial D} g(z) dz = 2\pi i \sum_{j=1}^k \text{res}(g(z), z_j),$$

where $\text{res}(g(z), z = z_j)$ is the *residue* of g corresponding to z_j . \diamond

When a complex contour integral cannot be evaluated, it is useful to bound its magnitude in terms of the values taken by the integrand and the length of the contour; these types of bounds are called ML-estimates.

Theorem 2.15 (ML-estimates [81]). Suppose that $\Gamma \subset \mathbb{C}$ is a piecewise smooth curve. If $g: \mathbb{C} \rightarrow \mathbb{C}$ is a continuous function on Γ , then

$$\left| \int_{\Gamma} g(z) dz \right| \leq \int_{\Gamma} |g(z)| dz.$$

Moreover, if Γ has length $L > 0$ and $|g(z)| \leq M$ for all $z \in \Gamma$, then

$$\left| \int_{\Gamma} g(z) dz \right| \leq M \cdot L.$$

\diamond

The theory of several complex variables for k -variate functions $g: D_1 \times \dots \times D_k \rightarrow \mathbb{C}$ is significantly more involved than the univariate case; cf. [99, 192, 193]. If such a function g is analytic on $D_1 \times \dots \times D_k \subseteq \mathbb{C}^n$, then it is analytic in each variable separately [193, Definition 1.1.3]. In this dissertation, it will suffice to deal with functions $g(s_1, \dots, s_k)$ that are analytic in each variable separately by applying the definitions and results introduced above to a single argument s_i , while taking the remaining $k - 1$ variables to be arbitrarily fixed.

2.2.3 \mathcal{L}_p and \mathcal{H}_p spaces

We introduce here the relevant \mathcal{L}_p and \mathcal{H}_p spaces of matrix-valued functions. Under some mild assumptions, the kernels or transfer functions that characterize the input-to-output response of a multitude of dynamical systems are elements of these spaces.

Definition 2.16 (\mathcal{H}_2 and \mathcal{H}_∞ norms [245, Chapter 4.3]). The \mathcal{H}_2 norm of a k -variate complex-matrix-valued function $\mathbf{H}: \mathbb{C}_{\geq 0} \times \dots \times \mathbb{C}_{\geq 0} \rightarrow \mathbb{C}^{n_1 \times n_2}$ is defined as

$$\|\mathbf{H}\|_{\mathcal{H}_2^{n_1 \times n_2}} \stackrel{\text{def}}{=} \left(\frac{1}{(2\pi)^k} \sup_{x_1 > 0, \dots, x_k > 0} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \|\mathbf{H}(x_1 + iy_1, \dots, x_k + iy_k)\|_{\mathbb{F}}^2 dy_1 \dots dy_k \right)^{\frac{1}{2}}. \quad (2.20)$$

The \mathcal{H}_∞ norm of \mathbf{H} is defined as

$$\|\mathbf{H}\|_{\mathcal{H}_\infty^{n_1 \times n_2}} \stackrel{\text{def}}{=} \sup_{\text{Re}(s_1) > 0, \dots, \text{Re}(s_k) > 0} \|\mathbf{H}(s_1, \dots, s_k)\|_2. \quad (2.21)$$

◇

The spaces of all functions $\mathbf{H}: \mathbb{C}_{\geq 0} \times \dots \times \mathbb{C}_{\geq 0} \rightarrow \mathbb{C}^{n_1 \times n_2}$ that are analytic in each variable on $\mathbb{C}_{\geq 0}$ with finite \mathcal{H}_2 norm (2.20) or \mathcal{H}_∞ norm (2.21) are denoted $\mathcal{H}_2^{n_1 \times n_2}(\mathbb{C}_{\geq 0}^k)$ and $\mathcal{H}_\infty^{n_1 \times n_2}(\mathbb{C}_{\geq 0}^k)$, respectively. The space $\mathcal{H}_2^{n_1 \times n_2}(\mathbb{C}_{\geq 0}^k)$ is a Hilbert space endowed with the inner product

$$\begin{aligned} \langle \mathbf{H}_1, \mathbf{H}_2 \rangle_{\mathcal{H}_2^{n_1 \times n_2}} &\stackrel{\text{def}}{=} \\ &\frac{1}{(2\pi)^k} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \text{tr}(\overline{\mathbf{H}}_1(-i\omega_1, \dots, -i\omega_k) \mathbf{H}_2(i\omega_1, \dots, i\omega_k)^\top) d\omega_1 \dots d\omega_k \end{aligned} \quad (2.22)$$

for $\mathbf{H}_1, \mathbf{H}_2 \in \mathcal{H}_2^{n_1 \times n_2}(\mathbb{C}_{\geq 0}^k)$. When it is obvious from the context, we drop the dependence of the space on $\mathbb{C}_{\geq 0}^k$.

Definition 2.17 (\mathcal{L}_2 norm [245, Chapter 4.3]). The \mathcal{L}_2 norm of a k -variate real-matrix-valued function $\mathbf{h}: \mathbb{R}_{\geq 0} \times \dots \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n_1 \times n_2}$ is defined as

$$\|\mathbf{h}\|_{\mathcal{L}_2^{n_1 \times n_2}} \stackrel{\text{def}}{=} \left(\int_0^\infty \dots \int_0^\infty \|\mathbf{h}(\tau_1, \dots, \tau_k)\|_{\mathbb{F}}^2 d\tau_1 \dots d\tau_k \right)^{\frac{1}{2}}. \quad (2.23)$$

The \mathcal{L}_∞ norm of \mathbf{h} is defined as

$$\|\mathbf{h}\|_{\mathcal{L}_\infty^{n_1 \times n_2}} \stackrel{\text{def}}{=} \sup_{t_1 > 0, \dots, t_k > 0} \|\mathbf{h}(t_1, \dots, t_k)\|_\infty. \quad (2.24)$$

◇

The \mathcal{L}_2 norm in (2.23) is in fact equivalent to the \mathcal{H}_2 norm in (2.20) under some mild assumptions; this result is known as Plancherel's Theorem in k -variables [43]. The space of all functions $\mathbf{h}: \mathbb{R}_{\geq 0} \times \dots \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n_1 \times n_2}$ with finite \mathcal{L}_2 norm (2.23) is denoted $\mathcal{L}_2^{n_1 \times n_2}(\mathbb{R}_{\geq 0}^k)$. Like the Hardy space $\mathcal{H}_2^{n_1 \times n_2}(\mathbb{C}_{\geq 0}^k)$, $\mathcal{L}_2^{n_1 \times n_2}(\mathbb{R}_{\geq 0}^k)$ is a Hilbert space endowed with the inner product

$$\langle \mathbf{h}_1, \mathbf{h}_2 \rangle_{\mathcal{L}_2^{n_1 \times n_2}} \stackrel{\text{def}}{=} \int_0^\infty \dots \int_0^\infty \text{tr}(\mathbf{h}_1(\tau_1, \dots, \tau_k) \mathbf{h}_2(\tau_1, \dots, \tau_k)^\top) d\tau_1 \dots d\tau_k$$

for $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{L}_2^{n_1 \times n_2}(\mathbb{R}_{\geq 0}^k)$.

A useful tool for analyzing systems of time-dependent differential equations is the *Laplace transform*. We will use this transformation in the subsequent chapters to derive frequency-domain (or Laplace domain) representations of the dynamical systems we encounter therein.

Definition 2.18 (Multivariate Laplace transformation [5, Chapter 7.3.1]). For a time-domain function $\mathbf{x}: \mathbb{R}_{\geq 0} \times \cdots \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n_1 \times n_2}$, the k -dimensional multivariate Laplace transformation $\mathbf{X}: \mathbb{C} \times \cdots \times \mathbb{C} \rightarrow \mathbb{C}^{n_1 \times n_2}$ of \mathbf{x} is defined to be

$$\mathbf{X}(s_1, \dots, s_k) \stackrel{\text{def}}{=} \int_0^\infty \cdots \int_0^\infty \mathbf{x}(t_1, \dots, t_k) e^{-s_1 t_1} \cdots e^{-s_k t_k} dt_1 \cdots dt_k$$

provided the integrals exist. \diamond

For the purposes of this dissertation, we will always have that $k = 1$ or $k = 2$.

2.3 Systems theory for linear dynamical systems

This section introduces the basic system-theoretic definitions and ideas for linear time-invariant (LTI) dynamical systems. Our presentation primarily follows that of [4] and occasionally [67, 245]; similar treatments can be found in other related texts, e.g., [5, 31, 32].

2.3.1 Basic notation and definitions

For the time being, we consider continuous-time LTI dynamical systems of the form

$$\mathcal{G}_{\text{lo}} : \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{x}_0 = \mathbf{x}(0), \\ \mathbf{y}_{\text{lo}}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \end{cases} \quad (2.25)$$

where $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{D} \in \mathbb{R}^{p \times m}$; unless otherwise specified, it is assumed that the *mass matrix* \mathbf{E} is nonsingular. While (2.25) is stated for a non-homogeneous initial condition, one can always assume that $\mathbf{x}_0 = \mathbf{0}_n$ by replacement of \mathbf{x} with $\mathbf{x} - \mathbf{x}_0$ without loss of generality. The differential equations in (2.25) model the influence of the external inputs $\mathbf{u}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m$ on the internal states $\mathbf{x}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^n$ and the system outputs $\mathbf{y}_{\text{lo}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ over time. We call \mathbb{R}^n the *state space* of (2.25) because the possible solution trajectories, or states \mathbf{x} of (2.25), take values in \mathbb{R}^n . The matrices $\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}$ and \mathbf{D} constitute a *state-space realization* of (2.25), and we use the shorthand

$$\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$$

when referring to a system \mathcal{G}_{lo} with the given realization (2.25); if ever $\mathbf{E} = \mathbf{I}_n$, then we take for granted that $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) = (\mathbf{I}_n, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$.

Remark 2.19. Henceforth, we refer to LTI systems of the type (2.25) interchangeably as *linear systems* or *linear-output systems*; the latter is to distinguish between the so-called *linear quadratic-output systems* that are the focus of Chapters 5 and 6. The choice of notation in \mathcal{G}_{lo} and \mathbf{y}_{lo} is made to reflect this distinction. \diamond

Definition 2.20 (Order of a LTI system [4, Definition 4.2]). The *dimension*, or *order* of the realization of a linear system (2.25) is defined to be the dimension n of the associated state space \mathbb{R}^n . \diamond

2.3.2 Solutions, input-to-output representations, and stability

For a given initial condition $\mathbf{x}_0 \in \mathbb{R}^n$ and input \mathbf{u} , the solution and output of the linear system \mathcal{G}_{lo} in (2.25) at time $t \geq 0$ are analytically given by

$$\mathbf{x}(t) = e^{\mathbf{E}^{-1}\mathbf{A}t}\mathbf{x}_0 + \int_0^\infty e^{\mathbf{E}^{-1}\mathbf{A}(t-\tau)}\mathbf{E}^{-1}\mathbf{B}\mathbf{u}(\tau)d\tau \quad (2.26)$$

$$\text{and } \mathbf{y}_{\text{lo}}(t) = \mathbf{C}e^{\mathbf{E}^{-1}\mathbf{A}t}\mathbf{x}_0 + \int_0^\infty \mathbf{C}e^{\mathbf{E}^{-1}\mathbf{A}(t-\tau)}\mathbf{E}^{-1}\mathbf{B}\mathbf{u}(\tau)d\tau + \mathbf{D}\mathbf{u}(t) \quad (2.27)$$

where $e^{\mathbf{X}t}$ denotes the usual matrix exponential [4, Equation 4.1.6]. For $\mathbf{x}_0 = \mathbf{0}_n$, the output equation (2.27) reveals an external description of the linear system (2.25) via the convolution

$$\mathbf{y}_{\text{lo}}(t) = \int_0^\infty \mathbf{g}_{\text{lo}}(t - \tau)\mathbf{u}(\tau)d\tau,$$

where $\mathbf{g}_{\text{lo}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$ is called the *impulse response* of \mathcal{G}_{lo} , and is defined as

$$\mathbf{g}_{\text{lo}}(t) = \begin{cases} \mathbf{C}e^{\mathbf{E}^{-1}\mathbf{A}t}\mathbf{E}^{-1}\mathbf{B} + \delta(t)\mathbf{D}, & \text{for } t \geq 0; \\ \mathbf{0}_{p \times m}, & \text{for } t < 0, \end{cases} \quad (2.28)$$

and $\delta: \mathbb{R} \rightarrow \{0, 1\}$ denotes the standard Dirac delta distribution. The impulse response returns the $\mathbf{y}_{\text{lo}}(t)$ in response to $\mathbf{u}(t) = \delta(t)$.

Oftentimes, it is more straightforward to deal with a system's (equivalent) formulation in the frequency (or, Laplace) domain. By applying the univariate Laplace transform in Definition 2.18 to (2.25), one obtains the system of frequency-dependent algebraic equations

$$\mathcal{G}_{\text{lo}}: \begin{cases} s\mathbf{E}\mathbf{X}(s) - \mathbf{E}\mathbf{x}_0 &= \mathbf{A}\mathbf{X}(s) + \mathbf{B}\mathbf{U}(s) \\ \mathbf{Y}_{\text{lo}}(s) &= \mathbf{C}\mathbf{X}(s) + \mathbf{D}\mathbf{U}(s), \end{cases} \quad (2.29)$$

where $\mathbf{X}: \mathbb{C} \rightarrow \mathbb{C}^n$, $\mathbf{U}: \mathbb{C} \rightarrow \mathbb{C}^m$, and $\mathbf{Y}_{\text{lo}}: \mathbb{C} \rightarrow \mathbb{C}^p$ are the Laplace transformations of the time-domain states, inputs, and outputs in (2.25). Because (2.25) and (2.29) are equivalent under the Laplace transform, we refer to them both as \mathcal{G}_{lo} by standard abuse of notation. Solving explicitly for $\mathbf{X}(s) = (s\mathbf{E} - \mathbf{A})^{-1}(\mathbf{B}\mathbf{U}(s) + \mathbf{E}\mathbf{x}_0)$ and substituting \mathbf{Y} in (2.29) reveals the relationship

$$\mathbf{Y}_{\text{lo}}(s) = (\mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D})\mathbf{U}(s) + (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}\mathbf{x}_0.$$

The complex-matrix-valued function $\mathbf{G}_{\text{lo}}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ defined as

$$\mathbf{G}_{\text{lo}}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \quad (2.30)$$

is called the *transfer function* of (2.25), and describes the system's input-to-output response via the relationship $\mathbf{Y}_{\text{lo}}(s) = \mathbf{G}_{\text{lo}}(s)\mathbf{U}(s)$ when $\mathbf{x}_0 = \mathbf{0}_n$, and is the Laplace transform of

the impulse response (2.28). Using Cramer's rule, the inverse of the resolvent $(s\mathbf{E} - \mathbf{A})^{-1}$ in (2.30) can be expressed as

$$(s\mathbf{E} - \mathbf{A})^{-1} = \frac{1}{\det(s\mathbf{E} - \mathbf{A})} \text{adj}(s\mathbf{E} - \mathbf{A}), \quad (2.31)$$

where $\text{adj}(s\mathbf{E} - \mathbf{A})$ is the adjugate matrix of the co-factors of $s\mathbf{E} - \mathbf{A}$. Thus, \mathbf{G}_{lo} is an order- n matrix-valued proper rational function; \mathbf{G}_{lo} is strictly proper whenever $\mathbf{D} = \mathbf{0}_{p \times m}$. Conversely, any univariate proper or strictly proper rational function is the transfer function of a linear state-space system (2.25); see [67, Lemma 2.30, Proposition 2.31].

Having defined the solution to the linear system (2.25), we can introduce the fundamental concept of asymptotic stability.

Definition 2.21 (Poles and asymptotic stability of a linear system [4, Definition 4.2]). The system \mathcal{G}_{lo} in (2.25) is said to be *asymptotically stable* if all the eigenvalues of the matrix pencil $s\mathbf{E} - \mathbf{A}$, i.e., all values $s \in \mathbb{C}$ such that $\det(s\mathbf{E} - \mathbf{A}) = 0$, have negative real parts. The eigenvalues of $s\mathbf{E} - \mathbf{A}$ are called the *poles* of the system (2.25). \diamond

Put differently, Definition 2.21 states that the system (2.25) is asymptotically stable if all its poles exist in the open left-half of the complex plane, $\mathbb{C}_{<0}$. Stability provides information about the long-term behavior of the autonomous system $\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t)$, i.e., the dynamics of (2.25) under zero external forcing $\mathbf{u} = \mathbf{0}_m$ with the initial condition \mathbf{x}_0 . Specifically, if the system (2.25) is asymptotically stable, then $\lim_{t \rightarrow \infty} \mathbf{x}(t) = \mathbf{0}_n$ for any solution trajectory of the autonomous system [4, Chapter 5.8]. We additionally note that the poles of the dynamical system (2.25) as defined above are precisely the poles of its rational transfer function \mathbf{G}_{lo} in (2.30), and vice versa.

For any pair of invertible matrices $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{n \times n}$, if we define a new state $\mathbf{z}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^n$ via the coordinate transformation $\mathbf{z}(t) = \mathbf{T}^{-1}\mathbf{x}(t)$ for all time $t \geq 0$, then the resulting system

$$\check{\mathcal{G}}_{\text{lo}} : \begin{cases} \mathbf{SET}\dot{\mathbf{z}}(t) = \mathbf{SAT}\mathbf{z}(t) + \mathbf{SB}\mathbf{u}(t), & \mathbf{z}_0 = \mathbf{S}\mathbf{x}(0), \\ \mathbf{y}_{\text{lo}}(t) = \mathbf{CT}\mathbf{z}(t) + \mathbf{D}\mathbf{u}(t) \end{cases} \quad (2.32)$$

is *equivalent* to \mathcal{G}_{lo} in (2.25), in the sense that the input-to-output mappings are the same. Indeed, let $\check{\mathbf{G}}_{\text{lo}}$ be the transfer function of the transformed system $\check{\mathcal{G}}_{\text{lo}}$ defined according to (2.30). It follows directly that

$$\check{\mathbf{G}}_{\text{lo}}(s) = \mathbf{CT}(s\mathbf{SET} - \mathbf{SAT})^{-1}\mathbf{SB} + \mathbf{D} = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \mathbf{G}_{\text{lo}}(s)$$

for all $s \in \mathbb{C}$. Thus, the transfer function \mathbf{G}_{lo} of a linear system (2.25) is *invariant* under state-space transformations, and we call it a *state-space invariant*. We thereby consider $\check{\mathcal{G}}_{\text{lo}}$ to be the same linear input-output system as \mathcal{G}_{lo} in (2.25) and write $\mathcal{G}_{\text{lo}} = \check{\mathcal{G}}_{\text{lo}}$. Although we highlight that the transient behavior of the state space can differ under change of coordinate transformations, see [210].

Definition 2.22 (Minimum phase system). Let \mathcal{G}_{lo} be a linear system as in (2.25) so that D has a left or right inverse. \mathcal{G}_{lo} is said to be *minimum phase* if it is asymptotically stable, and all the *zeros* of its transfer function \mathbf{G}_{lo} , i.e., all values $s \in \mathbb{C}$ for which $\mathbf{G}_{\text{lo}}(s) = \mathbf{0}_{p \times m}$, have negative real parts. \diamond

2.3.3 Algebraic operations on linear time-invariant systems

For the results of Chapter 4, we require some basic results regarding linear system interconnection and algebraic operations on systems. These are taken from [245, Section 3.6]. Throughout, suppose that \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ are respectively order- n_1 and order- n_2 linear systems formulated according to (2.25) and having the state-space realizations

$$\mathcal{G}_{\text{lo}} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \quad \text{and} \quad \check{\mathcal{G}}_{\text{lo}} = (\check{\mathbf{A}}, \check{\mathbf{B}}, \check{\mathbf{C}}, \check{\mathbf{D}}).$$

The input, output dimensions of \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ are (m_1, p_1) and (m_2, p_2) , respectively. We have assumed that $\mathbf{E} = \mathbf{I}_{n_1}$ and $\check{\mathbf{E}} = \mathbf{I}_{n_2}$ without loss of generality.

Proposition 2.23 (System cascade [245, Section 3.6]). The *cascade* $\mathcal{G}_{\text{lo}}\check{\mathcal{G}}_{\text{lo}}$ of the two linear systems \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ with $m_2 = p_1$ is an order- $(n_1 + n_2)$ linear system whose output is that of \mathcal{G}_{lo} and input that of $\check{\mathcal{G}}_{\text{lo}}$. Moreover, $\mathcal{G}_{\text{lo}}\check{\mathcal{G}}_{\text{lo}}$ has a pair of equivalent realizations satisfying

$$\begin{aligned} \mathcal{G}_{\text{lo}}\check{\mathcal{G}}_{\text{lo}} &= \left(\begin{bmatrix} \mathbf{A} & \mathbf{B}\check{\mathbf{C}} \\ \mathbf{0}_{n_2 \times n_1} & \check{\mathbf{A}} \end{bmatrix}, \begin{bmatrix} \mathbf{B}\check{\mathbf{D}} \\ \check{\mathbf{B}} \end{bmatrix}, [\mathbf{C} \quad \mathbf{D}\check{\mathbf{C}}], \mathbf{D}\check{\mathbf{D}} \right) \\ &= \left(\begin{bmatrix} \check{\mathbf{A}} & \mathbf{0}_{n_1 \times n_2} \\ \mathbf{B}\check{\mathbf{C}} & \mathbf{A} \end{bmatrix}, \begin{bmatrix} \check{\mathbf{B}} \\ \mathbf{B}\check{\mathbf{D}} \end{bmatrix}, [\mathbf{D}\check{\mathbf{C}} \quad \mathbf{C}], \mathbf{D}\check{\mathbf{D}} \right). \end{aligned} \quad (2.33)$$

The transfer function of $\mathcal{G}_{\text{lo}}\check{\mathcal{G}}_{\text{lo}}$ is given by $\mathbf{G}_{\text{lo}}\check{\mathbf{G}}_{\text{lo}}$. \diamond

The system cascade described by Proposition 2.23 essentially describes the creation of a new linear system by taking the output of \mathcal{G}_{lo} and feeding it as an input to $\check{\mathcal{G}}_{\text{lo}}$. Clearly, this is not a commutative operation. Algebraically, this corresponds to linear system multiplication.

Proposition 2.24 (System addition [245, Section 3.6]). The *parallel connection* or *addition* $\mathcal{G}_{\text{lo}} + \check{\mathcal{G}}_{\text{lo}}$ of the two linear systems \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ is an order- $(n_1 + n_2)$ linear system having the realization

$$\mathcal{G}_{\text{lo}} + \check{\mathcal{G}}_{\text{lo}} = \left(\begin{bmatrix} \mathbf{A} & \mathbf{0}_{n_1 \times n_2} \\ \mathbf{0}_{n_2 \times n_1} & \check{\mathbf{A}} \end{bmatrix}, \begin{bmatrix} \mathbf{B} \\ \check{\mathbf{B}} \end{bmatrix}, [\mathbf{C} \quad \check{\mathbf{C}}], \mathbf{D} + \check{\mathbf{D}} \right). \quad (2.34)$$

The transfer function of $\mathcal{G}_{\text{lo}} + \check{\mathcal{G}}_{\text{lo}}$ is given by $\mathbf{G}_{\text{lo}} + \check{\mathbf{G}}_{\text{lo}}$. \diamond

The inputs and outputs of the parallel connection are just the row- and column-concatenation of the inputs and outputs of \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$.

Proposition 2.25 (Inverse of an LTI system [245, Lemma 3.15]). Suppose that $\mathbf{D} \in \mathbb{R}^{p \times m}$ has full row (column) rank and let \mathbf{D}^\dagger denote a right (left) inverse of \mathbf{D} . Then, the system $\mathcal{G}_{\text{lo}}^\dagger$

$$\mathcal{G}_{\text{lo}}^\dagger = (\mathbf{A} - \mathbf{B}\mathbf{D}^\dagger\mathbf{C}, -\mathbf{B}\mathbf{D}^\dagger, \mathbf{D}^\dagger\mathbf{C}, \mathbf{D}^\dagger) \quad (2.35)$$

having the transfer function $\mathbf{G}_{\text{lo}}^\dagger$ is a *right (left) inverse* of \mathcal{G}_{lo} in the sense that $\mathbf{G}_{\text{lo}}\mathbf{G}_{\text{lo}}^\dagger = \mathbf{I}_p$ ($\mathbf{G}_{\text{lo}}^\dagger\mathbf{G}_{\text{lo}} = \mathbf{I}_m$). Moreover, if $m = p$ and $\mathbf{D}^\dagger = \mathbf{D}^{-1}$ is an inverse of \mathbf{D} , then we write $\mathcal{G}_{\text{lo}}^\dagger = \mathcal{G}_{\text{lo}}^{-1}$ and $\mathbf{G}_{\text{lo}}^\dagger = \mathbf{G}_{\text{lo}}^{-1}$, and call $\mathcal{G}_{\text{lo}}^{-1}$ an *inverse* of \mathcal{G}_{lo} . \diamond

Definition 2.26 (Dual of a linear system [4, Section 4.2.3], [245, Definition 3.9]). The *dual* of \mathcal{G}_{lo} is defined to be the linear system $\mathcal{G}_{\text{lo}}^*$ having the realization

$$\mathcal{G}_{\text{lo}}^* = (-\mathbf{A}^\top, -\mathbf{C}^\top, \mathbf{B}^\top, \mathbf{D}^\top) \quad (2.36)$$

and the transfer function $\mathbf{G}_{\text{lo}}^*(-s)^\top$. \diamond

2.3.4 Reachability, observability, and infinite Gramians

Two fundamental system-theoretic concepts are those of *reachability* and *observability*, which we introduce next. Ultimately, these provide a quantifiable energy-based characterization of how relevant a point $\check{\mathbf{x}} \in \mathbb{R}^n$ in state space is to the input-to-output dynamics of a linear system (2.25). For a complete treatment of these concepts, we refer the reader to [4, Chapter 4.2, Chapter 4.3] and [67, Chapter 2.2, Chapter 2.4].

Definition 2.27 (Reachability of a state [4, Definition 4.6, Definition 4.17], [67, Chapter 2.2]). Given the linear system $\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ in (2.25), we say a state $\check{\mathbf{x}} \in \mathbb{R}^n$ is *reachable* from the zero state if there exists a finite time $t_f > 0$, a finite-energy input $\mathbf{u} \in \mathcal{L}_2^m(0, t_f)$, and a solution trajectory $\mathbf{x}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^n$ such that

$$\mathbf{x}(0) = \mathbf{0}_n, \quad \mathbf{x}(t_f) = \check{\mathbf{x}}, \quad \text{and} \quad \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad \text{for all } t \in (0, t_f).$$

The *reachability subspace* $\mathcal{X}_{\text{reach}} \subseteq \mathbb{R}^n$ is the set of all such reachable states $\check{\mathbf{x}}$. We say the triplet $(\mathbf{E}, \mathbf{A}, \mathbf{B})$ is *completely reachable* if $\mathcal{X}_{\text{reach}} = \mathbb{R}^n$. \diamond

Definition 2.28 (Reachability matrix [4, Definition 4.6], [67, Chapter 2.2]). Given a linear system $\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ as in (2.25), the $n \times nm$ matrix

$$\mathcal{R}(\mathbf{E}, \mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{E}^{-1}\mathbf{B} & \mathbf{E}^{-1}\mathbf{A}\mathbf{E}^{-1}\mathbf{B} & \cdots & (\mathbf{E}^{-1}\mathbf{A})^{n-1}\mathbf{E}^{-1}\mathbf{B} \end{bmatrix} \quad (2.37)$$

is called the *reachability matrix* of \mathcal{G}_{lo} . \diamond

Theorem 2.29 (Characterization of the reachability subspace [4, Theorem 4.7]). Let $\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ be a linear system as in (2.25). The reachability subspace $\mathcal{X}_{\text{reach}}$ of \mathcal{G}_{lo} is given by

$$\mathcal{X}_{\text{reach}} = \text{Range}(\mathcal{R}(\mathbf{E}, \mathbf{A}, \mathbf{B})).$$

Hence, \mathcal{G}_{lo} is completely reachable if and only if $\text{rank}(\mathcal{R}(\mathbf{E}, \mathbf{A}, \mathbf{B})) = n$. \diamond

Definition 2.30 (Observability of a state [4, Definition 4.19], [67, Chapter 2.4]). Given the linear system $\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ in (2.25), a state $\check{\mathbf{x}} \in \mathbb{R}^n$ is *unobservable* if

$$\mathbf{y}_{\text{lo}}(t) = \mathbf{C}e^{\mathbf{E}^{-1}\mathbf{A}t}\check{\mathbf{x}} = \mathbf{0}_p \quad \text{for all } t \geq 0.$$

The *unobservable subspace* $\mathcal{X}_{\text{unobsv}} \subseteq \mathbb{R}^n$ is the set of all such unobservable states $\check{\mathbf{x}}$. We say the triplet $(\mathbf{E}, \mathbf{A}, \mathbf{C})$ is *completely observable* if $\mathcal{X}_{\text{unobsv}} = \{\mathbf{0}_n\}$. \diamond

Definition 2.31 (Observability matrix [4, Definition 4.19], [67, Chapter 2.4]). Given a linear system $\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ as in (2.25), the $np \times n$ matrix

$$\mathcal{O}(\mathbf{E}, \mathbf{A}, \mathbf{C}) \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{E}^{-1}\mathbf{A} \\ \vdots \\ \mathbf{C}(\mathbf{E}^{-1}\mathbf{A})^{n-1} \end{bmatrix} \quad (2.38)$$

is called the *observability matrix* of \mathcal{G}_{lo} . \diamond

Theorem 2.32 (Characterization of the unobservable subspace [4, Theorem 4.20]). Let $\mathcal{G}_{\text{lo}} = (\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ be a linear system as in (2.25). The unobservable subspace of \mathcal{G}_{lo} is given by

$$\mathcal{X}_{\text{unobsv}} = \text{Ker}(\mathcal{O}(\mathbf{E}, \mathbf{A}, \mathbf{C})).$$

Hence, \mathcal{G}_{lo} is completely observable if and only if $\text{rank}(\mathcal{O}(\mathbf{E}, \mathbf{A}, \mathbf{C})) = n$. \diamond

Reachability and observability are *dual* concepts.

Theorem 2.33 (Duality principle [4, Theorem 4.23]). Let \mathcal{G}_{lo} be a linear system as in (2.25) and $\mathcal{G}_{\text{lo}}^*$ be its dual defined according to Definition 2.26. Then, \mathcal{G}_{lo} is reachable (observable) if and only if $\mathcal{G}_{\text{lo}}^*$ is observable (reachable). \diamond

Two fundamental objects related to a linear system (2.25) are the so-called *infinite system Gramians*, or Gramians for short.

Definition 2.34 (Infinite system Gramians [4, Section 4.3]). Consider an asymptotically stable linear system \mathcal{G}_{lo} as in (2.25). The *infinite reachability Gramian* of \mathcal{G}_{lo} , $\mathbf{P} \in \mathbb{R}^{n \times n}$, is defined as

$$\mathbf{P} \stackrel{\text{def}}{=} \int_0^\infty e^{\mathbf{E}^{-1}\mathbf{A}\tau} \mathbf{E}^{-1} \mathbf{B} \left(e^{\mathbf{E}^{-1}\mathbf{A}\tau} \mathbf{E}^{-1} \mathbf{B} \right)^\top d\tau. \quad (2.39)$$

The *infinite observability Gramian* of \mathcal{G}_{lo} , $\mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} \in \mathbb{R}^{n \times n}$ is defined via

$$\mathbf{Q}_{\text{lo}} \stackrel{\text{def}}{=} \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{C}^\top \left(\mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{C}^\top \right)^\top d\tau. \quad (2.40)$$

\diamond

By their definition, it follows that both \mathbf{P} and $\mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E}$ are symmetric positive semi-definite (SPSD) matrices. Per Remark 2.19, the notation in \mathbf{Q}_{lo} is included to distinguish it from the quadratic-output system Gramian introduced in Chapter 5. The infinite system Gramians can also be expressed in the frequency domain [4, Section 4.3]. By applying Plancherel's Theorem [43] to the integrals in (2.39) and (2.40), we obtain the equivalent formulations

$$\mathbf{P} = \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\omega \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} ((i\omega \mathbf{E} - \mathbf{A})^{-1} \mathbf{B})^H d\omega, \quad (2.41)$$

$$\mathbf{Q}_{\text{lo}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbf{C} (i\omega \mathbf{E} - \mathbf{A})^{-1})^H \mathbf{C} (i\omega \mathbf{E} - \mathbf{A})^{-1} d\omega. \quad (2.42)$$

For non-asymptotically stable systems, the integrals in Definition 2.34 will *not* converge, while the contour integrals (2.41) and (2.42) will be well defined so long as a pole of the system does not lie on the imaginary axis. For asymptotically stable systems, the matrices \mathbf{P} and \mathbf{Q}_{lo} are uniquely characterized as solutions to dual generalized Lyapunov equations.

Proposition 2.35 (Gramians as solutions to generalized Lyapunov equations [4, Proposition 4.27]). Consider an asymptotically stable linear system \mathcal{G}_{lo} as in (2.25), and let the matrices $\mathbf{P}, \mathbf{Q}_{\text{lo}} \in \mathbb{R}^{n \times n}$ be defined according to (2.39) and (2.40). Then, \mathbf{P} and \mathbf{Q}_{lo} are the unique solutions to

$$\mathbf{A} \mathbf{P} \mathbf{E}^\top + \mathbf{E} \mathbf{P} \mathbf{A}^\top + \mathbf{B} \mathbf{B}^\top = \mathbf{0}_{n \times n}, \quad (2.43)$$

$$\mathbf{A}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} + \mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{A} + \mathbf{C}^\top \mathbf{C} = \mathbf{0}_{n \times n}. \quad (2.44)$$

◇

The infinite Gramians also characterize the reachable and unobservable subspaces of the corresponding system \mathcal{G}_{lo} according to the following result.

Lemma 2.36 (Reachable and unobservable subspaces in terms of Gramians [4, Theorem 4.15, Theorem 4.26]). Consider an asymptotically stable linear system \mathcal{G}_{lo} in (2.25) and let $\mathbf{P}, \mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} \in \mathbb{R}^{n \times n}$ be the reachability and observability Gramians of \mathcal{G}_{lo} defined according to (2.39) and (2.40). It holds that

$$\begin{aligned} \mathcal{X}_{\text{reach}} &= \text{Range}(\mathcal{R}(\mathbf{E}, \mathbf{A}, \mathbf{B})) = \text{Range}(\mathbf{P}), \\ \mathcal{X}_{\text{unobsv}} &= \text{Ker}(\mathcal{O}(\mathbf{E}, \mathbf{A}, \mathbf{C})) = \text{Ker}(\mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E}). \end{aligned}$$

Thus, \mathcal{G}_{lo} is completely reachable if and only if $\mathbf{P} = \mathbf{P}^\top \succ 0$ and completely observable if and only if $\mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} = \mathbf{E}^\top \mathbf{Q}_{\text{lo}}^\top \mathbf{E} \succ 0$. ◇

For results in Chapter 3, we will also use the *cross Gramian* of a linear system (2.25). This was first introduced in [71] to study the minimality of single-input, single-output systems, and then extended to multi-input, multi-output systems in [74, 123].

Definition 2.37 (Cross Gramian [4, Section 4.3.2]). Consider an asymptotically stable LTI system \mathcal{G}_{lo} as in (2.25) that is *square*, i.e., $m = p$. The *cross Gramian* of \mathcal{G}_{lo} , $\mathbf{X}_c \in \mathbb{R}^{n \times n}$, is defined as

$$\mathbf{X}_c \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} (i\omega \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{C} (i\omega \mathbf{E} - \mathbf{A})^{-1} d\omega. \quad (2.45)$$

◇

Similar to the controllability and observability Gramians, the cross Gramian (2.45) satisfies a generalized Sylvester equation [4, Section 4.3.2]:

$$\mathbf{A} \mathbf{X}_c \mathbf{E} + \mathbf{E} \mathbf{X}_c \mathbf{A} + \mathbf{B} \mathbf{C} = \mathbf{0}_{n \times n}. \quad (2.46)$$

It is shown in [75] that $\mathbf{X}_c^2 = \mathbf{P} \mathbf{Q}_{\text{lo}}$ for square *symmetric* linear systems (2.25), i.e., systems such that $\mathbf{G}_{\text{lo}}(s) = \mathbf{G}_{\text{lo}}(s)^{\text{T}}$ for all $s \in \mathbb{C}$.

As a final consideration of this section, we introduce the concept of *minimality*. This idea provides an answer to a fundamental question: When can a system have its order reduced with zero approximation error?

Definition 2.38 (Minimality of a linear system [4, Definition 4.36]). We say that a realization $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C})$ of the system \mathcal{G}_{lo} in (2.25) is *minimal* if, among all possible realizations, it has the smallest possible dimension n . ◇

There is a useful characterization of minimality in terms of reachability and observability.

Lemma 2.39 ([4, Lemma 4.42]). A realization $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ of the system \mathcal{G}_{lo} in (2.25) is minimal if and only if it is both reachable and observable. ◇

If an order- n realization $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ of the system (2.25) is *not* minimal, then one can always find a lower-order realization described by $r < n$ differential equations that exactly recovers the dynamics of the system. Thus, the model order of a non-minimal system can be reduced with zero approximation error by removing components of the state space that are neither controllable nor observable. One can obtain such a minimal realization by computing the *Kalman canonical decomposition* [245, Theorem 3.10] of a system (2.25).

When a system (2.25) is minimal, it is reasonable to consider that one can reduce the model order by instead truncating states which are, in some sense, difficult to reach or difficult to observe. This is the fundamental idea behind *balanced truncation* model reduction, which we discuss at length in Section 2.4.3.

2.3.5 Linear system norms

To assess the quality of a reduced-order model used to approximate (2.25), we require some notion of the distance between two dynamical systems. To this end, we introduce the \mathcal{H}_2

and \mathcal{H}_∞ norms of a dynamical system (2.25); these are defined as the Hardy space norms from Definition 2.16 of the system's transfer function (2.30).

Definition 2.40 (\mathcal{H}_2 and \mathcal{H}_∞ norms of a linear system [4, Chapter 5], [245, Chapter 4.3]). Consider a linear system \mathcal{G}_{lo} in (2.25). The \mathcal{H}_2 norm of \mathcal{G}_{lo} is defined to be

$$\|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_2} \stackrel{\text{def}}{=} \|\mathbf{G}_{\text{lo}}\|_{\mathcal{H}_2^{p \times m}(\mathbb{C}_{\geq 0})} = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathbf{G}_{\text{lo}}(i\omega)\|_{\text{F}}^2 d\omega \right)^{\frac{1}{2}}. \quad (2.47)$$

The \mathcal{H}_∞ norm of \mathcal{G}_{lo} is defined to be

$$\|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_\infty} \stackrel{\text{def}}{=} \|\mathbf{G}_{\text{lo}}\|_{\mathcal{H}_\infty^{p \times m}(\mathbb{C}_{\geq 0})} = \sup_{\omega \in \mathbb{R}} \|\mathbf{G}_{\text{lo}}(i\omega)\|_2. \quad (2.48)$$

◇

The norms provided in Definition 2.40 have equivalent formulations in the time domain; see [4, Section 5.1]. If a system (2.25) is asymptotically stable, then its transfer function (2.30) is analytic in $\mathbb{C}_{\geq 0}$, and the \mathcal{H}_∞ norm of the system is finite. Moreover, if $\mathbf{D} = \mathbf{0}_{p \times m}$, then the \mathcal{H}_2 norm of the system is finite as well. If a system \mathcal{G}_{lo} in (2.25) is *not* asymptotically stable, we write $\|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_2} = \|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_\infty} = \infty$ by convention. The Hilbert space structure of $\mathcal{H}_2^{p \times m}(\mathbb{C}_{\geq 0})$ can be extended to a pair of linear systems \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ via the inner product of their transfer functions, i.e.,

$$\langle \mathcal{G}_{\text{lo}}, \check{\mathcal{G}}_{\text{lo}} \rangle_{\mathcal{H}_2} \stackrel{\text{def}}{=} \langle \mathbf{G}_{\text{lo}}, \check{\mathbf{G}}_{\text{lo}} \rangle_{\mathcal{H}_2^{p \times m}(\mathbb{C}_{\geq 0})} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{tr} \left(\overline{\mathbf{G}}_{\text{lo}}(-i\omega) \check{\mathbf{G}}_{\text{lo}}(i\omega)^\top \right) d\omega, \quad (2.49)$$

where $\overline{\mathbf{G}}_{\text{lo}}(s) = \overline{\mathbf{C}}(s\mathbf{E} - \mathbf{A})\overline{\mathbf{B}} + \overline{\mathbf{D}}$. Note that (2.49) returns the \mathcal{H}_2 norm defined in (2.47) (up to a nonnegative square root) when $\check{\mathcal{G}}_{\text{lo}} = \mathcal{G}_{\text{lo}}$.

The formulations for the \mathcal{H}_2 and \mathcal{H}_∞ norms provided in Definition 2.40 are mostly of theoretical interest. There are two alternative (exact) expressions for the system \mathcal{H}_2 norm that are far more amenable to computation. These expressions are based on the linear system Gramians and the pole-residue form of the system's transfer function [5, Section 2.1].

Theorem 2.41 (Gramian-based formulae [97, Lemma 2.3]). Let \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ be asymptotically stable linear systems as in (2.25) of order- n and order- r , respectively. Suppose additionally that $\mathbf{D} = \check{\mathbf{D}} = \mathbf{0}_{p \times m}$. Let $\mathbf{X} \in \mathbb{R}^{n \times r}$ and $\mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ be solutions of the generalized Sylvester equations

$$\mathbf{A}\mathbf{X}\tilde{\mathbf{E}}^\top + \mathbf{E}\mathbf{X}\tilde{\mathbf{A}}^\top + \mathbf{B}\tilde{\mathbf{B}}^\top = \mathbf{0}_{n \times r} \quad \text{and} \quad \mathbf{A}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{A}} - \mathbf{C}^\top \tilde{\mathbf{C}} = \mathbf{0}_{n \times r}. \quad (2.50)$$

Then, \mathbf{X} and \mathbf{Z}_{lo} are unique, and the \mathcal{H}_2 inner product of \mathcal{G}_{lo} and $\check{\mathcal{G}}_{\text{lo}}$ is

$$\langle \mathcal{G}_{\text{lo}}, \check{\mathcal{G}}_{\text{lo}} \rangle_{\mathcal{H}_2} = -\text{tr} \left(\mathbf{B}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{B}} \right) = \text{tr} \left(\mathbf{C}\mathbf{X}\tilde{\mathbf{C}}^\top \right). \quad (2.51)$$

If $\mathcal{G}_{\text{lo}} = \tilde{\mathcal{G}}_{\text{lo}}$, then $\mathbf{X} = \mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{Z}_{\text{lo}} = -\mathbf{Q}_{\text{lo}} \in \mathbb{R}^{n \times n}$ according to (2.39) and (2.40). Thus, the \mathcal{H}_2 norm of \mathcal{G}_{lo} is given by

$$\|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_2}^2 = \text{tr}(\mathbf{B}^\top \mathbf{Q}_{\text{lo}} \mathbf{B}) = \text{tr}(\mathbf{C} \mathbf{P} \mathbf{C}^\top). \quad (2.52)$$

◇

Theorem 2.42 (Transfer function-based formulae [97, Lemma 2.4]). Let \mathcal{G}_{lo} and $\tilde{\mathcal{G}}_{\text{lo}}$ be asymptotically stable linear systems as in (2.25) of order- n and order- r , respectively, having the transfer functions \mathbf{G}_{lo} and $\tilde{\mathbf{G}}_{\text{lo}}$ defined according to (2.59). Suppose additionally that $\mathbf{D} = \tilde{\mathbf{D}} = \mathbf{0}_{p \times m}$. Suppose that \mathcal{G}_{lo} has simple poles μ_1, \dots, μ_n so that \mathbf{G}_{lo} can be expanded in pole-residue form as

$$\mathbf{G}_{\text{lo}}(s) = \sum_{k=1}^n \frac{\delta_k \boldsymbol{\beta}_k^\top}{s - \mu_k}, \quad \delta_k \in \mathbb{C}^p, \quad \boldsymbol{\beta}_k \in \mathbb{C}^m.$$

Then, the \mathcal{H}_2 inner product of \mathcal{G}_{lo} and $\tilde{\mathcal{G}}_{\text{lo}}$ is

$$\left\langle \mathcal{G}_{\text{lo}}, \tilde{\mathcal{G}}_{\text{lo}} \right\rangle_{\mathcal{H}_2} = \sum_{k=1}^n \delta_k^\top \tilde{\mathbf{G}}_{\text{lo}}(-\mu_k) \boldsymbol{\beta}_k. \quad (2.53)$$

If $\tilde{\mathcal{G}}_{\text{lo}} = \mathcal{G}_{\text{lo}}$, then the \mathcal{H}_2 norm of \mathcal{G}_{lo} is given by

$$\|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_2} = \left(\sum_{k=1}^n \delta_k^\top \mathbf{G}_{\text{lo}}(-\mu_k) \boldsymbol{\beta}_k \right)^{\frac{1}{2}}. \quad (2.54)$$

◇

2.4 Model reduction of linear time-invariant systems

We have already motivated the model reduction of generic dynamical systems in Chapter 1. Linear systems of the form (2.25) with high-dimensional state spaces, e.g., $n \sim 10^6$ and higher, are commonplace in applications. In this linear setting, the goal of model reduction is the construction of another, comparatively lower-order linear system, of the form

$$\tilde{\mathcal{G}}_{\text{lo}} : \begin{cases} \tilde{\mathbf{E}} \dot{\tilde{\mathbf{x}}}(t) &= \tilde{\mathbf{A}} \tilde{\mathbf{x}}(t) + \tilde{\mathbf{B}} \mathbf{u}(t), & \tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}_0, \\ \tilde{\mathbf{y}}_{\text{lo}}(t) &= \tilde{\mathbf{C}} \tilde{\mathbf{x}}(t) + \tilde{\mathbf{D}} \mathbf{u}(t), \end{cases} \quad (2.55)$$

where $\tilde{\mathbf{E}}, \tilde{\mathbf{A}} \in \mathbb{R}^{r \times r}$, $\tilde{\mathbf{B}} \in \mathbb{R}^{r \times m}$, $\tilde{\mathbf{C}} \in \mathbb{R}^{p \times r}$ and $\tilde{\mathbf{x}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^r$ for $r \ll n$. Henceforth and without loss of generality, we assume that $\mathbf{x}(0) = \mathbf{0}_n$ and $\tilde{\mathbf{x}}(0) = \tilde{\mathbf{x}}_0$. In order for (2.55) to

be an effective surrogate, we require that the reduced output $\tilde{\mathbf{y}}_{\text{lo}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ be a faithful recreation of the full output in (2.25) for arbitrary admissible inputs \mathbf{u} , i.e.,

$$\|\mathbf{y}_{\text{lo}} - \tilde{\mathbf{y}}_{\text{lo}}\| \leq \tau \cdot \|\mathbf{u}\|, \quad \tau > 0.$$

The choice of norms will inform the order-reduction algorithm used to compute $\tilde{\mathcal{G}}_{\text{lo}}$. In particular, both the \mathcal{H}_2 and \mathcal{H}_∞ system (transfer function) errors can be related to the corresponding output error in the time domain [5, Section 2.1]. Indeed, suppose that \mathcal{G}_{lo} and $\tilde{\mathcal{G}}_{\text{lo}}$ are asymptotically stable linear systems as in (2.25) of order- n and order- r , respectively, having the outputs \mathbf{y}_{lo} and $\tilde{\mathbf{y}}_{\text{lo}}$. Then for any $\mathbf{u} \in \mathcal{L}_2^m(\mathbb{R}_{\geq 0})$, it holds that

$$\begin{aligned} \|\mathbf{y}_{\text{lo}} - \tilde{\mathbf{y}}_{\text{lo}}\|_{\mathcal{L}_2^p(\mathbb{R}_{\geq 0})} &\leq \|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_2} \|\mathbf{u}\|_{\mathcal{L}_2^m(\mathbb{R}_{\geq 0})}, \\ \|\mathbf{y}_{\text{lo}} - \tilde{\mathbf{y}}_{\text{lo}}\|_{\mathcal{L}_2^p(\mathbb{R}_{\geq 0})} &\leq \|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_\infty} \|\mathbf{u}\|_{\mathcal{L}_2^m(\mathbb{R}_{\geq 0})}, \end{aligned} \quad (2.56)$$

where the \mathcal{L}_2 and \mathcal{L}_∞ norms are defined according to Definition 2.17. The \mathcal{H}_2 or \mathcal{H}_∞ model reduction error can thereby be used as *a posteriori* error estimators for the approximate output, or a minimization objective.

The general framework we consider for computing surrogate models of the form (2.55) is that of *Petrov-Galerkin projection*. Suppose that the state \mathbf{x} of the full-order model (2.25) evolves predominantly in an $r \ll n$ lower-dimensional subspace $\text{span}(\mathbf{V}) \subset \mathbb{R}^n$ spanned by the basis $\mathbf{V} \in \mathbb{R}^{n \times r}$, i.e., $\mathbf{x} \approx \mathbf{V}\tilde{\mathbf{x}}$. Substituting $\mathbf{x} \approx \mathbf{V}\tilde{\mathbf{x}}$ into (2.25), one then identifies a second r -dimensional subspace $\text{span}(\mathbf{W}) \subset \mathbb{R}^n$ spanned by the basis $\mathbf{W} \in \mathbb{R}^{n \times r}$ to enforce the Petrov-Galerkin orthogonality condition

$$\mathbf{W}^\top \left(\mathbf{E}\mathbf{V}\dot{\tilde{\mathbf{x}}}(t) - \mathbf{A}\mathbf{V}\tilde{\mathbf{x}}(t) - \mathbf{B}\mathbf{u}(t) \right) = \mathbf{0}_r, \quad t \geq 0. \quad (2.57)$$

This ensures that the reduced states exactly satisfy the dynamics of the reduced-order system. The order- r reduced model $\tilde{\mathcal{G}}_{\text{lo}}$ as in (2.55) resulting from this projection scheme is given by

$$\tilde{\mathbf{E}} = \mathbf{W}^\top \mathbf{E}\mathbf{V}, \quad \tilde{\mathbf{A}} = \mathbf{W}^\top \mathbf{A}\mathbf{V}, \quad \tilde{\mathbf{B}} = \mathbf{W}^\top \mathbf{B}, \quad \tilde{\mathbf{C}} = \mathbf{C}\mathbf{V}, \quad \text{and} \quad \tilde{\mathbf{D}} = \mathbf{D}. \quad (2.58)$$

The choice of $\tilde{\mathbf{D}} = \mathbf{D}$ is most common, although this is not required. Significantly, the reduced model depends upon the subspaces $\text{span}(\mathbf{V})$ and $\text{span}(\mathbf{W})$, *not* the particular choice of bases \mathbf{V} and \mathbf{W} . Thus, one can always replace \mathbf{V} and \mathbf{W} with well-conditioned matrices containing orthonormal columns that are computed via, e.g., a QR factorization.

In the regime of (2.58), computing a reduced-order model of the form (2.55) resolves to choosing left and right approximation subspaces spanned by $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{n \times r}$, and projecting the full-order matrix operators according to (2.58). There is a rich and well-established theory on the model reduction of LTI systems (2.25); see, e.g., the standard texts [4, 5, 31, 32]. Many system-theoretic model reduction techniques can be categorized as those based on the balancing of energy functionals and truncation of system states [46, 61, 69, 92, 96, 114, 151, 152],

or those based on the rational interpolation of the transfer function in (2.25) or the matching of its moments [5, 60, 80, 94, 97, 139, 195]. There is also a rich and interesting connection between rational transfer function interpolation and optimality in the \mathcal{H}_2 metric [97, 142, 217, 218, 233]. The references listed above are by no means comprehensive; in the remainder of this chapter, we describe the balancing-based, interpolatory, and \mathcal{H}_2 -optimal model reduction theory relevant to the remainder of this dissertation.

2.4.1 Interpolatory model reduction of linear systems

Recall that the input-to-output map of a linear system (2.25) is described (completely) by the order- n rational function (2.30). Because the system error, according to Definition 2.40, is the Hardy \mathcal{H}_2 or \mathcal{H}_∞ norm of the transfer function error, it is a reasonable strategy to design the reduced-order model in (2.55) so that its transfer function

$$\tilde{\mathbf{G}}_{\text{lo}}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m}, \quad \text{where } \tilde{\mathbf{G}}_{\text{lo}}(s) \stackrel{\text{def}}{=} \tilde{\mathbf{C}}(s\tilde{\mathbf{E}} - \tilde{\mathbf{A}})^{-1}\tilde{\mathbf{B}} + \tilde{\mathbf{D}}, \quad (2.59)$$

which is itself an order- r rational function, is a good approximation to \mathbf{G}_{lo} . One possibility is to construct $\tilde{\mathbf{G}}_{\text{lo}}$ as a rational interpolant of \mathbf{G}_{lo} ; this is the fundamental idea behind *interpolatory* model reduction, which we review in this section.

Roughly speaking, the interpolatory model reduction problem in a projection-based regime (2.58) resolves to designing specifically tailored model reduction bases $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{n \times r}$ so that the reduced-order transfer function (2.59) matches its full-order counterpart (2.30), or its derivatives, at selected points $\sigma_1, \dots, \sigma_k \in \mathbb{C}$. For multiple-input, multiple-output systems, instead of full-matrix interpolation, it is usually preferred to perform so-called *tangential interpolation*; that is, the interpolation of a matrix-valued function along specified vectors called tangential directions. The theoretical foundations of interpolatory model reduction have their roots in classical Padé approximants [14], which are rational Hermite interpolants that match the leading *moments* of a (scalar-valued) rational function $G(s)$, i.e., its derivatives about $s = 0$. The single-input, single-output interpolatory model reduction problem was related to a projection formulation by Skelton et al. [60, 239, 240], which was later developed into a numerically efficient framework by Grimme [94] using the rational Krylov method of Ruhe [195]. Gallivan et al. [80] established how to construct tangential interpolations using rational Krylov methods. These results are summarized in the following theorem.

Theorem 2.43 (Tangential interpolation of linear systems [5]). Consider \mathcal{G}_{lo} as in (2.25) and let $\tilde{\mathcal{G}}_{\text{lo}}$ be a reduced model (2.55) by projection (2.58) with $\mathbf{D} = \tilde{\mathbf{D}}$. Consider the interpolation points $\sigma, \mu \in \mathbb{C}$ such that $s\mathbf{E} - \mathbf{A}$ and $s\tilde{\mathbf{E}} - \tilde{\mathbf{A}}$ are nonsingular for all $s = \sigma, \mu$ and the left and right tangential direction vectors $\boldsymbol{\ell} \in \mathbb{C}^p$ and $\mathbf{r} \in \mathbb{C}^m$. Suppose that $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{n \times r}$ have full rank and satisfy

$$(\sigma\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{r} \in \text{Range}(\mathbf{V}) \quad \text{and} \quad (\mu\mathbf{E}^\top - \mathbf{A}^\top)^{-1}\mathbf{C}^\top\boldsymbol{\ell} \in \text{Range}(\mathbf{W}). \quad (2.60)$$

Then $\tilde{\mathcal{G}}_{\text{lo}}$ satisfies the tangential interpolation conditions

$$\mathbf{G}_{\text{lo}}(\sigma)\mathbf{r} = \tilde{\mathbf{G}}_{\text{lo}}(\sigma)\mathbf{r} \quad \text{and} \quad \boldsymbol{\ell}^\top \mathbf{G}_{\text{lo}}(\mu) = \boldsymbol{\ell}^\top \tilde{\mathbf{G}}_{\text{lo}}(\mu).$$

Moreover, if additionally $\sigma = \mu$, then $\boldsymbol{\ell}^\top \frac{d}{ds} \mathbf{G}_{\text{lo}}(\sigma)\mathbf{r} = \boldsymbol{\ell}^\top \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(\sigma)\mathbf{r}$. \diamond

Theorem 2.43 can be used to enforce interpolation at multiple points $\sigma_1, \dots, \sigma_k \in \mathbb{C}$ with multiple left- and right-tangential direction vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_k \in \mathbb{C}^p$ and $\mathbf{r}_1, \dots, \mathbf{r}_k \in \mathbb{C}^m$ in an obvious way. We refer to model reduction bases satisfying the conditions in (2.60) as *interpolatory bases*. Moreover, Theorem 2.43 can be extended to match higher-order interpolation conditions and even preserve internal system structures using constructions similar to (2.60); see [5, Theorem 3.3.3] and [22], [5, Theorem 3.4.1], respectively. For constructing interpolants with $\tilde{\mathbf{D}} \neq \mathbf{D}$, see [5, Theorem 3.3.3].

Interpolatory reduced models can be computed in a very numerically efficient manner using Theorem 2.43. Assuming the interpolation points and tangential directions are given, one simply needs to solve $2r$ shifted linear systems according to (2.60), then orthogonalize and project (2.58). Obviously, the choice of interpolation points and tangential directions will greatly affect the quality of the reduced model. Interestingly, it can be shown that \mathcal{H}_2 -optimal reduced models are tangential interpolants, and the optimal interpolation points are the mirror images of the reduced model's poles; we discuss this at length in Section 2.4.2. Other methods, which are based on adaptive point selection or greedy \mathcal{H}_∞ or \mathcal{L}_∞ error norm minimization, can be employed; cf. [1, 2, 6, 66].

2.4.2 \mathcal{H}_2 -optimal model reduction of linear systems

In this section, we review the optimal- \mathcal{H}_2 approximation of linear dynamical systems (2.25). This problem is stated precisely as follows: Given an order- n , asymptotically stable LTI system \mathcal{G}_{lo} as in (2.25), we wish to construct a reduced-order system $\tilde{\mathcal{G}}_{\text{lo}}$ as in (2.55) of a fixed approximation order $1 \leq r < n$ such that the \mathcal{H}_2 error in approximating (2.25) is minimized, i.e., $\tilde{\mathcal{G}}_{\text{lo}}$ solves

$$\|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_2} = \min_{\dim(\tilde{\mathcal{G}}_{\text{lo}})=r} \|\mathcal{G}_{\text{lo}} - \check{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_2} \quad \text{subj. to } \check{\mathcal{G}}_{\text{lo}} \text{ asymptotically stable.} \quad (2.61)$$

The \mathcal{H}_2 -optimal model reduction of linear systems is very well studied; see, e.g. [97, 142, 147, 148, 149, 150, 217, 218, 233, 237]. We refer the reader to [5, Chapter 5] for a comprehensive and in-depth discussion of this topic. The minimization problem in (2.61) is nonconvex, and global minimizers are hard to characterize. Thus, best practice in \mathcal{H}_2 -optimal model reduction instead is to identify reduced-order models that satisfy some first-order necessary conditions for (local) \mathcal{H}_2 optimality. The two most well-known optimality frameworks are the Sylvester equation-based, or Gramian-based framework attributed to Wilson [97, 217, 233], and the interpolation-based framework of Meier and Luenberger [97, 142, 217]. These were

shown to be equivalent using a structured orthogonality framework by Gugercin et al. [97]. We review both the Sylvester equation-based and interpolatory optimality frameworks for solving (2.61) to set the stage for the results in Chapter 6, which considers the optimal- \mathcal{H}_2 approximation problem for a class of weakly nonlinear systems. Because $\mathbf{D} = \tilde{\mathbf{D}}$ is necessary for a finite \mathcal{H}_2 error, we assume this throughout the subsequent discussion.

The Wilson (Sylvester equation-based) optimality framework.

The starting point for deriving the Wilson conditions is the so-called *error system* $\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}$. This is an order- $(n+r)$ linear system defined according to (2.25) and having the realization

$$\mathbf{E}_e = \begin{bmatrix} \mathbf{E} & \\ & \tilde{\mathbf{E}} \end{bmatrix}, \quad \mathbf{A}_e = \begin{bmatrix} \mathbf{A} & \\ & \tilde{\mathbf{A}} \end{bmatrix}, \quad \mathbf{B}_e = \begin{bmatrix} \mathbf{B} \\ \tilde{\mathbf{B}} \end{bmatrix}, \quad \mathbf{C}_e = \begin{bmatrix} \mathbf{C} & -\tilde{\mathbf{C}} \end{bmatrix}, \quad \text{and} \quad \mathbf{D}_e = \mathbf{D} - \tilde{\mathbf{D}}. \quad (2.62)$$

The Gramians of the error system $\mathbf{P}_e, \mathbf{E}_e^\top \mathbf{Q}_{\text{lo},e} \mathbf{E}_e \in \mathbb{R}^{(n+r) \times (n+r)}$ satisfy the generalized Lyapunov equations (2.43) and (2.44) for the realization above, and can be written in 2×2 block form as

$$\mathbf{P}_e = \begin{bmatrix} \mathbf{P} & \mathbf{X} \\ \mathbf{X}^\top & \tilde{\mathbf{P}} \end{bmatrix} \quad \text{and} \quad \mathbf{E}_e^\top \mathbf{Q}_{\text{lo},e} \mathbf{E}_e = \begin{bmatrix} \mathbf{E}_e^\top \mathbf{Q}_{\text{lo}} \mathbf{E}_e & \mathbf{E}_e^\top \mathbf{Z}_{\text{lo}} \\ \mathbf{Z}_{\text{lo}}^\top \mathbf{E}_e & \mathbf{E}_e^\top \tilde{\mathbf{Q}}_{\text{lo}} \mathbf{E}_e \end{bmatrix}, \quad (2.63)$$

where $\mathbf{P}, \mathbf{E}_e^\top \mathbf{Q}_{\text{lo}} \mathbf{E}_e \in \mathbb{R}^{n \times n}$ and $\tilde{\mathbf{P}}, \tilde{\mathbf{E}}^\top \tilde{\mathbf{Q}}_{\text{lo}} \tilde{\mathbf{E}} \in \mathbb{R}^{r \times r}$ are the full- and reduced-order system Gramians defined according to (2.39) and (2.40), while the matrices $\mathbf{X}, \mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ satisfy the generalized Sylvester equations in (2.50). Applying Theorem 2.41 to the realization (2.62) reveals the following expressions for the squared \mathcal{H}_2 error:

$$\begin{aligned} \|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2 &= \text{tr} \left(\mathbf{B}^\top \mathbf{Q}_{\text{lo}} \mathbf{B} + 2\mathbf{B}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{B}} + \tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lo}} \tilde{\mathbf{B}} \right) \\ &= \text{tr} \left(\mathbf{C} \mathbf{P} \tilde{\mathbf{C}}^\top - 2\mathbf{C} \mathbf{X} \tilde{\mathbf{C}}^\top + \tilde{\mathbf{C}} \tilde{\mathbf{P}} \tilde{\mathbf{C}}^\top \right). \end{aligned} \quad (2.64)$$

Thus, the error $\|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_2}$ can be interpreted as a cost function $\mathcal{J}: \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times m} \times \mathbb{R}^{p \times r} \rightarrow \mathbb{R}_{\geq 0}$ that takes the matrices of the reduced-order linear model (2.55) as arguments. Wilson [233] originally derived first-order optimality conditions for (2.61) by computing gradients of \mathcal{J} with respect to the reduced model matrices; this result is summarized next. Although we mention that in the original work Wilson assumed $\mathbf{E} = \mathbf{I}_n$; a proof of the conditions for a general nonsingular \mathbf{E} was given in [146, Theorem 2.44].

Theorem 2.44 (Sylvester-equation based \mathcal{H}_2 -optimality conditions for linear systems [97, 146, 217, 233]). Suppose that \mathcal{G}_{lo} and $\tilde{\mathcal{G}}_{\text{lo}}$ are asymptotically stable linear systems as in (2.25) and (2.55), and suppose additionally that $\tilde{\mathcal{G}}_{\text{lo}}$ minimizes the \mathcal{H}_2 error in (2.61). Then, the

following conditions hold:

$$\mathbf{0}_{r \times r} = \tilde{\mathbf{Q}}_{\text{lo}} \tilde{\mathbf{A}} \tilde{\mathbf{P}} + \mathbf{Z}_{\text{lo}}^{\top} \mathbf{A} \mathbf{X}, \quad (2.65\text{a})$$

$$\mathbf{0}_{r \times r} = \tilde{\mathbf{Q}}_{\text{lo}} \tilde{\mathbf{E}} \tilde{\mathbf{P}} + \mathbf{Z}_{\text{lo}}^{\top} \mathbf{E} \mathbf{X}, \quad (2.65\text{b})$$

$$\mathbf{0}_{r \times r} = \tilde{\mathbf{Q}}_{\text{lo}} \tilde{\mathbf{B}} + \mathbf{Z}_{\text{lo}}^{\top} \mathbf{B}, \quad (2.65\text{c})$$

$$\mathbf{0}_{r \times r} = \tilde{\mathbf{C}} \tilde{\mathbf{P}} - \mathbf{C} \mathbf{X}, \quad (2.65\text{d})$$

where $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}_{\text{lo}} \in \mathbb{R}^{r \times r}$ are the Gramians of (2.55) that satisfy (2.43) and (2.44), while $\mathbf{X}, \mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ satisfy the generalized matrix equations (2.50). Moreover, if $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}_{\text{lo}}$ are nonsingular, then the locally \mathcal{H}_2 -optimal reduced model $\tilde{\mathcal{G}}_{\text{lo}}$ is defined by Petrov-Galerkin projection (2.58), where the model reduction bases $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are given by

$$\mathbf{V} = \mathbf{X} \tilde{\mathbf{P}}^{-1} \quad \text{and} \quad \mathbf{W} = -\mathbf{Z}_{\text{lo}} \tilde{\mathbf{Q}}_{\text{lo}}^{-1}. \quad (2.66)$$

◇

We also mention the work of Hyland and Bernstein [111, 247] that offers conditions in terms of coupled Lyapunov equations. These are equivalent to the Wilson conditions, so we do not include them in our discussion.

Evidently, the optimality conditions in (2.65) depend explicitly on an \mathcal{H}_2 -optimal reduced model. As a necessary consequence, reduced models that satisfy the first-order optimality conditions in (2.65) must be computed using *iterative* algorithms. In [237], Xu and Zheng develop the so-called two-sided iterative algorithm (TSIA). The algorithm performs iteratively-corrected projection using the solutions to the Sylvester equations in (2.50). Computational considerations for TSIA, such as solving the matrix equations in (2.50) and general nonsingular \mathbf{E} , were addressed in [30]. TSIA is written down in Algorithm 2.4.1. In its description, we use $\mathcal{J}(\tilde{\mathcal{G}}_{\text{lo}}^{(i)})$ to denote the \mathcal{H}_2 error of the i -th model iterate.

The Meier-Luenberger (interpolation-based) optimality framework.

Interpolation-based optimality conditions for the \mathcal{H}_2 -minimization problem in (2.61) were derived for single-input, single-output systems by Meier and Luenberger [142]. These were extended to the multiple-input, multiple-output setting, in parallel in [97, 217]. For the sake of simplicity, we restrict our attention to reduced models (2.55) with *simple poles*; generalizations to the setting of higher-order poles can be found in [218]. Under this assumption, recall from Theorem 2.42 that the reduced-order transfer function $\tilde{\mathbf{G}}_{\text{lo}}$ in (2.59) can be expressed in pole-residue form as

$$\tilde{\mathbf{G}}_{\text{lo}}(s) = \sum_{k=1}^r \frac{\mathbf{c}_k \mathbf{b}_k^{\top}}{s - \lambda_k}, \quad \mathbf{c}_k \in \mathbb{C}^p, \quad \mathbf{b}_k \in \mathbb{C}^m. \quad (2.67)$$

Algorithm 2.4.1: Two-sided iterative algorithm (TSIA) [237].

Input: $\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}$ from (2.25), order r ($1 \leq r < n$), tolerance $\tau > 0$, maximum number of iteration steps $M \geq 1$, initial reduced model (2.55) given by $\tilde{\mathbf{E}}^{(0)}, \tilde{\mathbf{A}}^{(0)}, \tilde{\mathbf{B}}^{(0)}, \tilde{\mathbf{C}}^{(0)}$.

Output: $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ —state-space matrices of the converged model (2.55).

- 1 Iteration count $i = 0$.
 - 2 **while** $\mathcal{E}^{(i)} > \tau$ and $i \leq M$ **do**
 - 3 Solve generalized Sylvester equations (2.50) for $\mathbf{X}^{(i)}, \mathbf{Z}_{\text{lo}}^{(i)} \in \mathbb{R}^{n \times r}$:

$$\mathbf{A}^T \mathbf{Z}_{\text{lo}}^{(i)} \tilde{\mathbf{E}}^{(i)} + \mathbf{E}^T \mathbf{Z}_{\text{lo}}^{(i)} \tilde{\mathbf{A}}^{(i)} - \mathbf{C}^T \tilde{\mathbf{C}}^{(i)} = \mathbf{0} \text{ and } \mathbf{A} \mathbf{X}^{(i)} \tilde{\mathbf{E}}^{(i)T} + \mathbf{E} \mathbf{X}^{(i)} \tilde{\mathbf{A}}^{(i)T} + \mathbf{B} \tilde{\mathbf{B}}^{(i)T} = \mathbf{0}.$$
 - 4 Orthonormalize $\mathbf{X}^{(i)}$ and $\mathbf{Z}_{\text{lo}}^{(i)}$ to obtain \mathbf{V} and \mathbf{W} :

$$\mathbf{V} \leftarrow \text{orth}(\mathbf{X}^{(i)}), \quad \mathbf{W} \leftarrow \text{orth}(\mathbf{Z}_{\text{lo}}^{(i)}).$$
 - 5 Compute reduced-order matrices by Petrov-Galerkin projection using \mathbf{V} and \mathbf{W} :

$$\tilde{\mathbf{E}}^{(i+1)} = \mathbf{W}^T \mathbf{E} \mathbf{V}, \quad \tilde{\mathbf{A}}^{(i+1)} = \mathbf{W}^T \mathbf{A} \mathbf{V}, \quad \tilde{\mathbf{B}}^{(i+1)} = \mathbf{W}^T \mathbf{B}, \quad \tilde{\mathbf{C}}^{(i+1)} = \mathbf{C} \mathbf{V}.$$
 - 6 Compute the normalized \mathcal{H}_2 distance between model iterates:

$$\mathcal{E}^{(i+1)} = \frac{\left| \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lo}}^{(i)} \right) - \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lo}}^{(i+1)} \right) \right|}{\mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lo}}^{(0)} \right)}.$$
 - 7 Set $i \leftarrow i + 1$.
 - 8 **end**
-

The rank-1 matrices $\mathbf{c}_k \mathbf{b}_k^T \in \mathbb{C}^{p \times m}$ are the *residues* of $\tilde{\mathbf{G}}_{\text{lo}}$ corresponding to the pole at $s = \lambda_i$. We call the vectors $\mathbf{c}_k \in \mathbb{C}^p$ and $\mathbf{b}_k \in \mathbb{C}^m$ the *residue directions*. If the full-order transfer function (2.30) also has simple poles μ_1, \dots, μ_n , the squared \mathcal{H}_2 model reduction error $\|\mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2$ is given by

$$\|\mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^n \delta_i^T \left(\mathbf{G}_{\text{lo}}(-\mu_i) - \tilde{\mathbf{G}}_{\text{lo}}(-\mu_i) \right) \boldsymbol{\beta}_i - \sum_{i=1}^r \mathbf{c}_i^T \left(\mathbf{G}_{\text{lo}}(-\lambda_i) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_i) \right) \mathbf{b}_i, \quad (2.68)$$

where $\boldsymbol{\beta}_i \in \mathbb{C}^m$ and $\delta_i \in \mathbb{C}^p$ are the residue directions of \mathbf{G}_{lo} corresponding to $s = \mu_i$. Using the error expression in (2.68), one can arrive at interpolation-based necessary conditions for \mathcal{H}_2 optimality by taking appropriately defined perturbations of the reduced model poles and residues [23, 97], [5, Theorem 5.1.1]. These are summarized in the following result.

Algorithm 2.4.2: Iterative rational Krylov algorithm (IRKA) [97].

Input: $\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}$ from (2.25), order r ($1 \leq r < n$), tolerance $\epsilon > 0$, maximum number of iteration steps $M \geq 1$, initial interpolation points $\lambda_1^{(0)}, \dots, \lambda_r^{(0)} \in \mathbb{C}$, and directions $\mathbf{b}_1^{(0)}, \dots, \mathbf{b}_r^{(0)} \in \mathbb{C}^m$, $\mathbf{c}_1^{(0)}, \dots, \mathbf{c}_r^{(0)} \in \mathbb{C}^p$, closed under complex conjugation such that $\lambda_k^{(0)} \mathbf{E} - \mathbf{A}$ is invertible for all $k = 1, \dots, r$.

Output: $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ —state-space matrices of the converged model (2.55).

1 Iteration count $i = 0$.

2 **while** $\max_j |\lambda_j^{(i+1)} - \lambda_j^{(i)}| > \epsilon$ *and* $i \leq M$ **do**

3 Compute interpolatory model reduction bases $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times r}$ according to Theorem 2.45 such that

$$\left(\lambda_k^{(i)} \mathbf{E} - \mathbf{A} \right)^{-1} \mathbf{B} \mathbf{b}_k^{(i)} \in \text{Range}(\mathbf{V}) \quad \text{and} \quad \left(\lambda_k^{(i)} \mathbf{E}^\top - \mathbf{A}^\top \right)^{-1} \mathbf{C}^\top \mathbf{c}_k^{(i)} \in \text{Range}(\mathbf{W}).$$

4 Orthonormalize bases \mathbf{V} and \mathbf{W} :

$$\mathbf{V} \leftarrow \text{orth}(\mathbf{V}), \quad \mathbf{W} \leftarrow \text{orth}(\mathbf{W}).$$

5 Compute reduced-order matrices by Petrov-Galerkin projection:

$$\tilde{\mathbf{E}}^{(i+1)} = \mathbf{W}^\top \mathbf{E} \mathbf{V}, \quad \tilde{\mathbf{A}}^{(i+1)} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \tilde{\mathbf{B}}^{(i+1)} = \mathbf{W}^\top \mathbf{B}, \quad \tilde{\mathbf{C}}^{(i+1)} = \mathbf{C} \mathbf{V}.$$

6 Compute $\lambda_k^{(i+1)} \in \mathbb{C}$ and $\mathbf{b}_k^{(i+1)} \in \mathbb{C}^m$, $\mathbf{c}_k^{(i+1)} \in \mathbb{C}^p$ in (2.67) from the eigendecomposition of $s \tilde{\mathbf{E}} - \tilde{\mathbf{A}}$ and set $i \leftarrow i + 1$.

7 **end**

Theorem 2.45 (Interpolation-based \mathcal{H}_2 -optimality conditions for linear systems [23, 97, 142, 217]). Suppose that \mathcal{G}_{lo} and $\tilde{\mathcal{G}}_{\text{lo}}$ are asymptotically linear systems as in (2.25) and (2.55) with the transfer functions \mathbf{G}_{lo} and $\tilde{\mathbf{G}}_{\text{lo}}$ defined according to (2.30), and that $\tilde{\mathcal{G}}_{\text{lo}}$ has simple poles $\lambda_1, \dots, \lambda_r$. Let $\mathbf{b}_i \in \mathbb{C}^m$ and $\mathbf{c}_i \in \mathbb{C}^p$ be the corresponding residue directions in (2.67). If $\tilde{\mathcal{G}}_{\text{lo}}$ minimizes the \mathcal{H}_2 error in (2.61), then $\tilde{\mathcal{G}}_{\text{lo}}$ satisfies the tangential interpolation conditions:

$$\begin{aligned} \mathbf{G}_{\text{lo}}(-\lambda_k) \mathbf{b}_k &= \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \mathbf{b}_k, \\ \mathbf{c}_k^\top \mathbf{G}_{\text{lo}}(-\lambda_k) &= \mathbf{c}_k^\top \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k), \\ \text{and } \mathbf{c}_k^\top \frac{d}{ds} \mathbf{G}_{\text{lo}}(-\lambda_k) \mathbf{b}_k &= \mathbf{c}_k^\top \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \mathbf{b}_k. \end{aligned} \tag{2.69}$$

for all $k = 1, \dots, r$. ◇

In other words: The best linear system approximation is a bi-tangential Hermite interpolant and tangential Lagrange interpolant at the *mirror images of the reduced model poles*. The

corresponding tangential directions are the *residue directions* of the reduced model transfer function. Given the optimal interpolation data, one could compute a reduced model using the interpolatory bases in Theorem 2.43 with $\sigma_k = -\lambda_k$, $\mathbf{r}_k = \mathbf{b}_k$, and $\boldsymbol{\ell}_k = \mathbf{c}_k$ for all $k = 1, \dots, r$. As with the Wilson framework, the interpolation-based optimality conditions of Theorem 2.45 depend explicitly on the \mathcal{H}_2 -optimal reduced-order transfer function, which is not known *a priori*. To deal with this, Gugercin et al. [97] proposed the *iterative rational Krylov algorithm* (IRKA) for automatically determining the optimal interpolation data; the method is presented in Algorithm 2.4.2. The algorithm performs iteratively-corrected projection using the poles and residue directions of the previous reduced model iterate; it repeats until the poles stop changing, so that the interpolatory optimality conditions in (2.69) will be satisfied. It was shown recently in [147] that IRKA can be viewed as a Riemannian gradient descent method with a fixed step size; this perspective can be leveraged to guarantee convergence of the method under certain conditions. The works [148, 149, 150] establish interpolatory optimality conditions similar to (2.69) for more general settings, such as structured or parameter-dependent linear systems.

2.4.3 Balanced truncation model reduction

Balanced truncation model reduction [151, 152] and its variants [46, 61, 69, 92, 96, 114, 159, 244] are the gold standard for linear system approximation. The allure of balancing-based model reduction stems from the fact that these methods (i) preserve desirable qualitative features of the full-order system, e.g., asymptotic stability or passivity, and (ii) often provide error bounds. The original, so-called *Lyapunov* balanced truncation was first introduced by Mullis and Roberts [152], and later in the systems and control literature by Moore [151]. Popular variations of the original method are proposed in, e.g. [61, 69, 92, 114]. The essential ingredients to any balancing-based model reduction are the system Gramians; classically, these are the (infinite) reachability and observability Gramians introduced in Definition 2.34. The infinite Gramians satisfy generalized algebraic Lyapunov equations (2.43) and (2.44); in variants of balanced truncation, the Gramians satisfy generalized algebraic *Riccati* equations. Gramians define energy functionals that are thereby used to identify which system states contribute only marginally to the input-to-output dynamics of a system (2.25) in a pre-determined sense, i.e., those corresponding to small or large energies. Such states are deemed insignificant and can be truncated to reduce the model order.

Here, we review the basics behind Lyapunov balanced truncation [151, 152]. Other variants of balanced truncation—namely, those proposed in [61, 69, 92, 114, 181]—that are the focus of the results of Chapter 4 are presented therein. Portions of this discussion are taken from the author’s previous work [182].

The concept of balancing and Lyapunov balanced truncation

Recall the concepts of reachability and observability as well as the infinite Gramians of Section 2.3.4. The original formulation of balanced truncation model reduction theory from [151, 152] is based on the *reachability* and *observability energy functionals* $\ell_r: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ and $\ell_o: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ of the linear system (2.25), which are defined as

$$\begin{aligned}\ell_r(\tilde{\mathbf{x}}) &\stackrel{\text{def}}{=} \min_{\mathbf{x}(-\infty)=\mathbf{0}_n, \mathbf{x}(0)=\tilde{\mathbf{x}}} \frac{1}{2} \int_{-\infty}^0 \|\mathbf{u}(\tau)\|_2^2 d\tau, \\ \ell_o(\tilde{\mathbf{x}}) &\stackrel{\text{def}}{=} \frac{1}{2} \int_0^{\infty} \|\mathbf{y}(\tau)\|_2^2 d\tau, \quad \text{where } \mathbf{x}(0) = \tilde{\mathbf{x}}, \quad \mathbf{u}(t) = \mathbf{0}_m.\end{aligned}\tag{2.70}$$

For a point $\tilde{\mathbf{x}} \in \mathbb{R}^n$, $\ell_r(\tilde{\mathbf{x}})$ is the minimal energy required to drive the dynamics in (2.25) from $\mathbf{x}(-\infty) = \mathbf{0}_n$ to $\mathbf{x}(0) = \tilde{\mathbf{x}}$ using a finite energy input \mathbf{u} ; $\ell_o(\tilde{\mathbf{x}})$ is the energy produced by the system (2.25) with initial condition $\mathbf{x}(0) = \tilde{\mathbf{x}}$ under zero forcing $\mathbf{u}(t) = \mathbf{0}_m$. States that are “difficult to reach” are those that correspond to a large reachability energy $\ell_r(\tilde{\mathbf{x}})$, whereas states that are “difficult to observe” are those that correspond to a small observability energy $\ell_o(\tilde{\mathbf{x}})$. The cost functionals in (2.70) serve as a quantifiable characterization for how much each component of the state-space contributes to the input-to-output dynamics of the system. States that are difficult to reach or difficult to observe are expected to contribute little and can thus be truncated to reduce the model order with little consequence.

We already know that the reachability and observability Gramians of a system (2.25) characterize its reachable and unobservable subspaces from Lemma 2.36. Significantly, the infinite Gramians $\mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{E}^\top \mathbf{Q}_o \mathbf{E} \in \mathbb{R}^{n \times n}$ in (2.39) and (2.40) characterize the energy functionals in (2.70), as well.

Lemma 2.46 (Reachability and observability energy functionals [4, Lemma 4.29]). Consider an asymptotically stable, minimal linear system \mathcal{G}_o in (2.25) and let $\mathbf{P}, \mathbf{E}^\top \mathbf{Q}_o \mathbf{E} \in \mathbb{R}^{n \times n}$ be the reachability and observability Gramians of \mathcal{G}_o defined according to (2.39) and (2.40). Then, for any $\tilde{\mathbf{x}} \in \mathbb{R}^n$

$$\ell_r(\tilde{\mathbf{x}}) = \frac{1}{2} \tilde{\mathbf{x}}^\top \mathbf{P}^{-1} \tilde{\mathbf{x}} \quad \text{and} \quad \ell_o(\tilde{\mathbf{x}}) = \frac{1}{2} \tilde{\mathbf{x}}^\top \mathbf{E}^\top \mathbf{Q}_o \mathbf{E} \tilde{\mathbf{x}},$$

where ℓ_r and ℓ_o are the reachability and observability energy functionals defined in (2.70). \diamond

Lemma 2.46 says that (ii) states which are *difficult to reach* are those that are well approximated in the span of the eigenvectors of \mathbf{P} corresponding to *small* eigenvalues, and (ii) states which are *difficult to observe* are those that are well approximated in the span of the eigenvectors of $\mathbf{E}^\top \mathbf{Q}_o \mathbf{E}$ corresponding to *small* eigenvalues. Thus, one can reduce the model order by truncating states that are weakly reachable or weakly observable in this sense. However, the controllability and observability energies of a state are *basis dependent* quantities. Indeed, under a state-space transformation given by $\mathbf{z} = \mathbf{T}\mathbf{x}$ for $\mathbf{T} \in \mathbb{R}^{n \times n}$, the

matrices $\check{\mathbf{P}}$ and $\check{\mathbf{Q}}_{\text{lo}}$ that define the system Gramians obey the *contragradient* transformation laws

$$\check{\mathbf{P}} = \mathbf{T}\mathbf{P}\mathbf{T}^\top \quad \text{and} \quad \check{\mathbf{Q}}_{\text{lo}} = \mathbf{T}^{-\top}\mathbf{Q}_{\text{lo}}\mathbf{T}^{-1}. \quad (2.71)$$

Hence, states that are difficult to reach might simultaneously be easy to observe and vice versa, depending on the basis used to realize the Gramians; see [4, Ex. 7.1] for an illustration of this dilemma. It is thus natural to ask: Does there exist a basis of \mathbb{R}^n in which states that are difficult to reach are simultaneously difficult to observe? This is the fundamental question behind the concept of *balancing* and *balanced truncation model reduction*.

Definition 2.47 (Balancing and Hankel singular values [4, Definition 7.2]). Consider an asymptotically stable, minimal linear system \mathcal{G}_{lo} in (2.25) and let \mathbf{P} and $\mathbf{E}^\top\mathbf{Q}_{\text{lo}}\mathbf{E} \in \mathbb{R}^{n \times n}$ be the reachability and observability Gramians of \mathcal{G}_{lo} defined according to (2.39) and (2.40). We say that the system \mathcal{G}_{lo} is *balanced* if

$$\mathbf{P} = \mathbf{E}^\top\mathbf{Q}_{\text{lo}}\mathbf{E} = \Sigma \stackrel{\text{def}}{=} \text{diag}(\sigma_1\mathbf{I}_{m_1}, \sigma_2\mathbf{I}_{m_2}, \dots, \sigma_q\mathbf{I}_{m_q}), \quad (2.72)$$

where $\sigma_1 > \sigma_2 > \dots > \sigma_q$ are called the *Hankel singular values* of \mathcal{G}_{lo} , and their multiplicities m_k satisfy $m_1 + m_2 + \dots + m_q = n$. \diamond

We call the realization of a system (2.25) in the coordinate system where (2.72) a *balanced realization* of the system.

Remark 2.48. From a linear algebraic point of view, *balancing* means the *simultaneous diagonalization of two symmetric positive (semi) definite matrices* [4, Remark 7.1.1]. \diamond

Remark 2.49. The Hankel singular values are often introduced in the literature as the square roots of the eigenvalues of the products of the system Gramians, i.e.,

$$\sigma_i = \sqrt{\lambda_i(\mathbf{P}\mathbf{E}^\top\mathbf{Q}_{\text{lo}}\mathbf{E})}, \quad i = 1, \dots, n,$$

counted with multiplicity. While these definitions are equivalent, the one above reveals that the eigenvalues are *system invariants* because, under the change of coordinate transformation $\mathbf{z} = \mathbf{T}^{-1}\mathbf{x}$ by $\mathbf{T} \in \mathbb{R}^{n \times n}$, it holds that $\check{\mathbf{P}}\mathbf{E}^\top\check{\mathbf{Q}}_{\text{lo}}\mathbf{E} = \mathbf{T}\mathbf{P}\mathbf{E}^\top\mathbf{Q}_{\text{lo}}\mathbf{E}\mathbf{T}^{-1}$. In other words, *equivalent systems have the same Hankel singular values*. In fact, these are the singular values of the so-called *Hankel operator* of a linear system (2.25); see [4, Section 5.4] for further discussion. \diamond

The subsequent result states that every asymptotically stable and minimal linear system (2.25) is equivalent to a balanced system according to Definition 2.47.

Theorem 2.50 (Balancing transformation [4, Section 7.3]). Consider an asymptotically stable, minimal linear system \mathcal{G}_{lo} in (2.25) and let $\mathbf{P}, \mathbf{E}^\top\mathbf{Q}_{\text{lo}}\mathbf{E} \in \mathbb{R}^{n \times n}$ be the reachability

and observability Gramians of \mathcal{G}_{lo} defined according to (2.39) and (2.40). Let $\mathbf{R}, \mathbf{Q}_{\text{lo}} \in \mathbb{R}^{n \times n}$ be Cholesky factors of \mathbf{P} and \mathbf{Q}_{lo} so that

$$\mathbf{P} = \mathbf{R}\mathbf{R}^\top \quad \text{and} \quad \mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} = \mathbf{L}_{\text{lo}} \mathbf{L}_{\text{lo}}^\top.$$

Then, a balancing transformation that achieves (2.72) via (2.32) is given by

$$\mathbf{T} = \mathbf{\Sigma}^{-1/2} \mathbf{U}^\top \mathbf{L}_{\text{lo}}^\top \quad \text{and} \quad \mathbf{S} = \mathbf{T}^{-1} = \mathbf{R}\mathbf{Y}\mathbf{\Sigma}^{-1/2}, \quad (2.73)$$

where $\mathbf{L}_{\text{lo}}^\top \mathbf{E} \mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{Y}^\top$ is the singular value decomposition of $\mathbf{L}_{\text{lo}}^\top \mathbf{E} \mathbf{R} \in \mathbb{R}^{n \times n}$. Moreover, the singular values $\mathbf{\Sigma}$ of $\mathbf{L}_{\text{lo}}^\top \mathbf{E} \mathbf{R}$ are the Hankel singular values of \mathcal{G}_{lo} . \diamond

Proof of Theorem 2.50. Assume without loss of generality that $\mathbf{E} = \mathbf{I}_n$. By multiplying $\mathbf{T}\mathbf{S} = \mathbf{I}_n$, it is readily seen that $\mathbf{T} = \mathbf{S}^{-1}$ in (2.73). From (2.71), the transformed Gramians satisfy

$$\check{\mathbf{P}} = \mathbf{\Sigma}^{-1/2} \mathbf{U}^\top \mathbf{L}_{\text{lo}}^\top (\mathbf{R}\mathbf{R}^\top) \mathbf{L}_{\text{lo}} \mathbf{U} \mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1/2} \mathbf{U}^\top (\mathbf{U}\mathbf{\Sigma}\mathbf{Y}^\top) (\mathbf{U}\mathbf{\Sigma}\mathbf{Y}^\top)^\top \mathbf{U} \mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}.$$

Likewise, it follows readily that $\check{\mathbf{Q}}_{\text{lo}} = \mathbf{S}^\top \mathbf{Q}_{\text{lo}} \mathbf{S} = \mathbf{\Sigma}$, and that the singular values of $\mathbf{L}_{\text{lo}}^\top \mathbf{R}$ are the Hankel singular values of \mathcal{G}_{lo} . \square

It is obvious that, because the Gramians are equal and diagonal in a balanced basis, states that are weakly reachable are simultaneously weakly observable in this basis. Moreover, the i -th Hankel singular value σ_i serves as a precise characterization of how easily a point in state space $\check{\mathbf{x}} \in \mathbb{R}^n$ is to reach and observe. Indeed, in balanced coordinates (2.72), Lemma 2.46 reveals that states that are difficult to reach and difficult to observe are exactly those corresponding to the smallest Hankel singular values. The model order is thereby reduced by simply truncating the trailing $n - r$ components of the state space; this reduction can be expressed compactly by writing the system in block form, as given in the following theorem. The asymptotic stability result is due to Pernebo and Silverman [169], while the error bound (2.77) is due to Enns [69].

Theorem 2.51 (Balanced truncation model reduction [169, Theorem 3.2], [69]). Consider an asymptotically stable, minimal, and balanced linear system \mathcal{G}_{lo} in (2.25) having the balanced realization

$$\mathbf{E} = \mathbf{I}_n, \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \text{and} \quad [\mathbf{C}_1 \quad \mathbf{C}_2], \quad (2.74)$$

where the state-space matrices are partitioned according to the system Gramians $\mathbf{P} = \mathbf{Q}_{\text{lo}} = \mathbf{\Sigma} = \text{diag}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$, with

$$\mathbf{\Sigma}_1 = \text{diag}(\sigma_1 \mathbf{I}_{m_1}, \dots, \sigma_r \mathbf{I}_{m_k}) \quad \text{and} \quad \mathbf{\Sigma}_2 = \text{diag}(\sigma_{k+1} \mathbf{I}_{m_{r+1}}, \dots, \sigma_q \mathbf{I}_{m_q})$$

for $r = m_1 + \dots + m_k$. Then, the order- r reduced model

$$\tilde{\mathcal{G}}_{\text{lo,bt}} : \begin{cases} \dot{\mathbf{x}}_1(t) &= \tilde{\mathbf{A}}_{11}\mathbf{x}_1(t) + \tilde{\mathbf{B}}_1\mathbf{u}(t), \\ \tilde{\mathbf{y}}_{\text{lo,bt}}(t) &= \tilde{\mathbf{C}}_1\mathbf{x}_1(t) + \tilde{\mathbf{D}}\mathbf{u}(t), \end{cases} \quad (2.75)$$

obtained via balanced truncation and having the transfer function

$$\tilde{\mathcal{G}}_{\text{lo,bt}}(s) = \tilde{\mathbf{C}}_1(s\mathbf{I}_r - \tilde{\mathbf{A}}_{11})^{-1}\tilde{\mathbf{B}}_1 + \tilde{\mathbf{D}}, \quad (2.76)$$

is balanced, asymptotically stable, and satisfies the error bound

$$\|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,bt}}\|_{\mathcal{H}_\infty} \leq 2(\sigma_{k+1} + \dots + \sigma_q). \quad (2.77)$$

◇

By Theorem 2.51, if the trailing $q - (k + 1)$ Hankel singular values are small, then the \mathcal{H}_∞ model error, and thus the \mathcal{L}_2 output error by (2.56), is guaranteed to be small as well. We emphasize that (2.77) is an *a priori* error bound; in other words, one can compute the bound (2.77) for an order of reduction r without having to compute the corresponding BT reduced model. In a practical implementation, balancing and truncation are performed simultaneously. This approach, called the *square-root algorithm* for balanced truncation [124, 208], [4, Section 7.4], is presented in Algorithm 2.4.3. The dominant cost of the algorithm is solving the pair of generalized Lyapunov equations in (2.43) and (2.44). This can be done exactly using the Bartels-Stewart Algorithm [18], although this requires dense matrix operations and has $O(n^3)$ complexity, which is typically infeasible for systems of order $n \gtrsim 10^5$. Fortunately, the singular values of the system Gramians tend to decay quickly due to the low-rank right-hand sides in (2.43) and (2.44); see [7, 15, 168]. Thus, the exact Cholesky factors in Algorithm 2.4.3 can be replaced by approximate, *low-rank* factors $\tilde{\mathbf{R}} \in \mathbb{R}^{n \times k_1}$, $\tilde{\mathbf{L}}_{\text{lo}} \in \mathbb{R}^{n \times k_2}$ where $\mathbf{P} \approx \tilde{\mathbf{R}}\tilde{\mathbf{R}}^\top$, $\mathbf{Q}_{\text{lo}} \approx \tilde{\mathbf{L}}_{\text{lo}}\tilde{\mathbf{L}}_{\text{lo}}^\top$ and $k_1, k_2 \ll n$. These low-rank factors can be computed very efficiently using sparse matrix operations; see, e.g. [34, 121, 203].

Singular perturbation balancing.

A model reduction technique related to balanced truncation is the *singular perturbation approximation* [72, 73, 133, 134]. Let

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \quad \text{where } \mathbf{x}_1 \in \mathbb{R}^r, \quad \mathbf{x}_2 \in \mathbb{R}^{n-r},$$

be the state vector of a balanced linear system (2.25) that satisfies the hypotheses of Theorem 2.51. Because the system is balanced, we have $\mathbf{E} = \mathbf{I}_n$. Balanced truncation as described by Theorem 2.51 corresponds to directly truncating the $n - r$ trailing state components, i.e., setting $\mathbf{x}_2(t) = \mathbf{0}_{n-r}$ for all $t \geq 0$. By contrast, the singular perturbation approximation

Algorithm 2.4.3: Square-root algorithm for Lyapunov balanced truncation [4].

Input: E, A, B, C, D from (2.25), order r ($1 \leq r < n$).

Output: $\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}$ —state-space matrices of (2.55).

- 1 Compute Cholesky factors $R \in \mathbb{R}^{n \times n}$ and $L_{\text{lo}} \in \mathbb{R}^{n \times n}$ of $P \in \mathbb{R}^{n \times n}$ in (2.39) and $Q_{\text{lo}} \in \mathbb{R}^{n \times n}$ in (2.40) from the generalized Lyapunov equations (2.43) and (2.44).
- 2 Compute the singular value decomposition of $L_{\text{lo}}^T E R$ partitioned according to

$$L_{\text{lo}}^T E R = U \Sigma Y^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix},$$

for $\Sigma_1 \in \mathbb{R}^{r \times r}$, $U_1, Y_1 \in \mathbb{R}^{n \times r}$ and $\Sigma_2 \in \mathbb{R}^{(n-r) \times (n-r)}$, $U_2, Y_2 \in \mathbb{R}^{n \times (n-r)}$.

- 3 Compute the model reduction bases $W, V \in \mathbb{R}^{n \times r}$ as

$$W = L_{\text{lo}} U_1 \Sigma_1^{-1/2}, \quad V = R Y_1 \Sigma_1^{-1/2}.$$

- 4 Compute the reduced model $\tilde{\mathcal{G}}_{\text{lo}}$ by projection (2.58) using W and V :

$$\begin{aligned} \tilde{E} &= I_r, \\ \tilde{A} &= \Sigma_1^{-1/2} U_1^T (L_{\text{lo}}^T A R) Y_1 \Sigma_1^{-1/2}, \\ \tilde{B} &= \Sigma_1^{-1/2} U_1^T (L_{\text{lo}}^T B), \\ \tilde{C} &= (C R) Y_1 \Sigma_1^{-1/2}, \\ \text{and } \tilde{D} &= D. \end{aligned} \tag{2.78}$$

instead sets $\dot{\mathbf{x}}_2(t) = \mathbf{0}_{n-r}$ for all $t \geq 0$. Under this assumption, the trailing $n - r$ states in the resulting reduced model (2.55) can be rearranged to solve explicitly for \mathbf{x}_2 :

$$\mathbf{0}_{n-r} = \mathbf{A}_{21} \mathbf{x}_1(t) + \mathbf{A}_{22} \mathbf{x}_2(t) + \mathbf{B}_2 \mathbf{u}(t) \quad \implies \quad \mathbf{x}_2(t) = -\mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{x}_1(t) - \mathbf{B}_2 \mathbf{u}(t),$$

where \mathbf{A}_{22} is invertible by asymptotic stability. Substituting this expression for \mathbf{x}_2 in the equations for \mathbf{x}_1 leads to the *balanced singular perturbation approximation*

$$\tilde{\mathcal{G}}_{\text{lo,spa}} : \begin{cases} \dot{\mathbf{x}}_1(t) &= \tilde{\mathbf{A}}_{\text{spa}} \mathbf{x}_1(t) + \tilde{\mathbf{B}}_{\text{spa}} \mathbf{u}(t), \\ \tilde{\mathbf{y}}_{\text{lo,spa}}(t) &= \tilde{\mathbf{C}}_{\text{spa}} \mathbf{x}_1(t) + \tilde{\mathbf{D}}_{\text{spa}} \mathbf{u}(t), \end{cases} \tag{2.79}$$

where $\tilde{\mathbf{A}}_{\text{spa}} \in \mathbb{R}^{r \times r}$, $\tilde{\mathbf{B}}_{\text{spa}} \in \mathbb{R}^{r \times m}$, $\tilde{\mathbf{C}}_{\text{spa}} \in \mathbb{R}^{p \times r}$, and $\tilde{\mathbf{D}}_{\text{spa}} \in \mathbb{R}^{p \times m}$ are defined as

$$\begin{aligned}\tilde{\mathbf{A}}_{\text{spa}} &\stackrel{\text{def}}{=} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}), \\ \tilde{\mathbf{B}}_{\text{spa}} &\stackrel{\text{def}}{=} (\mathbf{B}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{B}_2), \\ \tilde{\mathbf{C}}_{\text{spa}} &\stackrel{\text{def}}{=} (\mathbf{C}_1 - \mathbf{C}_2\mathbf{A}_{22}^{-1}\mathbf{A}_{21}), \\ \tilde{\mathbf{D}}_{\text{spa}} &\stackrel{\text{def}}{=} (\mathbf{D} - \mathbf{C}_2\mathbf{A}_{22}^{-1}\mathbf{B}_2).\end{aligned}\tag{2.80}$$

The above approximation can be implemented for any coordinate system; it need not be balanced. When the original system is balanced, however, the *balanced singular perturbation approximation* (2.79) has the following properties.

Theorem 2.52 (Singular perturbation balancing [73]). Consider an asymptotically stable, minimal, and balanced linear system \mathcal{G}_{lo} in (2.25) having the balanced realization in (2.74) for $r = m_1 + \dots + m_k$. Then, the order- r reduced model $\tilde{\mathcal{G}}_{\text{lo,spa}}$ in (2.79) obtained via the balanced singular perturbation approximation is balanced, asymptotically stable, and satisfies the error bound (2.77). \diamond

One interesting additional property of the singular perturbation approximation is that it interpolates the full-order model at $s = 0$, while the reduced model obtained via balanced truncation interpolates at $s = \infty$. A square-root or low-rank implementation of the balanced singular perturbation approximation can also be achieved; see [127] for details.

Chapter 3

On the balanced truncation error bound

Equality is known to hold in the balanced truncation \mathcal{H}_∞ error bound (2.77) when only *one* Hankel singular value is truncated [69], i.e., $\Sigma_2 = \sigma_q \mathbf{I}_{m_q}$ in Theorem 2.51 with $r = n - m_q$, or when the full-order model \mathcal{G}_\circ is single-input, single-output and *state-space symmetric* [131], i.e., \mathcal{G}_\circ has a realization (2.25) satisfying $\mathbf{E} = \mathbf{I}_n$, $\mathbf{A} = \mathbf{A}^\top$, $\mathbf{B} = \mathbf{C}^\top$ with $m = p = 1$. In this chapter, we establish more general conditions for which the bound (2.77) holds with equality, thereby providing an *exact* formula for the error in the reduction.

The results of this chapter, as well as parts of the discussion, are taken from the author’s previous work [182].

[182] Reiter, S., Damm, T., Embree, M., and Gugercin, S. (2024a). [On the balanced truncation error bound and sign parameters from arrowhead realizations](#). *Advances in Computational Mathematics*, 50(1):10.

A preliminary version of the work in this chapter is also available in the author’s M. S. thesis [189].

First, we provide an example of a system from power systems modeling that motivated the initial investigation and led to the results in this section.

3.1 A motivating example from power systems modeling

A common technique for modeling the frequency response of an electrical power network with *coherent* generators—that is, generators that swing in a synchronized fashion in response to excitations—is to aggregate them into a single effective machine. It is shown in [144, 145] that for a network of $n - 1$ coherent generators modeled by the *swing equations* with first-order turbine control, the aggregate frequency dynamics are approximated well by an order- n

single-input, single-output linear system \mathcal{G}_{lo} as in (2.25) having the transfer function

$$\mathbf{G}_{\text{lo}}(s) = \frac{1}{\check{m} + \check{d} + \sum_{i=1}^{n-1} \frac{r_i^{-1}}{\tau_i s + 1}}. \quad (3.1)$$

Here, $\check{m} = \sum_{i=1}^{n-1} m_i \in \mathbb{R}$ and $\check{d} = \sum_{i=1}^{n-1} d_i \in \mathbb{R}$ denote the aggregate inertia and damping coefficients of the generators in the network, while $\tau_i \in \mathbb{R}$ and $r_i^{-1} \in \mathbb{R}$ denote the time constant and droop control coefficient of the i -th generator, $i = 1 \dots, n-1$. For the theoretical justification that \mathbf{G}_{lo} as defined in (3.1) sufficiently approximates the network response, see [144, Sec. 2].

There is a natural coordinate system for the model described by (3.1) in which its dynamics are described by simple expressions involving the physical parameters of the network; this is obtained by applying [144, Equation 1] to the dynamics [161, p. 3009, Example 2] and simplifying. In these coordinates, the realization of \mathcal{G}_{lo} having the transfer function (3.1) has an *arrowhead form*:

$$\mathbf{A} = \begin{bmatrix} -\check{d} & \sqrt{\frac{r_1^{-1}}{\check{m}\tau_1}} & \cdots & \sqrt{\frac{r_{n-1}^{-1}}{\check{m}\tau_{n-1}}} \\ -\sqrt{\frac{r_1^{-1}}{\check{m}\tau_1}} & -\frac{1}{\tau_1} & & \\ \vdots & & \ddots & \\ -\sqrt{\frac{r_{n-1}^{-1}}{\check{m}\tau_{n-1}}} & & & -\frac{1}{\tau_{n-1}} \end{bmatrix}, \quad \mathbf{B} = \frac{1}{\sqrt{\check{m}}} \mathbf{e}_1, \quad \mathbf{C} = \frac{1}{\sqrt{\check{m}}} \mathbf{e}_1^\top. \quad (3.2)$$

Apart from its first row, first column, and main diagonal, all of the entries of \mathbf{A} are zero. Note that \mathbf{A} in (3.2) is *not* state-space symmetric; rather, the realization satisfies a particular sign symmetry condition $\mathbf{A} = \mathbf{S} \mathbf{A}^\top \mathbf{S}$ and $\mathbf{B} = \mathbf{S} \mathbf{C}^\top$, where $\mathbf{S} \stackrel{\text{def}}{=} \text{diag}(+1, -1, \dots, -1) \in \mathbb{R}^{n \times n}$. While applying balanced truncation in Algorithm 2.4.3 to this model problem, we observed that the \mathcal{H}_∞ error bound (2.77) was *tight*—that is, the bound held with equality—for *all orders of reduction*. This motivated us to investigate whether the balanced truncation error bound holds with equality for a more general class of systems than those that are state-space symmetric. Ultimately, we prove that the bound (2.77) will hold with equality when the truncated part of the model in (2.74) is a slight generalization of state-space symmetric; this is a weaker condition than the state-space symmetric case, where the truncated part of the model is always state-space symmetric.

3.2 The sign symmetry of balanced realizations

We start our discussion by describing an important sign symmetry property of balanced systems originally characterized in [157, 232], which will be essential for the main result of this section.

Theorem 3.1 (Canonical balanced form and sign parameters [4, Sec. 7.4]). Suppose that \mathcal{G}_\circ is an asymptotically stable and minimal single-input, single-output linear system as in (2.25). Then, \mathcal{G}_\circ has a balanced realization satisfying the sign-symmetric form

$$\mathbf{E} = \mathbf{I}_n, \quad \mathbf{A} = \mathbf{S}\mathbf{A}^\top\mathbf{S} \quad \text{and} \quad \mathbf{B} = (\mathbf{C}\mathbf{S})^\top, \quad (3.3)$$

where $\mathbf{S} = \text{diag}(s_1, s_2, \dots, s_n)$ and $s_i \in \{\pm 1\}$ for $i = 1, \dots, n$ are the *sign parameters* of the linear system \mathcal{G}_\circ . Moreover, the sign parameters s_i correspond to σ_j for some $j = 1, \dots, q$, and are ordered such that the associated Hankel singular values are non-increasing. \diamond

Like the Hankel singular values, the sign parameters are system invariants. Ober [156, 157] shows that every asymptotically stable, minimal, single-input, single-output linear system is equivalent to a balanced system with a realization satisfying (3.3), which we call the *canonical form* of a balanced system. In the case of distinct Hankel singular values, i.e., $q = n$, we can explicitly apply the formula [4, Eq. (7.24)] to construct the canonical form of \mathcal{G}_\circ having the state-space matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} as

$$\mathbf{A}_{ij} = \frac{-\gamma_i\gamma_j}{s_i s_j \sigma_i + \sigma_j}, \quad \mathbf{B}_i = \gamma_i, \quad \mathbf{C}_i = s_i \gamma_i, \quad i, j = 1, \dots, n. \quad (3.4)$$

In fact, [157] shows that all single-input, single-output balanced systems are parameterized by the distinct Hankel singular values $\sigma_1, \dots, \sigma_n > 0$, the signs $s_1, \dots, s_n \in \{\pm 1\}$, and the entries $\gamma_1, \dots, \gamma_n$ of $\mathbf{B} \in \mathbb{R}^n$. A similar formula holds for systems with repeated Hankel singular values; see [4, Eq. (7.26)]. We include this formula to emphasize that the sign symmetry (3.3) is not merely an artifact of the canonical balanced realization; these signs arise explicitly in the parameterization of all balanced linear systems [157]. Consequently, multiple sign parameters can correspond to the same Hankel singular value, as is the case with repeated Hankel singular values; see [136, Section 2.7] for specific details. As we will show in this section, the importance of these sign parameters and, equivalently, the sign-symmetry of the canonical balanced form (3.3), is that they provide sufficient conditions for determining whether the \mathcal{H}_∞ error bound (2.77) holds with equality. We will also show how to determine these sign parameters from *any* (not necessarily balanced) realization of \mathcal{G}_\circ satisfying the sign-symmetry structure (3.3).

Remark 3.2. Any balanced realization of a system \mathcal{G}_\circ is unique up to orthogonal transformations of the state space [151], permitting multiple balanced realizations of \mathcal{G}_\circ that obey the same sign symmetry as (3.3) but with different permutations of the signs on the diagonal of \mathbf{S} . In these realizations, the associated balanced coordinates are generally *not* ordered in decreasing significance, in contrast to the canonical form. \diamond

Consider a balanced linear system \mathcal{G}_\circ as in (2.25) having the canonical form (3.3). Since the reduced model obtained via balanced truncation is independent of the initial system realization, for the time being, we will assume, without loss of generality, that \mathcal{G}_\circ is already

balanced with the realization given in (3.3), and so $\mathbf{E} = \mathbf{I}_n$. Define $r = m_1 + \cdots + m_k$ for $1 \leq k < q$, and partition the sign matrix as

$$\mathbf{S} = \text{diag}(\mathbf{S}_1, \mathbf{S}_2),$$

where

$$\mathbf{S}_1 = \text{diag}(s_1, \dots, s_k) \quad \text{and} \quad \mathbf{S}_2 = \text{diag}(s_{k+1}, \dots, s_n).$$

Partition \mathbf{A} , \mathbf{B} , and \mathbf{C} as in (2.74), revealing the sign symmetries

$$\begin{aligned} \mathbf{A}_{11} &= \mathbf{S}_1 \mathbf{A}_{11}^\top \mathbf{S}_1, & \mathbf{A}_{12} &= \mathbf{S}_1 \mathbf{A}_{21}^\top \mathbf{S}_2, & \mathbf{A}_{21} &= \mathbf{S}_2 \mathbf{A}_{12}^\top \mathbf{S}_1, \\ \mathbf{A}_{22} &= \mathbf{S}_2 \mathbf{A}_{22}^\top \mathbf{S}_2, & \mathbf{B}_1 &= (\mathbf{C}_1 \mathbf{S}_1)^\top, & \text{and } \mathbf{B}_2 &= (\mathbf{C}_2 \mathbf{S}_2)^\top, \end{aligned} \quad (3.5)$$

which follow from direct multiplication in (3.3). Therefore, the Lyapunov equations in (2.43) and (2.44) both have the solution $\mathbf{P} = \mathbf{Q}_{\text{lo}} = \mathbf{\Sigma}$, i.e.

$$\mathbf{A}\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A}^\top + \mathbf{B}\mathbf{B}^\top = \mathbf{0}_{n \times n} \quad \text{and} \quad \mathbf{A}^\top \mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A} + \mathbf{C}^\top \mathbf{C} = \mathbf{0}_{n \times n}. \quad (3.6)$$

A special case of systems is those that are *state-space symmetric*.

Definition 3.3 (State-space symmetric systems [131]). If a linear system \mathcal{G}_{lo} has a (not necessarily balanced) realization satisfying (3.3) with $\mathbf{S} = \mathbf{I}_n$, then we say \mathcal{G}_{lo} is *state-space symmetric*. In this case, $\mathbf{E} = \mathbf{I}_n$, $\mathbf{A} = \mathbf{A}^\top$ and $\mathbf{B} = \mathbf{C}^\top$. \diamond

As stated earlier, the \mathcal{H}_∞ error bound (2.77) for balanced truncation holds with equality if only one singular value is truncated [69]. State-space symmetric systems have the property that the error bound (2.77) holds with equality *for any truncation order* [131]. Next, we show that only the truncated part of the model in (2.74) need be a generalization of state-space symmetric for the bound (2.77) to be tight.

3.3 Generalized conditions for the balanced truncation error bound to hold with equality

In this section, we show that the \mathcal{H}_∞ error bound for balanced truncation holds with equality for a more general class of systems than the state-space symmetric ones just described. In accordance with the partitioned balanced realization (2.74), we define the *truncated system*

$$\tilde{\mathcal{G}}_t : \begin{cases} \dot{\mathbf{x}}_2(t) &= \mathbf{A}_{22} \mathbf{x}_2(t) + \mathbf{B}_2 \mathbf{u}(t) \\ \tilde{\mathbf{y}}_{\text{lo}}(t) &= \mathbf{C}_2 \mathbf{x}_2(t), \end{cases} \quad (3.7)$$

having transfer function $\tilde{\mathbf{G}}_t(s) = \mathbf{C}_2 (s\mathbf{I}_r - \mathbf{A}_{22})^{-1} \mathbf{B}_2$ and the balanced Gramian(s) $\mathbf{\Sigma}_2 = \text{diag}(\sigma_{k+1} \mathbf{I}_{m_{k+1}}, \dots, \sigma_q \mathbf{I}_{m_q})$. We refer to (3.7) as the *truncated system* because its state

$\mathbf{x}_2: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n-r}$ contains the state components that are removed in the process of balanced truncation. Note from (3.5) that the realization of $\tilde{\mathcal{G}}_t$ in (3.7) is also balanced, and satisfies the sign symmetry condition $\mathbf{A}_{22} = \mathbf{S}_2 \mathbf{A}_{22}^\top \mathbf{S}_2$ and $\mathbf{B}_2 = (\mathbf{C}_2 \mathbf{S}_2)^\top$. For the results that follow, we will allow the signs in \mathbf{S}_1 to vary, but assume that either $\mathbf{S}_2 = \mathbf{I}_{n-r}$ or $\mathbf{S}_2 = -\mathbf{I}_{n-r}$. In other words, we do *not* assume that \mathcal{G}_{lo} is state-space symmetric; we only assume that the sign parameters corresponding to the *truncated* Hankel singular values are the same. In such cases, the truncated system $\tilde{\mathcal{G}}_t$ obeys the state-space symmetry $\mathbf{A}_{22} = \mathbf{A}_{22}^\top$ and $\mathbf{B}_2 = \pm \mathbf{C}_2^\top$. We note that this is a slight relaxation of the truncated system (3.7) being state-space symmetric according to Definition 3.3.

Theorem 3.4 (Conditions for the error bound (2.77) to hold with equality [182, Theorem 3.1]). Suppose that \mathcal{G}_{lo} is an order- n asymptotically stable, minimal, and balanced single-input, single-output system according to (2.25). Partition the Hankel singular values of \mathcal{G}_{lo} as $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$ according to (2.74) for $r = m_1 + \dots + m_k$. Conformally partition the sign matrix, $\mathbf{S} = \text{diag}(\mathbf{S}_1, \mathbf{S}_2)$, with $\mathbf{S}_1 \in \mathbb{R}^{r \times r}$ and $\mathbf{S}_2 \in \mathbb{R}^{(n-r) \times (n-r)}$. Let $\tilde{\mathcal{G}}_{\text{lo,bt}}$ denote the order- r reduced model obtained from \mathcal{G}_{lo} via balanced truncation, as in (2.75). If all the signs in \mathbf{S}_2 are the same, i.e.,

$$\mathbf{S}_2 = \text{diag}(+1, \dots, +1) \quad \text{or} \quad \mathbf{S}_2 = \text{diag}(-1, \dots, -1), \quad (3.8)$$

then the truncated Hankel singular values are distinct,

$$\Sigma_2 = \text{diag}(\sigma_{k+1}, \dots, \sigma_q),$$

and $\tilde{\mathcal{G}}_{\text{lo,bt}}$ in (2.75) achieves the \mathcal{H}_∞ error bound (2.77), i.e.

$$\|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,bt}}\|_{\mathcal{H}_\infty} = 2(\sigma_{k+1} + \dots + \sigma_q).$$

◇

Proof of Theorem 3.4. Because $\tilde{\mathcal{G}}_{\text{lo,bt}}$ is obtained via balanced truncation, it satisfies the upper bound (2.77), and so it suffices to show that this bound is attained for the particular frequency $\omega = 0$, i.e.,

$$2 \text{tr}(\Sigma_2) = |\mathbf{G}_{\text{lo}}(0) - \tilde{\mathbf{G}}_{\text{lo,bt}}(0)| = |\mathbf{C} \mathbf{A}^{-1} \mathbf{B} - \mathbf{C}_1 \mathbf{A}_{11}^{-1} \mathbf{B}_1|.$$

Note that the absolute values come from the fact that \mathbf{G}_{lo} and $\tilde{\mathbf{G}}_{\text{lo,bt}}$ are scalar valued when $p = m = 1$. First, note that the difference $\tilde{\mathbf{G}}_{\text{lo,bt}}(0) - \mathbf{G}_{\text{lo}}(0) = \mathbf{C} \mathbf{A}^{-1} \mathbf{B} - \mathbf{C}_1 \mathbf{A}_{11}^{-1} \mathbf{B}_1$ in the right-hand side of the above error expression can be expressed in terms of the Schur complement

$$\mathbf{A} / \mathbf{A}_{11} \stackrel{\text{def}}{=} \mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12}.$$

To see this, define the matrices

$$\mathbf{M}_1 = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ -\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{I}_{n-r} \end{bmatrix} \quad \text{and} \quad \mathbf{M}_r = \begin{bmatrix} \mathbf{I}_r & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{0} & \mathbf{I}_{n-r} \end{bmatrix}.$$

One can verify via the formula for the inverse of a 2×2 block matrix [135, Theorem 2.1] that $\mathbf{A}^{-1} = \mathbf{M}_r \text{diag}(\mathbf{A}_{11}^{-1}, (\mathbf{A}/\mathbf{A}_{11})^{-1}) \mathbf{M}_1$, and so

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{lo,bt}}(0) - \mathbf{G}_{\text{lo}}(0) &= \mathbf{C} \left(\mathbf{A}^{-1} - \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \mathbf{B} \\ &= \mathbf{C} \left(\mathbf{M}_r \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}/\mathbf{A}_{11})^{-1} \end{bmatrix} \mathbf{M}_1 - \mathbf{M}_r \begin{bmatrix} \mathbf{A}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{M}_1 \right) \mathbf{B} \\ &= \mathbf{C} \mathbf{M}_r \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}/\mathbf{A}_{11})^{-1} \end{bmatrix} \mathbf{M}_1 \mathbf{B} \\ &= \text{tr} \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}/\mathbf{A}_{11})^{-1} \end{bmatrix} \mathbf{M}_1 \mathbf{B} \mathbf{C} \mathbf{M}_r \right). \end{aligned} \quad (3.9)$$

This last equality follows from the fact that $\tilde{\mathbf{G}}_{\text{lo,bt}}(0) - \mathbf{G}_{\text{lo}}(0)$ is a scalar, and that the trace of the product of matrices is invariant under cyclic permutation. By the sign symmetry (3.3), we can transform the Lyapunov equation (3.6) into

$$-\mathbf{B}\mathbf{C} = -\mathbf{B}\mathbf{B}^\top \mathbf{S} = \mathbf{A}\mathbf{\Sigma}\mathbf{S} + \mathbf{\Sigma}\mathbf{A}^\top \mathbf{S} = \mathbf{A}(\mathbf{\Sigma}\mathbf{S}) + \mathbf{\Sigma}(\mathbf{S}\mathbf{A}\mathbf{S})\mathbf{S} = \mathbf{A}(\mathbf{\Sigma}\mathbf{S}) + (\mathbf{\Sigma}\mathbf{S})\mathbf{A}. \quad (3.10)$$

Using the partitioning of \mathbf{A} as in (2.74), block matrix multiplication reveals $\mathbf{M}_1 \mathbf{A} \mathbf{M}_r = \text{diag}(\mathbf{A}_{11}, \mathbf{A}/\mathbf{A}_{11})$. Multiplying (3.10) on the left and right by \mathbf{M}_1 and \mathbf{M}_r respectively, we can exploit the triangular structure of these matrices to obtain

$$\begin{aligned} -\mathbf{M}_1 \mathbf{B} \mathbf{C} \mathbf{M}_r &= \mathbf{M}_1 (\mathbf{A}(\mathbf{M}_r \mathbf{M}_r^{-1})(\mathbf{\Sigma}\mathbf{S}) + (\mathbf{\Sigma}\mathbf{S})(\mathbf{M}_1^{-1} \mathbf{M}_1) \mathbf{A}) \mathbf{M}_r \\ &= (\mathbf{M}_1 \mathbf{A} \mathbf{M}_r) \mathbf{M}_r^{-1} (\mathbf{\Sigma}\mathbf{S}) \mathbf{M}_r + \mathbf{M}_1 (\mathbf{\Sigma}\mathbf{S}) \mathbf{M}_1^{-1} (\mathbf{M}_1 \mathbf{A} \mathbf{M}_r) \\ &= \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}/\mathbf{A}_{11} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \mathbf{\Sigma}_1 & \star \\ \mathbf{0} & \mathbf{S}_2 \mathbf{\Sigma}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{S}_1 \mathbf{\Sigma}_1 & \mathbf{0} \\ \star & \mathbf{S}_2 \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}/\mathbf{A}_{11} \end{bmatrix}. \end{aligned}$$

(Entries indicated by \star are irrelevant.) Inserting this expression into (3.9) and again using invariance of the trace under cyclic permutations, we obtain

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{lo,bt}}(0) - \mathbf{G}_{\text{lo}}(0) &= \text{tr} \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}/\mathbf{A}_{11})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}/\mathbf{A}_{11} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \mathbf{\Sigma}_1 & \star \\ \mathbf{0} & \mathbf{S}_2 \mathbf{\Sigma}_2 \end{bmatrix} \right) \\ &\quad + \text{tr} \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{A}/\mathbf{A}_{11})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \mathbf{\Sigma}_1 & \mathbf{0} \\ \star & \mathbf{S}_2 \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}/\mathbf{A}_{11} \end{bmatrix} \right) \\ &= 2 \text{tr} \left(\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \mathbf{\Sigma}_2 \end{bmatrix} \right) \\ &= 2 \text{tr}(\mathbf{S}_2 \mathbf{\Sigma}_2). \end{aligned}$$

Thus, if $\mathbf{S}_2 = \pm I_{n-r}$ then $|\mathbf{G}_{\text{lo}}(0) - \tilde{\mathbf{G}}_{\text{lo,bt}}(0)| = 2 \text{tr}(\boldsymbol{\Sigma}_2)$ and hence the bound (2.77) is *sharp*. To see that the Hankel singular values are distinct, note that the conclusion that $|\mathbf{G}_{\text{lo}}(0) - \tilde{\mathbf{G}}_{\text{lo,bt}}(0)| = 2 \text{tr}(\boldsymbol{\Sigma}_2)$, along with the fact that $\|\mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo,bt}}\|_{\mathcal{H}_\infty} \leq 2(\sigma_{k+1} + \dots + \sigma_q)$, necessarily implies that $m_{k+1} = \dots = m_q = 1$. Otherwise, the magnitude $2 \text{tr}(\boldsymbol{\Sigma}_2)$ would exceed the bound in (2.77). We thus conclude that

$$\|\mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo,bt}}\|_{\mathcal{H}_\infty} = 2(\sigma_{k+1} + \dots + \sigma_q),$$

completing the proof. □

This result neatly generalizes the sharpness of the balanced truncation error bound when only one Hankel singular value is truncated: in this case, the hypothesis of Theorem 3.4 is always satisfied, since the truncated system contains only one distinct sign parameter.

Remark 3.5. Theorem 3.4 can also be deduced by adapting results from [160], which proves an \mathcal{H}_∞ lower bound on the error in balanced truncation model reduction for systems with semi-definite Hankel operators. While [160, Proposition 13] is stated for systems that are semi-definite, its proof only requires that the *truncated system* is semi-definite. With this insight, we can apply [160, Corollary 14] to the systems in Theorem 3.4, showing that the lower bound on the error in [160, Corollary 14] holds with equality. ◇

We illustrate Theorem 3.4 with a synthetic example showing how the balanced truncation error bound holds with equality when the truncated system obeys the sign consistency in (3.8).

Example 3.6 (An example with flipped signs and a strict bound). We construct a linear system \mathcal{G}_{lo} of order $n = 4$ in its canonical form (3.4). Start by specifying the system's Hankel singular values

$$\boldsymbol{\Sigma} = \text{diag}(10^1, 10^0, 10^{-1}, 10^{-2}),$$

the corresponding sign parameters

$$\mathbf{S} = \text{diag}(1, 1, -1, -1),$$

and the entries γ_i of $\mathbf{B} \in \mathbb{R}^4$

$$\gamma_1 = 1, \gamma_2 = 2, \gamma_3 = 3, \gamma_4 = 4.$$

Since the Hankel singular values of the construction \mathcal{G}_{lo} are distinct, we can apply the formula (3.4) to construct its canonical balanced realization explicitly. The system \mathcal{G}_{lo} is asymptotically stable, minimal, and balanced by construction, having the canonical form

$$\mathbf{A} = \left[\begin{array}{cc|cc} -0.05 & -0.18 & 0.30 & 0.40 \\ -0.18 & -2.00 & 6.67 & 8.08 \\ \hline -0.30 & -6.67 & -45.00 & -109.09 \\ -0.40 & -8.08 & -109.09 & -800.00 \end{array} \right], \quad \mathbf{B} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{C}^\top = \begin{bmatrix} 1 \\ 2 \\ -3 \\ -4 \end{bmatrix}.$$

Table 3.1: \mathcal{H}_∞ norm of the error system, compared to the balanced truncation upper bound (2.77) for a system where the hypothesis (3.8) of Theorem 3.4 holds for $r = 2$ and $r = 3$, but *not* for $r = 1$.

	$\ \mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,bt}}\ _{\mathcal{H}_\infty}$	$2(\sigma_{r+1} + \cdots + \sigma_n)$
$r = 1$	1.780×10^0	2.220×10^0
$r = 2$	2.200×10^{-1}	2.200×10^{-1}
$r = 3$	2.000×10^{-2}	2.000×10^{-2}

We compute reduced order models via balanced truncation of orders $r = 1, 2, 3$. The partition of \mathcal{G}_{lo} above highlights the truncation order $r = 2$ to expose the sign symmetry of the truncated system. Table 3.1 compares the \mathcal{H}_∞ -norm of the error system to the balanced truncation upper bound (2.77). For reduction to orders $r = 2$ and $r = 3$, the condition (3.8) is met, and the balanced truncation bound holds with equality, as guaranteed by Theorem 3.4. However, for reduction to order $r = 1$, the truncated system does not obey the required sign consistency (3.8), and the upper bound (2.77) holds with a *strict inequality*. \diamond

As a consequence of Theorem 3.4, we will show that the \mathcal{H}_∞ error bound (2.77) also holds with equality when performing singular perturbation balancing (2.79), provided the sign parameters of the system \mathcal{G}_{lo} satisfy (3.8). Opmeer and Reis make this observation in [160, Section V.A], also using the concept of reciprocal system as in the proof below, but in a slightly different manner.

Theorem 3.7. Suppose that \mathcal{G}_{lo} is an order- n asymptotically stable, minimal, and balanced single-input, single-output system according to (2.25). Partition the Hankel singular values of \mathcal{G}_{lo} as $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$ according to (2.74) for $r = m_1 + \cdots + m_k$. Conformally partition the sign matrix, $\mathbf{S} = \text{diag}(\mathbf{S}_1, \mathbf{S}_2)$, with $\mathbf{S}_1 \in \mathbb{R}^{r \times r}$ and $\mathbf{S}_2 \in \mathbb{R}^{(n-r) \times (n-r)}$. Let $\tilde{\mathcal{G}}_{\text{lo,spa}}$ be the order- r balanced singular perturbation approximation of \mathcal{G}_{lo} defined according to (2.79). If all the signs in \mathbf{S}_2 are the same, as in (3.8), then $\tilde{\mathcal{G}}_{\text{lo,spa}}$ in (2.79) achieves the error bound (2.77):

$$\|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,spa}}\|_{\mathcal{H}_\infty} = 2(\sigma_{k+1} + \cdots + \sigma_q).$$

\diamond

Proof of Theorem 3.7. Without loss of generality, assume that \mathcal{G}_{lo} is balanced with the canonical form satisfying (3.3). It is shown in [134, Theorem 3.2] that the model reduction error from the r -th order balanced singular perturbation approximation $\tilde{\mathcal{G}}_{\text{lo,spa}}$ to \mathcal{G}_{lo} can be written as

$$\|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,spa}}\|_{\mathcal{H}_\infty} = \|\hat{\mathcal{G}}_{\text{lo}} - \hat{\mathcal{G}}_{\text{lo,bt}}\|_{\mathcal{H}_\infty},$$

where $\hat{\mathcal{G}}_{\text{lo}}$ is the *reciprocal system* of \mathcal{G}_{lo} , which is itself a linear system (2.25) having the realization $\hat{\mathcal{G}}_{\text{lo}} = (\mathbf{A}^{-1}, \mathbf{A}^{-1}\mathbf{B}, \mathbf{C}\mathbf{A}^{-1}, \tilde{\mathbf{D}}_{\text{spa}})$. In fact, by [134, Lemma 3.1], the given realization of $\hat{\mathcal{G}}_{\text{lo}}$ is balanced with the Gramian Σ , and so the Hankel singular values of $\hat{\mathcal{G}}_{\text{lo}}$ are the

same as those of the original system \mathcal{G}_{lo} . Moreover, it is obvious that the reciprocal system obeys the same sign symmetry condition (3.3) as the original, that is $\mathbf{A}^{-1} = \mathbf{S}\mathbf{A}^{-\top}\mathbf{S}$, where \mathbf{S} carries the sign parameters of \mathcal{G}_{lo} in (3.3), and

$$\mathbf{A}^{-1}\mathbf{B} = \mathbf{S}\mathbf{A}^{-\top}\mathbf{S}\mathbf{S}\mathbf{C}^{\top} = (\mathbf{C}\mathbf{A}^{-1}\mathbf{S})^{\top}.$$

Then, the submatrices of the reciprocal system partitioned according to (2.74) satisfy those in (3.5). This shows that $\widehat{\mathcal{G}}_{\text{lo}}$ satisfies the hypotheses of Theorem 3.4. Applying the result of Theorem 3.4 to $\widehat{\mathcal{G}}_{\text{lo}}$ and $\widehat{\mathcal{G}}_{\text{lo,bt}}$, we conclude that

$$\|\mathcal{G}_{\text{lo}} - \widetilde{\mathcal{G}}_{\text{lo,spa}}\|_{\mathcal{H}_{\infty}} = \|\widehat{\mathcal{G}}_{\text{lo}} - \widehat{\mathcal{G}}_{\text{lo,bt}}\|_{\mathcal{H}_{\infty}} = 2(\sigma_{k+1} + \cdots + \sigma_q),$$

thus completing the proof. \square

3.4 On the sign parameters of a linear system

Theorem 3.4 shows that the balanced truncation \mathcal{H}_{∞} error bound (2.77) holds with equality when the truncated sign parameters are *consistent*, i.e., $\mathbf{S}_2 = \pm\mathbf{I}_{n-r}$. One could check this condition by computing the canonical balanced form described in Theorem 3.1, although it may not always be feasible to compute a full balancing transformation. Recall the power systems example introduced in Section 3.1; the realization in (3.2) is *not* balanced, but satisfies a sign symmetry condition

$$\mathbf{E} = \mathbf{I}_n, \quad \mathbf{A} = \check{\mathbf{S}}\mathbf{A}^{\top}\check{\mathbf{S}}, \quad \mathbf{B} = (\mathbf{C}\check{\mathbf{S}})^{\top} \quad \text{for } \check{\mathbf{S}} = \text{diag}(\check{s}_1, \dots, \check{s}_n) \in \mathbb{R}^{n \times n}. \quad (3.11)$$

with $\check{s}_1 = +1$ and $\check{s}_i = -1$, $i = 2, \dots, n$ for this example. Without any further investigation, it does not follow that the signs $\check{s}_i = \pm 1$ in $\check{\mathbf{S}}$ are the sign parameters of the linear system because the corresponding realization is not balanced. However, one can verify computationally that $s_1 = \check{s}_1 = +1$ and $s_i = \check{s}_i = -1$, $i = 2, \dots, n$, suggesting that this *is* in fact the case. We prove this fact here; in other words, we show that the sign parameters of a single-input, single-output linear system satisfying the hypotheses of Theorem 3.1 can be ascertained (up to a permutation) from *any* (not necessarily balanced) realization of the system that satisfies a sign-symmetry condition such as (3.3). We then strengthen this result for a class of systems with arrowhead representations, such as (3.2).

To show this, we recall the *cross Gramian* $\mathbf{X}_c \in \mathbb{R}^{n \times n}$ of a single-input, single-output linear system (2.25) from Definition 2.37. Because such a system has a scalar-valued transfer function, the system is trivially square and symmetric, and so \mathbf{X}_c is well defined. From [75], it holds that $\mathbf{X}_c = (\mathbf{P}\mathbf{Q}_{\text{lo}})^{1/2}$, and so the eigenvalues of \mathbf{X}_c are real. This is because they are the square roots of the eigenvalues of a positive semi-definite matrix. As it turns out, the sign parameters of a linear system \mathcal{G}_{lo} in this setting are given analytically by the signs of the eigenvalues of the cross Gramian, i.e.

$$s_i = \text{sign}(\lambda_i(\mathbf{X}_c)), \quad i = 1, \dots, n, \quad \lambda_1(\mathbf{X}_c) \geq \lambda_2(\mathbf{X}_c) \geq \cdots \geq \lambda_n(\mathbf{X}_c), \quad (3.12)$$

where s_i is the i -th sign parameter of \mathcal{G}_{lo} in (3.3). This fact can be obtained by combining results from [136] with [4, Lemma 5.6]. It can be shown that the Hankel singular values of a system \mathcal{G}_{lo} are the unsigned eigenvalues of its cross Gramian [136, Section 2.7], so that the given ordering $\lambda_i(\mathbf{X}_c)$ is consistent with that of the system's Hankel singular values.

Using (3.12), we can now inspect the eigenvalues of the cross Gramian \mathbf{X}_c to determine whether a system's sign parameters obey the hypotheses of Theorem 3.4. This leads to the following result.

Theorem 3.8. Let \mathcal{G}_{lo} be an order- n asymptotically stable and minimal single-input, single-output system, as in (2.25). Suppose \mathcal{G}_{lo} has a realization satisfying the generalized sign symmetry condition (3.11) with signs $\check{s}_i \in \{\pm 1\}$ for $i = 1, \dots, n$. Then there exists a permutation $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ of $(1, 2, \dots, n)$ such that

$$s_{\pi_i} = \check{s}_i, \quad i = 1, \dots, n.$$

That is, the signs $\check{s}_1, \dots, \check{s}_n$ are a permutation of the sign parameters s_1, \dots, s_n of \mathcal{G}_{lo} . \diamond

Proof of Theorem 3.8. Consider the realization (3.11) of the linear system \mathcal{G}_{lo} . Applying the sign symmetry property and multiplying (2.46) on the left by $\check{\mathbf{S}}$ reveals

$$\begin{aligned} \mathbf{0}_{n \times n} &= \check{\mathbf{S}}(\mathbf{A}\mathbf{X}_c + \mathbf{X}_c\mathbf{A} + \mathbf{B}\mathbf{C}) = \check{\mathbf{S}}\mathbf{A}\mathbf{X}_c + \check{\mathbf{S}}\mathbf{X}_c(\check{\mathbf{S}}\mathbf{A}^\top\check{\mathbf{S}}) + (\check{\mathbf{S}}\mathbf{B})(\check{\mathbf{S}}\mathbf{B})^\top \\ &= (\check{\mathbf{S}}\mathbf{A}\check{\mathbf{S}})(\check{\mathbf{S}}\mathbf{X}_c) + (\check{\mathbf{S}}\mathbf{X}_c)(\check{\mathbf{S}}\mathbf{A}^\top\check{\mathbf{S}}) + (\check{\mathbf{S}}\mathbf{B})(\check{\mathbf{S}}\mathbf{B})^\top. \end{aligned}$$

Note that $(\check{\mathbf{S}}\mathbf{B})(\check{\mathbf{S}}\mathbf{B})^\top$ is symmetric positive semi-definite. Because \mathcal{G}_{lo} is asymptotically stable and minimal, Lyapunov's theorem [4, sect. 6.2] implies that $\check{\mathbf{X}}_c \stackrel{\text{def}}{=} \check{\mathbf{S}}\mathbf{X}_c$ is symmetric positive definite. Then, we claim that the sign pattern of $\check{\mathbf{S}}$ determines the signs of the eigenvalues of \mathbf{X}_c . To see this, let $\lambda(\mathbf{X}_c)$ denote the spectrum of \mathbf{X}_c and observe that

$$\lambda(\mathbf{X}_c) = \lambda(\check{\mathbf{S}}\check{\mathbf{X}}_c) = \lambda(\check{\mathbf{X}}_c^{1/2}\check{\mathbf{S}}\check{\mathbf{X}}_c\check{\mathbf{X}}_c^{-1/2}) = \lambda(\check{\mathbf{X}}_c^{1/2}\check{\mathbf{S}}\check{\mathbf{X}}_c^{1/2}).$$

Thus $\check{\mathbf{S}}$ is a congruence transformation of \mathbf{X}_c , and hence by Sylvester's law of inertia [4, Prop. 6.15] $\check{\mathbf{S}}$ and \mathbf{X}_c have the same number of positive and negative eigenvalues. Recalling (3.12), there must exist some permutation $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ of $(1, 2, \dots, n)$ such that $s_{\pi_i} = \check{s}_i$ for $i = 1, \dots, n$. The permutation stems from the fact that Sylvester's law of inertia determines the signs of the eigenvalues of \mathbf{X}_c ; it does not tell us anything about their magnitude. \square

Theorem 3.8 says that the sign pattern of $\check{\mathbf{S}}$ in the realization (3.11) reveals the sign parameters of the corresponding linear system (2.25). As illustrated with the example in Section 3.1, such a realization may be directly obtained via the modeling process or some underlying conservation laws. The realization (3.11) does not, however, indicate the ordering of the sign parameters with respect to the Hankel singular values. There exists a permuted realization

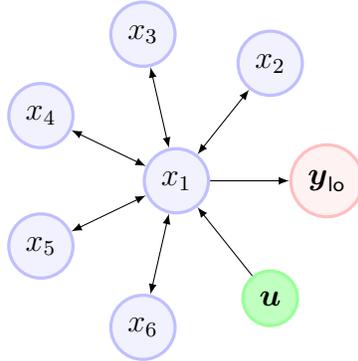


Figure 3.1: An arrowhead network with $n = 6$, with input \mathbf{u} and output \mathbf{y}_{1o} restricted to state x_1 .

that reveals the canonical ordering of the sign parameters of \mathcal{G}_{1o} ; that is, the Hankel singular values to which they correspond are non-increasing. Next, we study a special class of systems having an *arrowhead structure* for which, under some mild assumptions, the result of Theorem 3.8 can be strengthened.

3.5 Systems with arrowhead realizations

In this section, we apply the result of Theorem 3.8 to linear systems with *arrowhead realizations*. In general, we say a single-input, single-output linear system \mathcal{G}_{1o} as in (2.25) has an arrowhead realization if it has a realization $(\mathbf{I}_n, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ where

$$\mathbf{A} = \begin{bmatrix} \delta_1 & \alpha_2 & \cdots & \alpha_n \\ \beta_2 & \delta_2 & & \\ \vdots & & \ddots & \\ \beta_n & & & \delta_n \end{bmatrix}, \quad \mathbf{B} = \gamma \mathbf{e}_1, \quad \text{and} \quad \mathbf{C} = \mathbf{e}_1^\top, \quad (3.13)$$

where $\alpha_i, \beta_i, \delta_i \in \mathbb{R}$ and $\gamma \in \mathbb{R}$ is nonzero. Occasionally, we refer to systems having arrowhead realizations as *arrowhead systems*. Systems with arrowhead realizations arise in network models where the internal state variables interact as in Figure 3.1, with the input \mathbf{u} and output \mathbf{y}_{1o} directly interacting only with the first state variable, x_1 . The example from Section 3.1 having the realization (3.2) is a specific instance of the general structure in (3.13).

First, we present some useful facts regarding arrowhead systems. Let

$$\begin{aligned} \boldsymbol{\alpha} &= [\alpha_2 \quad \cdots \quad \alpha_n] \in \mathbb{R}^{1 \times (n-1)}, \\ \boldsymbol{\beta} &= [\beta_2 \quad \cdots \quad \beta_n]^\top \in \mathbb{R}^{(n-1) \times 1}, \\ \text{and } \boldsymbol{\Delta} &= \text{diag}(\delta_2, \dots, \delta_n) \in \mathbb{R}^{(n-1) \times (n-1)}. \end{aligned}$$

If $\mathbf{A} \in \mathbb{R}^{n \times n}$ has the arrowhead form (3.13) with $\delta_i \neq 0$ for all $i = 2, \dots, n$, and $\delta_1 - \boldsymbol{\alpha} \boldsymbol{\Delta}^{-1} \boldsymbol{\beta} \neq 0$, then the inverse of the arrowhead matrix \mathbf{A} can be expressed as a diagonal matrix plus a rank-one update [197]:

$$\mathbf{A}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{\Delta}^{-1} \end{bmatrix} + \rho \begin{bmatrix} -1 \\ \boldsymbol{\Delta}^{-1} \boldsymbol{\beta} \end{bmatrix} \begin{bmatrix} -1 & \boldsymbol{\alpha} \boldsymbol{\Delta}^{-1} \end{bmatrix}, \quad (3.14)$$

where

$$\rho = \frac{1}{\delta_1 - \boldsymbol{\alpha} \boldsymbol{\Delta}^{-1} \boldsymbol{\beta}}.$$

This expression follows from the Sherman–Morrison–Woodbury formula [86, p. 65]; though most easily derived when $\delta_1 \neq 0$, the formula holds even when $\delta_1 = 0$. While we have stated the formula for arrowhead matrices having real entries, (3.14) holds when the nonzero entries of \mathbf{A} are complex-valued, as well. Applying (3.14) to the transfer function (2.30) of a system having an arrowhead realization (3.13) leads to the revealing form

$$\mathbf{G}_{\text{lo}}(s) = \frac{\gamma}{s - \delta_1 - \sum_{i=2}^n \frac{\alpha_i \beta_i}{s - \delta_i}}. \quad (3.15)$$

Any system having a transfer function \mathbf{G}_{lo} with the general form (3.15) has an arrowhead realization given by (3.13). Conversely, given any system having an arrowhead realization (3.13), the transfer function \mathbf{G}_{lo} can be expressed in the form (3.15).

The entries of the arrowhead matrix in (3.13) reveal whether or not the corresponding arrowhead realization is minimal.

Lemma 3.9. Let \mathcal{G}_{lo} be an order- n asymptotically stable and minimal single-input, single-output system, as in (2.25) having the arrowhead realization in (3.13). Then

1. The pair (\mathbf{A}, \mathbf{C}) is observable if and only if $\delta_i \neq \delta_j$ for $i \neq j$ and $\boldsymbol{\alpha}^\top \neq \mathbf{0}_{n-1}$.
2. The pair (\mathbf{A}, \mathbf{B}) is reachable if and only if $\delta_i \neq \delta_j$ for $i \neq j$ and $\boldsymbol{\beta} \neq \mathbf{0}_{n-1}$. ◇

Proof. By the duality principle, Theorem 2.33, it suffices to prove the statement about observability, since the dual of \mathcal{G}_{lo} is also an arrowhead system. By the Hautus test [245, Theorem 4.15], the pair (\mathbf{A}, \mathbf{C}) is not observable if and only if there exists an eigenvector $\mathbf{v} \neq \mathbf{0}_n$ of \mathbf{A} such that $\mathbf{C}\mathbf{v} = \mathbf{e}_1^\top \mathbf{v} = 0$ due to the structure of the realization (3.13), i.e., $v_1 = 0$. We prove the necessary condition by the contrapositive implication. If $\alpha_j = 0$, for some $j = 2, \dots, n$, then $\mathbf{v} = \mathbf{e}_j \in \mathbb{R}^n$ is an eigenvector of \mathbf{A} with $v_1 = 0$. If $\delta_i = \delta_j$ and $\alpha_i \neq 0$, $\alpha_j \neq 0$ for $i \neq j$, then $\alpha_j \mathbf{e}_i - \alpha_i \mathbf{e}_j$ is an eigenvector of \mathbf{A} with $v_1 = 0$. In both cases, (\mathbf{A}, \mathbf{C}) is not observable by the previous application of the Hautus test.

We now prove sufficiency of the conditions by contradiction. Suppose that $\delta_i \neq \delta_j$ for $i \neq j$, $\alpha_j \neq 0$ for $j = 2, \dots, n$, and for the sake of contradiction, that (\mathbf{A}, \mathbf{C}) is not observable. Again by the Hautus test, there exists an eigenvector $\mathbf{v} \neq \mathbf{0}_n$ of \mathbf{A} such that $\mathbf{C}\mathbf{v} = \mathbf{0}_n$, and

so $v_1 = 0$. The eigenvector condition $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ with $v_1 = 0$ implies that $\delta_j v_j = \lambda v_j$ for $j = 2, \dots, n$. Hence, $v_i \neq 0$ for exactly one $i \in \{2, \dots, n\}$, otherwise we would have $\mathbf{v} = \mathbf{0}_n$. Without loss of generality, take $v_i = 1$ so that $\mathbf{v} = \mathbf{e}_i$. But then $\mathbf{A}\mathbf{v} = \mathbf{A}\mathbf{e}_i = \alpha_i \mathbf{e}_1 + \delta_i \mathbf{e}_i$, and so \mathbf{v} is not an eigenvector of \mathbf{A} since $\alpha_i \neq 0$. It follows that (\mathbf{A}, \mathbf{C}) is observable by contradiction. \square

From Lemma 3.9, it follows straightforwardly that an arrowhead system is minimal if and only if α_i and β_i are nonzero for all $i = 2, \dots, n$ and $\delta_i \neq \delta_j$ for all $i \neq j$. We are now in a position to apply Theorem 3.8 to these arrowhead systems.

Corollary 3.10 (Sign parameters from arrowhead realizations). Let \mathcal{G}_{lo} be an order- n , asymptotically stable, and minimal single-input, single-output system with an arrowhead realization (3.13). Then there exists a permutation $\pi = (\pi_1, \pi_2, \dots, \pi_n)$ of $(1, 2, \dots, n)$ such that the sign parameters of \mathcal{G}_{lo} are given by

$$s_{\pi_1} = \text{sign}(\gamma), \quad \text{and} \quad s_{\pi_i} = \text{sign}(\gamma\alpha_i\beta_i), \quad i = 2, \dots, n.$$

\diamond

Proof. First, note that any arrowhead realization of the form (3.13) can be made to satisfy the generalized sign symmetry condition (3.11) via an appropriate diagonal change of basis. In particular, let

$$\mathbf{T} = \sqrt{|\gamma|} \text{diag} \left(1, \sqrt{\frac{|\beta_2|}{|\alpha_2|}}, \dots, \sqrt{\frac{|\beta_n|}{|\alpha_n|}} \right). \quad (3.16)$$

Then it follows that the transformed system

$$\mathbf{A}_s = \mathbf{T}^{-1} \mathbf{A} \mathbf{T} = \begin{bmatrix} \delta_1 & \text{sign}(\alpha_2) \sqrt{|\alpha_2 \beta_2|} & \cdots & \text{sign}(\alpha_n) \sqrt{|\alpha_n \beta_n|} \\ \text{sign}(\beta_2) \sqrt{|\alpha_2 \beta_2|} & \delta_2 & & \\ \vdots & & \ddots & \\ \text{sign}(\beta_n) \sqrt{|\alpha_n \beta_n|} & & & \delta_n \end{bmatrix},$$

$$\mathbf{B}_s = \mathbf{T}^{-1} \mathbf{B} = \text{sign}(\gamma) \sqrt{|\gamma|} \mathbf{e}_1, \quad \mathbf{C}_s = \mathbf{C} \mathbf{T} = \sqrt{|\gamma|} \mathbf{e}_1^\top$$

satisfies (3.11) with the sign matrix

$$\check{\mathbf{S}} = \text{diag} (\text{sign}(\gamma), \text{sign}(\gamma\alpha_2\beta_2), \dots, \text{sign}(\gamma\alpha_n\beta_n)).$$

In other words, the transformed arrowhead realization is such that $\mathbf{A}_s = \check{\mathbf{S}} \mathbf{A}_s^\top \check{\mathbf{S}}$ and $\mathbf{B}_s = (\mathbf{C}_s \check{\mathbf{S}})^\top$. Therefore, the entries of the sign matrix $\check{\mathbf{S}}$ above are entirely determined by the signs of γ and the off-diagonal entries $\{\alpha_i\}$ and $\{\beta_i\}$ of the arrowhead matrix. Since \mathcal{G}_{lo} is asymptotically stable and minimal by hypothesis, we can apply Theorem 3.8 to obtain the stated result. \square

Corollary 3.10 is stated for systems having the particular structure in (3.13). While this structure differs slightly from the power system model in (3.2), the two models are related by the simple diagonal state space transformation (3.16), and the corollary applies to both.

3.6 A special case of arrowhead systems

We return now to the power systems example from Section 3.1 that motivated our study. For these systems, we first observed numerically that the balanced truncation error bound (2.77) was tight for all orders of reduction, and then noticed that the truncated part of the model in these systems' canonical balanced forms obeyed the sign pattern hypothesis (3.8) in Theorem 3.4. Applying Corollary 3.10 to this power system model \mathcal{G}_{lo} with realization (3.2) guarantees that

$$s_{\pi_1} = \text{sign}\left(\frac{1}{\sqrt{m}}\right) = +1, \quad s_{\pi_i} = \text{sign}\left(-\sqrt{\frac{r_{i-1}^{-1}}{m\tau_{i-1}}}\right) = -1, \quad i = 2, \dots, n.$$

However, numerically we observed that $s_1 = +1$ and $s_i = -1$ for all $i = 2, \dots, n$, suggesting that the permutation π given in the statement of Corollary 3.10 is the identity. This turns out to be true; we next show that this pattern holds in a more general setting. Specifically, if a system \mathcal{G}_{lo} has an arrowhead realization (3.13) such that the diagonal entries as well as the signs of the products of the off-diagonal entries are *negative*, i.e., $\delta_i < 0$ for all $i = 1, 2, \dots, n$ and $\text{sign}(\alpha_i\beta_i) = -1$ for all $i = 2, \dots, n$, then the trailing $n - 1$ sign parameters of the arrowhead system \mathcal{G}_{lo} are identical.

Corollary 3.11 (Conditions for a single sign flip after the first parameter). Let \mathcal{G}_{lo} be an order- n asymptotically stable, minimal, single-input, single-output system as in (2.25) with an arrowhead realization as in (3.13). Suppose further that the arrowhead realization (3.13) is such that $\text{sign}(\alpha_i\beta_i) = -1$ for all $i = 2, \dots, n$ and $\delta_i < 0$ for all $i = 1, \dots, n$. Then the sign parameters of \mathcal{G}_{lo} are given by

$$s_1 = \text{sign}(\gamma), \quad s_i = -\text{sign}(\gamma), \quad i = 2, \dots, n.$$

Moreover, the Hankel singular values of \mathcal{G}_{lo} are distinct. \diamond

Proof. Without loss of generality, suppose that $\text{sign}(\gamma) = +1$. If the given arrowhead realization of \mathcal{G}_{lo} is such that the permutation π in Corollary 3.10 is the identity, then the result follows trivially. Otherwise, Corollary 3.10 only guarantees that \mathcal{G}_{lo} has one positive sign parameter and $n - 1$ negative sign parameters. We claim that the positive sign parameter corresponds to the dominant eigenvalue of the cross Gramian, i.e., $s_1 = +1$. As noted in [4, Remark 5.4.3], $2 \text{tr}(\mathbf{X}_c) = -\text{tr}(\mathbf{C}\mathbf{A}^{-1}\mathbf{B})$. Since $\mathbf{B} = \gamma\mathbf{e}_1$ and $\mathbf{C} = \mathbf{e}_1$ in the given realization (3.13), the formula (3.14) reveals

$$\mathbf{C}\mathbf{A}^{-1}\mathbf{B} = \frac{\gamma}{\delta_1 - \sum_{i=2}^n \frac{\alpha_i\beta_i}{\delta_i}}.$$

Since this quantity is a scalar, $\mathbf{CA}^{-1}\mathbf{B} = \text{tr}(\mathbf{CA}^{-1}\mathbf{B})$. The hypothesis that $\delta_i < 0$ for all $i = 1, \dots, n$, along with the assumption that $\text{sign}(\gamma) = +1$ implies

$$\text{tr}(\mathbf{CA}^{-1}\mathbf{B}) = \mathbf{CA}^{-1}\mathbf{B} = \frac{\gamma}{\delta_1 - \sum_{i=2}^n \frac{\alpha_i \beta_i}{\delta_i}} < 0.$$

Thus, $\text{tr}(\mathbf{X}_c) = -\frac{1}{2} \text{tr}(\mathbf{CA}^{-1}\mathbf{B}) > 0$. Because the trace of \mathbf{X}_c is equal to the sum of its eigenvalues and \mathbf{X}_c has only one positive eigenvalue, the dominant eigenvalue of \mathbf{X}_c must be the positive one. Thus we conclude that $s_1 = +1$ and $s_i = -1$, for all $i = 2, \dots, n$. Theorem 3.4 further implies that the $n - 1$ trailing Hankel singular values of \mathcal{G}_{lo} , and thus all Hankel singular values of \mathcal{G}_{lo} , are distinct. For the case of $\text{sign}(\gamma) = -1$, the proof follows nearly identically by noting that $\text{tr}(\mathbf{X}_c) = -\frac{1}{2} \text{tr}(\mathbf{CA}^{-1}\mathbf{B}) < 0$, which shows that $s_1 = -1$ and $s_i = +1$ for all $i = 2, \dots, n$ by the same logic as above. \square

Corollary 3.11 implies that the trailing sign parameters of the systems described here are consistent, and can be determined directly from the off-diagonal entries of the arrowhead matrix using the formula

$$s_1 = \text{sign}(\gamma), \quad s_i = \text{sign}(\gamma \alpha_i \beta_i), \quad i = 2, \dots, n.$$

In other words, the sign parameters of systems having an arrowhead realization satisfying the hypotheses of Corollary 3.11 exhibit *a single sign flip after the first parameter*, and so these systems obey the necessary sign consistency conditions in Theorem 3.4 for *all* orders of reduction: the balanced truncation \mathcal{H}_∞ error bound (2.77) *will always hold with equality*.

Example 3.12. We illustrate Theorem 3.4 and Corollary 3.11 with a particular example of the power system model presented in Subsection 3.1. Consider a network with $n - 1 = 4$ coherent generators. Take $\check{m} = 0.044$, $\check{d} = 0.038$,

$$(r_1^{-1}, r_2^{-1}, r_3^{-1}, r_4^{-1}) = (0.013, 0.014, 0.022, 0.025),$$

and

$$(\tau_1, \tau_2, \tau_3, \tau_4) = (5.01, 6.82, 7.38, 7.79).$$

Since the physical parameters associated with the system are all positive, Corollary 3.11 implies, for the realization of \mathcal{G}_{lo} given by (3.2), that the sign parameters are

$$s_1 = \text{sign}\left(\frac{1}{\sqrt{\check{m}}}\right) = +1, \quad s_i = \text{sign}\left(-\sqrt{\frac{r_{i-1}^{-1}}{\check{m}\tau_{i-1}}}\right) = -1, \quad i = 2, \dots, 5. \quad (3.17)$$

These signs can be verified by computing a balanced realization of \mathcal{G}_{lo} satisfying the symmetry condition (3.3) with the sign matrix

$$\mathbf{S} = \text{diag}(1, -1, -1, -1, -1).$$

Table 3.2: \mathcal{H}_∞ norm of the error system, compared to the balanced truncation upper bound (2.77) for a power system, $n = 5$.

	$\ \mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,bt}}\ _{\mathcal{H}_\infty}$	$2(\sigma_{r+1} + \cdots + \sigma_n)$
$r = 1$	1.747×10^1	1.747×10^1
$r = 2$	7.067×10^{-2}	7.067×10^{-2}
$r = 3$	1.697×10^{-4}	1.697×10^{-4}
$r = 4$	8.248×10^{-8}	8.248×10^{-8}

The Hankel singular values of \mathcal{G}_{lo} with transfer function \mathbf{G}_{lo} in (3.1) are

$$\Sigma = \text{diag}(11.63, 7.13, 3.53 \times 10^{-2}, 8.48 \times 10^{-5}, 4.12 \times 10^{-8}).$$

We construct the canonical balanced realization of \mathcal{G}_{lo} using the formula (3.4), then compute order- r approximations via balanced truncation to \mathcal{G}_{lo} for $r = 2, 3$, and 4. Under these conditions, the truncated systems obey the sign consistency in (3.8) for each r . As in Example 3.6, we highlight this symmetry by partitioning the system for $r = 3$:

$$\mathbf{A} = \left[\begin{array}{ccc|cc} -0.9913 & 0.5924 & -0.0467 & 0.0020 & 0.0000 \\ -0.5924 & -0.0216 & 0.0087 & -0.0004 & -0.0000 \\ 0.0467 & 0.0087 & -0.1800 & 0.0157 & 0.0003 \\ \hline -0.0020 & -0.0004 & 0.0157 & -0.1437 & -0.0062 \\ -0.0000 & -0.0000 & 0.0003 & -0.0062 & -0.1372 \end{array} \right],$$

$$\mathbf{B} = \left[\begin{array}{c} -4.8009 \\ -0.5552 \\ 0.1126 \\ -0.0049 \\ -0.0001 \end{array} \right], \quad \mathbf{C}^T = \left[\begin{array}{c} -4.8021 \\ 0.5552 \\ -0.1126 \\ 0.0049 \\ 0.0001 \end{array} \right], \quad \mathbf{D} = 0.$$

Table 3.2 compares the \mathcal{H}_∞ norm of the error system to the balanced truncation upper bound (2.77). Because the trailing sign parameters of \mathcal{G}_{lo} are all -1 , we can perform truncation at any order $r \geq 1$ and the truncated system will satisfy the sign requirements (3.8) of Theorem 3.4. Thus, the balanced truncation error bound holds with equality for approximations of all orders. \diamond

3.7 Conclusions

In this chapter, it is shown that the balanced truncation error bound (2.77) holds with equality for SISO systems satisfying the sign consistency condition (3.8), providing an explicit formula for the approximation error in terms of the system's Hankel singular values. This analysis generalizes an earlier result for state-space symmetric systems from [131]. It is

additionally proven that the sign parameters corresponding to a system's Hankel singular values are determined by the generalized state-space symmetry property (3.11) of the system. This result is strengthened for a special class of arrowhead systems, illustrated by a model of coherent generators in power systems. From these results, one can verify whether the sign consistency condition (3.8) in Theorem 3.4 holds and thus whether or not the corresponding order- r balanced truncation approximation achieves the \mathcal{H}_∞ error bound (2.77).

Chapter 4

Data-driven balancing: What to sample for various balancing-based reduced models

4.1 Introduction

We shift our attention in this chapter to *data-driven* formulations of balancing-based model reduction. The classical Lyapunov balanced truncation (BT), which was discussed in the previous chapter, and its variants are *intrusive*; that is, they require an explicit state-space realization $(\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C})$ of the full-order model dynamics (2.25) to compute a reduced-order model (2.55) by Petrov-Galerkin projection. In complex applications, however, explicit computational models of the form (2.25) may be either difficult to obtain or wholly unavailable. Rather, the underlying system is only accessible in the form of *data*. Certain data-driven methodologies for reduced-order modeling are *non-intrusive* insofar as they construct low-dimensional surrogates such as (2.55) from input-to-output invariants, e.g., evaluations of the transfer function (2.30). The recent work [89] develops a data-driven reformulation of Lyapunov BT for LTI systems (2.25), called *quadrature-based balanced truncation* (QuadBT), that is able to construct approximate BT reduced models using only evaluations of the transfer function (2.30). As the major theoretical contribution of this chapter, we generalize the QuadBT framework to other types of balanced truncation model reduction. Specifically, we develop data-driven (quadrature-based) reformulations of balanced stochastic truncation (BST) [61, 92, 93], positive-real or passivity-preserving balanced truncation (PRBT) [61], bounded-real balanced truncation (BRBT) [159], and frequency-weighted balanced truncation (FWBT) [69, 116, 244] for *first-order* linear systems (2.25). We also develop a specially tailored data-driven formulation of position-velocity balanced truncation (sopvBT) [181] for the structured surrogate modeling of *second-order* linear systems. These results lay the theoretical foundation for the construction of various BT reduced models directly from input-output invariant frequency-response data.

The results of Sections 4.2 and 4.3 in this chapter are also available in the preprint [184], which has been provisionally accepted for publication in *Automatica*.

[184] Reiter, S., Gosea, I. V., and Gugercin, S. (2023). [Generalizations of data-driven balanc-](#)

ing: what to sample for different balancing-based reduced models. arXiv, 2312.12561.

The results of Section 4.5 will be available in a preprint that is in preparation [187].

[187] Reiter, S. and Werner, S. W. R. (2025a). Data-driven balanced truncation for second-order systems with generalized proportional damping. In preparation.

4.1.1 Literature review

A variety of methods for data-driven modeling based on different types of data and problem formulations have emerged in recent years. Not all are based on balanced truncation; nonetheless, we briefly overview a collection of them here for completeness. For the ease of presentation, we assume henceforth that $\mathbf{E} = \mathbf{I}_n$ without loss of generality.

For time-domain data such as discretely-sampled state trajectories, one can apply the dynamic mode decomposition [10, 174, 200, 213], which is a special instance of the more general framework of operator inference [27, 164, 167, 179, 202]. Roughly speaking, these methods perform a least squares fit to trajectory data to learn some reduced-order matrix operators that explain the underlying dynamics. For frequency-domain data such as transfer function evaluations, one can apply the Loewner interpolation framework [139, 166, 172], the vector-fitting method for rational least-squares fitting of the data [65, 100, 227], or methods based on barycentric rational forms that combine interpolation and least-squares [90, 153].

This chapter will focus on data-driven formulations of balanced truncation. In the original BT paper [151], Moore motivated a data-based implementation of (approximate) balanced truncation using time-domain trajectories. The balanced proper orthogonal decomposition (POD) method [191, 204, 229] approximately recovers the reachability and observability Gramians \mathbf{P} and \mathbf{Q}_{lo} via numerical quadrature rules constructed from time- or frequency-domain snapshots of the state. Although it is a data-driven method, balanced POD is still intrusive insofar as it is projection-based and uses state trajectories that depend on a particular realization of the full-order model. Other approximate methods for BT that use numerical quadrature rules for approximating the system Gramians are proposed in [35, 45]. We also mention the empirical Gramian framework of [107, 108, 122]. The so-called empirical Gramians that are used in these works can be viewed as an extension of the classical system Gramians for nonlinear and parameter-dependent systems that facilitate data-driven computation. As with balanced POD, these methods use realization-dependent state trajectories, not input-to-output invariants. The eigensystem realization algorithm [115] uses input-to-output invariant impulse responses, which are based on the time domain, but we are primarily interested in frequency-domain methods. By contrast, the QuadBT method [89] constructs approximate Lyapunov BT reduced models entirely from input-to-output invariant transfer function data. The fundamental innovation of [89] is to implicitly use low-rank factors derived from numerical quadrature rules applied to \mathbf{P} in (2.41) and \mathbf{Q}_{lo} in (2.42)

in place of the exact Cholesky factors \mathbf{R} and \mathbf{L}_{lo} in Algorithm 2.4.3. These low-rank factors are never explicitly formed; rather, the quadrature-based approximations to (2.78) that result from this replacement are computable *entirely from data*. Using similar ideas, a data-driven implementation of the balanced singular perturbation approximation presented in Section 2.4.3 was developed in [127].

4.1.2 Chapter contents

The essential quantities in any balancing-based model reduction are *Gramians*: A pair of symmetric positive (semi-)definite matrices that obey the contragradient transformation laws in (2.71). *Balancing* is just the simultaneous diagonalization of two such matrices. Recall from Section 2.4.3 that in the classical (Lyapunov) BT due to [151, 152], these are the reachability and observability Gramians that uniquely satisfy the dual algebraic *Lyapunov* equations (ALEs) in (2.43) and (2.44). In variants of balanced truncation, the Gramians often satisfy algebraic *Riccati* equations (AREs). Once the *relevant* Gramians—that is, relevant to the type of BT being performed—are specified, any variant of BT proceeds identically according to Algorithm 2.4.3. The same can be said for the BT-MOR of second-order systems; see Section 4.5. By exploiting this insight, we develop here theoretical extensions of the (quadrature-based) data-driven balancing framework from [89] to other types of BT-MOR.

In the first part of this chapter, extensions of QuadBT for *first-order* linear systems (2.25) are considered. Section 4.2 presents an abstract framework for (intrusive) Lyapunov balanced truncation wherein the relevant Gramians satisfy a generic pair of ALEs. Unlike the traditional BT, these Gramians can be respectively interpreted as the reachability and observability Gramians of a pair of related linear systems that are *not* necessarily the underlying full-order model \mathcal{G}_{lo} . At least each of the first-order BT variants covered in Section 4.3—namely BST, PRBT, and BRBT—can be interpreted under this abstract formulation. By replacing the exact square-root factors of the abstract Gramians with appropriately chosen *quadrature-based* factors, we show how to compute the reduced-order quantities that determine a BT-ROM from certain transfer function data. This leads to a generalized framework for quadrature-based balancing, which we call *generalized QuadBT* (GenQuadBT). The theoretical and algorithmic formulations of GenQuadBT are presented in Theorem 4.1 and Algorithm 4.2.2. In Section 4.3 we interpret GenQuadBT in the settings of BST, PRBT, and BRBT; see Theorems 4.5, 4.9, and 4.13. Unlike [89], the data required to recover each type of BT reduced model are not necessarily evaluations of the underlying system’s transfer function; rather, they are determined by the pair of Gramians being balanced. Specifically, these data are evaluations of various *spectral factors* associated with the full-order model transfer function. In contrast to QuadBT, it is not clear how to evaluate these spectral factors in practice. Nonetheless, our results lay the theoretical foundation for data-driven implementations of BST, PRBT, and BRBT. Section 4.4 presents a quadrature-based formulation of the frequency-weighted balanced truncation (FWBT) developed by Enns [69, 116] and the self-weighted balanced truncation of Zhou [244] as a special case. We show that computing a

frequency-weighted BT-ROM requires sampling the underlying transfer function \mathbf{G}_{lo} as well as the associated input- and output-frequency weights \mathbf{G}_{i} and \mathbf{G}_{o} . Finally, in Section 4.5, we derive a specially tailored extension of QuadBT for *second-order* dynamical systems. Specifically, this is a data-driven reformulation of the position-velocity BT (sopvBT) from [181]. The method applies to any system that exhibits a type of generalized proportional damping. Throughout the chapter, numerical experiments serve as a proof of concept and validate the data-based BT-ROMs.

4.2 Generalized quadrature-based balancing

In this section, we develop a generalized framework for quadrature-based (data-driven) balancing that is applicable to each of the BT-variants for LTI systems (2.25) mentioned in Section 4.1, namely BST, PRBT, and BRBT. Traditional formulations of these variants are intrusive insofar as they require a state-space realization of a linear system (2.25) to compute a balanced reduced model by projecting the full-order matrix operators \mathbf{A} , \mathbf{B} , and \mathbf{C} according to (2.58). While the QuadBT algorithm [89] is non-intrusive, it is currently limited to the usual Lyapunov setting where the Gramians being balanced are \mathbf{P} and \mathbf{Q}_{lo} .

As the first step towards generalizing QuadBT to other types of balancing, we present here an abstract formulation of Lyapunov-based BT where the Gramians are the unique solutions to a generic pair of ALEs; the methods of BST, PRBT, and BRBT can all be interpreted from this perspective. Based on this abstract formulation, we then derive a generalized version of QuadBT that is applicable to each of the aforementioned variants.

4.2.1 An abstract framework for Lyapunov balanced truncation

Consider the solutions $\mathbf{P}_{\mathcal{X}} \in \mathbb{R}^{n \times n}$ and $\mathbf{Q}_{\mathcal{Y}} \in \mathbb{R}^{n \times n}$ to the pair of ALEs:

$$\mathbf{A}\mathbf{P}_{\mathcal{X}} + \mathbf{P}_{\mathcal{X}}\mathbf{A}^{\top} + \mathbf{B}_{\mathcal{X}}\mathbf{B}_{\mathcal{X}}^{\top} = \mathbf{0}_{n \times n}, \quad (4.1a)$$

$$\mathbf{A}^{\top}\mathbf{Q}_{\mathcal{Y}} + \mathbf{Q}_{\mathcal{Y}}\mathbf{A} + \mathbf{C}_{\mathcal{Y}}^{\top}\mathbf{C}_{\mathcal{Y}} = \mathbf{0}_{n \times n}, \quad (4.1b)$$

where $\mathbf{B}_{\mathcal{X}} \in \mathbb{R}^{n \times m_x}$ and $\mathbf{C}_{\mathcal{Y}} \in \mathbb{R}^{p_y \times n}$ for $m_x, p_y \in \mathbb{Z}_{>0}$. Because \mathbf{A} is assumed Hurwitz, the solutions $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ to (4.1a) and (4.1b) are unique. Moreover, they can respectively be viewed as the reachability and observability Gramians of a pair of asymptotically stable linear systems \mathcal{X} and \mathcal{Y} defined according to (2.25):

$$\mathcal{X} = (\mathbf{A}, \mathbf{B}_{\mathcal{X}}, \mathbf{C}_{\mathcal{X}}, \mathbf{D}_{\mathcal{X}}) \quad \text{and} \quad \mathcal{Y} = (\mathbf{A}, \mathbf{B}_{\mathcal{Y}}, \mathbf{C}_{\mathcal{Y}}, \mathbf{D}_{\mathcal{Y}}), \quad (4.2)$$

where $\mathbf{B}_{\mathcal{Y}} \in \mathbb{R}^{n \times m_y}$, $\mathbf{D}_{\mathcal{Y}} \in \mathbb{R}^{p_y \times m_y}$ and $\mathbf{C}_{\mathcal{X}} \in \mathbb{R}^{p_x \times n}$, $\mathbf{D}_{\mathcal{X}} \in \mathbb{R}^{p_x \times m_x}$ for $m_y, p_x \in \mathbb{Z}_{>0}$. Exact formulations of these systems, along with the corresponding state-space quadruples, in the context of each BT variant we consider, are provided in Section 4.3. For ease of

reference later on, we refer to (4.1a) as the *reachability Lyapunov equation* of the system \mathcal{X} , and (4.1b) as the *observability Lyapunov equation* of the system \mathcal{Y} . We also refer to $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ as *agnostic Gramians* because they are agnostic to any particular kind of BT we will consider. We further assume that the given realizations of \mathcal{X} and \mathcal{Y} are minimal. Under these assumptions, $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ are symmetric positive definite (SPD) and thus admit the Cholesky factorizations

$$\mathbf{P}_{\mathcal{X}} = \mathbf{R}_{\mathcal{X}}\mathbf{R}_{\mathcal{X}}^{\top} \quad \text{and} \quad \mathbf{Q}_{\mathcal{Y}} = \mathbf{L}_{\mathcal{Y}}\mathbf{L}_{\mathcal{Y}}^{\top},$$

where $\mathbf{L}_{\mathcal{Y}}, \mathbf{R}_{\mathcal{X}} \in \mathbb{R}^{n \times n}$. The agnostic Gramians $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ obey the transformation laws in (2.71). Thus, one can reduce the model order of a linear system (2.25) by:

1. Computing a balancing transformation for $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ as laid out in Theorem 2.50;
2. Applying this transformation to the system (2.25) being approximated and truncating.

This procedure effectively neglects the components of the state space that correspond to the smallest singular values of the joint Cholesky factors $\mathbf{L}_{\mathcal{Y}}^{\top}\mathbf{R}_{\mathcal{X}}$. Algorithmically, the balancing and truncation steps occur simultaneously using the same square-root algorithm from [124, 208]. This is written down in Algorithm 4.2.1. Note that this is nearly identical to Algorithm 2.4.3, with the Cholesky factors $\mathbf{L}_{\mathcal{Y}}$ and $\mathbf{R}_{\mathcal{X}}$ of the agnostic Gramians taking the place of \mathbf{L}_{lo} and \mathbf{R} . In fact, Algorithm 4.2.1 contains Algorithm 2.4.3 for the special case of $\mathcal{X} = \mathcal{Y} = \mathcal{G}_{\text{lo}}$.

Once the Cholesky factors $\mathbf{L}_{\mathcal{Y}}$ and $\mathbf{R}_{\mathcal{Y}}$ of the relevant Gramians are specified, Steps 2–4 of Algorithm 4.2.1 proceed identically. The same can be said for a quadrature-based implementation; one only needs that the Gramians being balanced elicit exploitable integral representations, and the structure of the abstract Lyapunov equations in (4.1) provides precisely this. This insight enables us to derive a generalized formulation of QuadBT that is applicable to any type of balancing that can be cast into the previously described abstract formulation, including BST, PRBT, and BRBT.

4.2.2 Theoretical formulation for data-driven balancing

Recall that the projection matrices $\mathbf{W} \in \mathbb{R}^{n \times r}$ and $\mathbf{V} \in \mathbb{R}^{n \times r}$ given in Algorithm 4.2.1 are defined as

$$\mathbf{W} = \mathbf{L}_{\mathcal{Y}}\mathbf{U}_1\mathbf{\Sigma}_1^{-1/2} \quad \text{and} \quad \mathbf{V} = \mathbf{R}_{\mathcal{X}}\mathbf{Y}_1\mathbf{\Sigma}_1^{-1/2}.$$

Because $\mathbf{U}_1, \mathbf{Y}_1$, and $\mathbf{\Sigma}_1$ are extracted from the singular valued decomposition (SVD) of $\mathbf{L}_{\mathcal{Y}}^{\top}\mathbf{R}_{\mathcal{X}}$, the BT reduced model given by (4.3) is entirely specified by the four key quantities

$$\mathbf{L}_{\mathcal{Y}}^{\top}\mathbf{R}_{\mathcal{X}}, \quad \mathbf{L}_{\mathcal{Y}}^{\top}\mathbf{A}\mathbf{R}_{\mathcal{X}}, \quad \mathbf{L}_{\mathcal{Y}}^{\top}\mathbf{B}, \quad \text{and} \quad \mathbf{C}\mathbf{R}_{\mathcal{X}}. \quad (4.4)$$

By replacing the exact Cholesky factors above with (inexact) quadrature-based factors derived from numerical quadrature rules that applied to $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$, we derive approximations

Algorithm 4.2.1: Square-root algorithm for abstract Lyapunov balanced truncation [184].

Input: $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ from (2.25), order r ($1 \leq r < n$).

Output: $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{D}}$ —state-space matrices of (2.55).

- 1 Compute Cholesky factors $\mathbf{R}_\mathcal{X} \in \mathbb{R}^{n \times n}$ and $\mathbf{L}_\mathcal{Y} \in \mathbb{R}^{n \times n}$ of $\mathbf{P}_\mathcal{X} \in \mathbb{R}^{n \times n}$ and $\mathbf{Q}_\mathcal{Y} \in \mathbb{R}^{n \times n}$ that solve the Lyapunov equations in (4.1).
- 2 Compute the singular value decomposition of $\mathbf{L}_\mathcal{Y}^\top \mathbf{R}_\mathcal{X}$ partitioned according to

$$\mathbf{L}_\mathcal{Y}^\top \mathbf{R}_\mathcal{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{Y}^\top = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1^\top \\ \mathbf{Y}_2^\top \end{bmatrix}$$

for $\mathbf{\Sigma}_1 \in \mathbb{R}^{r \times r}$, $\mathbf{U}_1, \mathbf{Y}_1 \in \mathbb{R}^{n \times r}$ and $\mathbf{\Sigma}_2 \in \mathbb{R}^{(n-r) \times (n-r)}$, $\mathbf{U}_2, \mathbf{Y}_2 \in \mathbb{R}^{n \times (n-r)}$.

- 3 Compute the model reduction bases $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{n \times r}$ as

$$\mathbf{W} = \mathbf{L}_\mathcal{Y} \mathbf{U}_1 \mathbf{\Sigma}_1^{-1/2}, \quad \mathbf{V} = \mathbf{R}_\mathcal{X} \mathbf{Y}_1 \mathbf{\Sigma}_1^{-1/2}.$$

- 4 Compute the reduced model $\tilde{\mathcal{G}}_{\text{lo}}$ by projection (2.58) using \mathbf{W} and \mathbf{V} :

$$\begin{aligned} \tilde{\mathbf{E}} &= \mathbf{I}_r, \\ \tilde{\mathbf{A}} &= \mathbf{\Sigma}_1^{-1/2} \mathbf{U}_1^\top (\mathbf{L}_\mathcal{Y}^\top \mathbf{A} \mathbf{R}_\mathcal{X}) \mathbf{Y}_1 \mathbf{\Sigma}_1^{-1/2}, \\ \tilde{\mathbf{B}} &= \mathbf{\Sigma}_1^{-1/2} \mathbf{U}_1^\top (\mathbf{L}_\mathcal{Y}^\top \mathbf{B}), \\ \tilde{\mathbf{C}} &= (\mathbf{C} \mathbf{R}_\mathcal{X}) \mathbf{Y}_1 \mathbf{\Sigma}_1^{-1/2}, \\ \text{and } \tilde{\mathbf{D}} &= \mathbf{D}. \end{aligned} \tag{4.3}$$

to the deterministic quantities in (4.4) that are entirely computable (up to quadrature errors) from various state-space invariant input-output data; see Theorem 4.1. These quadrature rules are never explicitly formed but rather form the basis for our analysis. This generalized presentation contains QuadBT as a special case; see Remark 4.2. We note that the results that follow from this insight are not significantly different from those in [89] at face value. Rather, the power of this abstraction lies in its generality. Indeed, multiple BT variants based on balancing the solutions to algebraic Riccati equations can be framed in this abstract Lyapunov setting, and it provides the foundation for the data-driven formulations of BST, PRBT, and BRBT that are presented in Section 4.3.

Recall the linear systems \mathcal{X} and \mathcal{Y} introduced in (4.2). Because the Gramians $\mathbf{P}_\mathcal{X}, \mathbf{Q}_\mathcal{Y} \in \mathbb{R}^{n \times n}$ being balanced in this abstract formulation are solutions to Lyapunov equations (4.1), they can be uniquely expressed as contour integrals and are amenable to low-rank factoriza-

tions via (implicit) numerical quadrature rules. Specifically, we have:

$$\mathbf{P}_X = \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\omega \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X ((i\omega \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X)^H d\omega, \quad (4.5)$$

$$\mathbf{Q}_Y = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbf{C}_Y (i\omega \mathbf{I}_n - \mathbf{A})^{-1})^H \mathbf{C}_Y (i\omega \mathbf{I}_n - \mathbf{A})^{-1} d\omega. \quad (4.6)$$

Consider a numerical quadrature rule defined by the nodes $i\zeta_1, \dots, i\zeta_J \in i\mathbb{R}$ and weights $\varrho_1^2, \dots, \varrho_J^2 \in \mathbb{R}$. Applying this rule to \mathbf{P}_X in (4.5) reveals the approximate factorization

$$\mathbf{P}_X \approx \sum_{j=1}^J \varrho_j^2 (i\zeta_j \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X ((i\zeta_j \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X)^H = \check{\mathbf{R}}_X \check{\mathbf{R}}_X^H,$$

where $\check{\mathbf{R}}_X \in \mathbb{C}^{n \times m_x J}$ is defined as

$$\check{\mathbf{R}}_X \stackrel{\text{def}}{=} [\varrho_1 (i\zeta_1 \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X \quad \varrho_2 (i\zeta_2 \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X \quad \cdots \quad \varrho_J (i\zeta_J \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_X]. \quad (4.7)$$

We assume that the factor $1/2\pi$ is included in each of the quadrature weights. Likewise, applying a numerical quadrature rule defined by the nodes $i\vartheta_1, \dots, i\vartheta_K \in i\mathbb{R}$ and weights $\varphi_1^2, \dots, \varphi_K^2 \in \mathbb{R}$ to \mathbf{Q}_Y produces the approximate factorization

$$\mathbf{Q}_Y \approx \sum_{k=1}^K \varphi_k^2 (\mathbf{C}_Y (i\vartheta_k \mathbf{I}_n - \mathbf{A})^{-1})^H \mathbf{C}_Y (i\vartheta_k \mathbf{I}_n - \mathbf{A})^{-1} = \check{\mathbf{L}}_Y \check{\mathbf{L}}_Y^H,$$

where $\check{\mathbf{L}}_Y \in \mathbb{C}^{n \times p_y K}$ is defined according to

$$\check{\mathbf{L}}_Y^H \stackrel{\text{def}}{=} \begin{bmatrix} \varphi_1 \mathbf{C}_Y (i\vartheta_1 \mathbf{I}_n - \mathbf{A})^{-1} \\ \varphi_2 \mathbf{C}_Y (i\vartheta_2 \mathbf{I}_n - \mathbf{A})^{-1} \\ \vdots \\ \varphi_K \mathbf{C}_Y (i\vartheta_K \mathbf{I}_n - \mathbf{A})^{-1} \end{bmatrix}. \quad (4.8)$$

Replacing the exact Cholesky factors \mathbf{R}_X and \mathbf{L}_Y in (4.4) with the quadrature-based factors in (4.7) and (4.8) already yields a low-rank implementation of Algorithm 4.2.1. The resulting (approximate) BT-ROM is determined by the quadrature-based approximations to (4.4), namely

$$\check{\mathbf{L}}_Y^H \check{\mathbf{R}}_X, \quad \check{\mathbf{L}}_Y^H \mathbf{A} \check{\mathbf{R}}_X, \quad \check{\mathbf{L}}_Y^H \mathbf{B}, \quad \text{and} \quad \mathbf{C} \check{\mathbf{R}}_X. \quad (4.9)$$

We recall from our previous discussion that (4.9) provides everything one needs to compute a BT-ROM. Significantly, the modified quantities in (4.9) can be computed *entirely from transfer function data*. We prove this next. For the results that follow, we recall the notation described in (2.1) for block matrices: For a (block) matrix $\mathbf{X} \in \mathbb{C}^{p_y K \times m_x J}$, we use

$$\mathbf{X}_{j,k} \stackrel{\text{def}}{=} \mathbf{X}_{(k-1)p_y+1:kp_y, (j-1)m_x+1:jm_x} \in \mathbb{C}^{K \times J},$$

to denote the block submatrix of \mathbf{X} containing the rows $(k-1)p_y+1, \dots, kp_y$ and the columns $(j-1)m_x+1, \dots, jm_x$. If $p_y = 1$ (or $m_x = 1$) we instead write $\mathbf{X}_{:,j}$ ($\mathbf{X}_{k,:}$) to denote the j -th (k -th) block entry of \mathbf{X} .

Theorem 4.1. Define $\mathbf{G}_{\sigma, \mathbf{A}}: \mathbb{C} \rightarrow \mathbb{C}^{p_y \times m_x}$, $\mathbf{G}_{\mathbf{B}}: \mathbb{C} \rightarrow \mathbb{C}^{p_y \times m}$, and $\mathbf{G}_{\mathbf{C}}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m_x}$ by

$$\mathbf{G}_{\sigma, \mathbf{A}}(s) \stackrel{\text{def}}{=} \mathbf{C}_{\mathcal{Y}} (s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{X}}, \quad (4.10a)$$

$$\mathbf{G}_{\mathbf{B}}(s) \stackrel{\text{def}}{=} \mathbf{C}_{\mathcal{Y}} (s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}, \quad (4.10b)$$

$$\text{and } \mathbf{G}_{\mathbf{C}}(s) \stackrel{\text{def}}{=} \mathbf{C} (s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{X}}. \quad (4.10c)$$

Let the quadrature-based square-root factors $\check{\mathbf{R}}_{\mathcal{X}} \in \mathbb{C}^{n \times m_x J}$ and $\check{\mathbf{L}}_{\mathcal{Y}} \in \mathbb{C}^{n \times p_y K}$ be defined according to (4.7) and (4.8). Define the so-called data matrices

$$\begin{aligned} \mathbb{E} &\stackrel{\text{def}}{=} \check{\mathbf{L}}_{\mathcal{Y}}^H \check{\mathbf{R}}_{\mathcal{X}} \in \mathbb{C}^{p_y K \times m_x J}, & \mathbb{A} &\stackrel{\text{def}}{=} \check{\mathbf{L}}_{\mathcal{Y}}^H \mathbf{A} \check{\mathbf{R}}_{\mathcal{X}} \in \mathbb{C}^{p_y K \times m_x J}, \\ \mathbb{B} &\stackrel{\text{def}}{=} \check{\mathbf{L}}_{\mathcal{Y}}^H \mathbf{B} \in \mathbb{C}^{p_y K \times m}, & \mathbb{C} &\stackrel{\text{def}}{=} \mathbf{C} \check{\mathbf{R}}_{\mathcal{X}} \in \mathbb{C}^{p \times m_x J}. \end{aligned} \quad (4.11)$$

Then, the entries of the data matrices in (4.11) are defined by

$$\mathbb{E}_{\mathbf{k}, \mathbf{j}} = -\varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \frac{\mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\vartheta_{\mathbf{k}}) - \mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\zeta_{\mathbf{j}})}{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} - \dot{\mathbf{i}}\zeta_{\mathbf{j}}}, \quad (4.12a)$$

$$\mathbb{A}_{\mathbf{k}, \mathbf{j}} = -\varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \frac{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} \mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\vartheta_{\mathbf{k}}) - \dot{\mathbf{i}}\zeta_{\mathbf{j}} \mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\zeta_{\mathbf{j}})}{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} - \dot{\mathbf{i}}\zeta_{\mathbf{j}}}, \quad (4.12b)$$

$$\mathbb{B}_{\mathbf{k}, \cdot} = \varphi_{\mathbf{k}} \mathbf{G}_{\mathbf{B}}(\dot{\mathbf{i}}\vartheta_{\mathbf{k}}) \text{ and } \mathbb{C}_{\cdot, \mathbf{j}} = \varrho_{\mathbf{j}} \mathbf{G}_{\mathbf{C}}(\dot{\mathbf{i}}\zeta_{\mathbf{j}}), \quad (4.12c)$$

for all $\mathbf{j} = 1, \dots, J$ and $\mathbf{k} = 1, \dots, K$. \diamond

Proof of Theorem 4.1. For any $i, \ell \in \mathbb{N}$, we introduce the matrix

$$\mathbf{I}_{i, \ell} = [\mathbf{e}_{(i-1)\ell+1} \quad \mathbf{e}_{(i-1)\ell+2} \quad \cdots \quad \mathbf{e}_{i\ell}] \in \mathbb{R}^{n \times \ell}. \quad (4.13)$$

Note $\mathbf{I}_{i, \ell}$ contains a subset of the columns of the $n \times n$ identity and has the effect of retrieving columns $(i-1)\ell+1$ through $i\ell$ of a matrix by right multiplication. The proof is a generalization of standard algebraic manipulations found in, e.g., [89], and heavily exploits the two resolvent identities in Lemma 2.7. First, by the definition of \mathbb{E} in (4.11), $\check{\mathbf{L}}_{\mathcal{Y}}$ in (4.8), and $\check{\mathbf{R}}_{\mathcal{X}}$ in (4.7), applying the first resolvent identity (2.16) gives

$$\begin{aligned} \mathbb{E}_{\mathbf{k}, \mathbf{j}} &= \mathbf{I}_{k, p_y}^T \mathbb{E} \mathbf{I}_{j, m_x} = \mathbf{I}_{k, p_y}^T \check{\mathbf{L}}_{\mathcal{Y}}^H \check{\mathbf{R}}_{\mathcal{X}} \mathbf{I}_{j, m_x} = \varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \mathbf{C}_{\mathcal{Y}} (\dot{\mathbf{i}}\vartheta_{\mathbf{k}} \mathbf{I}_n - \mathbf{A})^{-1} (\dot{\mathbf{i}}\zeta_{\mathbf{j}} \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{X}} \\ &= \varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \mathbf{C}_{\mathcal{Y}} \left(\frac{(\dot{\mathbf{i}}\zeta_{\mathbf{j}} \mathbf{I}_n - \mathbf{A})^{-1} - (\dot{\mathbf{i}}\vartheta_{\mathbf{k}} \mathbf{I}_n - \mathbf{A})^{-1}}{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} - \dot{\mathbf{i}}\zeta_{\mathbf{j}}} \right) \mathbf{B}_{\mathcal{X}} = -\varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \frac{\mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\vartheta_{\mathbf{k}}) - \mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\zeta_{\mathbf{j}})}{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} - \dot{\mathbf{i}}\zeta_{\mathbf{j}}}, \end{aligned}$$

where the last line follows from the definition of $\mathbf{G}_{\sigma, \mathbf{A}}(s)$ in (4.10a). This proves (4.12a). Similarly, by the definition of \mathbb{A} in (4.11) and $\check{\mathbf{L}}_{\mathcal{Y}}$ and $\check{\mathbf{R}}_{\mathcal{X}}$, applying the second resolvent identity in (2.17) gives

$$\begin{aligned} \mathbb{A}_{\mathbf{k}, \mathbf{j}} &= \mathbf{I}_{k, p_y}^T \mathbb{A} \mathbf{I}_{j, m_x} = \mathbf{I}_{k, p_y}^T \check{\mathbf{L}}_{\mathcal{Y}}^H \mathbf{A} \check{\mathbf{R}}_{\mathcal{X}} \mathbf{I}_{j, m_x} = \varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \mathbf{C}_{\mathcal{Y}} (\dot{\mathbf{i}}\vartheta_{\mathbf{k}} \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{A} (\dot{\mathbf{i}}\zeta_{\mathbf{j}} \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{X}} \\ &= \varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \mathbf{C}_{\mathcal{Y}} \left(\frac{\dot{\mathbf{i}}\zeta_{\mathbf{j}} (\dot{\mathbf{i}}\zeta_{\mathbf{j}} \mathbf{I}_n - \mathbf{A})^{-1} - \dot{\mathbf{i}}\vartheta_{\mathbf{k}} (\dot{\mathbf{i}}\vartheta_{\mathbf{k}} \mathbf{I}_n - \mathbf{A})^{-1}}{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} - \dot{\mathbf{i}}\zeta_{\mathbf{j}}} \right) \mathbf{B}_{\mathcal{X}} \\ &= -\varphi_{\mathbf{k}} \varrho_{\mathbf{j}} \frac{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} \mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\vartheta_{\mathbf{k}}) - \dot{\mathbf{i}}\zeta_{\mathbf{j}} \mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\zeta_{\mathbf{j}})}{\dot{\mathbf{i}}\vartheta_{\mathbf{k}} - \dot{\mathbf{i}}\zeta_{\mathbf{j}}}, \end{aligned}$$

which proves (4.12b). The formulae in (4.12c) for \mathbb{B} and \mathbb{C} follow directly from their definition: Observe that

$$\mathbb{B}_{k,:} = \mathbf{I}_{k,p_y}^\top \mathbb{B} = \mathbf{I}_{k,p_y}^\top \check{\mathbf{L}}_y^H \mathbf{B} = \varphi_k \mathbf{C}_y (\dot{\mathbf{v}}_k \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} = \varphi_k \mathbf{G}_B(\dot{\mathbf{v}}_k),$$

and likewise

$$\mathbb{C}_{:j} = \mathbb{C} \mathbf{I}_{j,m_x} = \mathbf{C} \check{\mathbf{R}}_x \mathbf{I}_{j,m_x} = \varrho_j \mathbf{C} (\dot{\mathbf{i}}_j \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_x = \varrho_j \mathbf{G}_C(\dot{\mathbf{i}}_j),$$

thus completing the proof. \square

By replacing the exact quantities in (4.4) with the (approximate) quadrature-based quantities in (4.11) that are computed from data, we obtain a completely data-driven formulation of Algorithm 4.2.1. We refer to this as *generalized quadrature-based balanced truncation* (GenQuadBT); its algorithmic formulation is presented in Algorithm 4.2.2. The choice of notation $\mathbf{G}_{\sigma,A}$, \mathbf{G}_B , and \mathbf{G}_C for the transfer functions in (4.10a)–(4.10c) is intentional: *the underscored quantities in each transfer function are the quantities in the data-driven reduced model that require samples of that transfer function*. Put differently, Theorem 4.1 and the associated notation can be interpreted as follows.

1. The construction of \mathbb{E} and hence its SVD, as well as the reduced-order $\tilde{\mathbf{A}}$ in Steps 2 and 3 of Algorithm 4.2.2 require samples of $\mathbf{G}_{\sigma,A}(s)$ at the left and right quadrature nodes;
2. The construction of the reduced-order $\tilde{\mathbf{B}}$ in Step 3 of Algorithm 4.2.2 requires samples of $\mathbf{G}_B(s)$ at the left quadrature nodes;
3. The construction of the reduced-order $\tilde{\mathbf{C}}$ in Step 3 of Algorithm 4.2.2 requires samples of $\mathbf{G}_C(s)$ at the right quadrature nodes.

Thus, in principle, Algorithm 4.2.2 requires only the left and right quadrature weights and nodes used to implicitly approximate \mathbf{P}_x and \mathbf{Q}_y , and samples of the transfer functions $\mathbf{G}_{\sigma,A}$, \mathbf{G}_B and \mathbf{G}_C given in (4.10a)–(4.10c) evaluated at these nodes (or, at least the ability to evaluate them). We emphasize that at no point do we explicitly compute the quadrature-based approximations of the Gramians \mathbf{P}_x and \mathbf{Q}_y . Indeed, these are leveraged *implicitly* to derive the quadrature-based square-root factors in (4.7) and (4.8) and subsequently realize the quantities in (4.11) from input-to-output data. There is also an error analysis in [89, Proposition 3.2] that translates directly to this generalized setting. The key deviation of Theorem 4.1 and Algorithm 4.2.2 from the work of [89] is that the transfer function evaluations required in this generalized setting are not necessarily those of $\mathbf{G}_{\mathbf{I}_o}$, the transfer function of the underlying system. Nonetheless, Algorithm 4.2.2 avoids any explicit reference to internal quantities; e.g., a state-space realization of $\mathcal{G}_{\mathbf{I}_o}$, or any other linear model. Because $\mathbf{D} = \lim_{s \rightarrow \infty} \mathbf{G}_{\mathbf{I}_o}(s)$, the feedthrough term \mathbf{D} included as an input to Algorithm 4.2.2 can also be obtained from transfer function data.

Algorithm 4.2.2: Generalized quadrature-based balanced truncation [184].

Input: Mappings $\mathbf{G}_{\sigma, \mathbf{A}}$, \mathbf{G}_B , \mathbf{G}_C , left, right quadrature nodes and weights

$$\{\dot{\mathbf{i}}\vartheta_k, \varphi_k\}_{k=1}^K, \{\dot{\mathbf{i}}\zeta_j, \varrho_j\}_{j=1}^J, \text{ order } r \ (1 \leq r < n), \text{ and } \mathbf{D} \in \mathbb{R}^{p \times m}.$$

Output: $\tilde{\mathbf{A}}$, $\tilde{\mathbf{B}}$, $\tilde{\mathbf{C}}$, $\tilde{\mathbf{D}}$ —state-space matrices of (2.55).

- 1 Evaluate the mappings to obtain the data $\{\mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\vartheta_k)\}_{k=1}^K$, $\{\mathbf{G}_{\sigma, \mathbf{A}}(\dot{\mathbf{i}}\zeta_j)\}_{j=1}^J$, $\{\mathbf{G}_B(\dot{\mathbf{i}}\vartheta_k)\}_{k=1}^K$ and $\{\mathbf{G}_C(\dot{\mathbf{i}}\zeta_j)\}_{j=1}^J$ and construct the data matrices in (4.11) according to Theorem 4.1.
- 2 Compute the singular value decomposition of $\mathbb{E} \in \mathbb{C}^{p_y K \times m_x J}$ partitioned according to

$$\mathbb{E} = \check{\mathbf{U}} \check{\mathbf{\Sigma}} \check{\mathbf{Y}}^\top = \begin{bmatrix} \check{\mathbf{U}}_1 & \check{\mathbf{U}}_2 \end{bmatrix} \begin{bmatrix} \check{\mathbf{\Sigma}}_1 & \\ & \check{\mathbf{\Sigma}}_2 \end{bmatrix} \begin{bmatrix} \check{\mathbf{Y}}_1^\top \\ \check{\mathbf{Y}}_2^\top \end{bmatrix},$$

for $\check{\mathbf{\Sigma}}_1 \in \mathbb{R}^{r \times r}$, $\check{\mathbf{\Sigma}}_2 \in \mathbb{R}^{(p_y K - r) \times (m_x J - r)}$, and $\check{\mathbf{U}}_1, \check{\mathbf{U}}_2, \check{\mathbf{Y}}_1, \check{\mathbf{Y}}_2$ partitioned conformally.

- 3 Compute the data-driven reduced model according to:

$$\begin{aligned} \tilde{\mathbf{E}} &= \mathbf{I}_r, \\ \tilde{\mathbf{A}} &= \check{\mathbf{\Sigma}}_1^{-1/2} \check{\mathbf{U}}_1^\top (\mathbb{A}) \check{\mathbf{Y}}_1 \check{\mathbf{\Sigma}}_1^{-1/2}, \\ \tilde{\mathbf{B}} &= \check{\mathbf{\Sigma}}_1^{-1/2} \check{\mathbf{U}}_1^\top (\mathbb{B}), \\ \tilde{\mathbf{C}} &= (\mathbb{C}) \check{\mathbf{Y}}_1 \check{\mathbf{\Sigma}}_1^{-1/2}, \\ \text{and } \tilde{\mathbf{D}} &= \mathbf{D}. \end{aligned} \tag{4.14}$$

Remark 4.2. Theorem 4.1 and Algorithm 4.2.2 contain the original QuadBT of [89] as a special case. Indeed, in the Lyapunov setting, we simply have that $\mathbf{Q}_y = \mathbf{Q}_{\mathbf{I}_o}$ and $\mathbf{P}_x = \mathbf{P}$, and so the generalized equations (4.1) are the dual ALEs (2.44) and (2.43) corresponding to $\mathcal{G}_{\mathbf{I}_o}$. Then, $\mathbf{B}_x = \mathbf{B}$, $\mathbf{C}_y = \mathbf{C}$, and the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}$, \mathbf{G}_B , and \mathbf{G}_C all equal $\mathbf{G}_{\mathbf{I}_o} - \mathbf{D}$. This is precisely the aggregate result of [89, Proposition 3.1 and 3.3]. \diamond

4.3 What to sample for balanced truncation variants

What remains to be seen is what the contrived transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}$, \mathbf{G}_B , and \mathbf{G}_C in (4.10a)–(4.10c) correspond to for different BT variants. We investigate this question next for BST, PRBT, and BRBT. Applying the generalized result of Theorem 4.1 to each of these variants, we answer the question: What do you need to sample for different (data-driven) balanced reduced models? Unlike in the Lyapunov setting, the to-be-sampled data are not necessarily measurements of $\mathbf{G}_{\mathbf{I}_o}$, the transfer function of the underlying system. We

ultimately show that for each of BST, PRBT, and BRBT, the general quantities in (4.10a)–(4.10c) can be interpreted in terms of certain *spectral factorizations*.

In the rest of this section, we sequentially derive data-based formulations of BST, PRBT, and BRBT according to the following structure.

1. First, we review the key details of these methods and introduce the relevant Gramians, as well as the spectral factorizations that arise from the matrix equations that determine the said Gramians.
2. Second, we describe how each variant fits the abstract Lyapunov framework of Section 4.2.1. In particular, we provide specific formulations for the linear systems \mathcal{X} and \mathcal{Y} in (4.2) as well as the agnostic Gramians $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ that solve (4.1). This enables us to apply the generalized framework of Section 4.2.
3. Finally, we interpret the result of Theorem 4.1 applied to the type of BT in question to derive explicit expressions of the transfer functions (4.10a)–(4.10c).

Our review of the BT variants that we consider uses [96] as its primary source. For a comprehensive overview of balancing-based model reduction, we refer to the surveys and book chapters [25, 46, 96], [4, Ch. 7]. Many variants of BT, including a subset of those introduced here, can be framed in the language of dissipative systems and supply rates [230, 231]; a very nice presentation from this perspective can be found in [46, Sec. 2.3]. Our treatment of spectral factorizations follows [245, Chapter 13.4]; we refer the reader there for a more detailed study.

For the results in this section, we use the following notation and definitions: When $\mathcal{G}_{\text{lo}} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ does not fit in a single line, we represent the corresponding system by

$$\mathcal{G}_{\text{lo}} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array} \right).$$

For linear systems (2.25) with a nontrivial \mathbf{D} term, and thus a proper transfer function \mathbf{G}_{lo} , we use the notation $\mathbf{G}_{\text{lo},\infty}$ to denote the *strictly proper* part of \mathbf{G}_{lo} , i.e.,

$$\mathbf{G}_{\text{lo},\infty}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} = \mathbf{G}_{\text{lo}}(s) - \mathbf{D}, \quad s \in \mathbb{C}. \quad (4.15)$$

If a system \mathcal{G}_{lo} of the form (2.25) is *not* asymptotically stable, we may decompose its transfer function (2.30) as $\mathbf{G}_{\text{lo}} = \mathbf{G}_{\text{lo}}^- + \mathbf{G}_{\text{lo}}^+$, where \mathbf{G}_{lo}^- has poles in $\mathbb{C}_{<0}$ (and is thus analytic in $\mathbb{C}_{\geq 0}$) and \mathbf{G}_{lo}^+ has poles in $\mathbb{C}_{\geq 0}$ (and is thus analytic in $\mathbb{C}_{<0}$). For a linear system \mathcal{G}_{lo} in (2.25) with the transfer function \mathbf{G}_{lo} , we refer to the linear subsystem $\mathcal{G}_{\text{lo}}^-$ corresponding to \mathbf{G}_{lo}^- in the decomposition $\mathbf{G}_{\text{lo}} = \mathbf{G}_{\text{lo}}^- + \mathbf{G}_{\text{lo}}^+$ as the *purely stable* part of \mathcal{G}_{lo} . Likewise, we call \mathbf{G}_{lo}^+ the *purely stable* part of the transfer function \mathbf{G}_{lo} .

4.3.1 Data-driven balanced stochastic truncation

Intrusive balanced stochastic truncation.

Model reduction by balanced stochastic truncation (BST) can be viewed as a *spectral factor-based* algorithm for approximating a linear system (2.25). BST was first introduced in [61] for the model reduction of stochastic processes, and further studied in [33, 92, 93, 105, 219]. Compared to Lyapunov balanced truncation, BST provides an *a priori* error bound on the relative approximation error, and preserves asymptotic stability as well as the minimum phase property (Definition 2.22) of a linear system.

Consider an asymptotically stable, minimal, and square ($m = p$) linear system \mathcal{G}_{lo} as in (2.25) with a nonsingular feedthrough term \mathbf{D} . By [245, Corollary 13.28] there exists a *minimum phase right spectral factor* $\mathbf{W} : \mathbb{C} \rightarrow \mathbb{C}^{m \times m}$ of $\mathbf{G}_{\text{lo}} \mathbf{G}_{\text{lo}}^{\text{H}}$, i.e., \mathbf{W} satisfies

$$\mathbf{W}(-s)^{\text{T}} \mathbf{W}(s) = \mathbf{G}_{\text{lo}}(s) \mathbf{G}_{\text{lo}}(-s)^{\text{T}}, \quad s \in \mathbb{C}.$$

The spectral factor \mathbf{W} is the transfer function of a minimal, asymptotically stable linear system (2.25), denoted \mathcal{W} , having the realization

$$\begin{aligned} \mathcal{W} &= \left(\mathbf{A}, \mathbf{B}_{\mathcal{W}}, \mathbf{C}_{\mathcal{W}}, (\mathbf{D}\mathbf{D}^{\text{T}})^{1/2} \right) \\ \text{for } \mathbf{B}_{\mathcal{W}} &\stackrel{\text{def}}{=} \mathbf{P}\mathbf{C}^{\text{T}} + \mathbf{B}\mathbf{D}^{\text{T}} \quad \text{and} \quad \mathbf{C}_{\mathcal{W}} \stackrel{\text{def}}{=} (\mathbf{D}\mathbf{D}^{\text{T}})^{-1/2} (\mathbf{C} - \mathbf{B}_{\mathcal{W}}^{\text{T}} \mathbf{Q}_{\mathcal{W}}^{-}), \end{aligned} \quad (4.16)$$

where $\mathbf{P} \in \mathbb{R}^{n \times n}$ is the reachability Gramian of \mathcal{G}_{lo} that solves (2.43), while $\mathbf{Q}_{\mathcal{W}}^{-} \in \mathbb{R}^{n \times n}$ is the *minimal* (stabilizing) solution to the ARE

$$\mathbf{A}^{\text{T}} \mathbf{Q}_{\mathcal{W}} + \mathbf{Q}_{\mathcal{W}} \mathbf{A} + (\mathbf{C} - \mathbf{B}_{\mathcal{W}}^{\text{T}} \mathbf{Q}_{\mathcal{W}})^{\text{T}} (\mathbf{D}\mathbf{D}^{\text{T}})^{-1} (\mathbf{C} - \mathbf{B}_{\mathcal{W}}^{\text{T}} \mathbf{Q}_{\mathcal{W}}) = \mathbf{0}_{n \times n}. \quad (4.17)$$

The *minimal* solution to (4.17) is the unique SPD matrix $\mathbf{Q}_{\mathcal{W}}^{-}$ that obeys the partial order $\mathbf{0} \prec \mathbf{Q}_{\mathcal{W}}^{-} \preceq \mathbf{Q}_{\mathcal{W}}$ for all symmetric solutions $\mathbf{Q}_{\mathcal{W}}$ of (4.17); see [245, Theorem 13.11]. Explicitly, \mathbf{W} is written as

$$\mathbf{W}(s) = \mathbf{C}_{\mathcal{W}}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{W}} + \mathbf{D}\mathbf{D}^{\text{T}}, \quad s \in \mathbb{C}.$$

We call the factor \mathbf{W} *minimum phase* because the solution $\mathbf{Q}_{\mathcal{W}}$ to (4.17) used in its construction is the minimal solution to (4.17). By [245, Thm. 13.11], $\mathbf{Q}_{\mathcal{W}}^{-}$ is unique and SPD. In BST, $\mathbf{Q}_{\mathcal{W}}^{-}$ takes the place of the usual observability Gramian $\mathbf{Q}_{\text{lo}} \in \mathbb{R}^{n \times n}$ and is balanced against the reachability Gramian $\mathbf{P} \in \mathbb{R}^{n \times n}$. The model order is then reduced by truncating the trailing $n - r$ components of the state space in these so-called balanced stochastic coordinates.

Definition 4.3 (Balanced stochastic realization [219]). We say that a state-space realization of the minimal system \mathcal{G}_{lo} in (2.25) is a *balanced stochastic realization* if

$$\mathbf{P} = \mathbf{Q}_{\mathcal{W}}^{-} = \Sigma^{\text{bst}} = \text{diag}(\varsigma_1 \mathbf{I}_{m_1}, \varsigma_2 \mathbf{I}_{m_2}, \dots, \varsigma_q \mathbf{I}_{m_q}), \quad (4.18)$$

where $1 \geq \varsigma_1 > \varsigma_2 > \dots > \varsigma_q > 0$ and their multiplicities satisfy $m_1 + \dots + m_q = n$. The values ς_i are called the *stochastic singular values* of \mathcal{G}_{lo} . \diamond

Theorem 4.4 (Balanced stochastic truncation model reduction [61, 92, 93, 105]). Consider an asymptotically stable, minimal linear system \mathcal{G}_{lo} in (2.25) having the balanced stochastic realization

$$\mathbf{A}_{\text{bst}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B}_{\text{bst}} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \text{and} \quad \mathbf{C}_{\text{bst}} = [\mathbf{C}_1 \quad \mathbf{C}_2]$$

according to Definition 4.3. The matrices are partitioned with respect to $\mathbf{P} = \mathbf{Q}_{\mathcal{W}}^- = \boldsymbol{\Sigma}^{\text{bst}} = \text{diag}(\boldsymbol{\Sigma}_1^{\text{bst}}, \boldsymbol{\Sigma}_2^{\text{bst}})$, where

$$\boldsymbol{\Sigma}_1^{\text{bst}} = \text{diag}(\varsigma_1 \mathbf{I}_{m_1}, \dots, \varsigma_k \mathbf{I}_{m_k}) \quad \text{and} \quad \boldsymbol{\Sigma}_2^{\text{bst}} = \text{diag}(\varsigma_{k+1} \mathbf{I}_{m_{k+1}}, \dots, \varsigma_q \mathbf{I}_{m_q})$$

for $r = m_1 + \dots + m_k$ and $1 \leq k < q$. Then the order- r reduced model $\tilde{\mathcal{G}}_{\text{lo,bst}} = (\mathbf{A}_{11}, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D})$ obtained via balanced stochastic truncation is balanced in the sense of (4.18), asymptotically stable, minimal, and satisfies the relative \mathcal{H}_∞ error bound

$$\|\mathcal{G}_{\text{lo}}^{-1}(\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo,bst}})\|_{\mathcal{H}_\infty} \leq \prod_{i=k+1}^q \frac{1 + \varsigma_i}{1 - \varsigma_i} - 1.$$

Moreover, if \mathcal{G}_{lo} is minimum phase, then $\tilde{\mathcal{G}}_{\text{lo,bst}}$ is as well. \diamond

The inverse system $\mathcal{G}_{\text{lo}}^{-1}$ is well-defined according to Proposition 2.25. The fact that BST preserves asymptotic stability and minimality was proven in [105], while preservation of the minimum phase property was shown in [92]. The relative \mathcal{H}_∞ error bound is due to Green [93]. Both [33, 219] investigate numerically reliable algorithms for BST.

Quadrature-based balanced stochastic truncation.

To derive a data-driven formulation of BST, we describe how it fits in the GenQuadBT framework of Section 4.2. Specifically, we need to show that the Gramians in BST, $\mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{Q}_{\mathcal{W}}^- \in \mathbb{R}^{n \times n}$, are the reachability and observability Gramians $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{Q}_{\mathcal{Y}}$ of some linear systems \mathcal{X} and \mathcal{Y} . Obviously, $\mathbf{P}_{\mathcal{X}} = \mathbf{P}$ is the reachability Gramian of the system being approximated, and so

$$\mathcal{X} = (\mathbf{A}, \mathbf{B}_{\mathcal{X}}, \mathbf{C}_{\mathcal{X}}, \mathbf{D}_{\mathcal{X}}) = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) = \mathcal{G}_{\text{lo}},$$

with \mathbf{P} solving (2.43). For \mathcal{Y} , recall the definition of $\mathbf{C}_{\mathcal{W}} = (\mathbf{D}\mathbf{D}^\top)^{-1/2}(\mathbf{C} - \mathbf{B}_{\mathcal{W}}\mathbf{Q}_{\mathcal{W}}^-)$ from (4.16). Fixing the right-hand side of the ARE (4.17) with $\mathbf{Q}_{\mathcal{W}} = \mathbf{Q}_{\mathcal{W}}^-$ produces an ALE:

$$\begin{aligned} \mathbf{A}^\top \mathbf{Q}_{\mathcal{W}} + \mathbf{Q}_{\mathcal{W}} \mathbf{A} + \underbrace{(\mathbf{C} - \mathbf{B}_{\mathcal{W}}^\top \mathbf{Q}_{\mathcal{W}}^-)^\top (\mathbf{D}\mathbf{D}^\top)^{-1} (\mathbf{C} - \mathbf{B}_{\mathcal{W}}^\top \mathbf{Q}_{\mathcal{W}}^-)}_{= \mathbf{C}_{\mathcal{W}}^\top \mathbf{C}_{\mathcal{W}}} &= \mathbf{0}_{n \times n} \\ \implies \mathbf{A}^\top \mathbf{Q}_{\mathcal{W}} + \mathbf{Q}_{\mathcal{W}} \mathbf{A} + \mathbf{C}_{\mathcal{W}}^\top \mathbf{C}_{\mathcal{W}} &= \mathbf{0}_{n \times n}. \end{aligned} \quad (4.19)$$

Clearly, the ALE (4.19) is uniquely solved by $\mathbf{Q}_{\mathcal{W}}^-$. It follows from the definition of $\mathbf{C}_{\mathcal{W}}$ that (4.19) is the *observability Lyapunov equation* of $\mathcal{Y} = \mathcal{W}$, the system (4.16) associated with the spectral factor \mathbf{W} , and $\mathbf{Q}_{\mathcal{Y}} = \mathbf{Q}_{\mathcal{W}}^-$ is the *observability Gramian* of $\mathcal{Y} = \mathcal{W}$, i.e.,

$$\mathcal{Y} = (\mathbf{A}, \mathbf{B}_{\mathcal{Y}}, \mathbf{C}_{\mathcal{Y}}, \mathbf{D}_{\mathcal{Y}}) = (\mathbf{A}, \mathbf{B}_{\mathcal{W}}, \mathbf{C}_{\mathcal{W}}, (\mathbf{D}\mathbf{D}^\top)^{1/2}) = \mathcal{W}.$$

We are now at a point where we can apply the GenQuadBT framework of Section 4.2 to $\mathcal{X} = \mathcal{G}_{\text{lo}}$ and $\mathcal{Y} = \mathcal{W}$ with $\mathbf{P}_{\mathcal{X}} = \mathbf{P}$ and $\mathbf{Q}_{\mathcal{Y}} = \mathbf{Q}_{\mathcal{W}}^-$. The relevant Gramians are decomposed into the quadrature-based factors $\check{\mathbf{R}}_{\mathcal{X}}$ and $\check{\mathbf{L}}_{\mathcal{Y}}$ defined according to (4.7) and (4.8) with $\mathbf{B}_{\mathcal{X}} = \mathbf{B}$ and $\mathbf{C}_{\mathcal{Y}} = \mathbf{C}_{\mathcal{W}}$ in (4.16). From this particular choice of Gramians, we derive explicit expressions for the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}$, $\mathbf{G}_{\mathbf{B}}$, and $\mathbf{G}_{\mathbf{C}}$ (4.10a)–(4.10c) for the setting of data-driven BST.

Theorem 4.5 (Balanced stochastic truncation from data). Let $\mathbf{Q}_{\mathcal{W}}^- \in \mathbb{R}^{n \times n}$ be the stabilizing solution to (4.19) and $\mathbf{P} \in \mathbb{R}^{n \times n}$ be the reachability Gramian of \mathcal{G} that solves (2.43). Then, for BST the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}$, $\mathbf{G}_{\mathbf{B}}$, and $\mathbf{G}_{\mathbf{C}}$ defined as in (4.10a)–(4.10c) of Theorem 4.1 are given by

$$\mathbf{G}_{\mathbf{C}}(s) = \mathbf{G}_{\text{lo}, \infty}(s), \text{ and} \quad (4.20)$$

$$\mathbf{G}_{\sigma, \mathbf{A}}(s) = \mathbf{G}_{\mathbf{B}}(s) = \left((\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo}, \infty}(s) \right)^-, \quad (4.21)$$

where $\mathbf{G}_{\text{lo}, \infty}$ is the strictly proper part of the transfer function of the linear system \mathcal{G}_{lo} in (2.30), and \mathbf{W} is the transfer function of \mathcal{W} in (4.16). \diamond

Proof of Theorem 4.5. In the setting of BST, it holds that $\mathbf{B}_{\mathcal{X}} = \mathbf{B}$. Thus, by definition of $\mathbf{G}_{\mathbf{C}}$ in (4.10c),

$$\mathbf{G}_{\mathbf{C}}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{X}} = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} = \mathbf{G}_{\text{lo}, \infty}(s),$$

proving (4.20). To prove (4.21), first observe that

$$\mathbf{G}_{\sigma, \mathbf{A}}(s) = \mathbf{G}_{\mathbf{B}}(s) = \mathbf{C}_{\mathcal{W}}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B},$$

because $\mathbf{C}_{\mathcal{Y}} = \mathbf{C}_{\mathcal{W}}$ in (4.16). Thus, (4.21) amounts to proving

$$\mathbf{C}_{\mathcal{W}}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} = \left((\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo}, \infty}(s) \right)^-.$$

Because \mathbf{D} is nonsingular, the inverse system corresponding to $(\mathbf{W}(-s)^\top)^{-1}$ is well defined by Proposition 2.25. Using the given state-space realizations of \mathcal{G}_{lo} in (2.25) and \mathcal{W} in (4.16), we calculate a realization of the cascaded system having the transfer function $(\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo}, \infty}(s)$ using Proposition 2.23:

$$\left(\begin{array}{cc|c} -\mathbf{A}^\top + \mathbf{C}_{\mathcal{W}}^\top \mathbf{D}^{-1} \mathbf{B}_{\mathcal{W}}^\top & \mathbf{C}_{\mathcal{W}}^\top \mathbf{D}^{-1} \mathbf{C} & \mathbf{0}_{n \times m} \\ \mathbf{0}_{n \times n} & \mathbf{A} & \mathbf{B} \\ \hline \mathbf{D}^{-1} \mathbf{B}_{\mathcal{W}}^\top & \mathbf{D}^{-1} \mathbf{C} & \mathbf{0}_{m \times m} \end{array} \right). \quad (4.22)$$

The ARE in (4.17) can be rearranged into the form

$$(-\mathbf{A}^\top + \mathbf{C}_\mathcal{W}^\top \mathbf{D}^{-1} \mathbf{B}_\mathcal{W}^\top) (-\mathbf{Q}_\mathcal{W}^-) + \mathbf{A} \mathbf{Q}_\mathcal{W}^- + \mathbf{C}_\mathcal{W}^\top \mathbf{D}^{-1} \mathbf{C} = \mathbf{0}_{n \times n}.$$

Using this reformulation, it becomes clear that the state-space transformation

$$\mathbf{T} = \begin{bmatrix} \mathbf{I}_n & -\mathbf{Q}_\mathcal{W}^- \\ \mathbf{0}_{n \times n} & \mathbf{I}_n \end{bmatrix} \in \mathbb{R}^{2n \times 2n}$$

decouples the cascaded system realization (4.22), i.e., the transformed realization satisfies

$$\left(\begin{array}{cc|c} -\mathbf{A}^\top + \mathbf{C}_\mathcal{W}^\top \mathbf{D}^{-1} \mathbf{B}_\mathcal{W}^\top & \mathbf{0}_{n \times n} & \mathbf{Q}_\mathcal{W}^- \mathbf{B} \\ \mathbf{0}_{n \times n} & \mathbf{A} & \mathbf{B} \\ \hline \mathbf{D}^{-1} \mathbf{B}_\mathcal{W}^\top & \mathbf{C}_\mathcal{W} & \mathbf{0}_{m \times m} \end{array} \right). \quad (4.23)$$

Note that $-\mathbf{A}^\top + \mathbf{C}_\mathcal{W}^\top \mathbf{D}^{-1} \mathbf{B}_\mathcal{W}^\top$ is purely anti-stable while \mathbf{A} is purely stable. Because $(\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo},\infty}(s)$ is the transfer function of (4.23), it can be decomposed into its stable and anti-stable parts

$$\begin{aligned} (\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo},\infty}(s) &= \left((\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo},\infty}(s) \right)^+ + \left((\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo},\infty}(s) \right)^- \\ &= \left((\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo},\infty}(s) \right)^+ + \underbrace{\mathbf{C}_\mathcal{W} (s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}}_{= \mathbf{G}_{\sigma, \mathbf{A}}(s) = \mathbf{G}_\mathbf{B}(s)}. \end{aligned}$$

Thus, both $\mathbf{G}_{\sigma, \mathbf{A}}(s)$ and $\mathbf{G}_\mathbf{B}(s)$ are equivalent to $\left((\mathbf{W}(-s)^\top)^{-1} \mathbf{G}_{\text{lo},\infty}(s) \right)^-$, as claimed in (4.21). \square

In aggregate, Theorems 4.1 and 4.5 provide the theoretical foundation for a data-driven formulation of BST. We refer to this as *quadrature-based BST* (QuadBST); Algorithm 4.2.2 yields QuadBST when the transfer functions $\mathbf{G}_\mathbf{C}$ and $\mathbf{G}_{\sigma, \mathbf{A}}$, $\mathbf{G}_\mathbf{B}$ to be sampled are taken to be (4.20) and (4.21) in Theorem 4.5. We emphasize that these results form the theoretical formulation for QuadBST; it is not clear how to measure, e.g., the spectral factor \mathbf{W} , from actual samples of \mathbf{G}_{lo} . One could accomplish this by using a realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ of a system (2.25) to compute $\mathbf{Q}_\mathcal{W}^-$ from (4.17), form \mathcal{W} in (4.16), and then directly evaluate $\mathbf{W}(s)$ at points $s \in \mathbb{C}$ using the computed realization. This, however, is an intrusive process as it requires a realization of the linear system being approximated. This will be the case for the data-driven formulations of PRBT and BRBT that we present next, as well.

4.3.2 Data-driven positive-real balanced truncation

Intrusive positive-real balanced truncation.

A closely-related technique to BST is that of *positive-real balanced truncation* (PRBT) or *passivity-preserving* balanced truncation, also introduced in [61] and studied further in [170].

As its name suggests, PRBT is used for the approximation of positive-real or passive systems. We say a system is *passive* if it has the property that the energy produced by the system can never exceed the energy supplied to it. Mathematically, this property is expressed via the dissipation inequality

$$\int_0^t \mathbf{y}_{\text{lo}}(\tau) \mathbf{u}(\tau) d\tau \geq 0 \text{ for all } t > 0 \text{ and } \mathbf{u} \in \mathcal{L}_2^m(\mathbb{R}_{\geq 0}).$$

Passive systems are ubiquitous in applications of physics and engineering; electrical circuits are one such example. Passive systems can always be expressed as *port-Hamiltonian* systems [141, 216], and vice versa. A system being passive is equivalent to its transfer function being *positive real*; see, e.g. [4, Theorem 5.30].

Definition 4.6 (Positive-real systems [4, Ch. 5]). We say the asymptotically stable linear system \mathcal{G}_{lo} in (2.25) is *positive real* if it is square ($m = p$) and its transfer function \mathbf{G}_{lo} in (2.30) satisfies

$$\Phi(s) \stackrel{\text{def}}{=} \mathbf{G}_{\text{lo}}(s) + \mathbf{G}_{\text{lo}}(-s)^{\text{T}} \succeq 0, \text{ for all } s \in i\mathbb{R}. \quad (4.24)$$

We say \mathcal{G}_{lo} is *strictly positive real* if the inequality in (4.24) is strict. \diamond

The function $\Phi: \mathbb{C} \rightarrow \mathbb{C}^{m \times m}$ is called the *Popov function*. Because asymptotic stability is a prerequisite for positive realness in Definition 4.6, we henceforth assume that any positive real system is also asymptotically stable. Although in general, we mention that the notion of positive-realness can be defined for systems with poles on the imaginary axis, as well. Henceforth, we take the term positive real to mean *strictly positive real* in the sense of Definition 4.6.

Suppose that $\Phi(0) = \mathbf{D} + \mathbf{D}^{\text{T}} \succ 0$. By [245, Corollary 13.27], an asymptotically stable, minimal, and square system (2.25) is (strictly) positive real if and only if there exists a stabilizing solution $\mathbf{Q}_{\mathcal{M}}^- \in \mathbb{R}^{n \times n}$ to the *positive-real algebraic Riccati equation* (PR-ARE):

$$\mathbf{A}^{\text{T}} \mathbf{Q}_{\mathcal{M}} + \mathbf{Q}_{\mathcal{M}} \mathbf{A} + (\mathbf{C} - \mathbf{B}^{\text{T}} \mathbf{Q}_{\mathcal{M}})^{\text{T}} (\mathbf{D} + \mathbf{D}^{\text{T}})^{-1} (\mathbf{C} - \mathbf{B}^{\text{T}} \mathbf{Q}_{\mathcal{M}}) = \mathbf{0}_{n \times n}. \quad (4.25)$$

It follows directly from (4.24) that if a system is positive real, then so too is its dual (2.36). The dual statement of [245, Corollary 13.27] says that the system (2.25) is positive real if and only if there exists a stabilizing solution $\mathbf{P}_{\mathcal{N}}^- \in \mathbb{R}^{n \times n}$ to the dual PR-ARE

$$\mathbf{A} \mathbf{P}_{\mathcal{N}} + \mathbf{P}_{\mathcal{N}} \mathbf{A}^{\text{T}} + (\mathbf{B} - \mathbf{P}_{\mathcal{N}} \mathbf{C}^{\text{T}}) (\mathbf{D} + \mathbf{D}^{\text{T}})^{-1} (\mathbf{B} - \mathbf{P}_{\mathcal{N}} \mathbf{C}^{\text{T}})^{\text{T}} = \mathbf{0}_{n \times n}. \quad (4.26)$$

By [245, Corollary 13.27], both $\mathbf{P}_{\mathcal{N}}^-$ and $\mathbf{Q}_{\mathcal{M}}^-$ are unique and SPD. In PRBT, $\mathbf{P}_{\mathcal{N}}^-$ and $\mathbf{Q}_{\mathcal{M}}^-$ are balanced and take the place of \mathbf{P} and \mathbf{Q}_{lo} , respectively.

Definition 4.7 (Positive-real balanced realization). We say that a state-space realization of a minimal, positive-real system \mathcal{G}_{lo} in (2.25) is a *positive-real balanced realization* if

$$\mathbf{P}_{\mathcal{N}}^- = \mathbf{Q}_{\mathcal{M}}^- = \Sigma^{\text{prbt}} = \text{diag}(\pi_1 \mathbf{I}_{m_1}, \pi_2 \mathbf{I}_{m_2}, \dots, \pi_q \mathbf{I}_{m_q}), \quad (4.27)$$

where $1 \geq \pi_1 > \pi_2 > \dots > \pi_q > 0$ and their multiplicities satisfy $m_1 + \dots + m_q = n$. The values π_i are called the *positive-real singular values* of \mathcal{G}_{lo} . \diamond

Theorem 4.8 (Positive-real balanced truncation [61], [96, Theorem 5]). Consider an asymptotically stable, minimal, positive-real linear system \mathcal{G}_{lo} in (2.25) having the positive-real balanced realization

$$\mathbf{A}_{\text{prbt}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B}_{\text{prbt}} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \text{and} \quad \mathbf{C}_{\text{prbt}} = [\mathbf{C}_1 \quad \mathbf{C}_2]$$

according to Definition 4.7. The matrices are partitioned with respect to $\mathbf{P}_{\mathcal{N}}^- = \mathbf{Q}_{\mathcal{M}}^- = \boldsymbol{\Sigma}^{\text{prbt}} = \text{diag}(\boldsymbol{\Sigma}_1^{\text{prbt}}, \boldsymbol{\Sigma}_2^{\text{prbt}})$, where

$$\boldsymbol{\Sigma}_1^{\text{prbt}} = \text{diag}(\pi_1 \mathbf{I}_{m_1}, \dots, \pi_k \mathbf{I}_{m_k}) \quad \text{and} \quad \boldsymbol{\Sigma}_2^{\text{prbt}} = \text{diag}(\pi_{k+1} \mathbf{I}_{m_{k+1}}, \dots, \pi_q \mathbf{I}_{m_q})$$

for $r = m_1 + \dots + m_k$ and $1 \leq k < q$. Then, the order- r reduced model $\tilde{\mathcal{G}}_{\text{lo,prbt}} = (\mathbf{A}_{11}, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D})$ obtained via positive-real balanced truncation and having the transfer function $\tilde{\mathbf{G}}_{\text{lo,prbt}}$ is balanced in the sense of (4.27), asymptotically stable, minimal, and positive real. Moreover, $\tilde{\mathcal{G}}_{\text{lo,prbt}}$ satisfies the multiplicative-type error bound

$$\|(\mathbf{D}^\top + \mathbf{G}_{\text{lo}})^{-1} - (\mathbf{D}^\top + \tilde{\mathbf{G}}_{\text{lo,prbt}})^{-1}\|_{\mathcal{H}m \times m_\infty} \leq 2\|\mathbf{D} + \mathbf{D}^\top\|_2^2 \sum_{i=k+1}^q \pi_i. \quad (4.28)$$

◇

To summarize, PRBT preserves positive-realness *in addition to* asymptotic stability [61, 105]. The multiplicative-type error bound was not derived until much later in [96].

Quadrature-based positive-real balanced truncation.

The PR-AREs in (4.25) and (4.26) are closely related to a spectral factorization of the Popov function appearing in (4.24). Consider an asymptotically stable, minimal system \mathcal{G}_{lo} in (2.25) with $\mathbf{D} + \mathbf{D}^\top \succ 0$. By [245, Corollary 13.27], \mathcal{G}_{lo} is strictly positive real if and only if there exists a minimum-phase *right* spectral factor $\mathbf{M}: \mathbb{C} \rightarrow \mathbb{C}^{m \times m}$ of the Popov function, i.e.,

$$\Phi(s) = \mathbf{G}_{\text{lo}}(s) + \mathbf{G}_{\text{lo}}(-s)^\top = \mathbf{M}(-s)^\top \mathbf{M}(s), \quad s \in \mathbb{C}.$$

The spectral factor \mathbf{M} is the transfer function of a minimal and asymptotically stable linear system \mathcal{M} , defined by the realization

$$\mathcal{M} = (\mathbf{A}, \mathbf{B}, \mathbf{C}_{\mathcal{M}}, (\mathbf{D} + \mathbf{D}^\top)^{1/2}) \quad \text{for} \quad \mathbf{C}_{\mathcal{M}} \stackrel{\text{def}}{=} (\mathbf{D} + \mathbf{D}^\top)^{-1/2} (\mathbf{C} - \mathbf{B}^\top \mathbf{Q}_{\mathcal{M}}^-), \quad (4.29)$$

where $\mathbf{Q}_{\mathcal{M}}^- \in \mathbb{R}^{n \times n}$ is the unique minimal solution to (4.25). Explicitly, \mathbf{M} is written as

$$\mathbf{M}(s) = (\mathbf{D} + \mathbf{D}^\top)^{-1/2} (\mathbf{C} - \mathbf{B}^\top \mathbf{Q}_{\mathcal{M}}^-) (s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} + (\mathbf{D} + \mathbf{D}^\top)^{1/2}.$$

Applying [245, Corollary 13.27] to the dual of \mathcal{G}_{lo} provides the existence of a minimum-phase *left* spectral factor $\mathbf{N}: \mathbb{C} \rightarrow \mathbb{C}^{m \times m}$ of the Popov function, i.e.,

$$\Phi(s) = \mathbf{G}_{\text{lo}}(s) + \mathbf{G}_{\text{lo}}(-s)^\top = \mathbf{N}(s)\mathbf{N}(-s)^\top, \quad s \in \mathbb{C}.$$

Moreover, \mathbf{N} is the transfer function of a minimal and asymptotically stable system

$$\mathcal{N} = (\mathbf{A}, \mathbf{B}_{\mathcal{N}}, \mathbf{C}, (\mathbf{D} + \mathbf{D}^\top)^{1/2}) \quad \text{for} \quad \mathbf{B}_{\mathcal{N}} \stackrel{\text{def}}{=} (\mathbf{B} - \mathbf{P}_{\mathcal{N}}^- \mathbf{C}^\top) (\mathbf{D} + \mathbf{D}^\top)^{-1/2}, \quad (4.30)$$

where $\mathbf{P}_{\mathcal{N}}^- \in \mathbb{R}^{n \times n}$ is the unique minimal solution to (4.26). Explicitly, \mathbf{N} is given by

$$\mathbf{N}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1} (\mathbf{B} - \mathbf{P}_{\mathcal{N}}^- \mathbf{C}^\top) (\mathbf{D} + \mathbf{D}^\top)^{-1/2} + (\mathbf{D} + \mathbf{D}^\top)^{1/2}.$$

At this point, we are prepared to show how PRBT fits under the GenQuadBT framework of Section 4.2. Fixing $\mathbf{Q}_{\mathcal{M}}^-$ and $\mathbf{P}_{\mathcal{N}}^-$ in the right-hand sides of (4.25) and (4.26) yields a pair of ALEs:

$$\begin{aligned} \mathbf{A}^\top \mathbf{Q}_{\mathcal{M}} + \mathbf{Q}_{\mathcal{M}} \mathbf{A} + \underbrace{(\mathbf{C} - \mathbf{B}^\top \mathbf{Q}_{\mathcal{M}}^-)^\top (\mathbf{D} + \mathbf{D}^\top)^{-1} (\mathbf{C} - \mathbf{B}^\top \mathbf{Q}_{\mathcal{M}}^-)}_{= \mathbf{C}_{\mathcal{M}}^\top \mathbf{C}_{\mathcal{M}}} &= \mathbf{0}_{n \times n}, \\ \implies \mathbf{A}^\top \mathbf{Q}_{\mathcal{M}} + \mathbf{Q}_{\mathcal{M}} \mathbf{A} + \mathbf{C}_{\mathcal{M}}^\top \mathbf{C}_{\mathcal{M}} &= \mathbf{0}_{n \times n}, \end{aligned} \quad (4.31)$$

$$\begin{aligned} \mathbf{A} \mathbf{P}_{\mathcal{N}} + \mathbf{P}_{\mathcal{N}} \mathbf{A}^\top + \underbrace{(\mathbf{B} - \mathbf{P}_{\mathcal{N}}^- \mathbf{C}^\top) (\mathbf{D} + \mathbf{D}^\top)^{-1} (\mathbf{B} - \mathbf{P}_{\mathcal{N}}^- \mathbf{C}^\top)^\top}_{= \mathbf{B}_{\mathcal{N}} \mathbf{B}_{\mathcal{N}}^\top} &= \mathbf{0}_{n \times n}, \\ \implies \mathbf{A} \mathbf{P}_{\mathcal{N}} + \mathbf{P}_{\mathcal{N}} \mathbf{A}^\top + \mathbf{B}_{\mathcal{N}} \mathbf{B}_{\mathcal{N}}^\top &= \mathbf{0}_{n \times n}. \end{aligned} \quad (4.32)$$

Because the stabilizing solution to (4.25) is unique, (4.31) is uniquely solved by $\mathbf{Q}_{\mathcal{M}} = \mathbf{Q}_{\mathcal{M}}^-$. Moreover, it is clear from the definition of $\mathbf{C}_{\mathcal{M}}$ in (4.29) that (4.31) is the *observability Lyapunov equation* of $\mathcal{Y} = \mathcal{M}$ in (4.29), and so $\mathbf{Q}_{\mathcal{Y}} = \mathbf{Q}_{\mathcal{M}}^-$ is the *observability Gramian* of $\mathcal{Y} = \mathcal{M}$. In other words, $\mathcal{Y} = \mathcal{M}$ in this instance with

$$\mathcal{Y} = (\mathbf{A}, \mathbf{B}_{\mathcal{Y}}, \mathbf{C}_{\mathcal{Y}}, \mathbf{D}_{\mathcal{Y}}) = (\mathbf{A}, \mathbf{B}, \mathbf{C}_{\mathcal{M}}, (\mathbf{D} + \mathbf{D}^\top)^{1/2}) = \mathcal{M}.$$

Likewise, (4.32) is uniquely solved by $\mathbf{P}_{\mathcal{N}} = \mathbf{P}_{\mathcal{N}}^-$. It follows from the definition of $\mathbf{B}_{\mathcal{N}}$ in (4.30) that (4.32) is the *reachability Lyapunov equation* of $\mathcal{X} = \mathcal{N}$ in (4.30), and so $\mathbf{P}_{\mathcal{X}} = \mathbf{P}_{\mathcal{N}}^-$ is the *reachability Gramian* of $\mathcal{X} = \mathcal{N}$, i.e.,

$$\mathcal{X} = (\mathbf{A}, \mathbf{B}_{\mathcal{X}}, \mathbf{C}_{\mathcal{X}}, \mathbf{D}_{\mathcal{X}}) = (\mathbf{A}, \mathbf{B}_{\mathcal{N}}, \mathbf{C}, (\mathbf{D} + \mathbf{D}^\top)^{1/2}) = \mathcal{N}.$$

The relevant Gramians $\mathbf{P}_{\mathcal{N}}^-$ and $\mathbf{Q}_{\mathcal{M}}^-$ can thereby be decomposed into the quadrature-based factors (4.7) and (4.8) with $\mathbf{B}_{\mathcal{X}} = \mathbf{B}_{\mathcal{N}}$ in (4.30) and $\mathbf{C}_{\mathcal{Y}} = \mathbf{C}_{\mathcal{M}}$ in (4.29). Applying Theorem 4.1 in this setting allows to interpret $\mathbf{G}_{\sigma, \mathbf{A}}$, $\mathbf{G}_{\mathbf{B}}$ and $\mathbf{G}_{\mathbf{C}}$ in (4.10a)–(4.10c) in terms of the spectral factors \mathbf{M} and \mathbf{N} .

Theorem 4.9 (Positive-real balanced truncation from data). Let $\mathbf{Q}_{\mathcal{M}}^- \in \mathbb{R}^{n \times n}$ and $\mathbf{P}_{\mathcal{N}}^- \in \mathbb{R}^{n \times n}$ be the stabilizing solutions to (4.25) and (4.26). Then, for PRBT the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}(s)$, $\mathbf{G}_{\mathbf{B}}(s)$, and $\mathbf{G}_{\mathbf{C}}(s)$ defined as in (4.10a)–(4.10c) of Theorem 4.1 are given by

$$\mathbf{G}_{\sigma, \mathbf{A}}(s) = \left((\mathbf{M}(-s)^\top)^{-1} \mathbf{N}_\infty(s) \right)^-, \quad (4.33)$$

$$\mathbf{G}_{\mathbf{B}}(s) = \mathbf{M}_\infty(s), \quad \text{and} \quad \mathbf{G}_{\mathbf{C}}(s) = \mathbf{N}_\infty(s), \quad (4.34)$$

where \mathbf{M} and \mathbf{N} are the spectral factors associated with the linear systems \mathcal{M} and \mathcal{N} defined in (4.29) and (4.30). \diamond

Proof of Theorem 4.9. In this case, $\mathbf{B}_{\mathcal{X}} = \mathbf{B}_{\mathcal{N}}$, and $\mathbf{C}_{\mathcal{Y}} = \mathbf{C}_{\mathcal{M}}$ as in (4.30) and (4.29). So, (4.10b) and (4.10c) are given by

$$\begin{aligned} \mathbf{G}_{\mathbf{B}}(s) &= \mathbf{C}_{\mathcal{M}}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} = \mathbf{M}_\infty(s), \\ \mathbf{G}_{\mathbf{C}}(s) &= \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}_{\mathcal{N}} = \mathbf{N}_\infty(s), \end{aligned}$$

thus proving (4.34). The claim in (4.33) follows identically from the argument of Theorem 4.5 by replacing $\mathbf{W}(s)$ with $\mathbf{M}(s)$ and $\mathbf{G}_{\text{lo}, \infty}(s)$ with $\mathbf{N}_\infty(s)$. \square

As was the case with BST, Theorems 4.1 and 4.9 provide the ingredients for a data-driven implementation of PRBT, that we call *quadrature-based PRBT* (QuadPRBT). In this case, the requisite data are given by the *spectral factors \mathbf{M} and \mathbf{N} of the Popov function*. Algorithm 4.2.2 yields QuadPRBT when the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}$ and $\mathbf{G}_{\mathbf{B}}$, $\mathbf{G}_{\mathbf{C}}$ to be sampled are replaced with those in (4.33) and (4.34).

4.3.3 Data-driven bounded-real balanced truncation

Intrusive bounded-real balanced truncation.

An important class of systems is those having transfer functions that are bounded along the imaginary axis. We call these systems bounded real; they are used in parameterizing all stabilizing controllers of a system such that the closed-loop system satisfies a particular \mathcal{H}_∞ constraint [85].

Definition 4.10 (Bounded-real systems [4, Ch. 5]). We say the asymptotically stable linear system \mathcal{G}_{lo} in (2.25) is *bounded real* if its transfer function \mathbf{G}_{lo} in (2.30) satisfies

$$\gamma^2 \mathbf{I}_m + \mathbf{G}_{\text{lo}}(-s)^\top \mathbf{G}_{\text{lo}}(s) \succeq 0, \quad s \in i\mathbb{R}, \quad (4.35)$$

where $\gamma \stackrel{\text{def}}{=} \|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_\infty}$. We say \mathcal{G}_{lo} is *strictly bounded real* if the inequality in (4.35) is strict. \diamond

Henceforth, since we only deal with *strictly* bounded-real systems, we take systems to be bounded-real in the strict sense. Since it is always possible to normalize \mathcal{G}_{lo} such that $\|\mathcal{G}_{\text{lo}}\|_{\mathcal{H}_\infty} \leq 1$, we take $\gamma = 1$ without loss of generality.

Define $\mathbf{R}_{\mathcal{J}} \stackrel{\text{def}}{=} \mathbf{I}_m - \mathbf{D}^\top \mathbf{D} \in \mathbb{R}^{m \times m}$, and suppose $\mathbf{R}_{\mathcal{J}} \succ 0$. By the Bounded Real Lemma [245, Corollary 13.24], an asymptotically stable and minimal system (2.25) is (strictly) bounded real if and only if there exists a stabilizing solution $\mathbf{Q}_{\mathcal{J}}^- \in \mathbb{R}^{n \times n}$ to the *bounded-real algebraic Riccati equation* (BR-ARE)

$$\mathbf{A}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{Q}_{\mathcal{J}} \mathbf{A} + \mathbf{C}^\top \mathbf{C} + (\mathbf{B}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{D}^\top \mathbf{C})^\top \mathbf{R}_{\mathcal{J}}^{-1} (\mathbf{B}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{D}^\top \mathbf{C}) = \mathbf{0}_{n \times n}. \quad (4.36)$$

Define $\mathbf{R}_{\mathcal{K}} \stackrel{\text{def}}{=} \mathbf{I}_p - \mathbf{D} \mathbf{D}^\top \in \mathbb{R}^{p \times p}$, and suppose $\mathbf{R}_{\mathcal{K}} \succ 0$. Applying the dual result to [245, Corollary 13.24] implies that a system is bounded-real if and only if there exists a stabilizing solution $\mathbf{P}_{\mathcal{K}}^- \in \mathbb{R}^{n \times n}$ to the dual BR-ARE, i.e.

$$\mathbf{A} \mathbf{P}_{\mathcal{K}} + \mathbf{P}_{\mathcal{K}} \mathbf{A}^\top + \mathbf{B} \mathbf{B}^\top + (\mathbf{P}_{\mathcal{K}} \mathbf{C}^\top + \mathbf{B} \mathbf{D}^\top) \mathbf{R}_{\mathcal{K}}^{-1} (\mathbf{P}_{\mathcal{K}} \mathbf{C}^\top + \mathbf{B} \mathbf{D}^\top)^\top = \mathbf{0}_{n \times n}. \quad (4.37)$$

By [245, Corollary 13.24], both $\mathbf{P}_{\mathcal{K}}^-$ and $\mathbf{Q}_{\mathcal{J}}^-$ are unique and SPD. In BRBT, the stabilizing solutions $\mathbf{P}_{\mathcal{K}}^-$ and $\mathbf{Q}_{\mathcal{J}}^-$ to the BR-AREs take the place of the usual reachability and observability Gramians.

Definition 4.11 (Bounded-real balanced realization). We say that a state-space realization of a minimal, bounded-real system \mathcal{G}_{lo} in (2.25) is a *bounded-real balanced realization* if

$$\mathbf{P}_{\mathcal{K}}^- = \mathbf{Q}_{\mathcal{J}}^- = \Sigma^{\text{brbt}} = \text{diag}(\xi_1 \mathbf{I}_{m_1}, \xi_2 \mathbf{I}_{m_2}, \dots, \xi_q \mathbf{I}_{m_q}), \quad (4.38)$$

where $1 \geq \xi_1 > \xi_2 > \dots > \xi_q > 0$ and their multiplicities satisfy $m_1 + \dots + m_q = n$. The values ξ_i are called the *bounded-real singular values* of \mathcal{G}_{lo} . \diamond

Theorem 4.12 (Bounded-real balanced truncation [159]). Consider an asymptotically stable, minimal, bounded-real linear system \mathcal{G}_{lo} in (2.25) having the bounded-real balanced realization

$$\mathbf{A}_{\text{brbt}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B}_{\text{brbt}} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \text{and} \quad \mathbf{C}_{\text{brbt}} = [\mathbf{C}_1 \quad \mathbf{C}_2]$$

according to Definition 4.11. The matrices are partitioned with respect to $\mathbf{P}_{\mathcal{K}}^- = \mathbf{Q}_{\mathcal{J}}^- = \Sigma^{\text{brbt}} = \text{diag}(\Sigma_1^{\text{brbt}}, \Sigma_2^{\text{brbt}})$, where

$$\Sigma_1^{\text{brbt}} = \text{diag}(\xi_1 \mathbf{I}_{m_1}, \dots, \xi_k \mathbf{I}_{m_k}) \quad \text{and} \quad \Sigma_2^{\text{brbt}} = \text{diag}(\xi_{k+1} \mathbf{I}_{m_{k+1}}, \dots, \xi_q \mathbf{I}_{m_q})$$

for $r = m_1 + \dots + m_k$ and $1 \leq k < q$. Then, the order- r reduced model $\tilde{\mathcal{G}}_{\text{lo,brbt}} = (\mathbf{A}_{11}, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D})$ obtained via bounded-real balanced truncation and having the transfer function $\tilde{\mathbf{G}}_{\text{lo,brbt}}$ is balanced in the sense of (4.38), asymptotically stable, minimal, and bounded real. Moreover, suppose that $\mathbf{K}(s)$ and $\mathbf{J}(s)$ are left and right spectral factors

of $\mathbf{I}_p - \mathbf{G}_{\text{lo}}(s)\mathbf{G}_{\text{lo}}(-s)^\top$ and $\mathbf{I}_m - \mathbf{G}_{\text{lo}}(-s)^\top\mathbf{G}_{\text{lo}}(s)$, and let $\widetilde{\mathbf{K}}(s)$ and $\widetilde{\mathbf{J}}(s)$ be the analogous spectral factors for the reduced $\widetilde{\mathcal{G}}_{\text{lo,brbt}}$. Then, $\widetilde{\mathcal{G}}_{\text{lo,brbt}}$ satisfies the error bound

$$\max \left\{ \left\| \begin{bmatrix} \mathbf{G}_{\text{lo}} - \widetilde{\mathbf{G}}_{\text{lo,brbt}} \\ \mathbf{K} - \widetilde{\mathbf{K}} \end{bmatrix} \right\|_{\mathcal{H}_\infty}, \left\| \begin{bmatrix} \mathbf{G}_{\text{lo}} - \widetilde{\mathbf{G}}_{\text{lo,brbt}} \\ \mathbf{J} - \widetilde{\mathbf{J}} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \right\} \leq 2 \sum_{i=k+1}^q \xi_i. \quad (4.39)$$

◇

Thus, BRBT preserves asymptotic stability and the bounded-real property. Moreover, if the singular values ξ_i are small, then the reduced-order transfer function *and* spectral factors are guaranteed to be close to their full-order counterparts in the \mathcal{H}_∞ sense.

Quadrature-based bounded-real balanced truncation.

We have already mentioned the relevant spectral factorizations in Theorem 4.12, but we introduce them formally here. Consider an asymptotically stable, minimal system \mathcal{G}_{lo} in (2.25) and assume that $\mathbf{R}_{\mathcal{J}} = \mathbf{I}_m - \mathbf{D}^\top\mathbf{D} \succ 0$. By [245, Corollary 13.21] there exists a minimum-phase *right* spectral factor $\mathbf{J}: \mathbb{C} \rightarrow \mathbb{C}^{m \times m}$ of $\mathbf{I}_m - \mathbf{G}_{\text{lo}}(-s)^\top\mathbf{G}_{\text{lo}}(s)$, i.e.,

$$\mathbf{J}(-s)^\top\mathbf{J}(s) = \mathbf{I}_m - \mathbf{G}_{\text{lo}}(-s)^\top\mathbf{G}_{\text{lo}}(s), \quad s \in \mathbb{C}.$$

The spectral factor \mathbf{J} is the transfer function for a minimal and asymptotically stable linear system \mathcal{J} , defined by the realization

$$\mathcal{J} = (\mathbf{A}, \mathbf{B}, \mathbf{C}_{\mathcal{J}}, \mathbf{R}_{\mathcal{J}}^{1/2}) \quad \text{for} \quad \mathbf{C}_{\mathcal{J}} \stackrel{\text{def}}{=} \mathbf{R}_{\mathcal{J}}^{-1/2} (\mathbf{B}^\top\mathbf{Q}_{\mathcal{J}}^- + \mathbf{D}^\top\mathbf{C}), \quad (4.40)$$

where $\mathbf{Q}_{\mathcal{J}}^- \in \mathbb{R}^{n \times n}$ is the unique minimal solution to (4.36). Likewise, if $\mathbf{R}_{\mathcal{K}} = \mathbf{I}_p - \mathbf{D}\mathbf{D}^\top \succ 0$, by the dual result [245, Corollary 13.27] there exists a minimum-phase *left* spectral factor $\mathbf{K}: \mathbb{C} \rightarrow \mathbb{C}^{m \times m}$ of $\mathbf{I}_p - \mathbf{G}_{\text{lo}}(s)\mathbf{G}_{\text{lo}}(-s)^\top$, i.e.,

$$\mathbf{I}_p - \mathbf{G}_{\text{lo}}(s)\mathbf{G}_{\text{lo}}(-s)^\top = \mathbf{K}(s)\mathbf{K}(-s)^\top, \quad s \in \mathbb{C},$$

where \mathbf{K} is the transfer function for a minimal and asymptotically stable system \mathcal{K} :

$$\mathcal{K} = (\mathbf{A}, \mathbf{B}_{\mathcal{K}}, \mathbf{C}, \mathbf{R}_{\mathcal{K}}^{1/2}) \quad \text{for} \quad \mathbf{B}_{\mathcal{K}} \stackrel{\text{def}}{=} (\mathbf{P}_{\mathcal{K}}^- \mathbf{C}^\top + \mathbf{B}\mathbf{D}^\top) \mathbf{R}_{\mathcal{K}}^{-1/2}, \quad (4.41)$$

and $\mathbf{P}_{\mathcal{K}}^- \in \mathbb{R}^{n \times n}$ is the unique minimal solution to (4.37). Using these factors, we interpret the dual BR-AREs as the reachability and observability Lyapunov equations of some linear systems. To this end, we introduce the notation

$$\begin{aligned} \widehat{\mathbf{B}}_{\mathcal{K}} &\stackrel{\text{def}}{=} [\mathbf{B} \quad \mathbf{B}_{\mathcal{K}}] \in \mathbb{R}^{n \times 2m}, & \widehat{\mathbf{R}}_{\mathcal{K}} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{D} & \mathbf{R}_{\mathcal{K}}^{1/2} \end{bmatrix} \in \mathbb{R}^{p \times 2m}, \\ \widehat{\mathbf{C}}_{\mathcal{J}} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{C} \\ \mathbf{C}_{\mathcal{J}} \end{bmatrix} \in \mathbb{R}^{2p \times n}, & \widehat{\mathbf{R}}_{\mathcal{J}} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{D} \\ \mathbf{R}_{\mathcal{J}}^{1/2} \end{bmatrix} \in \mathbb{R}^{2p \times m}. \end{aligned} \quad (4.42)$$

Then, fixing $\mathbf{Q}_{\mathcal{J}}^-$ and $\mathbf{P}_{\mathcal{K}}^-$ in the right-hand sides of (4.36) and (4.37), we see that

$$\begin{aligned} \mathbf{A}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{Q}_{\mathcal{J}} \mathbf{A} + \underbrace{\mathbf{C}^\top \mathbf{C} + (\mathbf{B}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{D}^\top \mathbf{C})^\top \mathbf{R}_{\mathcal{J}}^{-1} (\mathbf{B}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{D}^\top \mathbf{C})}_{= \widehat{\mathbf{C}}_{\mathcal{J}}^\top \widehat{\mathbf{C}}_{\mathcal{J}}} &= \mathbf{0}_{n \times n} \\ &\implies \mathbf{A}^\top \mathbf{Q}_{\mathcal{J}} + \mathbf{Q}_{\mathcal{J}} \mathbf{A} + \widehat{\mathbf{C}}_{\mathcal{J}}^\top \widehat{\mathbf{C}}_{\mathcal{J}} = \mathbf{0}_{n \times n}, \end{aligned} \quad (4.43)$$

$$\begin{aligned} \mathbf{A} \mathbf{P}_{\mathcal{K}} + \mathbf{P}_{\mathcal{K}} \mathbf{A}^\top + \underbrace{\mathbf{B} \mathbf{B}^\top + (\mathbf{P}_{\mathcal{K}} \mathbf{C}^\top + \mathbf{B} \mathbf{D}^\top) \mathbf{R}_{\mathcal{K}}^{-1} (\mathbf{P}_{\mathcal{K}} \mathbf{C}^\top + \mathbf{B} \mathbf{D}^\top)^\top}_{= \widehat{\mathbf{B}}_{\mathcal{K}} \widehat{\mathbf{B}}_{\mathcal{K}}^\top} &= \mathbf{0}_{n \times n} \\ &\implies \mathbf{A} \mathbf{P}_{\mathcal{K}} + \mathbf{P}_{\mathcal{K}} \mathbf{A}^\top + \widehat{\mathbf{B}}_{\mathcal{K}} \widehat{\mathbf{B}}_{\mathcal{K}}^\top = \mathbf{0}_{n \times n}. \end{aligned} \quad (4.44)$$

Because the stabilizing solution to (4.36) is unique, (4.43) is uniquely solved by $\mathbf{Q}_{\mathcal{J}} = \mathbf{Q}_{\mathcal{J}}^-$. Thus, (4.43) is the *observability Lyapunov equation* of the linear system $\mathcal{Y} = \widehat{\mathcal{J}}$ defined as

$$\mathcal{Y} = (\mathbf{A}, \mathbf{B}_{\mathcal{Y}}, \mathbf{C}_{\mathcal{Y}}, \mathbf{D}_{\mathcal{Y}}) = (\mathbf{A}, \mathbf{B}, \widehat{\mathbf{C}}_{\mathcal{J}}, \widehat{\mathbf{R}}_{\mathcal{J}}) = \widehat{\mathcal{J}}, \quad (4.45)$$

where $\mathbf{Q}_{\mathcal{J}}^-$ is the *observability Gramian* of $\mathcal{Y} = \widehat{\mathcal{J}}$. By the definition of the matrices in (4.42), the transfer function $\widehat{\mathcal{J}}: \mathbb{C} \rightarrow \mathbb{C}^{2p \times m}$ of the system $\widehat{\mathcal{J}}$ is given by

$$\widehat{\mathcal{J}}(s) = \widehat{\mathbf{C}}_{\mathcal{J}}(s\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} + \widehat{\mathbf{R}}_{\mathcal{J}} = \begin{bmatrix} \mathbf{G}_{\text{lo}}(s) \\ \mathbf{J}(s) \end{bmatrix}, \quad s \in \mathbb{C},$$

where $\mathbf{J}(s)$ is the right spectral factor of $\mathbf{I}_m - \mathbf{G}_{\text{lo}}(-s)^\top \mathbf{G}_{\text{lo}}(s)$ defined by (4.40). Likewise, the stabilizing solution to (4.37) is unique, and so (4.44) is uniquely solved by $\mathbf{P}_{\mathcal{K}} = \mathbf{P}_{\mathcal{K}}^-$. Thus, (4.44) is the *reachability Lyapunov equation* of the linear system $\mathcal{X} = \widehat{\mathcal{K}}$ defined as

$$\mathcal{X} = (\mathbf{A}, \mathbf{B}_{\mathcal{X}}, \mathbf{C}_{\mathcal{X}}, \mathbf{D}_{\mathcal{X}}) = (\mathbf{A}, \widehat{\mathbf{B}}_{\mathcal{K}}, \mathbf{C}, \widehat{\mathbf{R}}_{\mathcal{K}}) = \widehat{\mathcal{K}}, \quad (4.46)$$

where $\mathbf{P}_{\mathcal{K}}^-$ is the *reachability Gramian* of $\mathcal{X} = \widehat{\mathcal{K}}$. By (4.42), the transfer function $\widehat{\mathcal{K}}: \mathbb{C} \rightarrow \mathbb{C}^{p \times 2m}$ of the system $\widehat{\mathcal{K}}$ is given by

$$\widehat{\mathcal{K}}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1} \widehat{\mathbf{B}}_{\mathcal{K}} + \widehat{\mathbf{R}}_{\mathcal{K}} = [\mathbf{G}_{\text{lo}}(s) \quad \mathbf{K}(s)], \quad s \in \mathbb{C},$$

where $\mathbf{K}(s)$ is the left spectral factor of $\mathbf{I}_p - \mathbf{G}_{\text{lo}}(s) \mathbf{G}_{\text{lo}}(-s)^\top$ defined by (4.40). Thus, $\mathbf{Q}_{\mathcal{Y}} = \mathbf{Q}_{\mathcal{J}}^-$ and $\mathbf{P}_{\mathcal{X}} = \mathbf{P}_{\mathcal{K}}^-$ can be decomposed into quadrature-based factors (4.7) and (4.8) with $\mathbf{B}_{\mathcal{X}} = \widehat{\mathbf{B}}_{\mathcal{K}}$ and $\mathbf{C}_{\mathcal{Y}} = \widehat{\mathbf{C}}_{\mathcal{J}}$. Applying Theorem 4.1 in this setting allows us to interpret $\mathbf{G}_{\sigma, \mathbf{A}}$, $\mathbf{G}_{\mathbf{B}}$ and $\mathbf{G}_{\mathbf{C}}$ in (4.10a)–(4.10c) in terms of $\widehat{\mathcal{J}}$ and $\widehat{\mathcal{K}}$.

Theorem 4.13 (Bounded-real balanced truncation from data.). Let $\mathbf{Q}_{\mathcal{J}}^- \in \mathbb{R}^{n \times n}$ and $\mathbf{P}_{\mathcal{K}}^- \in \mathbb{R}^{n \times n}$ be the minimal solutions to (4.36) and (4.37). Then, for BRBT the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}(s)$, $\mathbf{G}_{\mathbf{B}}(s)$, and $\mathbf{G}_{\mathbf{C}}(s)$ defined in (4.10a)–(4.10c) of Theorem 4.1 are given by:

$$\mathbf{G}_{\sigma, \mathbf{A}}(s) = \left(\left[\begin{array}{cc} \mathbf{I}_p & \mathbf{0}_{p \times m} \\ \mathbf{G}_{\text{lo}, \infty}(-s)^\top & \mathbf{J}_{\infty}(-s)^\top \end{array} \right]^{-1} \left[\begin{array}{cc} \mathbf{G}_{\text{lo}, \infty}(s) & \mathbf{K}_{\infty}(s) \\ -\mathbf{D}^\top \mathbf{G}_{\text{lo}, \infty}(s) & -\mathbf{D}^\top \mathbf{K}_{\infty}(s) \end{array} \right] \right)^{-}, \quad (4.47)$$

$$\mathbf{G}_{\mathbf{B}}(s) = \widehat{\mathcal{J}}_{\infty}(s), \quad \text{and} \quad \mathbf{G}_{\mathbf{C}}(s) = \widehat{\mathcal{K}}_{\infty}(s), \quad (4.48)$$

where \mathbf{J} , $\widehat{\mathbf{J}}$ and \mathbf{K} , $\widehat{\mathbf{K}}$ are the transfer functions of the linear systems defined in (4.40), (4.45), (4.41), and (4.46). \diamond

Proof of Theorem 4.13. In this setting, we have $\mathbf{B}_\mathcal{X} = \widehat{\mathbf{B}}_\mathcal{K}$ and $\mathbf{C}_\mathcal{Y} = \widehat{\mathbf{C}}_\mathcal{J}$ defined in (4.42). So, from the definition of $\mathbf{G}_\mathbf{B}$ and $\mathbf{G}_\mathbf{C}$ in (4.10b) and (4.10c), it follows immediately that

$$\begin{aligned}\mathbf{G}_\mathbf{B}(s) &= \widehat{\mathbf{C}}_\mathcal{J}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B} = \widehat{\mathbf{J}}_\infty(s), \\ \mathbf{G}_\mathbf{C}(s) &= \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\widehat{\mathbf{B}}_\mathcal{K} = \widehat{\mathbf{K}}_\infty(s),\end{aligned}$$

proving (4.48). For the cascaded system in (4.47), we introduce the notation

$$\mathbf{Z}(s) = \begin{bmatrix} \mathbf{I}_p & \mathbf{0}_{p \times m} \\ \mathbf{G}_{\text{lo},\infty}(-s)^\top & \mathbf{J}_\infty(-s)^\top \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{G}_{\text{lo},\infty}(s) & \mathbf{K}_\infty(s) \\ -\mathbf{D}^\top \mathbf{G}_{\text{lo},\infty}(s) & -\mathbf{D}^\top \mathbf{K}_\infty(s) \end{bmatrix}.$$

Thus, $\mathbf{Z}: \mathbb{C} \rightarrow \mathbb{C}^{2p \times 2m}$ is the transfer function of some linear system \mathcal{Z} . Thus, to prove (4.47) it suffices to show that

$$\mathbf{Z}(s)^- = \widehat{\mathbf{C}}_\mathcal{J}(s\mathbf{I}_n - \mathbf{A})^{-1}\widehat{\mathbf{B}}_\mathcal{K}.$$

To this end, we introduce the matrices:

$$\begin{aligned}\widehat{\mathbf{B}} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{0}_{n \times p} & \mathbf{B} \end{bmatrix} \in \mathbb{R}^{n \times (p+m)}, & \widehat{\mathbf{C}} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{C} \\ -\mathbf{D}^\top \mathbf{C} \end{bmatrix} \in \mathbb{R}^{(p+m) \times n}, \\ \widehat{\mathbf{R}} &\stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{I}_p & \mathbf{0}_{p \times m} \\ \mathbf{0}_{m \times p} & \mathbf{R}_\mathcal{J}^{1/2} \end{bmatrix} \in \mathbb{R}^{(p+m) \times (p+m)}.\end{aligned}$$

One can verify that the linear systems corresponding to the transfer functions

$$\begin{bmatrix} \mathbf{I}_p & \mathbf{0}_{p \times m} \\ \mathbf{G}_{\text{lo},\infty}(-s)^\top & \mathbf{J}_\infty(-s)^\top \end{bmatrix}^{-1} \quad \text{and} \quad \begin{bmatrix} \mathbf{G}_{\text{lo},\infty}(s) & \mathbf{K}_\infty(s) \\ -\mathbf{D}^\top \mathbf{G}_{\text{lo},\infty}(s) & -\mathbf{D}^\top \mathbf{K}_\infty(s) \end{bmatrix}$$

have realizations given by

$$(-\mathbf{A}^\top + \widehat{\mathbf{C}}_\mathcal{J}^\top \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{B}}^\top, -\widehat{\mathbf{C}}_\mathcal{J}^\top \mathbf{R}^{-1}, \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{B}}^\top, \widehat{\mathbf{R}}^{-1}) \quad \text{and} \quad (\mathbf{A}, \widehat{\mathbf{B}}_\mathcal{K}, \widehat{\mathbf{C}}, \mathbf{0}_{p+m}),$$

respectively. Then, the realization of the cascaded system \mathcal{Z} can be computed according to Proposition 2.23 to be

$$\mathcal{Z} = \left(\begin{array}{cc|c} -\mathbf{A}^\top + \widehat{\mathbf{C}}_\mathcal{J}^\top \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{B}}^\top & \widehat{\mathbf{C}}_\mathcal{J}^\top \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{C}} & \mathbf{0}_{n \times (p+m)} \\ \mathbf{0}_n & \mathbf{A} & \widehat{\mathbf{B}}_\mathcal{K} \\ \hline \widehat{\mathbf{R}}^{-1} \mathbf{B}^\top & \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{C}} & \mathbf{0}_{p+m} \end{array} \right). \quad (4.49)$$

By manipulating the ARE (4.36) into the form

$$(-\mathbf{A}^\top + \widehat{\mathbf{C}}_\mathcal{J}^\top \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{B}}^\top)(-\mathbf{Q}_\mathcal{J}^-) + \mathbf{A} \mathbf{Q}_\mathcal{J}^- + \widehat{\mathbf{C}}_\mathcal{J}^\top \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{C}} = \mathbf{0}_{n \times n},$$

it becomes clear that the state-space transformation defined as

$$\mathbf{T} = \begin{bmatrix} \mathbf{I}_n & -\mathbf{Q}_{\mathcal{J}}^- \\ \mathbf{0}_n & \mathbf{I}_n \end{bmatrix} \in \mathbb{R}^{2n \times 2n}$$

decouples the cascaded system in (4.49). In other words, the transformed state space realization of \mathcal{Z} is given by

$$\mathcal{Z} = \left(\begin{array}{cc|c} -\mathbf{A}^\top + \widehat{\mathbf{C}}_{\mathcal{J}}^\top \widehat{\mathbf{R}}^{-1} \widehat{\mathbf{B}}^\top & \mathbf{0}_n & \mathbf{Q}_{\mathcal{J}}^- \widehat{\mathbf{B}}_{\mathcal{K}} \\ \mathbf{0}_n & \mathbf{A} & \widehat{\mathbf{B}}_{\mathcal{K}} \\ \hline \widehat{\mathbf{R}}^{-1} \mathbf{B}^\top & \widehat{\mathbf{C}}_{\mathcal{J}} & \mathbf{0}_{p+m} \end{array} \right).$$

Evidently, the stable part of \mathcal{Z} has the transfer function $\mathbf{Z}(s)^- = \widehat{\mathbf{C}}_{\mathcal{J}}(s\mathbf{I}_n - \mathbf{A})^{-1} \widehat{\mathbf{B}}_{\mathcal{K}}$, thus proving previous claim. \square

Theorems 4.1 and 4.13 provide the foundation for a data-driven implementation of BRBT, which we call *quadrature-based* BRBT (QuadBRBT). Algorithm 4.2.2 yields QuadBRBT when the transfer functions $\mathbf{G}_{\sigma, \mathbf{A}}$ and $\mathbf{G}_{\mathbf{B}}, \mathbf{G}_{\mathbf{C}}$ to be sampled are replaced with those in (4.47) and (4.48).

4.3.4 Common setup for numerical experiments

Before presenting the numerical results for this section, we discuss here common aspects of the setup and environment for the numerical experiments performed in the subsequent Section 4.3.5, as well as the later Sections 4.4.3, and 4.5.4.

In each section, we use an exponential trapezoidal quadrature rule [211] to implicitly approximate the relevant Gramians. For simplicity, we use an even and equal number $N = J = K$ of left and right quadrature nodes interweaved along the imaginary axis so that the imaginary parts of the nodes satisfy

$$\vartheta_1 < \zeta_1 < \vartheta_2 < \zeta_2 < \cdots < \vartheta_N < \zeta_N \quad (4.50)$$

and that these points are closed under complex conjugation

$$\{-i\vartheta_i\}_{i=1}^N = \{i\vartheta_i\}_{i=1}^N \quad \text{and} \quad \{-i\zeta_j\}_{j=1}^N = \{i\zeta_j\}_{j=1}^N. \quad (4.51)$$

For each example, the quadrature rules will be applied in chosen intervals along the imaginary axis $[i\omega_{\min}, i\omega_{\max}]$, for real-valued $0 < \omega_{\min} < \omega_{\max}$. The points will be linearly spaced or logarithmically spaced in the interval $[i\omega_{\min}, i\omega_{\max}]$; we make the choice clear whenever it arises.

Table 4.1: Relative Frobenius errors (4.53) in the first 20 (stochastic (4.18), positive-real (4.27), or bounded-real (4.38)) singular values of the RLC circuit benchmark. The smallest error for each set of quadrature-based reduced models is highlighted in **boldface**.

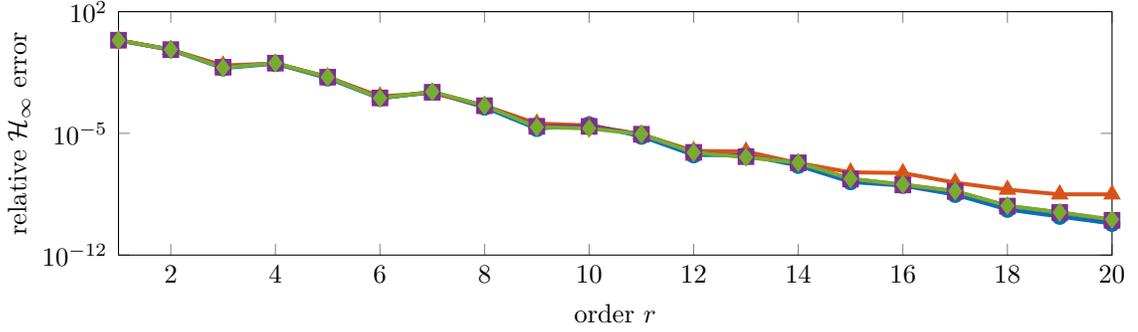
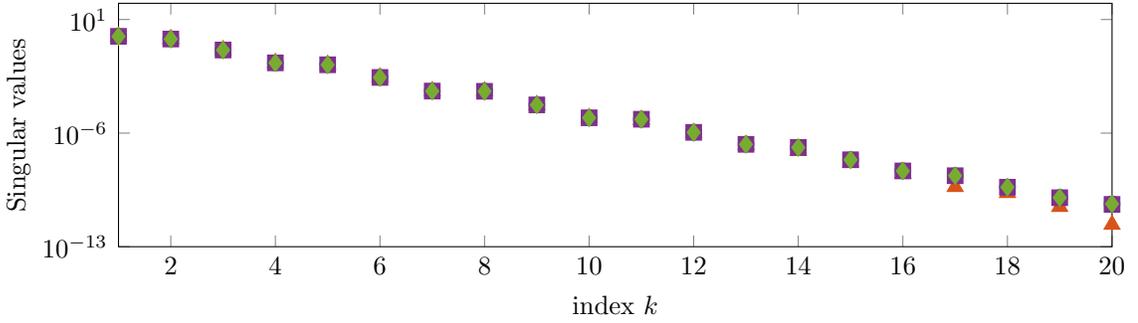
	$N = 50$	$N = 100$	$N = 200$
$\ \Sigma^{\text{bst}} - \check{\Sigma}^{\text{bst}}\ _{\text{F}} / \ \Sigma^{\text{bst}}\ _{\text{F}}$	3.3872e-2	2.3464e-2	2.6343e-2
$\ \Sigma^{\text{prbt}} - \check{\Sigma}^{\text{prbt}}\ _{\text{F}} / \ \Sigma^{\text{prbt}}\ _{\text{F}}$	3.3603e-2	1.6474e-2	1.8390e-2
$\ \Sigma^{\text{brbt}} - \check{\Sigma}^{\text{brbt}}\ _{\text{F}} / \ \Sigma^{\text{brbt}}\ _{\text{F}}$	5.0034e-1	4.6219e-1	4.5352e-1

Each of the data-based methods used in this chapter requires evaluating some sort of transfer function to obtain the relevant data. The data used to generate these data-driven BT-ROMs are all computed *synthetically*. That is, an explicit computational model such as (2.25) or (4.72) is used to evaluate the relevant transfer function data. For benchmarking, we compare each of the *data-driven* BT-ROMs to their *intrusive* counterparts. These intrusive benchmarks are computed using the MATLAB toolbox MORLAB [36].

4.3.5 Numerical results

In this section, we provide a numerical proof of concept for the data-driven reduced models computed by QuadBST, QuadPRBT, and QuadBRBT. We assume the common setup for numerical experiments outlined in Section 4.3.4. The data-driven approaches are compared to their intrusive counterparts BST, PRBT, and BRBT. We test the methods on a single-input, single-output RLC circuit model of order $n = 400$ taken from [96] with the choice of physical parameters $R = C = L = 0.1$ and $\bar{R} = 1$. We use the circuit model because it is asymptotically stable, passive, square, and contains a nonsingular feedthrough term by construction. Moreover, we normalize the model so that it is bounded real (4.35). Thus, all of the BT-variants discussed in this section can be reasonably applied to this benchmark.

We compute the spectral factor data used to construct the quadrature-based reduced models intrusively by directly constructing a state-space realization of each of the spectral factors, and then evaluating its transfer function at the required points. The built-in MATLAB routine ‘icare’ is used to compute the stabilizing solution of the associated AREs. In some practical scenarios, solely evaluations of \mathbf{G}_{lo} are available. A similar approach as in [29] can then be used: first, construct a reduced-order surrogate using these data via QuadBT, and then use this surrogate to obtain numerical evaluations of the spectral factors required for QuadBST, QuadPRBT, and QuadBRBT. However, at this point one could just as well compute a reduced model *intrusively* using BST, PRBT, or BRBT from the data-driven intermediate.

(a) Absolute \mathcal{H}_∞ errors (4.52) for the order $r = 1, 2, \dots, 20$ BST and QuadBST reduced models.

(b) True stochastic singular values (4.18) compared against the approximate ones.



Figure 4.1: Results for the quadrature-based and intrusive BST reduced models of the RLC circuit benchmark.

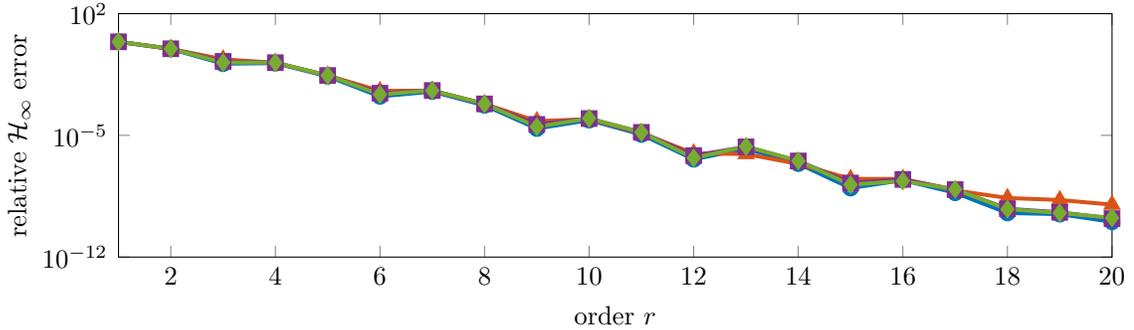
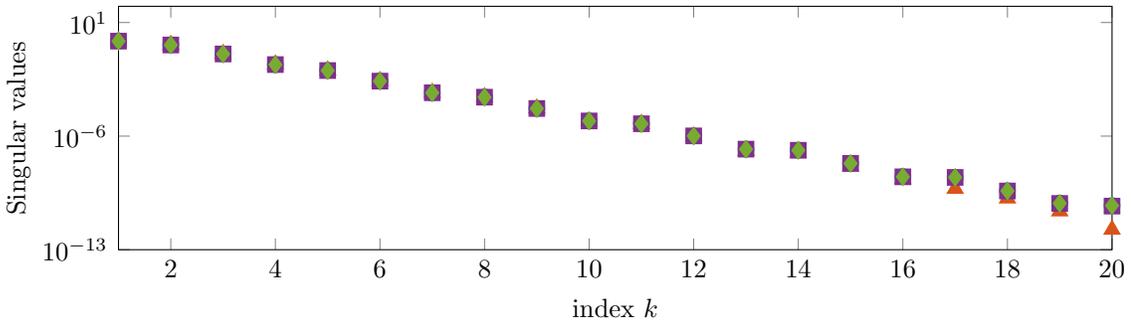
For comparing the computed reduced models, we use the absolute \mathcal{H}_∞ error:

$$\text{abserr}_{\mathcal{H}_\infty} \stackrel{\text{def}}{=} \|\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo}}\|_{\mathcal{H}_\infty}, \quad (4.52)$$

which is computed using the ‘norm’ command in MATLAB’s Control Systems Toolbox. The algorithm used to compute the \mathcal{H}_∞ norm is from [48]. For each type of balancing, we also compute the error in the *true* (stochastic (4.18), positive-real (4.27), or bounded-real (4.38)) singular values $\sigma(\mathbf{L}_y^T \mathbf{R}_x)$ against the *data-based* singular values $\sigma(\mathbf{E})$ via the relative error in the Frobenius norm:

$$\text{relerr}_{\mathbf{F}} \stackrel{\text{def}}{=} \frac{\|\Sigma - \check{\Sigma}\|_{\mathbf{F}}}{\|\Sigma\|_{\mathbf{F}}}. \quad (4.53)$$

For each of QuadBST, QuadPRBT, and QuadBST, reduced models of orders $r = 1, \dots, 20$ are computed according to Algorithm 4.2.2 using the appropriate data sampled along $N = 50, 100, 200$ points. The left and right points are constructed by choosing N logarithmically spaced points from $i10^{-1}$ to $i10^4$, interweaving these points according to (4.50), and then

(a) Absolute \mathcal{H}_∞ errors (4.52) for the order $r = 1, 2, \dots, 20$ PRBT and QuadPRBT reduced models.

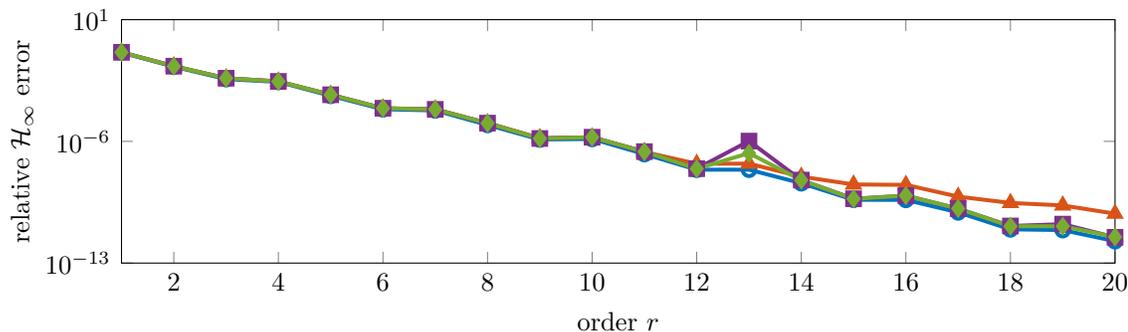
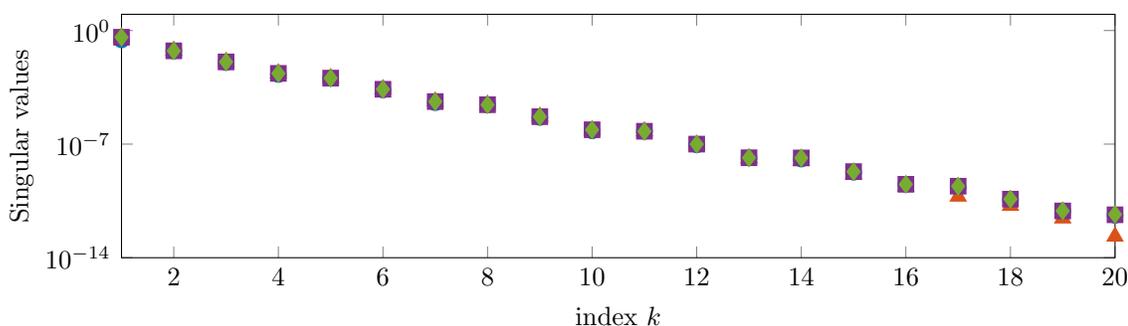
(b) True positive-real singular values (4.27) compared against the approximate ones.



Figure 4.2: Results for the quadrature-based and intrusive PRBT reduced models of the RLC circuit benchmark.

closing the two sets of points under conjugation according to (4.51). We also compute (intrusive) order $r = 1, \dots, 30$ reduced models using MORLAB’s implementation of BST, PRBT, and BRBT for benchmarking.

The results are recorded in Figures 4.1, 4.2, and 4.3. For each of the intrusive and quadrature-based BST, PRBT, and BRBT reduced models, the top plots contain the absolute \mathcal{H}_∞ errors, whereas the bottom plots contain the (relevant) singular values Σ of the matrix $L_y^T R_x$ plotted against the singular values $\tilde{\Sigma}$ of the data-based \mathbb{E} . The relative errors in the singular values (4.53) are reported in Table 4.1. As each figure illustrates, the quadrature-based reduced models produce nearly the same \mathcal{H}_∞ error as the intrusive reduced models for all orders of reduction, and the relevant singular values in each instance are very close to the true singular values. From $N = 50$ to $N = 100$ and $N = 200$, the approximation quality in the \mathcal{H}_∞ error of the quadrature-based reduced models increases. The only exceptions are the order $r = 13$ QuadBRBT reduced models, which exhibit a marginally higher absolute

(a) Absolute \mathcal{H}_∞ errors (4.52) for the order $r = 1, 2, \dots, 20$ BRBT and QuadBRBT reduced models.

(b) True bounded-real singular values (4.38) compared against the approximate ones.

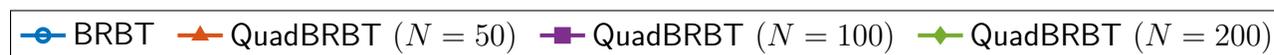


Figure 4.3: Results for the quadrature-based and intrusive BRBT reduced models of the RLC circuit benchmark.

\mathcal{H}_∞ error for $N = 100$ and $N = 200$. From $N = 100$ to $N = 200$, the \mathcal{H}_∞ errors and the error in the Hankel singular values do not follow a specific trend. Surprisingly, the errors in Table 4.1 show that the errors in the stochastic and bounded-real singular get slightly larger from $N = 100$ to $N = 200$. This might suggest that the choice of quadrature nodes is not optimal.

We emphasize that these numerical results are intended to be a *proof of concept*. They are included to illustrate that, if one had the requisite data prescribed by Theorems 4.5, 4.9, and 4.13, then the data-driven reduced models computed by QuadBST, QuadPRBT, and QuadBRBT perform very closely to their intrusive counterparts.

4.4 Frequency-weighted balanced truncation

Each of the previously discussed variants of balanced truncation aims to reproduce a system (2.25) or its transfer function (2.30), over the *entire* frequency range; that is, by weighing all frequencies equally. In numerous applications, however, one would like to weigh certain frequencies more than others. This motivates *frequency-weighted* model reduction: Given a linear system \mathcal{G}_{lo} as in (2.25) having the transfer function \mathbf{G}_{lo} , an *input weight* $\mathbf{G}_i: \mathbb{C} \rightarrow \mathbb{C}^{m \times m_i}$, and an *output weight* $\mathbf{G}_o: \mathbb{C} \rightarrow \mathbb{C}^{p_o \times p}$, the problem is to compute a reduced-order system $\tilde{\mathcal{G}}_{\text{lo}}$ so that the *frequency-weighted error* is small, i.e.:

$$\text{Find } \tilde{\mathbf{G}}_{\text{lo}} \text{ so that } \left\| \mathbf{G}_o \left(\mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}} \right) \mathbf{G}_i \right\|_{\mathcal{H}_\infty^{p_o \times m_i}} \text{ is small for given } \mathbf{G}_o \text{ and } \mathbf{G}_i. \quad (4.54)$$

For this section, \mathbf{G}_i and \mathbf{G}_o will be order- n_i and order- n_o rational transfer functions of some underlying linear systems \mathcal{G}_i and \mathcal{G}_o formulated according to (2.25). Several attempts have been made at the frequency-weighted problem (4.54); see [69, 116, 128, 221, 244]. Enns [69] was the first to consider the frequency-weighted problem in the context of balanced realizations using input and output weights. The goal of this section is to derive a *data-driven* formulation of the frequency-weighted balanced truncation (FWBT) of Enns [69], and extend the quadrature-based framework of [89] to this setting. Instead of doing so with the generalized framework of Section 4.2, we address the frequency-weighted problem (4.54) in a vacuum. Although it is possible to derive a data-driven formulation of FWBT using the generalized approach, it turns out to be more natural to do so directly from the problem formulation given in (4.54). In addition to the general framework of Enns [69], we will also discuss the self-weighted variation of Zhou [244] as a special case. It needs to be mentioned that usually, the weights \mathbf{G}_i and \mathbf{G}_o are not provided; rather, the user is interested in the reduced model performance over a known frequency band $\Omega = [\omega_1, \omega_2]$. The frequency-limited method by Gawronski and Juang [84] aims to address this problem, although we do not consider their approach here. There are instances in which the weights in (4.54) are known; e.g., in controller reduction or in the *self-weighted* case considered in [244], where $\mathbf{G}_i = \mathbf{I}_m$ and $\mathbf{G}_o = \mathbf{G}_{\text{lo}}^{-1}$. In this case, the frequency-weighted problem (4.54) becomes a *relative error* problem.

4.4.1 Intrusive frequency-weighted balanced truncation

The general formulation of Enns.

Consider an asymptotically stable and minimal linear system \mathcal{G}_{lo} as in (2.25), and a pair of minimal linear systems \mathcal{G}_i and \mathcal{G}_o defined by the realizations

$$\mathcal{G}_i = (\mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i, \mathbf{D}_i) \quad \text{and} \quad \mathcal{G}_o = (\mathbf{A}_o, \mathbf{B}_o, \mathbf{C}_o, \mathbf{D}_o), \quad (4.55)$$

where $\mathbf{A}_i \in \mathbb{R}^{n_i \times n_i}$, $\mathbf{B}_i \in \mathbb{R}^{n_i \times m_i}$, $\mathbf{C}_i \in \mathbb{R}^{m \times n_i}$, $\mathbf{D}_i \in \mathbb{R}^{m \times m_i}$, and $\mathbf{A}_o \in \mathbb{R}^{n_o \times n_o}$, $\mathbf{B}_o \in \mathbb{R}^{n_o \times p}$, $\mathbf{C}_o \in \mathbb{R}^{p_o \times n_i}$, $\mathbf{D}_o \in \mathbb{R}^{p_o \times p}$ for $n_i, m_i, p_i, n_o, m_o, p_o \in \mathbb{Z}_{>0}$. In order to guarantee the frequency

error in (4.54) is finite—barring a possibly unstable reduced model—we assume that \mathcal{G}_i and \mathcal{G}_o are asymptotically stable. The systems (4.55) respectively model the *input weight* and *output weight*

$$\mathbf{G}_i(s) = \mathbf{C}_i(s\mathbf{I}_{n_i} - \mathbf{A}_i)^{-1}\mathbf{B}_i + \mathbf{D}_i \quad \text{and} \quad \mathbf{G}_o(s) = \mathbf{C}_o(s\mathbf{I}_{n_o} - \mathbf{A}_o)^{-1}\mathbf{B}_o + \mathbf{D}_o. \quad (4.56)$$

The relevant Gramians in FWBT can be straightforwardly derived from the behavior of the input-to-output response of (2.25) under the weights (4.56). Recall the integral formulations of the reachability Gramian (2.41) and observability Gramian (2.42). Consider the input-to-state and state-to-output behavior of the weighted transfer function $\mathbf{G}_o\mathbf{G}_i$: An input $\mathbf{U}(s)$ applied to the weighted system first goes through the input weight \mathbf{G}_i , leading to a weighted input which is then fed to \mathbf{G}_o . Alternatively, we can view \mathbf{G}_o as acting on inputs of the form $\mathbf{G}_i(s)\mathbf{U}(s)$, leading to the *weighted input-to-state map* $(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}\mathbf{G}_i(s)$. Replacing the usual state-to-input map in the integral formulation (2.41) of the reachability Gramian $\mathbf{P} \in \mathbb{R}^{n \times n}$ with this weighted map leads to the *weighted reachability Gramian*

$$\mathbf{P}_i = \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\omega\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}\mathbf{G}_i(i\omega) ((i\omega\mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B}\mathbf{G}_i(i\omega))^H d\omega. \quad (4.57)$$

Consider the state-to-output map of the weighted transfer function: The state of the system passes through the output weight \mathcal{G}_o , leading to a weighted output. This leads to a *weighted state-to-output map* $\mathbf{G}_o(s)\mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}$. Replacing the usual state-to-output map in (2.42) with this weighted map leads to the *weighted observability Gramian*

$$\mathbf{Q}_o = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbf{G}_o(i\omega)\mathbf{C}(i\omega\mathbf{E} - \mathbf{A})^{-1})^H \mathbf{G}_o(i\omega)\mathbf{C}(i\omega\mathbf{E} - \mathbf{A})^{-1} d\omega. \quad (4.58)$$

The weighted Gramians $\mathbf{P}_i \in \mathbb{R}^{n \times n}$ and $\mathbf{Q}_o \in \mathbb{R}^{n \times n}$ can also be obtained from solutions to ALEs as follows. Using Proposition 2.23, we compute realizations of the cascaded systems $\mathcal{G}_i\mathcal{G}_o$ and $\mathcal{G}_o\mathcal{G}_i$ with the transfer functions $\mathbf{G}_i\mathbf{G}_o$ and $\mathbf{G}_o\mathbf{G}_i$:

$$\mathcal{G}_i\mathcal{G}_o = (\widehat{\mathbf{A}}_i, \widehat{\mathbf{B}}_i, \widehat{\mathbf{C}}_i, \widehat{\mathbf{D}}_i) = \left(\left[\begin{array}{cc} \mathbf{A} & \mathbf{B}\mathbf{C}_i \\ \mathbf{0}_{n_i \times n} & \mathbf{A}_i \end{array} \right], \left[\begin{array}{c} \mathbf{B}\mathbf{D}_i \\ \mathbf{B}_i \end{array} \right], [\mathbf{C} \quad \mathbf{D}\mathbf{C}_i], \mathbf{D}\mathbf{D}_i \right), \quad (4.59)$$

$$\mathcal{G}_o\mathcal{G}_i = (\widehat{\mathbf{A}}_o, \widehat{\mathbf{B}}_o, \widehat{\mathbf{C}}_o, \widehat{\mathbf{D}}_o) = \left(\left[\begin{array}{cc} \mathbf{A} & \mathbf{0}_{n \times n_o} \\ \mathbf{B}_o\mathbf{C} & \mathbf{A}_o \end{array} \right], \left[\begin{array}{c} \mathbf{B} \\ \mathbf{B}_o\mathbf{D} \end{array} \right], [\mathbf{D}_o\mathbf{C} \quad \mathbf{C}_o], \mathbf{D}_o\mathbf{D} \right). \quad (4.60)$$

Assuming no pole-zero cancellation in the transfer functions $\mathbf{G}_i\mathbf{G}_o$ and $\mathbf{G}_o\mathbf{G}_i$, the realizations in (4.59) and (4.60) are minimal. Moreover, the cascaded systems are asymptotically stable, and so the reachability Gramian $\widehat{\mathbf{P}}_i \in \mathbb{R}^{n \times n}$ of (4.59) and the observability Gramian $\widehat{\mathbf{Q}}_o \in \mathbb{R}^{n \times n}$ of (4.60) uniquely satisfy the Lyapunov equations

$$\widehat{\mathbf{A}}_i\widehat{\mathbf{P}}_i + \widehat{\mathbf{P}}_i\widehat{\mathbf{A}}_i^\top + \widehat{\mathbf{B}}_i\widehat{\mathbf{B}}_i^\top = \mathbf{0}_{n \times n} \quad \text{and} \quad \widehat{\mathbf{A}}_o^\top\widehat{\mathbf{Q}}_o + \widehat{\mathbf{Q}}_o\widehat{\mathbf{A}}_o + \widehat{\mathbf{C}}_o^\top\widehat{\mathbf{C}}_o = \mathbf{0}_{n \times n}. \quad (4.61)$$

It can be shown [4, Proposition 7.19] that the (1,1)-blocks of $\widehat{\mathbf{P}}_i$ and $\widehat{\mathbf{Q}}_o$ of dimension $n \times n$ are the weighted Gramians \mathbf{P}_i and \mathbf{Q}_o in (4.57) and (4.58). Thus, under the given

assumptions, the weighted Gramians \mathbf{P}_i and \mathbf{Q}_o are SPD, and uniquely determined by the ALEs (4.61).

In FWBT, the weighted Gramians (4.57) and (4.58) are balanced, in place of the usual reachability and observability Gramians.

Definition 4.14 (Frequency-weighted balanced realization [69]). We say that a state-space realization of the minimal system \mathcal{G}_{lo} in (2.25) is a *frequency-weighted balanced realization* if

$$\mathbf{P}_i = \mathbf{Q}_o = \Sigma^{\text{fwbt}} = \text{diag}(\hat{\sigma}_1 \mathbf{I}_{m_1}, \hat{\sigma}_2 \mathbf{I}_{m_2}, \dots, \hat{\sigma}_q \mathbf{I}_{m_q}), \quad (4.62)$$

where $\hat{\sigma}_1 > \hat{\sigma}_2 > \dots > \hat{\sigma}_q > 0$ and their multiplicities satisfy $m_1 + \dots + m_q = n$. The values $\hat{\sigma}_i$ are called the *frequency-weighted singular values*, and are dependent upon \mathcal{G}_{lo} , \mathcal{G}_i , and \mathcal{G}_o . \diamond

Theorem 4.15 (Frequency-weighted balanced truncation [69, 116]). Consider an asymptotically stable, minimal, linear system \mathcal{G}_{lo} in (2.25) having the frequency-weighted balanced realization

$$\mathbf{A}_{\text{fwbt}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{B}_{\text{fwbt}} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \text{and} \quad \mathbf{C}_{\text{fwbt}} = [\mathbf{C}_1 \quad \mathbf{C}_2]$$

according to Definition 4.14. The matrices are partitioned with respect to $\mathbf{P}_i = \mathbf{Q}_o = \Sigma^{\text{fwbt}} = \text{diag}(\Sigma_1^{\text{fwbt}}, \Sigma_2^{\text{fwbt}})$, where

$$\Sigma_1^{\text{fwbt}} = \text{diag}(\hat{\sigma}_1 \mathbf{I}_{m_1}, \dots, \hat{\sigma}_k \mathbf{I}_{m_k}) \quad \text{and} \quad \Sigma_2^{\text{fwbt}} = \text{diag}(\hat{\sigma}_{k+1} \mathbf{I}_{m_{k+1}}, \dots, \hat{\sigma}_q \mathbf{I}_{m_q})$$

for $r = m_1 + \dots + m_k$ and $1 \leq k < q$. Then, the order- r reduced model $\tilde{\mathcal{G}}_{lo, \text{fwbt}} = (\mathbf{A}_{11}, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D})$ obtained via frequency-weighted balanced truncation is balanced in the sense of (4.62). If either $\mathbf{G}_i = \mathbf{I}_{m_i}$ or $\mathbf{G}_o = \mathbf{I}_{p_o}$, then $\tilde{\mathcal{G}}_{lo, \text{fwbt}}$ is asymptotically stable, and the frequency-weighted error bound holds:

$$\left\| \mathcal{G}_o \left(\mathcal{G}_{lo} - \tilde{\mathcal{G}}_{lo, \text{fwbt}} \right) \mathcal{G}_i \right\|_{\mathcal{H}_\infty} \leq 2 \sum_{i=k+1}^q \left(\hat{\sigma}_i^2 + (\alpha_i + \beta_i) \hat{\sigma}_i^{3/2} + \alpha_i \beta_i \hat{\sigma}_i \right)^{1/2}, \quad (4.63)$$

where α_i and β_i are the \mathcal{H}_∞ norms of some transfer functions that depend upon the weights \mathbf{G}_i and \mathbf{G}_o , as well as the transfer functions of the reduced-order models obtained by FWBT of order $j = 1, \dots, k$. \diamond

The original formulation of Theorem 4.15 and the stability result are due to Enns [69], while the error bound is due to Kim et al. [116]. To guarantee asymptotic stability in the reduced model for the case of two-sided weighting, the method by Lin and Chiu [128] instead balances the Schur complements of the Gramians $\hat{\mathbf{P}}_i$ and $\hat{\mathbf{Q}}_o$ that satisfy (4.61). A similarly complicated error bound to (4.63) holds in this case, as well. A simpler error bound is provided by the variation of [221]. We do not consider either of the approaches in [128, 221] further. In a practical implementation, FWBT according to Theorem 4.15 is achieved using a square-root implementation such as Algorithm 4.2.1, where $\mathbf{R}_x = \mathbf{R}_i$ and $\mathbf{L}_y = \mathbf{L}_o$ are Cholesky factors of \mathbf{P}_i and \mathbf{Q}_o .

The self-weighted approach of Zhou.

A special case of the frequency-weighted balancing of Enns is the *self-weighted* method of Zhou [244]. Suppose the linear system \mathcal{G}_{lo} in (2.25) has a nonsingular feedthrough term \mathbf{D} so that its inverse system is well defined according to Proposition 2.25. If (2.25) is additionally assumed to be minimum phase, i.e., the zeros of its transfer function lie in $\mathbb{C}_{<0}$, then $\mathcal{G}_{\text{lo}}^{-1}$ computed according to (2.35) will be asymptotically stable, as well. Under these assumptions, and when the input- and output-weights are taken to be

$$\mathbf{G}_i(s) = \mathbf{I}_m \quad \text{and} \quad \mathbf{G}_o(s) = \mathbf{G}_{\text{lo}}(s)^{-1},$$

the frequency-weighted balancing described by Theorem 4.15 becomes a *relative* error method, i.e., it attempts to minimize the relative \mathcal{H}_∞ approximation error $\|\mathbf{G}_{\text{lo}}^{-1}(\mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}})\|_{\mathcal{H}_\infty^{p \times m}}$. In this special case, the weighted reachability Gramian (4.57) is simply the usual reachability Gramian $\mathbf{P}_i = \mathbf{P}$ in (2.41), and the weighted observability Gramian \mathbf{Q}_o is defined as in (4.58) with $\mathbf{G}_o = \mathbf{G}_{\text{lo}}^{-1}$. Alternatively, \mathbf{Q}_o can be computed as the solution to the observability Lyapunov equation of the inverse system (2.35):

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^\top \mathbf{Q}_o + \mathbf{Q}_o (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) + (\mathbf{D}^{-1}\mathbf{C})^\top (\mathbf{D}^{-1}\mathbf{C}) = \mathbf{0}_{n \times n}. \quad (4.64)$$

See [244, Lemma 2] for details. In the self-weighted approach, \mathbf{P} is balanced with \mathbf{Q}_o that solves (4.64) to yield a frequency-weighted balanced realization according to Definition 4.14. The reduced model obtained by truncating this balanced realization has the following properties.

Theorem 4.16 (Self-weighted balanced truncation [244]). Consider an asymptotically stable, minimal, square, and minimum-phase linear system \mathcal{G}_{lo} in (2.25). Suppose that $\tilde{\mathcal{G}}_{\text{lo},\text{fwbt}}$ is obtained via Zhou's self-weighted balanced truncation according to Theorem 4.15 with $\mathbf{G}_i = \mathbf{I}_m$ and $\mathbf{G}_o = \mathbf{G}_{\text{lo}}^{-1}$. Then, $\tilde{\mathcal{G}}_{\text{lo},\text{fwbt}}$ is balanced in the sense of (4.62), asymptotically stable, and satisfies the relative error bounds

$$\begin{aligned} \left\| \mathcal{G}_{\text{lo}}^{-1} \left(\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo},\text{fwbt}} \right) \right\|_{\mathcal{H}_\infty} &\leq \prod_{i=k+1}^q \left(1 + 2\hat{\sigma}_i \sqrt{1 + \hat{\sigma}_i^2} + 2\hat{\sigma}_i^2 \right) - 1, \\ \left\| \tilde{\mathcal{G}}_{\text{lo}}^{-1} \left(\mathcal{G}_{\text{lo}} - \tilde{\mathcal{G}}_{\text{lo},\text{fwbt}} \right) \right\|_{\mathcal{H}_\infty} &\leq \prod_{i=k+1}^q \left(1 + 2\hat{\sigma}_i \sqrt{1 + \hat{\sigma}_i^2} + 2\hat{\sigma}_i^2 \right) - 1. \end{aligned} \quad (4.65)$$

◇

4.4.2 Quadrature-based frequency-weighted balanced truncation

Following the ideas of Sections 4.2 and 4.3, we derive here a quadrature-based/data-driven formulation of the frequency-weighted balanced truncation [69, 244] just introduced. As

our starting point, we use the readily available integral formulations for the to-be-balanced weighted Gramians $\mathbf{P}_i \in \mathbb{R}^{n \times n}$ in (4.57) and $\mathbf{Q}_o \in \mathbb{R}^{n \times n}$ in (4.58). Let \mathbf{G}_i and \mathbf{G}_o be the input- and output-weights defined in (4.56). Consider a numerical quadrature rule defined by the nodes $i\zeta_1, \dots, i\zeta_J \in i\mathbb{R}$ and weights $\varrho_1^2, \dots, \varrho_J^2 \in \mathbb{R}$. Applying this rule to \mathbf{P}_i in (4.57) reveals the approximate factorization

$$\mathbf{P}_i \approx \sum_{j=1}^J \varrho_j^2 (i\zeta_j \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} \mathbf{G}_i(i\zeta_j) ((i\zeta_j \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} \mathbf{G}_i(i\zeta_j))^H = \check{\mathbf{R}}_i \check{\mathbf{R}}_i^H,$$

where $\check{\mathbf{R}}_i \in \mathbb{C}^{n \times m_i J}$ is defined as

$$\check{\mathbf{R}}_i \stackrel{\text{def}}{=} [\varrho_1 (i\zeta_1 \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} \mathbf{G}_i(i\zeta_1) \quad \cdots \quad \varrho_J (i\zeta_J \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} \mathbf{G}_i(i\zeta_J)]. \quad (4.66)$$

We assume that the factor $1/2\pi$ is included in each of the quadrature weights. Likewise, applying a numerical quadrature rule defined by the nodes $i\vartheta_1, \dots, i\vartheta_K \in i\mathbb{R}$ and weights $\varphi_1^2, \dots, \varphi_K^2 \in \mathbb{R}$ to \mathbf{Q}_o in (4.58) produces the approximate factorization

$$\mathbf{Q}_o \approx \sum_{k=1}^K \varphi_k^2 (\mathbf{G}_o(i\vartheta_k) \mathbf{C} (i\vartheta_k \mathbf{I}_n - \mathbf{A})^{-1})^H \mathbf{G}_o(i\vartheta_k) \mathbf{C} (i\vartheta_k \mathbf{I}_n - \mathbf{A})^{-1} = \check{\mathbf{L}}_o \check{\mathbf{L}}_o^H,$$

where $\check{\mathbf{L}}_o \in \mathbb{C}^{n \times p_o K}$ is defined according to

$$\check{\mathbf{L}}_o^H \stackrel{\text{def}}{=} \begin{bmatrix} \varphi_1 \mathbf{G}_o(i\vartheta_1) \mathbf{C} (i\vartheta_1 \mathbf{I}_n - \mathbf{A})^{-1} \\ \vdots \\ \varphi_K \mathbf{G}_o(i\vartheta_K) \mathbf{C} (i\vartheta_K \mathbf{I}_n - \mathbf{A})^{-1} \end{bmatrix}. \quad (4.67)$$

An exact (intrusive) implementation of the FWBT described by Theorem 4.15 can be achieved using Algorithm 4.2.1, where the factors $\mathbf{R}_\mathcal{X} = \mathbf{R}_i$ and $\mathbf{L}_\mathcal{Y} = \mathbf{L}_o$ are the Cholesky factors of $\mathbf{P}_i = \mathbf{R}_i \mathbf{R}_i^T$ and $\mathbf{Q}_o = \mathbf{L}_o \mathbf{L}_o^T$. As in Section 4.2, we arrive at a quadrature-based implementation by replacing the exact factors \mathbf{R}_i and \mathbf{L}_o with $\check{\mathbf{R}}_i$ and $\check{\mathbf{L}}_o$ in (4.66) and (4.67). The resulting (approximate) reduced model is thereby determined by the quadrature-based approximations to (4.4):

$$\check{\mathbf{L}}_o^H \check{\mathbf{R}}_i, \quad \check{\mathbf{L}}_o^H \mathbf{A} \check{\mathbf{R}}_i, \quad \check{\mathbf{L}}_o^H \mathbf{B}, \quad \text{and} \quad \mathbf{C} \check{\mathbf{R}}_i. \quad (4.68)$$

As with (4.9) in Theorem 4.1, the approximations in (4.68) are computable directly from data. In contrast to before, these data are *evaluations of the linear model transfer function* \mathbf{G}_o , and the *frequency weights* \mathbf{G}_i and \mathbf{G}_o .

Theorem 4.17 (Frequency-weighted balanced truncation from data). Let the quadrature-based square root factors $\check{\mathbf{R}}_i \in \mathbb{C}^{n \times m_i J}$ and $\check{\mathbf{L}}_o \in \mathbb{C}^{n \times p_o K}$ be defined according to (4.66)

and (4.67). Then, for FWBT the data-matrices defined in (4.11) for $\mathbf{R}_{\mathcal{X}} = \check{\mathbf{R}}_i$ and $\mathbf{L}_{\mathcal{Y}} = \check{\mathbf{L}}_o$ are given by

$$\mathbb{E}_{k,j} = -\varphi_k \varrho_j \frac{\mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\vartheta_k) \mathbf{G}_i(i\zeta_j) - \mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\zeta_j) \mathbf{G}_i(i\zeta_j)}{i\vartheta_k - i\zeta_j}, \quad (4.69a)$$

$$\mathbb{A}_{k,j} = -\varphi_k \varrho_j \frac{i\vartheta_k \mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\vartheta_k) \mathbf{G}_i(i\zeta_j) - i\zeta_j \mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\zeta_j) \mathbf{G}_i(i\zeta_j)}{i\vartheta_k - i\zeta_j}, \quad (4.69b)$$

$$\mathbb{B}_{k,:} = \varphi_k \mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\vartheta_k) \quad \text{and} \quad \mathbb{C}_{:,j} = \varrho_j \mathbf{G}_{lo,\infty}(i\zeta_j) \mathbf{G}_i(i\zeta_j), \quad (4.69c)$$

for all $j = 1, \dots, J$ and $k = 1, \dots, K$. \diamond

Proof of Theorem 4.17. The argument uses the same technology as the proof of Theorem 4.1. In fact, the result follows directly by running through the arguments used to prove Theorem 4.1 with $\mathbf{R}_{\mathcal{X}} = \check{\mathbf{R}}_i$ and $\mathbf{L}_{\mathcal{Y}} = \check{\mathbf{L}}_o$. For illustrative purposes, we still prove (4.69a) directly here. Using $\mathbf{R}_{\mathcal{X}} = \check{\mathbf{R}}_i$ and $\mathbf{L}_{\mathcal{Y}} = \check{\mathbf{L}}_o$ in (4.66) and (4.67) in the definition of \mathbb{E} in (4.11), the first resolvent identity (2.16) gives

$$\begin{aligned} \mathbb{E}_{k,j} &= \mathbf{I}_{k,p_y}^T \mathbb{E} \mathbf{I}_{j,m_x} = \mathbf{I}_{k,p_y}^T \check{\mathbf{L}}_o^H \check{\mathbf{R}}_i \mathbf{I}_{j,m_x} \\ &= \varphi_k \varrho_j \mathbf{G}_o(i\vartheta_k) \mathbf{C} (i\vartheta_k \mathbf{I}_n - \mathbf{A})^{-1} (i\zeta_j \mathbf{I}_n - \mathbf{A})^{-1} \mathbf{B} \mathbf{G}_i(i\zeta_j) \\ &= \varphi_k \varrho_j \mathbf{G}_o(i\vartheta_k) \mathbf{C} \left(\frac{(i\zeta_j \mathbf{I}_n - \mathbf{A})^{-1} - (i\vartheta_k \mathbf{I}_n - \mathbf{A})^{-1}}{i\vartheta_k - i\zeta_j} \right) \mathbf{B} \mathbf{G}_i(i\zeta_j) \\ &= -\varphi_k \varrho_j \frac{\mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\vartheta_k) \mathbf{G}_i(i\zeta_j) - \mathbf{G}_o(i\vartheta_k) \mathbf{G}_{lo,\infty}(i\zeta_j) \mathbf{G}_i(i\zeta_j)}{i\vartheta_k - i\zeta_j}, \end{aligned}$$

thus proving (4.69a). The other claims for (4.69b) and (4.69c) follow nearly identically. \square

Applying Theorem 4.17 for the special case of $\mathbf{G}_i = \mathbf{I}_m$ and $\mathbf{G}_o = \mathbf{G}_{lo}^{-1}$ results in a quadrature-based formulation of Zhou's self-weighted approach [244].

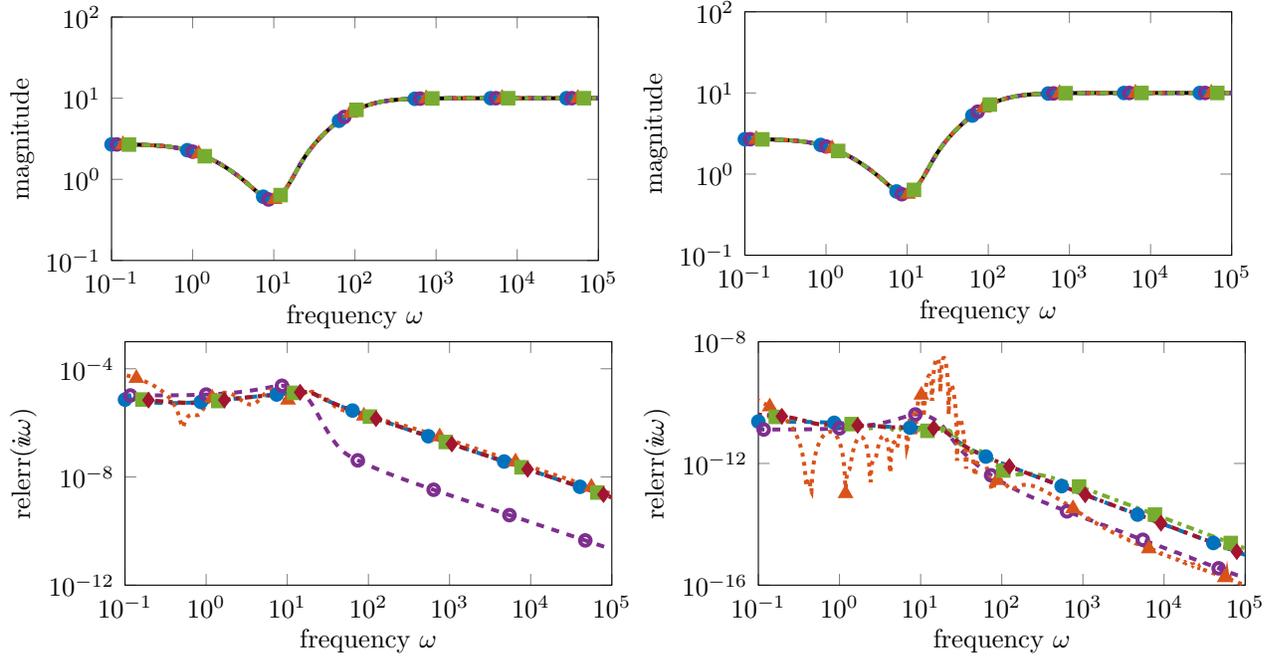
Corollary 4.18 (Self-weighted balanced truncation from data). Let the quadrature-based square root factors $\check{\mathbf{R}}_i \in \mathbb{C}^{n \times m_i J}$ and $\check{\mathbf{L}}_o \in \mathbb{C}^{n \times p_o K}$ be defined according to (4.66) and (4.67) for the choice of frequency weights $\mathbf{G}_i = \mathbf{I}_m$ and $\mathbf{G}_o = \mathbf{G}_{lo}^{-1}$. Then, the data-matrices defined in (4.11) for $\mathbf{R}_{\mathcal{X}} = \check{\mathbf{R}}_i$ and $\mathbf{L}_{\mathcal{Y}} = \check{\mathbf{L}}_o$ are given by

$$\mathbb{E}_{k,j} = -\varphi_k \varrho_j \frac{\mathbf{G}_{lo}(i\vartheta_k)^{-1} \mathbf{G}_{lo,\infty}(i\vartheta_k) - \mathbf{G}_{lo}(i\vartheta_k)^{-1} \mathbf{G}_{lo,\infty}(i\zeta_j)}{i\vartheta_k - i\zeta_j}, \quad (4.70a)$$

$$\mathbb{A}_{k,j} = -\varphi_k \varrho_j \frac{i\vartheta_k \mathbf{G}_{lo}(i\vartheta_k)^{-1} \mathbf{G}_{lo,\infty}(i\vartheta_k) - i\zeta_j \mathbf{G}_{lo}(i\vartheta_k)^{-1} \mathbf{G}_{lo,\infty}(i\zeta_j)}{i\vartheta_k - i\zeta_j}, \quad (4.70b)$$

$$\mathbb{B}_{k,:} = \varphi_k \mathbf{G}_{lo}(i\vartheta_k)^{-1} \mathbf{G}_{lo,\infty}(i\vartheta_k) \quad \text{and} \quad \mathbb{C}_{:,j} = \varrho_j \mathbf{G}_{lo,\infty}(i\zeta_j), \quad (4.70c)$$

for all $j = 1, \dots, J$ and $k = 1, \dots, K$. \diamond

(a) Transfer function magnitudes and relative errors for $r = 10$.(b) Transfer function magnitudes and relative errors for $r = 20$.Figure 4.4: Frequency response results for order $r = 10$ and $r = 20$ reduced-order models of the RLC circuit benchmark.

4.4.3 Numerical results

Here, we investigate QuadFWBT on a simple test problem. Specifically, we compare our quadrature-based implementation of Zhou’s self-weighted approach against an intrusive implementation. We choose to test this specific case of our method because, as already discussed, the input and output weight functions (4.56) are usually unknown outside of special instances. The common setup for numerical experiments outlined in Section 4.3.4 is assumed. We again use the RLC circuit model described in Section 4.3.5 as a test model, but with order $n = 1000$ and the physical parameters $R = 0.1$, $C = L = 0.001$. The RLC circuit model has a nonsingular \mathbf{D} term, so that the inverse transfer function $\mathbf{G}_{\text{lo}}^{-1}$ is well defined.

For the model reduction of the RLC circuit model, we compute reduced models of order $r = 10$ and $r = 20$ intrusively using BT and FWBT with self-weighting, and non-intrusively using QuadFWBT according to Corollary 4.18 for $N = 50, 100$, and 200 quadrature nodes.

Table 4.2: Relative \mathcal{H}_∞ errors according to (4.71) for the BT and frequency-weighted reduced-order models of the RLC circuit benchmark for $r = 10$ and 20. The smallest error for each order is highlighted in **boldface**.

	$r = 10$	$r = 20$
FWBT	1.3903e-5	2.3830e-12
BT	2.3535e-5	4.1638e-12
QuadFWBT ($N = 50$)	9.7923e-5	2.9343e-10
QuadFWBT ($N = 100$)	1.7401e-5	3.6533e-12
QuadFWBT ($N = 200$)	1.9475e-5	4.2062e-11

As before, the nodes are taken to be $N/2$ logarithmically spaced points from $i10^{-1}$ to $i10^4$, which are then closed under complex conjugation. The left and right points are interwoven according to the setup described in Section 4.3.4. We include BT in the results to compare the behavior of the reduced models computed using a frequency-weighted approach against one computed using an absolute error approach. Because MORLAB does not have an implementation of frequency-limited BT with weights, we have implemented the method ourselves.

For the presentation of the results, we plot the magnitude of the full- and reduced-order model transfer functions, as well as the pointwise relative error

$$\text{relerr}(i\omega_k) \stackrel{\text{def}}{=} \left| \mathbf{G}_{\text{lo}}(i\omega_k)^{-1} \left(\mathbf{G}_{\text{lo}}(i\omega_k) - \tilde{\mathbf{G}}_{\text{lo}}(i\omega_k) \right) \right|,$$

for $\omega_k \in \Omega \stackrel{\text{def}}{=} [10^{-1}, 10^4]$ are $k = 1, 2, \dots, 750$ logarithmically spaced points. The absolute error is used because the RLC circuit is a single-input, single-output system. We also score the reduced models using an approximation to the relative \mathcal{H}_∞ error

$$\text{relerr}_{\mathcal{H}_\infty} \stackrel{\text{def}}{=} \max_{\omega_k \in \Omega} \left| \mathbf{G}_{\text{lo}}(i\omega_k)^{-1} \left(\mathbf{G}_{\text{lo}}(i\omega_k) - \tilde{\mathbf{G}}_{\text{lo}}(i\omega_k) \right) \right|, \quad (4.71)$$

where Ω is defined as above.

The performance of the order $r = 10$ and $r = 20$ reduced models is recorded in Figure 4.4. Both sets of reduced models produce high-fidelity approximations that are very close to the full-order frequency response. The quadrature-based reduced models more closely mimic the behavior of the intrusive (relative error) FWBT reduced models than the (absolute error) BT reduced models, particularly as the number of quadrature nodes N increases. The relative \mathcal{H}_∞ errors are recorded in Table 4.2; the BT reduced models provide the lowest error in this metric, but all the other computed reduced models perform very closely. As was the case in Section 4.3.5, the relative \mathcal{H}_∞ errors due to the QuadFWBT reduced models increase from $N = 100$ to $N = 200$. In contrast to the methods tested in Section 4.3.5, the data required for (self-weighted) FWBT *can* be obtained non-intrusively. Indeed, so long as one

can evaluate the underlying transfer function \mathbf{G}_{lo} , then the transfer function of the inverse system $\mathbf{G}_{\text{lo}}^{-1}$ can also be inferred from it.

4.5 Data-driven balancing for second-order systems

The mathematical modeling of, e.g., mechanical or electrical structures [226, Ch. 1], typically results in linear dynamical systems described by *second-order* differential equations of the form

$$\mathcal{G}_{\text{so}} : \begin{cases} \mathbf{M}_{\text{so}}\ddot{\mathbf{p}}(t) + \mathbf{D}_{\text{so}}\dot{\mathbf{p}}(t) + \mathbf{K}_{\text{so}}\mathbf{p}(t) = \mathbf{B}_{\text{u}}\mathbf{u}_{\text{so}}(t), \\ \mathbf{y}_{\text{so}}(t) = \mathbf{C}_{\text{p}}\mathbf{p}(t) + \mathbf{C}_{\text{v}}\dot{\mathbf{p}}(t), \end{cases} \quad (4.72)$$

where $\mathbf{M}_{\text{so}}, \mathbf{D}_{\text{so}}, \mathbf{K}_{\text{so}} \in \mathbb{R}^{n_{\text{so}} \times n_{\text{so}}}$ describe the differential equations, $\mathbf{B}_{\text{u}} \in \mathbb{R}^{n_{\text{so}} \times m}$ describes the input term, and $\mathbf{C}_{\text{p}}, \mathbf{C}_{\text{v}} \in \mathbb{R}^{p \times n_{\text{so}}}$ are the position- and velocity-output terms, respectively. The three matrices $\mathbf{M}_{\text{so}}, \mathbf{D}_{\text{so}}, \mathbf{K}_{\text{so}}$ are typically referred to as *mass*, *damping*, and *stiffness* terms. Overall, the resulting differential equations model the behavior of the system over time, where the external inputs $\mathbf{u}_{\text{so}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m$ influence the internal states $\mathbf{p}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n_{\text{so}}}$ and the quantities of interest $\mathbf{y}_{\text{so}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ are observed as outputs. The matrices $\mathbf{M}_{\text{so}}, \mathbf{D}_{\text{so}}, \mathbf{K}_{\text{so}}, \mathbf{B}_{\text{u}}, \mathbf{C}_{\text{p}}$, and \mathbf{C}_{v} form a state-space realization of (4.72), and we use the shorthand

$$\mathcal{G}_{\text{so}} = (\mathbf{M}_{\text{so}}, \mathbf{D}_{\text{so}}, \mathbf{K}_{\text{so}}, \mathbf{B}_{\text{u}}, \mathbf{C}_{\text{p}}, \mathbf{C}_{\text{v}})$$

when referring to a second-order system with the given realization (4.72). Throughout this section, we assume the mass \mathbf{M}_{so} to be nonsingular, that the initial conditions satisfy $\dot{\mathbf{p}}(0) = \mathbf{p}(0) = \mathbf{0}_n$, and the dynamics of (4.72) to be asymptotically stable; see Definition 4.20. The input-to-output mapping of (4.72) is equivalently described in the frequency domain via the corresponding $2n_{\text{so}}$ -degree rational transfer function

$$\mathbf{G}_{\text{so}}(s) = (s\mathbf{C}_{\text{v}} + \mathbf{C}_{\text{p}}) (s^2\mathbf{M}_{\text{so}} + s\mathbf{D}_{\text{so}} + \mathbf{K}_{\text{so}})^{-1} \mathbf{B}_{\text{u}}, \quad s \in \mathbb{C}. \quad (4.73)$$

One can derive (4.73) by applying the Laplace transform to (4.72) and solving for the input-to-output relationship. Systems with an internal structure of the form (4.72) arise in a variety of applications, ranging from the vibrational analysis of mechanical and acoustical structures [9, 226, 241] to the modeling of electro mechanical [40, 42] and particle systems.

In this section, we consider the problem of identifying *structured* representations of second-order system dynamics (4.72) directly from input-output data. In other words, our goal is to learn a reduced-order model of the form

$$\tilde{\mathcal{G}}_{\text{so}} : \begin{cases} \tilde{\mathbf{M}}_{\text{so}}\ddot{\tilde{\mathbf{p}}}(t) + \tilde{\mathbf{D}}_{\text{so}}\dot{\tilde{\mathbf{p}}}(t) + \tilde{\mathbf{K}}_{\text{so}}\tilde{\mathbf{p}}(t) = \tilde{\mathbf{B}}_{\text{u}}\mathbf{u}(t), \\ \tilde{\mathbf{y}}_{\text{so}}(t) = \tilde{\mathbf{C}}_{\text{p}}\tilde{\mathbf{p}}(t) + \tilde{\mathbf{C}}_{\text{v}}\dot{\tilde{\mathbf{p}}}(t), \end{cases} \quad (4.74)$$

from, e.g., evaluations of the transfer function (4.73), where $\tilde{\mathbf{M}}_{\text{so}}, \tilde{\mathbf{D}}_{\text{so}}, \tilde{\mathbf{K}}_{\text{so}} \in \mathbb{R}^{r_{\text{so}} \times r_{\text{so}}}$, $\tilde{\mathbf{B}}_{\text{u}} \in \mathbb{R}^{r_{\text{so}} \times m}$, $\tilde{\mathbf{C}}_{\text{p}}, \tilde{\mathbf{C}}_{\text{v}} \in \mathbb{R}^{p \times r_{\text{so}}}$, $\tilde{\mathbf{p}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{r_{\text{so}}}$, and $\tilde{\mathbf{y}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ for $1 \leq r_{\text{so}} < n_{\text{so}}$. The transfer

function $\widetilde{\mathbf{H}}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ of the learned model (4.74) has the same structure as (4.72) and should fit the given data in an appropriate measure, e.g., as an interpolant or in a least-squares sense. One can always rewrite the second-order system (4.72) as an equivalent $n = 2n_{\text{so}}$ -dimensional *first-order* system (2.25). This is accomplished by introducing the first-order state vector $\mathbf{x}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{2n_{\text{so}}}$ defined by $\mathbf{x}^\top \stackrel{\text{def}}{=} [\mathbf{p}^\top \quad \dot{\mathbf{p}}^\top]$, and re-organizing the second-order dynamics in (4.72) accordingly. The resulting $n = 2n_{\text{so}}$ -dimensional system (2.25), in the so-called *first-order companion form*, is defined by the realization parameters:

$$\mathbf{E} = \begin{bmatrix} \mathbf{I}_{n_{\text{so}}} & \mathbf{0}_{n_{\text{so}} \times n_{\text{so}}} \\ \mathbf{0}_{n_{\text{so}} \times n_{\text{so}}} & \mathbf{M}_{\text{so}} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{0}_{n_{\text{so}} \times n_{\text{so}}} & \mathbf{I}_{n_{\text{so}}} \\ -\mathbf{K}_{\text{so}} & -\mathbf{D}_{\text{so}} \end{bmatrix}, \quad \mathbf{C} = [\mathbf{C}_p \quad \mathbf{C}_v], \quad \mathbf{B} = \begin{bmatrix} \mathbf{0}_{n_{\text{so}} \times m} \\ \mathbf{B}_u \end{bmatrix}. \quad (4.75)$$

The input-to-output behavior of the n_{so} -dimension second-order system (4.72) and the $2n_{\text{so}}$ -dimensional first-order system described by (4.75) are identical. Therefore, it is in principle always possible to apply any data-driven technique for the reduced-order modeling of (generic) first-order systems (2.25), e.g., those highlighted in Section 4.1.1, to realize a first-order model of the underlying dynamics in (4.72). However, it is usually desirable to compute reduced-order surrogates that *preserve* the underlying second-order structure. While it is always possible to lift a second-order system to the first-order companion form, the converse is not true; see [143, Section III]. Moreover, twice the number of degrees of freedom is required to realize a first-order system that matches the behavior of the underlying second-order dynamics. Because of these facts, structured surrogates typically produce more accurate approximations compared to unstructured ones having the same number of internal degrees of freedom. Structured surrogates can also be plugged directly into available computational tools, e.g., solvers or optimizers, developed for structured models, and the system quantities in (4.74) derived from the modeling procedure may provide valuable physical insight into the underlying problem. We refer the reader to [196, 226] for a comparison among methods for structured and non-structured surrogate modeling in the setting of (4.72).

In recent years, a variety of methods developed for the intrusive *and* non-intrusive (data-driven) reduced modeling of generic *first-order* linear systems of the form (2.55) have been extended to the setting of (4.72). In the intrusive realm, notable examples include those based on generalizations of classical balanced truncation model reduction [36, 51, 143, 181] and transfer function interpolation or moment matching [11, 12, 20, 22, 234]. Balancing-based methods are particularly desirable because of their interpretation as truncating states associated with small reachability and observability energies. Moreover, for second-order dynamics, the free-velocity balanced truncation of [143] and position-velocity balanced truncation [181] preserve asymptotic stability for *symmetric* second-order systems. For the data-driven reduced modeling of systems (4.72), extensions of the interpolatory Loewner framework [139, 172], rational vector-fitting [65, 100, 227], operator inference [167, 179, 202], and methods based on barycentric rational forms [90] have been developed.

Continuing with the central focus of this chapter, we develop here an extension of the QuadBT framework of [89] for the structured surrogate modeling of second-order systems (4.72). In its original formulation, QuadBT is only applicable to *first-order* dynamical systems (2.25),

and so a tailored extension suitable for second-order dynamics is needed. Specifically, our method is a data-driven reformulation of the position-velocity balanced truncation (**sopvBT**) for second-order systems (4.72) proposed by Reis and Stykel [181]. In Theorem 4.24, we show how to derive the deterministic quantities in **sopvBT** from frequency-response data; namely, evaluations of the transfer function (4.73) and its position- and velocity-output subsystems. The proposed method is applicable to any system that satisfies a generalized proportional damping hypothesis. The resulting data-based reduced models are nearly indistinguishable from those computed intrusively by **sopvBT**, as illustrated by the numerical experiments in Section 4.5.4.

4.5.1 Second-order linear systems theory

In this section, we introduce the general damping model that we consider, as well as review the essential systems theory details of balanced truncation model reduction for second-order systems.

Generalized proportional damping model.

While mass and stiffness in (4.72) are typically given as constant matrices, e.g., those resulting from the discretization of a PDE, the modeling of damping, i.e., the dissipation and conservation of energy in the system, can be significantly more complex. Moreover, mathematical models of certain problems, such as those in vibro-acoustics that involve structural dynamics, acoustic wave propagation, and frequency-dependent material properties, only exist in the frequency (Laplace) domain, and are described by systems of frequency-dependent algebraic equations

$$\mathcal{G}_{\text{so}}^f : \begin{cases} (s^2 \mathbf{M}_{\text{so}} + s \mathbf{D}_{\text{so}}(s) + \mathbf{K}_{\text{so}}) \mathbf{P}(s) = \mathbf{B}_u \mathbf{U}(s), \\ \mathbf{Y}_{\text{so}}(s) = \mathbf{C}_p \mathbf{P}(s) + \mathbf{C}_v \mathbf{P}(s), \end{cases} \quad (4.76)$$

where $\mathbf{P}: \mathbb{C} \rightarrow \mathbb{C}^{n_{\text{so}}}$, $\mathbf{U}: \mathbb{C} \rightarrow \mathbb{C}^m$, and $\mathbf{Y}: \mathbb{C} \rightarrow \mathbb{C}^p$ are the Laplace transformations of the state, input, and output vectors from (4.72), and the internal damping $\mathbf{D}_{\text{so}}: \mathbb{C} \rightarrow \mathbb{C}^{n_{\text{so}} \times n_{\text{so}}}$ is frequency-dependent. See, for example [9, Section 2], [163] for a more detailed discussion of the different types of internal damping that appear in the modeling of second-order dynamical systems. We note that in the case where \mathbf{D}_{so} is constant, the time- and frequency-domain systems \mathcal{G}_{so} in (4.72) and $\mathcal{G}_{\text{so}}^f$ in (4.76) are in fact equivalent formulations of the same system under the Laplace transformation. In this work, we allow the damping term to be frequency-dependent and consider a generalized form of the proportional damping model:

$$\mathbf{D}_{\text{so}}(s) = f(s) \mathbf{M}_{\text{so}} + g(s) \mathbf{K}_{\text{so}}. \quad (4.77)$$

That is, we require \mathbf{D}_{so} to be a linear combination of the mass and stiffness matrices, where the weights $f: \mathbb{C} \rightarrow \mathbb{C}$ and $g: \mathbb{C} \rightarrow \mathbb{C}$ are scalar complex-valued frequency-dependent

functions. This framework covers, in particular, two classical damping models as described in the following.

1. **Rayleigh damping**, also known as proportional damping, is given via the constant linear combination of mass and stiffness matrices

$$\mathbf{D}_{\text{so}}(s) = \alpha \mathbf{M}_{\text{so}} + \beta \mathbf{K}_{\text{so}}. \quad (4.78)$$

This is covered in the proposed damping model (4.77) by setting $f(s) = \alpha \geq 0$ and $g(s) = \beta \geq 0$, with $\alpha, \beta \in \mathbb{R}$. The parameters α and β allow for choosing the frequency range in which the structure is damped, as well as the intensity of the damping.

2. **Structural damping**, also known as hysteretic damping, describes a constant damping effect over the full frequency range via

$$\mathbf{D}_{\text{so}}(s) = \dot{i} \frac{\eta}{s} \mathbf{K}_{\text{so}}, \quad (4.79)$$

where $\eta \in \mathbb{C}$ is a (material-dependent) structural loss factor. In the generalized damping model (4.77), this is given by $f(s) = 0$ and $g(s) = \dot{i} \frac{\eta}{s}$.

Second-order systems.

Here, we review the necessary facts and definitions for second-order systems.

Definition 4.19 (Symmetric second-order systems [181, Section 3.1]). We call the realization of the system *symmetric* if it satisfies

$$\mathbf{M}_{\text{so}} = \mathbf{M}_{\text{so}}^\top, \quad \mathbf{D}_{\text{so}} = \mathbf{D}_{\text{so}}^\top, \quad \mathbf{K}_{\text{so}} = \mathbf{K}_{\text{so}}^\top, \quad \mathbf{B}_{\text{u}} = \mathbf{C}_{\text{p}}^\top, \quad \text{and} \quad \mathbf{C}_{\text{v}} = \mathbf{0}_{p \times n_{\text{so}}}. \quad (4.80)$$

◇

Definition 4.20 (Asymptotic stability of a second-order system [181, Section 2]). The system \mathcal{G}_{so} in (4.72) is said to be *asymptotically stable* if all of the eigenvalues of the matrix pencil $s^2 \mathbf{M}_{\text{so}} + s \mathbf{D}_{\text{so}} + \mathbf{K}_{\text{so}}$, i.e., all values $s \in \mathbb{C}$ for which $\det(s^2 \mathbf{M}_{\text{so}} + s \mathbf{D}_{\text{so}} + \mathbf{K}_{\text{so}}) = 0$ have negative real parts. The eigenvalues of $s^2 \mathbf{M}_{\text{so}} + s \mathbf{D}_{\text{so}} + \mathbf{K}_{\text{so}}$ are called the *poles* of the system (4.72). ◇

In many applications, such as the case of mechanical systems, the mass, damping, and stiffness matrices \mathbf{M}_{so} , \mathbf{D}_{so} , and \mathbf{K}_{so} are further symmetric positive-definite. If a system (4.72) is symmetric with positive-definite mass, damping, and stiffness matrices, then it is automatically asymptotically stable.

Definition 4.21 (Reachability and observability of a second-order system [181, Section 2]). We say the system \mathcal{G}_{so} in (4.72) is *reachable* if $\text{rank} [s^2 \mathbf{M}_{\text{so}} + s \mathbf{D}_{\text{so}} + \mathbf{K}_{\text{so}}, \mathbf{B}_{\text{u}}] = n_{\text{so}}$ for all $s \in \mathbb{C}$. We say the system \mathcal{G}_{so} is *observable* if $\text{rank} [s^2 \mathbf{M}_{\text{so}}^{\text{T}} + s \mathbf{D}_{\text{so}}^{\text{T}} + \mathbf{K}_{\text{so}}^{\text{T}}, s \mathbf{C}_{\text{v}}^{\text{T}} + \mathbf{C}_{\text{p}}^{\text{T}}] = n_{\text{so}}$ for all $s \in \mathbb{C}$. The realization (4.72) of the system \mathcal{G}_{so} is *minimal* if it is both reachable and observable. \diamond

It can be shown that the second-order system (4.72) is reachable and observable if and only if the equivalent first-order system (2.25) defined by (4.75) is reachable and observable according to Definitions 2.27 and 2.30.

Just like the linear case (2.25), for any pair of invertible matrices $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{n_{\text{so}} \times n_{\text{so}}}$, if we define a new state $\mathbf{q}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{n_{\text{so}}}$ via the coordinate transformation $\mathbf{q}(t) = \mathbf{T}^{-1} \mathbf{p}(t)$ for all time $t \geq 0$, then the resulting system

$$\tilde{\mathcal{G}}_{\text{so}} : \begin{cases} \mathbf{S} \mathbf{M}_{\text{so}} \mathbf{T} \ddot{\mathbf{q}}(t) + \mathbf{S} \mathbf{D}_{\text{so}} \mathbf{T} \dot{\mathbf{q}}(t) + \mathbf{S} \mathbf{K}_{\text{so}} \mathbf{T} \mathbf{q}(t) = \mathbf{S} \mathbf{B}_{\text{u}} \mathbf{u}(t), \\ \mathbf{y}_{\text{so}}(t) = \mathbf{C}_{\text{p}} \mathbf{T} \mathbf{q}(t) + \mathbf{C}_{\text{v}} \mathbf{T} \dot{\mathbf{q}}(t) \end{cases} \quad (4.81)$$

is *equivalent* to \mathcal{G}_{so} in (4.72) in the sense that the input-to-output mappings are the same. With regard to the *intrusive* construction of second-order surrogate models (4.74), as with the first-order case, we consider reduced models computed by projection. In fact, this projection will underpin the data-driven method that we derive in this section. Given a pair of model reduction bases $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n_{\text{so}} \times r_{\text{so}}}$, the reduced-order model (4.74) constructed via *projection* is given by

$$\begin{aligned} \tilde{\mathbf{M}}_{\text{so}} &= \mathbf{W}^{\text{T}} \mathbf{M}_{\text{so}} \mathbf{V}, & \tilde{\mathbf{D}}_{\text{so}} &= \mathbf{W}^{\text{T}} \mathbf{D}_{\text{so}} \mathbf{V}, & \tilde{\mathbf{K}}_{\text{so}} &= \mathbf{W}^{\text{T}} \mathbf{K}_{\text{so}} \mathbf{V}, & \tilde{\mathbf{B}}_{\text{u}} &= \mathbf{W}^{\text{T}} \mathbf{B}_{\text{u}}, \\ \tilde{\mathbf{C}}_{\text{p}} &= \mathbf{C}_{\text{p}} \mathbf{V}, & \tilde{\mathbf{C}}_{\text{v}} &= \mathbf{C}_{\text{v}} \mathbf{V}. \end{aligned} \quad (4.82)$$

This projection can be derived just as the linear case was in Section 2.4.

4.5.2 Second-order balanced truncation

There have been multiple attempts to generalize the ideas of classical balanced truncation to second-order dynamical systems [51, 143, 181, 226]. The unifying feature of the proposed methods is the reformulation of the second-order system in (4.72) as the equivalent order- $2n_{\text{so}}$ first-order linear system (2.25) given by (4.75). Suppose that the second-order system (4.72) is asymptotically stable and minimal according to Definitions 4.20 and 4.21. Then, the equivalent system defined by (4.75) will be asymptotically stable and minimal as well, so the first-order Gramians (2.39) and (2.40) are well defined. Let the first-order system Gramians be partitioned according to the block structure in the first-order companion form (4.75) as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{\text{p}} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^{\text{T}} & \mathbf{P}_{\text{v}} \end{bmatrix}, \quad \mathbf{E}^{\text{H}} \mathbf{Q}_{\text{lo}} \mathbf{E} = \begin{bmatrix} \mathbf{Q}_{\text{p}} & \mathbf{Q}_{12} \mathbf{M}_{\text{so}} \\ \mathbf{M}_{\text{so}}^{\text{H}} \mathbf{Q}_{12}^{\text{H}} & \mathbf{M}_{\text{so}}^{\text{H}} \mathbf{Q}_{\text{v}} \mathbf{M}_{\text{so}} \end{bmatrix}. \quad (4.83)$$

The submatrices $\mathbf{P}_p, \mathbf{Q}_p \in \mathbb{C}^{n \times n}$ are called the *position-reachability and -observability Gramians*, while $\mathbf{P}_v, \mathbf{M}_{so}^H \mathbf{Q}_v \mathbf{M}_{so} \in \mathbb{C}^{n \times n}$ are defined as the *velocity-reachability and -observability Gramians*; these names are due to the Gramians' correspondence with the position \mathbf{p} and velocity $\dot{\mathbf{p}}$ states of the underlying second-order dynamics (4.72). Because \mathbf{P} and \mathbf{Q}_{lo} are symmetric positive semi-definite, so too are their diagonal submatrices. The Gramians \mathbf{P}_p and $\mathbf{M}_{so}^H \mathbf{Q}_v \mathbf{M}_{so}$ can alternatively be formulated as contour integrals in terms of the input-to-state and state-to-output mappings of the system (4.72) when $\mathbf{C}_v = \mathbf{0}_{p \times n_{so}}$:

$$\mathbf{P}_p = \frac{1}{2\pi} \int_{-\infty}^{\infty} (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so} + \mathbf{K}_{so})^{-1} \mathbf{B}_u \mathbf{B}_u^H (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so} + \mathbf{K}_{so})^{-H} d\omega, \quad (4.84)$$

$$\mathbf{Q}_v = \frac{1}{2\pi} \int_{-\infty}^{\infty} (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so} + \mathbf{K}_{so})^{-H} \mathbf{C}_p^H \mathbf{C}_p (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so} + \mathbf{K}_{so})^{-1} d\omega, \quad (4.85)$$

see, e.g. [175, Prop. 2.1], [45, Section 4.4]. The proposed balanced truncation methods in [143, 181] rely on balancing selected combinations of the second-order reachability and observability Gramians in (4.83), and subsequently truncating. This can be achieved simultaneously by using an appropriate generalization of the square-root algorithm for first-order BT [124, 209]. In particular, the so-called *free-velocity balanced truncation* of Meyer and Srinivasan [143] and *position-velocity balanced truncation* of Reis and Stykel [181] correspond to the simultaneous diagonalization of $\mathbf{P}_p, \mathbf{Q}_p$, and $\mathbf{P}_p, \mathbf{M}_{so}^H \mathbf{Q}_v \mathbf{M}_{so}$, respectively. We also mention the work [51], which can be viewed as a projection technique for the first-order system described by (4.75) combined with a recovery of the second-order structure. For a more detailed overview of balanced truncation model reduction for second-order systems, we refer the reader to [226, Section 3.4.3], [36, 196].

The primary consideration of this work is the position-velocity balanced truncation for second-order systems (*sopvBT*) from [181]. The main steps of the method are summarized in Algorithm 4.5.1. If the system (4.72) being approximated is symmetric according to Definition 4.19 with positive-definite mass, damping, and stiffness matrices, asymptotic stability is preserved by *sopvBT* [181, Corollary 3.2]. In fact, $\mathbf{P}_p = \mathbf{Q}_v$ in the symmetric case [181, Theorem 3.1]. Note also that Algorithm 4.5.1 can still be applied if either the position- or velocity-output terms in (4.72) are zero.

Remark 4.22. In the more complex case of frequency-dependent damping such as (4.76), one can still formulate structured Gramians via the integral representations in (4.84) and (4.85) by replacing the constant \mathbf{D}_{so} with $\mathbf{D}_{so}(s)$. Then, if the integrals converge, the structured Gramians are approximated by numerical quadrature rules, and Algorithm 4.5.1 can be applied. See, for instance, the work on structure-preserving model reduction of integral-differential equations [45] for further discussion. \diamond

Remark 4.23. The reason only *sopvBT* is considered here is due to the integral representations of the second-order system Gramians. Those in (4.84) and (4.85) are “natural” in a

Algorithm 4.5.1: Second-order position-velocity balanced truncation (sopvBT) [181].

Input: Second-order system (4.72) $\mathcal{G}_{\text{so}} = (\mathbf{M}_{\text{so}}, \mathbf{D}_{\text{so}}, \mathbf{K}_{\text{so}}, \mathbf{B}_{\text{u}}, \mathbf{C}_{\text{p}}, \mathbf{C}_{\text{v}})$, order r_{so}
 $(1 \leq r_{\text{so}} < n_{\text{so}})$.

Output: $\tilde{\mathcal{G}}_{\text{so}} = (\tilde{\mathbf{M}}_{\text{so}}, \tilde{\mathbf{D}}_{\text{so}}, \tilde{\mathbf{K}}_{\text{so}}, \tilde{\mathbf{B}}_{\text{u}}, \tilde{\mathbf{C}}_{\text{p}}, \tilde{\mathbf{C}}_{\text{v}})$ —state-space matrices of (4.74)

- 1 Compute Cholesky factorizations $\mathbf{P} = \mathbf{R}\mathbf{R}^{\text{H}}$ and $\mathbf{Q}_{\text{lo}} = \mathbf{L}_{\text{lo}}\mathbf{L}_{\text{lo}}^{\text{H}}$ using the first-order companion form (4.75), where the factors are partitioned as

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{\text{p}} \\ \star \end{bmatrix} \quad \text{and} \quad \mathbf{L}_{\text{lo}} = \begin{bmatrix} \star \\ \mathbf{L}_{\text{v}} \end{bmatrix}$$

for $\mathbf{R}_{\text{p}}, \mathbf{L}_{\text{v}} \in \mathbb{C}^{n_{\text{so}} \times n_{\text{so}}}$.

- 2 Compute the singular value decomposition:

$$\mathbf{L}_{\text{v}}^{\text{H}} \mathbf{M}_{\text{so}} \mathbf{R}_{\text{p}} = [\mathbf{U}_1 \quad \mathbf{U}_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1^{\text{H}} \\ \mathbf{Y}_2^{\text{H}} \end{bmatrix},$$

for $\Sigma_1 \in \mathbb{R}^{r_{\text{so}} \times r_{\text{so}}}$, $\Sigma_2 \in \mathbb{R}^{(n_{\text{so}}-r_{\text{so}}) \times (n_{\text{so}}-r_{\text{so}})}$ diagonal, and $\mathbf{U}_1, \mathbf{U}_2, \mathbf{Y}_1, \mathbf{Y}_2$ partitioned conformally.

- 3 Compute the reduced-order model (4.74) by projecting the full-order system matrices:

$$\begin{aligned} \tilde{\mathbf{M}}_{\text{so}} &= \mathbf{W}^{\text{H}} \mathbf{M}_{\text{so}} \mathbf{V} &= \mathbf{I}_{r_{\text{so}}}, \\ \tilde{\mathbf{D}}_{\text{so}} &= \mathbf{W}^{\text{H}} \mathbf{D}_{\text{so}} \mathbf{V} &= \Sigma_1^{-1/2} \mathbf{U}_1^{\text{H}} (\mathbf{L}_{\text{v}}^{\text{H}} \mathbf{D}_{\text{so}} \mathbf{R}_{\text{p}}) \mathbf{Y}_1 \Sigma_1^{-1/2}, \\ \tilde{\mathbf{K}}_{\text{so}} &= \mathbf{W}^{\text{H}} \mathbf{K}_{\text{so}} \mathbf{V} &= \Sigma_1^{-1/2} \mathbf{U}_1^{\text{H}} (\mathbf{L}_{\text{v}}^{\text{H}} \mathbf{K}_{\text{so}} \mathbf{R}_{\text{p}}) \mathbf{Y}_1 \Sigma_1^{-1/2}, \\ \tilde{\mathbf{B}}_{\text{u}} &= \mathbf{W}^{\text{H}} \mathbf{B}_{\text{u}} &= \Sigma_1^{-1/2} \mathbf{U}_1^{\text{H}} (\mathbf{L}_{\text{v}}^{\text{H}} \mathbf{B}_{\text{u}}), \\ \tilde{\mathbf{C}}_{\text{p}} &= \mathbf{C}_{\text{p}} \mathbf{V} &= (\mathbf{C}_{\text{p}} \mathbf{R}_{\text{p}}) \mathbf{V}_1 \Sigma_1^{-1/2}, \\ \tilde{\mathbf{C}}_{\text{v}} &= \mathbf{C}_{\text{v}} \mathbf{V} &= (\mathbf{C}_{\text{v}} \mathbf{R}_{\text{p}}) \mathbf{V}_1 \Sigma_1^{-1/2}. \end{aligned}$$

sense, given that they are defined in terms of the input-to-state and state-to-output maps of the system (4.72). These representations facilitate a data-driven reformulation of the method expressed in terms of transfer function data (4.73). Even though one can also write down integral formulations of \mathbf{P}_{v} and \mathbf{Q}_{p} , the structure of these integrals is not as amenable to the tools used in Section 4.5.3 to derive a data-based formulation of sopvBT. In particular, it is not clear what the resulting “data” correspond to. \diamond

4.5.3 Quadrature-based position-velocity balanced truncation

Algorithm 4.5.1 is intrusive, in the sense that it requires an explicit state-space model of the second-order system (4.72) in order to compute the Cholesky factors $\mathbf{R}_p \in \mathbb{C}^{n_{so} \times n_{so}}$, $\mathbf{L}_v \in \mathbb{C}^{n_{so} \times n_{so}}$, and a reduced-order system (4.74) by projection (4.82). Here, we present a data-driven reformulation of `sopvBT` in Algorithm 4.5.1. Our only assumption is that we are able to sample the full-order transfer function (4.73) at a prescribed range of frequencies along the imaginary axis, as well as the coefficient functions in (4.77), which we assume are known *a priori* based on empirical evidence. The resulting method, which we call *quadrature-based sopvBT* (`soQuadpvBT`), is a generalization of the `QuadBT` framework for *first-order* systems from [89]. This allows for the computation of (approximate) `sopvBT`-based second-order surrogate models (4.74) directly from state-invariant frequency-response data.

Deterministic quantities from data.

We begin with the following observation: The matrices \mathbf{U}_1 , \mathbf{Y}_1 , and $\mathbf{\Sigma}_1$ are derived from the truncated singular value decomposition of $\mathbf{L}_v^H \mathbf{M} \mathbf{R}_p$. Under the assumption that $\mathbf{D}_{so}(s) = f(s)\mathbf{M}_{so} + g(s)\mathbf{K}_{so}$, it follows that the balanced truncation reduced-order model produced by Algorithm 4.5.1 is fully specified by the five quantities

$$\mathbf{L}_v^H \mathbf{M}_{so} \mathbf{R}_p, \quad \mathbf{L}_v^H \mathbf{K}_{so} \mathbf{R}_p, \quad \mathbf{L}_v^H \mathbf{B}_u, \quad \mathbf{C}_p \mathbf{R}_p \quad \text{and} \quad \mathbf{C}_v \mathbf{R}_p. \quad (4.86)$$

Note that this is highly similar to the first-order setting of Sections 4.2–4.4. This observation, along with the integral formulations of the position-reachability and velocity-observability Gramians $\mathbf{P}_p \in \mathbb{R}^{n_{so} \times n_{so}}$, $\mathbf{M}_{so}^T \mathbf{Q}_v \mathbf{M}_{so} \in \mathbb{R}^{n_{so} \times n_{so}}$ in (4.84) and (4.85), suggests a natural extension of `QuadBT` for the second-order system (4.72). Specifically, we will use implicit numerical quadrature rules to derive low-rank approximations to the exact factors \mathbf{R}_p and \mathbf{L}_v , and ultimately show how to realize the quadrature-based approximations to the intrusive quantities in (4.86) from *transfer function data*.

Recall the integral representations of the Gramians (4.84) and (4.85), but with the frequency-dependent $\mathbf{D}_{so}(s)$ in place of \mathbf{D}_{so} . For compactness of the presentation, we denote the second-order matrix pencil by $\varphi: \mathbb{C} \rightarrow \mathbb{C}^{n_{so} \times n_{so}}$, i.e.,

$$\varphi(s) \stackrel{\text{def}}{=} (s^2 \mathbf{M}_{so} + s \mathbf{D}_{so}(s) + \mathbf{K}_{so}), \quad s \in \mathbb{C}.$$

Consider a numerical quadrature rule defined by the nodes $i\zeta_1, \dots, i\zeta_J \in i\mathbb{R}$ and weights $\varrho_1^2, \dots, \varrho_J^2 \in \mathbb{R}$. Applying this quadrature rule to \mathbf{P}_p in (4.84) reveals the approximate factorization

$$\mathbf{P}_p \approx \sum_{j=1}^J \varrho_j^2 \varphi(i\zeta_j)^{-1} \mathbf{B}_u (\varphi(i\zeta_j)^{-1} \mathbf{B}_u)^H = \check{\mathbf{R}}_p \check{\mathbf{R}}_p^H,$$

where $\check{\mathbf{R}}_p \in \mathbb{C}^{n_{so} \times mJ}$ is defined as

$$\check{\mathbf{R}}_p \stackrel{\text{def}}{=} [\varrho_1 \varphi(i\zeta_1)^{-1} \mathbf{B}_u \quad \varrho_2 \varphi(i\zeta_2)^{-1} \mathbf{B}_u \quad \cdots \quad \varrho_J \varphi(i\zeta_J)^{-1} \mathbf{B}_u]. \quad (4.87)$$

A similar factorization can be derived readily from \mathbf{Q}_v in (4.85). However, this representation of the Gramian assumes that $\mathbf{C}_v = \mathbf{0}_{p \times n_{so}}$. To incorporate velocity outputs into quadrature-based factors, we derive an alternative expression for the integrand in (4.85). Because \mathbf{Q}_v is derived from the (2, 2)-block of $\mathbf{E}^H \mathbf{Q}_{lo} \mathbf{E}$ in (4.83), it holds that

$$\mathbf{Q}_v = \mathbf{J}^T \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbf{C} (i\omega \mathbf{E} - \mathbf{A}(i\omega))^{-1})^H \mathbf{C} (i\omega \mathbf{E} - \mathbf{A}(i\omega))^{-1} d\omega \right) \mathbf{J},$$

where $\mathbf{J}^T \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{0}_{n_{so} \times n_{so}} & \mathbf{I}_{n_{so}} \end{bmatrix} \in \mathbb{R}^{n_{so} \times 2n_{so}}$. The ω -dependence in \mathbf{A} stems from the underlying frequency-dependent damping that appears in (4.75). Using the formula for the inverse of a block 2×2 matrix [38, Sec. 2.17], as long as $\omega \neq 0$ we have

$$\begin{aligned} \mathbf{C} (i\omega \mathbf{E} - \mathbf{A})^{-1} \mathbf{J} &= [\mathbf{C}_p \quad \mathbf{C}_v] \begin{bmatrix} i\omega \mathbf{I}_{n_{so}} & -\mathbf{I}_{n_{so}} \\ \mathbf{K}_{so} & i\omega \mathbf{M}_{so} + \mathbf{D}_{so}(i\omega) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0}_{n_{so} \times n_{so}} \\ \mathbf{I}_{n_{so}} \end{bmatrix} \\ &= [\mathbf{C}_p \quad \mathbf{C}_v] \begin{bmatrix} \star & (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so}(i\omega) + \mathbf{K}_{so})^{-1} \\ \star & i\omega (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so}(i\omega) + \mathbf{K}_{so})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{n_{so} \times n_{so}} \\ \mathbf{I}_{n_{so}} \end{bmatrix} \\ &= (\mathbf{C}_p + i\omega \mathbf{C}_v) (-\omega^2 \mathbf{M}_{so} + i\omega \mathbf{D}_{so}(i\omega) + \mathbf{K}_{so})^{-1} \\ &= (\mathbf{C}_p + i\omega \mathbf{C}_v) \varphi(i\omega)^{-1}. \end{aligned}$$

And so, the (2, 2)-block of $(\mathbf{C} (i\omega \mathbf{E} - \mathbf{A}(i\omega))^{-1})^H \mathbf{C} (i\omega \mathbf{E} - \mathbf{A}(i\omega))^{-1}$ can be written as

$$((\mathbf{C}_p + i\omega \mathbf{C}_v) \varphi(i\omega)^{-1})^H (\mathbf{C}_p + i\omega \mathbf{C}_v) \varphi(i\omega)^{-1}, \quad (4.88)$$

for all $\omega \neq 0$. This allows us to incorporate velocity outputs into a quadrature-based factorization of \mathbf{Q}_v , so long as the corresponding quadrature nodes are all nonzero. With this representation (4.88) in hand, applying a numerical quadrature rule defined by the nodes $i\vartheta_1, \dots, i\vartheta_K \in i\mathbb{R} \setminus \{0\}$ and weights $\varphi_1^2, \dots, \varphi_K^2 \in \mathbb{R}$ to \mathbf{Q}_v with the integrand (4.88) produces the approximate factorization

$$\mathbf{Q}_v \approx \sum_{k=1}^K \varphi_k^2 ((\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \varphi(i\vartheta_k)^{-1})^H (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \varphi(i\vartheta_k)^{-1} = \check{\mathbf{L}}_v \check{\mathbf{L}}_v^H,$$

where $\check{\mathbf{L}}_v \in \mathbb{C}^{n_{so} \times pK}$ is defined according to

$$\check{\mathbf{L}}_v^H \stackrel{\text{def}}{=} \begin{bmatrix} \varphi_1 (\mathbf{C}_p + i\vartheta_1 \mathbf{C}_v) \varphi(i\vartheta_1)^{-1} \\ \varphi_2 (\mathbf{C}_p + i\vartheta_2 \mathbf{C}_v) \varphi(i\vartheta_2)^{-1} \\ \vdots \\ \varphi_K (\mathbf{C}_p + i\vartheta_K \mathbf{C}_v) \varphi(i\vartheta_K)^{-1} \end{bmatrix}. \quad (4.89)$$

We note that, in the case of $\mathbf{C}_v = \mathbf{0}_{p \times n_{so}}$, the quadrature-based factor (4.89) returns precisely what we would get by applying the same numerical quadrature rule to \mathbf{Q}_v as defined in (4.85), and factoring.

Replacing the exact Cholesky factors \mathbf{R}_p and \mathbf{L}_v from Algorithm 4.5.1 with the quadrature-based approximations (4.87) and (4.89) in forming the matrices (4.86) already yields a low-rank implementation of `sopvBT`. As we show next, the significant result of this low-rank formulation is that the quadrature-based approximations to the deterministic quantities in (4.86) can be computed non-intrusively from transfer function evaluations, as well as the weights $f(s)$ and $g(s)$ in (4.77). Before stating the result, we introduce the transfer functions $\mathbf{G}_p: \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ and $\mathbf{G}_v: \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ defined as

$$\mathbf{G}_p(s) \stackrel{\text{def}}{=} \mathbf{C}_p \boldsymbol{\varphi}(s)^{-1} \mathbf{B}_u \quad \text{and} \quad \mathbf{G}_v(s) = s \mathbf{C}_v \boldsymbol{\varphi}(s)^{-1} \mathbf{B}_u. \quad (4.90)$$

Note that these are each the transfer function of a second-order system (4.72) with purely position and velocity outputs, respectively. In the general case, we have that $\mathbf{G}_{\text{so}}(s) = \mathbf{G}_p(s) + s \mathbf{G}_v(s)$.

Theorem 4.24 (Position-velocity balanced truncation from data). Define the functions $d: \mathbb{C} \rightarrow \mathbb{C}$, $n: \mathbb{C} \rightarrow \mathbb{C}$, and $h: \mathbb{C} \rightarrow \mathbb{C}$ by

$$d(s) \stackrel{\text{def}}{=} 1 + sg(s), \quad n(s) \stackrel{\text{def}}{=} s^2 + sf(s), \quad h(s) \stackrel{\text{def}}{=} \frac{n(s)}{d(s)}, \quad (4.91)$$

where $f(s)$ and $g(s)$ are from (4.77). Suppose the left and right quadrature nodes $\dot{\imath}\zeta_1, \dots, \dot{\imath}\zeta_J$ and $\dot{\imath}\vartheta_1, \dots, \dot{\imath}\vartheta_K$ in (4.87) and (4.89) are such that

$$\vartheta_k, \zeta_j \neq 0, \quad h(\dot{\imath}\vartheta_k) \neq h(\dot{\imath}\zeta_j), \quad \text{and} \quad d(\dot{\imath}\vartheta_k), d(\dot{\imath}\zeta_j) \neq 0,$$

for all $j = 1, \dots, J$ and $k = 1, \dots, K$. Let the quadrature-based factors $\check{\mathbf{R}}_p$ and $\check{\mathbf{L}}_v$ be given as in (4.87) and (4.89). Define the matrices

$$\begin{aligned} \mathbb{M} &\stackrel{\text{def}}{=} \check{\mathbf{L}}_v^H \mathbf{M}_{\text{so}} \check{\mathbf{R}}_p \in \mathbb{C}^{pK \times mJ}, & \mathbb{K} &\stackrel{\text{def}}{=} \check{\mathbf{L}}_v^H \mathbf{K}_{\text{so}} \check{\mathbf{R}}_p \in \mathbb{C}^{pK \times mJ}, & \mathbb{B}_u &\stackrel{\text{def}}{=} \check{\mathbf{L}}_v^H \mathbf{B}_u \in \mathbb{C}^{pK \times m}, \\ \mathbb{C}_p &\stackrel{\text{def}}{=} \mathbf{C}_p \check{\mathbf{R}}_p \in \mathbb{C}^{p \times mJ}, & \mathbb{C}_v &\stackrel{\text{def}}{=} \mathbf{C}_v \check{\mathbf{R}}_p \in \mathbb{C}^{p \times mJ}. \end{aligned} \quad (4.92)$$

Then, for all $1 \leq k \leq K$ and $1 \leq j \leq J$, the matrices in (4.92) are given by

$$\mathbb{M}_{k,j} = -\varphi_k \varrho_j \frac{d(\dot{\imath}\zeta_j)^{-1} \mathbf{G}_{\text{so}}(\dot{\imath}\vartheta_k) - d(\dot{\imath}\vartheta_k)^{-1} (\mathbf{G}_p(\dot{\imath}\zeta_j) + (\vartheta_k/\zeta_j) \mathbf{G}_v(\dot{\imath}\zeta_j))}{h(\dot{\imath}\vartheta_k) - h(\dot{\imath}\zeta_j)}, \quad (4.93a)$$

$$\mathbb{K}_{k,j} = -\varphi_k \varrho_j \frac{h(\dot{\imath}\vartheta_k) d(\dot{\imath}\zeta_j)^{-1} \mathbf{G}_{\text{so}}(\dot{\imath}\vartheta_k) - h(\dot{\imath}\zeta_j) d(\dot{\imath}\vartheta_k)^{-1} (\mathbf{G}_p(\dot{\imath}\zeta_j) + (\vartheta_k/\zeta_j) \mathbf{G}_v(\dot{\imath}\zeta_j))}{h(\dot{\imath}\vartheta_k) - h(\dot{\imath}\zeta_j)}, \quad (4.93b)$$

$$(\mathbb{B}_u)_{k,:} = \varphi_k \mathbf{G}_{\text{so}}(\dot{\imath}\vartheta_k), \quad (\mathbb{C}_p)_{:,j} = \varrho_j \mathbf{G}_p(\dot{\imath}\zeta_j), \quad \text{and} \quad (\mathbb{C}_v)_{:,j} = \frac{\varrho_j}{\dot{\imath}\zeta_j} \mathbf{G}_v(\dot{\imath}\zeta_j), \quad (4.93c)$$

where \mathbf{G}_p and \mathbf{G}_v are defined as in (4.90). \diamond

Proof of Theorem 4.24. The formulae for \mathbb{B}_u , \mathbb{C}_p and \mathbb{C}_v in (4.93c) are a direct consequence of their construction in (4.92) and the definitions of the quadrature-based factors $\check{\mathbf{R}}_p$ and $\check{\mathbf{L}}_v$. Observe that

$$(\mathbb{B}_u)_{k,:} = \mathbf{I}_{k,p}^\top \mathbb{B}_u = \mathbf{I}_{k,p}^\top \check{\mathbf{L}}_v^H \mathbf{B}_u = \varphi_k (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \varphi(i\vartheta_k)^{-1} \mathbf{B}_u = \varphi_k \mathbf{G}_{so}(i\vartheta_k),$$

where $\mathbf{I}_{k,p}$ is defined according to (4.13). Similarly:

$$\begin{aligned} (\mathbb{C}_p)_{:,j} &= \mathbf{C}_p \check{\mathbf{R}}_p \mathbf{I}_{j,m} = \varrho_j \mathbf{C}_p \varphi(i\zeta_j)^{-1} \mathbf{B}_u = \varrho_j \mathbf{G}_p(i\zeta_j), \\ (\mathbb{C}_v)_{:,j} &= \mathbf{C}_v \check{\mathbf{R}}_p \mathbf{I}_{j,m} = \varrho_j \mathbf{C}_v \varphi(i\zeta_j)^{-1} \mathbf{B}_u = \frac{\varrho_j}{i\zeta_j} \mathbf{G}_v(i\zeta_j). \end{aligned}$$

For (4.93a) and (4.93b), first note that so long as $d(s) \neq 0$

$$\begin{aligned} \varphi(s)^{-1} &= (s^2 \mathbf{M}_{so} + s \mathbf{D}_{so}(s) + \mathbf{K}_{so})^{-1} \\ &= ((s^2 + sf(s)) \mathbf{M}_{so} + (1 + sg(s)) \mathbf{K}_{so})^{-1} \quad (\text{by (4.77)}) \\ &= \frac{1}{1 + sg(s)} \left(\frac{s^2 + sf(s)}{1 + sg(s)} \mathbf{M}_{so} + \mathbf{K}_{so} \right)^{-1} \\ &= \frac{1}{d(s)} (h(s) \mathbf{M}_{so} + \mathbf{K}_{so})^{-1} = \frac{1}{d(s)} \phi(h(s))^{-1}, \end{aligned}$$

where $\phi(s) \stackrel{\text{def}}{=} (s \mathbf{M}_{so} + \mathbf{K}_{so})$. To show (4.93a) and (4.93b), we use the resolvent identities in Lemma 2.7. Under the hypothesis that $h(i\vartheta_k) \neq h(i\zeta_j)$ for all k and j , applying the first resolvent identity in (2.16), we have that

$$\begin{aligned} \mathbb{M}_{k,j} &= \mathbf{I}_{k,p}^\top (\check{\mathbf{L}}_v^H \mathbf{M}_{so} \check{\mathbf{R}}_p) \mathbf{I}_{j,m} = \varphi_k \varrho_j (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \varphi(i\vartheta_k)^{-1} \mathbf{M}_{so} \varphi(i\zeta_j)^{-1} \mathbf{B}_u \\ &= \varphi_k \varrho_j (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \frac{\phi(h(i\vartheta_k)) \mathbf{M}_{so} \phi(h(i\zeta_j))}{d(i\vartheta_k) d(i\zeta_j)} \mathbf{B}_u \\ &= -\varphi_k \varrho_j \frac{(\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \phi(h(i\vartheta_k)) \mathbf{B}_u - (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \phi(h(i\zeta_j)) \mathbf{B}_u}{d(i\vartheta_k) d(i\zeta_j) (h(i\vartheta_k) - h(i\zeta_j))} \\ &= -\varphi_k \varrho_j \frac{d(i\zeta_j)^{-1} \mathbf{G}_{so}(i\vartheta_k) - d(i\vartheta_k)^{-1} (\mathbf{G}_p(i\zeta_j) + (\vartheta_k/\zeta_j) \mathbf{G}_v(i\zeta_j))}{h(i\vartheta_k) - h(i\zeta_j)}, \end{aligned}$$

where the last line follows from the relationship $\varphi(s)^{-1} = d(s)^{-1} \phi(h(s))^{-1}$, and the definitions \mathbf{G}_p and \mathbf{G}_v in (4.90). Likewise, applying the second resolvent identity in (2.17), we have that

$$\begin{aligned} \mathbb{K}_{k,j} &= \mathbf{I}_{k,p}^\top (\check{\mathbf{L}}_v^H \mathbf{K}_{so} \check{\mathbf{R}}_p) \mathbf{I}_{j,m} = \varphi_k \varrho_j (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \varphi(i\vartheta_k)^{-1} \mathbf{K}_{so} \varphi(i\zeta_j)^{-1} \mathbf{B}_u \\ &= \varphi_k \varrho_j (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \frac{\phi(h(i\vartheta_k)) \mathbf{M}_{so} \phi(h(i\zeta_j))}{d(i\vartheta_k) d(i\zeta_j)} \mathbf{B}_u \\ &= -\varphi_k \varrho_j \frac{h(i\vartheta_k) (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \phi(h(i\vartheta_k)) \mathbf{B}_u - h(i\zeta_j) (\mathbf{C}_p + i\vartheta_k \mathbf{C}_v) \phi(h(i\zeta_j)) \mathbf{B}_u}{d(i\vartheta_k) d(i\zeta_j) (h(i\vartheta_k) - h(i\zeta_j))} \\ &= -\varphi_k \varrho_j \frac{h(i\vartheta_k) d(i\zeta_j)^{-1} \mathbf{G}_{so}(i\vartheta_k) - h(i\zeta_j) d(i\vartheta_k)^{-1} (\mathbf{G}_p(i\zeta_j) + (\vartheta_k/\zeta_j) \mathbf{G}_v(i\zeta_j))}{h(i\vartheta_k) - h(i\zeta_j)}, \end{aligned}$$

thus proving (4.93b). \square

In the formulae (4.93a) and (4.93b), the quantity $\mathbf{G}_p(\dot{\zeta}_j) + (\vartheta_k/\zeta_j)\mathbf{G}_v(\dot{\zeta}_j)$ that appears looks contrived at first glance. However, this is an artifact of Theorem 4.24 being stated as generally as possible to include position and velocity outputs; both are not necessarily required. In any realistic application, it is usually the case that either \mathbf{C}_p or \mathbf{C}_v is identically zero. So, this quantity will resolve to either

$$\mathbf{G}_p(\dot{\zeta}_j) + (\vartheta_k/\zeta_j)\mathbf{G}_v(\dot{\zeta}_j) = \mathbf{G}_p(\dot{\zeta}_j) = \mathbf{G}_{so}(\dot{\zeta}_j) \quad \text{if } \mathbf{C}_v = \mathbf{0}_{p \times n_{so}},$$

which is the second-order transfer function (4.73), or, a re-scaling of the second-order transfer function:

$$\mathbf{G}_p(\dot{\zeta}_j) + (\vartheta_k/\zeta_j)\mathbf{G}_v(\dot{\zeta}_j) = (\vartheta_k/\zeta_j)\mathbf{G}_v(\dot{\zeta}_j) = (\vartheta_k/\zeta_j)\mathbf{G}_{so}(\dot{\zeta}_j) \quad \text{if } \mathbf{C}_p = \mathbf{0}_{p \times n_{so}}.$$

Replacing the intrusive quantities (4.86) in Algorithm 4.5.1 with the data-based approximations (4.92) results in a data-driven reformulation of `sopvBT`, that we call *quadrature-based sopvBT* (`soQuadpvBT`) and present it in Algorithm 4.5.2. Outside of empirical knowledge used to design the damping coefficient functions $f(s)$ and $g(s)$, the method is entirely non-intrusive. As in [89], the quadrature-based approximations to the Gramians are never formed; they are only invoked implicitly to derive the data-based approximations (4.92). While Algorithm 4.5.2 is formulated generally to include position and velocity outputs, as is the case with Theorem 4.24, both are not necessarily required; depending on the application of interest, at minimum one of these is required.

4.5.4 Numerical results

We assume the common setup for numerical experiments outlined in Section 4.3.4. For the approximation of the benchmark problems presented in this section, we compare three different approaches against the proposed `soQuadpvBT` introduced in Section 4.5.3.

`soLoewner` is the Loewner identification framework for second-order Rayleigh-damped systems from [172].

`QuadBT` is the data-driven Lyapunov BT for first-order systems discussed in Sections 4.2 and 4.3 from [89], which fits a first-order reduced model from transfer function data.

`sopvBT` is the *intrusive* position-velocity BT for second-order systems from [181] and discussed in Section 4.5.2.

The intrusive `sopvBT` is included as a point of comparison for the non-intrusive, data-driven approaches. `QuadBT` constructs a first-order linear approximation of the form (2.25), and

Algorithm 4.5.2: Quadrature-based `sopvBT` (`soQuadpvBT`).

Input: Mappings \mathbf{G}_p or \mathbf{G}_v , f and g in (4.77), left, right quadrature nodes and weights $\{\dot{i}\vartheta_k, \varphi_k\}_{k=1}^K$, $\{\dot{i}\zeta_j, \varrho_j\}_{j=1}^J$, and order r_{so} ($1 \leq r_{so} < n_{so}$).

Output: $\tilde{\mathcal{G}}_{so} = (\tilde{\mathbf{M}}_{so}, \tilde{\mathbf{D}}_{so}, \tilde{\mathbf{K}}_{so}, \tilde{\mathbf{B}}_u, \tilde{\mathbf{C}}_p, \tilde{\mathbf{C}}_v)$ —state-space matrices of (4.74)

- 1 Evaluate the mappings to obtain the data $\{\mathbf{G}_p(\dot{i}\zeta_j), \mathbf{G}_v(\dot{i}\zeta_j), f(\dot{i}\zeta_j), g(\dot{i}\zeta_j)\}_{j=1}^J$, $\{\mathbf{G}_p(\dot{i}\vartheta_k), \mathbf{G}_v(\dot{i}\vartheta_k), f(\dot{i}\vartheta_k), g(\dot{i}\vartheta_k)\}_{k=1}^K$, and construct the data matrices in (4.92) according to Theorem 4.24.
- 2 Compute the singular value decomposition:

$$\mathbb{M} = [\check{\mathbf{U}}_1 \quad \check{\mathbf{U}}_2] \begin{bmatrix} \check{\Sigma}_1 & \\ & \check{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \check{\mathbf{Y}}_1^H \\ \check{\mathbf{Y}}_2^H \end{bmatrix},$$

for $\check{\Sigma}_1 \in \mathbb{R}^{r_{so} \times r_{so}}$, $\check{\Sigma}_2 \in \mathbb{R}^{(pK-r_{so}) \times (mJ-r_{so})}$ diagonal, and $\check{\mathbf{U}}_1, \check{\mathbf{U}}_2, \check{\mathbf{Y}}_1, \check{\mathbf{Y}}_2$ partitioned conformally.

- 3 Compute the data-driven reduced-order model (4.74) according to:

$$\begin{aligned} \tilde{\mathbf{M}}_{so} &= \mathbf{I}_{r_{so}}, \\ \tilde{\mathbf{K}}_{so} &= \Sigma_1^{-1/2} \mathbf{U}_1^H (\mathbb{K}) \mathbf{Y}_1 \Sigma_1^{-1/2}, \\ \tilde{\mathbf{D}}_{so}(s) &= f(s) \mathbf{I}_{r_{so}} + g(s) \tilde{\mathbf{K}}_{so} \\ \tilde{\mathbf{B}}_u &= \Sigma_1^{-1/2} \mathbf{U}_1^H (\mathbb{B}_u), \\ \tilde{\mathbf{C}}_p &= (\mathbb{C}_p) \mathbf{V}_1 \Sigma_1^{-1/2}, \\ \tilde{\mathbf{C}}_v &= (\mathbb{C}_v) \mathbf{V}_1 \Sigma_1^{-1/2}. \end{aligned}$$

is included to compare data-driven approaches that respect the underlying second-order structure against those that do not. The intrusive comparison `sopvBT` is implemented using the software package `MORLAB` version 6.0 [37].

For the presentation of the numerical results, we use the following error measures. To visibly compare different approximations, we plot the magnitude of the second-order transfer function (4.72) at discrete points. We also use the pointwise relative approximation errors of the transfer functions as

$$\text{relerr}(\dot{i}\omega_k) = \frac{\|\mathbf{G}_{so}(\dot{i}\omega_k) - \tilde{\mathbf{G}}_{so}(\dot{i}\omega_k)\|_2}{\|\mathbf{G}_{so}(\dot{i}\omega_k)\|_2}, \quad (4.94)$$

for frequencies $\omega_k \in \Omega$ in plots alongside the magnitude of the transfer functions, where Ω is a collection of discrete points in an interval $[\omega_{\min}, \omega_{\max}] \subset \mathbb{R}_{\geq 0}$. The specific choice of Ω varies from example to example. Additionally, we score the performance of the proposed

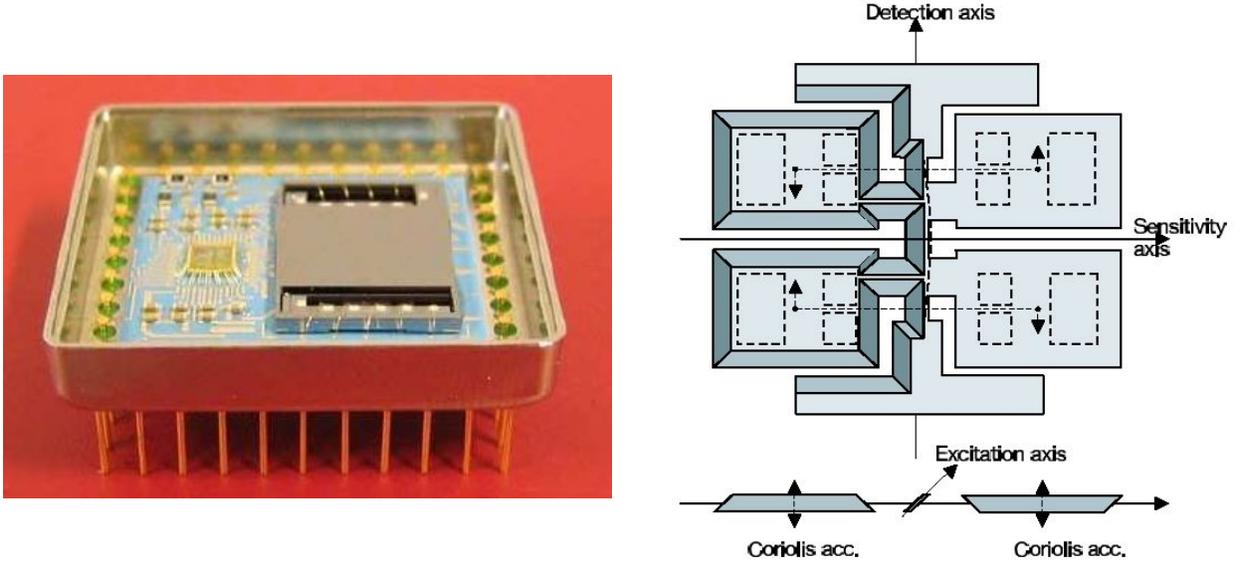


Figure 4.5: Visual representation of the butterfly gyroscope [40, 158].

methods using the maximum values of the relative pointwise errors

$$\text{relerr}_{\mathcal{H}_\infty} = \frac{\max_{\omega_k \in \Omega} \|\mathbf{G}_{\text{so}}(\dot{i}\omega_k) - \tilde{\mathbf{G}}_{\text{so}}(\dot{i}\omega_k)\|_2}{\max_{\omega_k \in \Omega} \|\mathbf{G}_{\text{so}}(\dot{i}\omega_k)\|_2} \quad (4.95)$$

to compare the worst-case performance of the computed reduced models over the frequency range Ω . The metric (4.95) serves to approximate the relative \mathcal{H}_∞ error. To compare the average performance of the computed reduced models over the frequency range Ω , we compute a discrete approximation to the relative \mathcal{H}_2 error as

$$\text{relerr}_{\mathcal{H}_2} = \frac{\sum_{\omega_k \in \Omega} \|\mathbf{G}_{\text{so}}(\dot{i}\omega_k) - \tilde{\mathbf{G}}_{\text{so}}(\dot{i}\omega_k)\|_F}{\sum_{\omega_k \in \Omega} \|\mathbf{G}_{\text{so}}(\dot{i}\omega_k)\|_F}. \quad (4.96)$$

4.5.5 Butterfly gyroscope

As a first example, we consider the butterfly gyroscope benchmark taken from the Oberwolfach Benchmark Collection [40, 158]. This benchmark problem models a vibrating mechanical gyroscope, used for navigational purposes. The model itself is a second-order system (4.72) with $n_{\text{so}} = 17\,361$ states, $m = 1$ input, and $p = 12$ outputs. The outputs model the displacement of the four electrodes in the x -, y -, and z -directions. The mass and stiffness matrices are symmetric and $\mathbf{C}_v = \mathbf{0}_{p \times n_{\text{so}}}$. The internal damping is described by Rayleigh damping

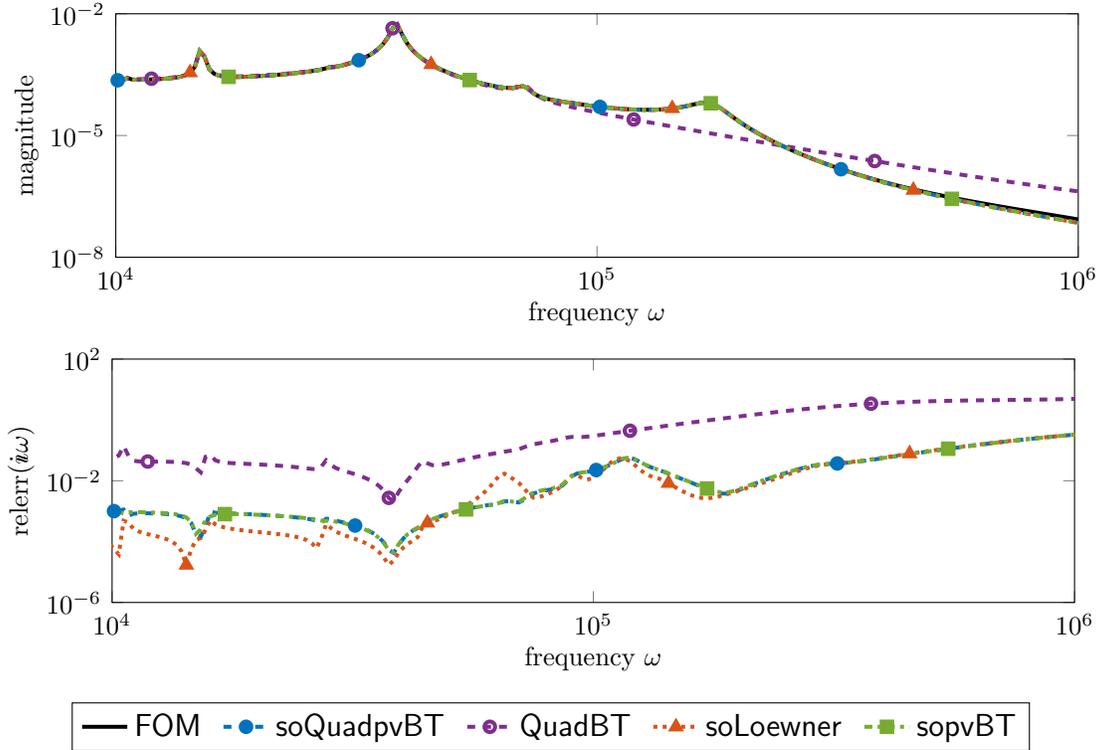


Figure 4.6: Frequency response results for reduced-order models of the butterfly gyroscope benchmark for order $r = 10$.

Table 4.3: Relative \mathcal{H}_∞ errors (4.95) and \mathcal{H}_2 errors (4.96) for the order $r = 10$ reduced models of the butterfly gyroscope benchmark. The smallest error is highlighted in **boldface**.

	soQuadpvBT	soLoewner	QuadBT	sopvBT
$\text{relerr}_{\mathcal{H}_\infty}$	4.5805e-4	4.6865e-4	1.0993e-2	4.6028e-4
$\text{relerr}_{\mathcal{H}_2}$	1.1148e-3	6.6473e-4	4.9337e-2	1.1259e-3

with $f(s) = 0$ and $g(s) = 10^{-6}$ according to (4.77). The interesting response behavior occurs in the high frequencies; figure 4.5 shows the basic setup of the system. We study this benchmark to illustrate how our method applies to a system with Rayleigh damping (4.78).

We compute reduced models of order $r = 10$ using soQuadpvBT, soLoewner, QuadBT and sopvBT. For the data-based approaches, we use $N = 200$ quadrature nodes each for the left and right points; these $N/2$ logarithmically spaced points are selected in the interval $i\Omega = [i10^4, i10^6]$ and closed under conjugation. In Figure 4.6 we plot the frequency-response of the full- and reduced-order transfer functions, as well as their pointwise-relative errors (4.94) in the frequency range $i\Omega$. We observe that each of the structure-preserving approaches (soQuadpvBT, soLoewner, and sopvBT) provide satisfactory approximations and are able

Table 4.4: Relative \mathcal{H}_∞ errors (4.95) and \mathcal{H}_2 errors (4.96) for the order $r = 50$ reduced models of the plate with TVAs. The smallest error is highlighted in **boldface**.

	soQuadpvBT	soLoewner	QuadBT
relerr $_{\mathcal{H}_\infty}$	6.4783e-8	6.5539e-8	1.6984e-7
relerr $_{\mathcal{H}_2}$	6.5947e-2	8.1058e-2	2.5217e-1

to capture the response peaks. By comparison, QuadBT performs roughly two orders of magnitude worse throughout the entire frequency range, and misses the response peak about 10^5 . Moreover, the non-intrusive soQuadpvBT very accurately mimics the behavior of the intrusive sopvBT. This finding is supported by Table 4.3, which reports the relative error measures (4.95) and (4.96). The relatively worse performance of QuadBT has more to do with the BT than the quadrature approximations. Applying intrusive (first-order) BT according to Algorithm 2.4.3 produces results that are nearly identical to those of QuadBT shown here. The reason is that a first-order approximation of dimension $r = 10$ cannot keep up with a structured second-order approximation of the same dimension.

4.5.6 Plate with tuned vibration absorbers

Next, we consider the model of the vibrational response of a strutted plate from [9, Sec. 4.2]. A slightly different version of the plate model is considered in Section 5.2.1 of this dissertation; a visual schematic of the plate is depicted therein in Figure 5.1a. We refer to Section 5.2.1 for the associated physical and material parameters of the plate being modeled. The model is a second-order system (4.72) with $n_{\text{so}} = 209100$ states, $m = 1$ input, and $p = 1$ output. For this example, $\mathbf{C}_v = \mathbf{0}_{p \times n_{\text{so}}}$ and $\mathbf{C}_p = \frac{1}{n_{\text{so}}} \mathbf{1}_{n_{\text{so}}}$, where $\mathbf{1}_{n_{\text{so}}} \in \mathbb{R}^{n_{\text{so}}}$ is the vector of all ones. The damping is assumed to be hysteretic (structural damping) with $g(s) = (i\eta)/s$ for $\eta = .001$. The tuned vibration absorbers (TVAs) connected to the plate are used to reduce the vibrational response in the frequency region about 48 Hz. These TVAs are modeled as discrete mass-spring-damper systems; matrices for the computational model itself are available at [8].

We compute reduced models of order $r = 50$ using soQuadpvBT, soLoewner, and QuadBT. For this example, sopvBT is not implemented due to the lack of available matrix equation solvers for complex-valued coefficient matrices. For computing the data-driven reduced models, $N = 250$ quadrature nodes are used; these are chosen to be $N/2$ linearly spaced points in the frequency range of $\Omega = [1, 250]$ Hz closed under complex conjugation. The frequency response of the full and reduced-order transfer functions, as well as their pointwise relative errors (4.94), are plotted in Figure 4.7. All three approaches (soQuadpvBT, soLoewner, and QuadBT) provide satisfactory approximations. Table 4.4 illustrates that soQuadpvBT slightly outperforms soLoewner, and both are marginally better than QuadBT. Visually, we observe

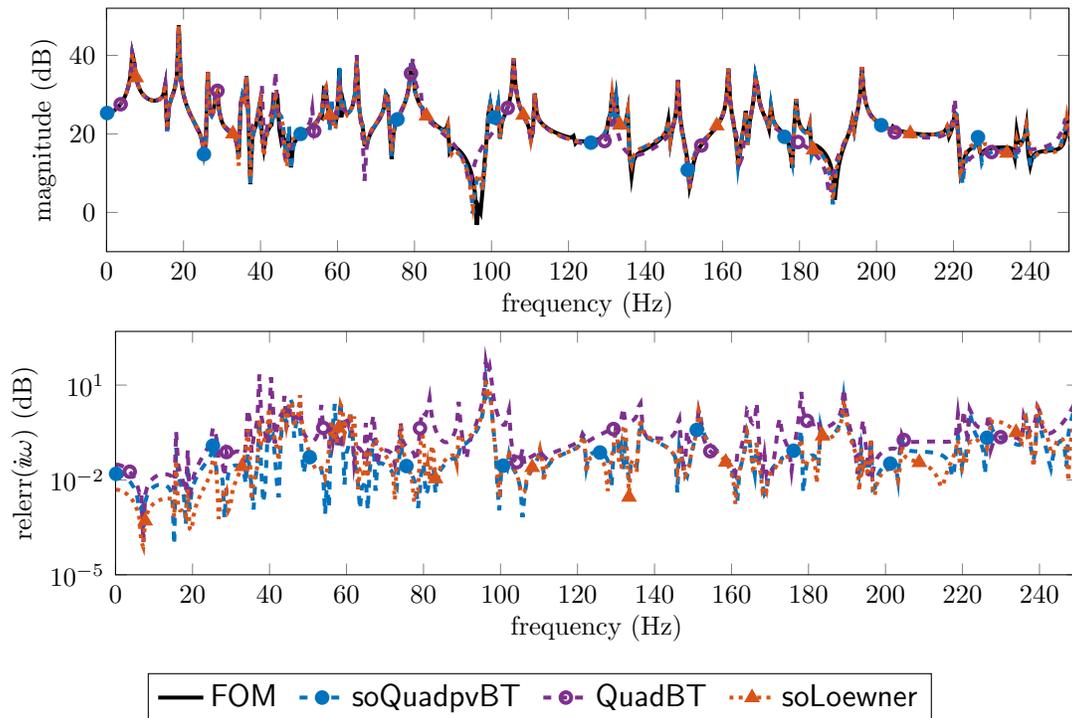


Figure 4.7: Frequency response results for reduced-order models of the plate with TVAs for order $r = 50$.

that QuadBT produces a slightly larger pointwise relative error over the whole frequency range, fails to capture the peaks of the full-order response near 180 Hz, and struggles to capture the dip around 100 Hz.

4.5.7 Mass-spring-damper network with velocity outputs

As a final example, we consider the example of a mass-spring damper (MSD) chain from [212, Example 2]. Specifically, the example is a damped linear vibration system consisting of three rows of d masses all connected to the right-hand side of a mass m_0 . The masses in the individual rows are connected by d springs, and the mass m_0 is connected to the fixed left-hand side of the base. The system order is $n_{\text{so}} = 3d + 1$. We impose Rayleigh damping on the system for the parameters $f(s) = g(s) = .002$, the input $\mathbf{B}_u = \mathbf{1}_{n_{\text{so}}} \in \mathbb{R}^{n_{\text{so}}}$ is the vector of all ones, $\mathbf{C}_p = \mathbf{0}_{p \times n_{\text{so}}}$, and $\mathbf{C}_v = \mathbf{1}_{n_{\text{so}}} \in \mathbb{R}^{n_{\text{so}}}$ is the vector of all ones. We include this example to illustrate how our method performs when velocity outputs are incorporated. One reason that the mass-spring-damper network with velocity outputs is interesting is because the second-order dynamics can be rewritten as an equivalent first-order port-Hamiltonian system in this case; see [98, Sec. 6.11].

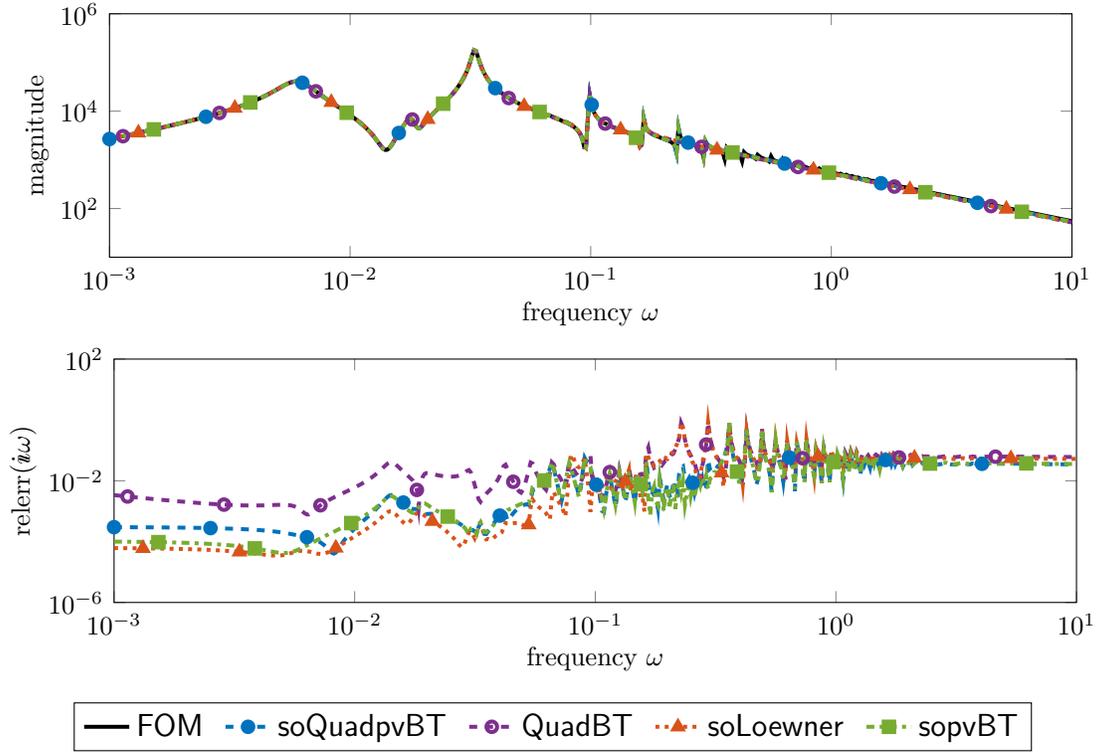


Figure 4.8: Frequency response results for reduced-order models of the coupled mass-spring-damper network with velocity outputs for order $r = 10$.

Table 4.5: Relative \mathcal{H}_∞ errors (4.95) and \mathcal{H}_2 errors (4.96) for the order $r = 10$ reduced models of the MSD chain. The smallest error is highlighted in **boldface**.

	soQuadpvBT	soLoewner	QuadBT	sopvBT
$\text{relerr}_{\mathcal{H}_\infty}$	4.5897e-3	1.4044e-2	1.4096e-2	4.5771e-3
$\text{relerr}_{\mathcal{H}_2}$	2.8771e-3	5.1830e-3	1.1213e-2	2.7849e-3

We compute reduced models of order $r = 10$ using soQuadpvBT, soLoewner, QuadBT, and sopvBT. For computing the data-driven reduced models, $N = 200$ quadrature nodes are used; these are chosen to be $N/2$ linearly spaced points in the frequency range of $i\Omega = [i10^{-3}, i10^1]$ closed under complex conjugation. The frequency response of the full and reduced-order transfer functions, as well as their pointwise relative errors (4.94), are plotted in Figure 4.8. As was the case for the previous two benchmarks, we again observe that the structure-preserving approaches outperform QuadBT overall. This is particularly evident at the lower frequencies; all four methods exhibit very similar behavior from $i10e0$ onward. The relative error measures are reported in Table 4.5; sopvBT performs the best overall, although soQuadpvBT performs very similarly.

4.6 Conclusions

In this chapter, several new methods for the data-driven balancing of linear dynamical systems of the form (2.25) and (4.72) are presented. These results lay the theoretical foundation for data-driven implementations of numerous BT variants, and effectively generalize the quadrature-based balancing framework of [89] to other kinds of BT-MOR tailored for various internal structures. In the setting of linear first-order systems (2.25), four different variants were considered. For the cases of BST, PRBT, and BRBT, it is shown that evaluating certain spectral factors associated with the underlying full-order model transfer function is required to do these kinds of BT from data. In the case of FWBT with weights, a data-driven formulation requires sampling the underlying full-order model transfer function, as well as the input and output frequency weights that dictate the frequency band of interest. The proposed quadrature-based/data-driven methods are applied to an RLC circuit model in order to validate the data-driven BT-ROMs. In each case, it was observed that the data-based reduced models perform very similarly to their intrusive counterparts. In the setting of linear second-order systems (4.72), a data-driven reformulation of `sopvBT` is developed. The resulting method (`soQuadpvBT`) is applicable to any second-order system that exhibits generalized proportional damping (4.77), and can incorporate both position and velocity outputs. The proposed approach is tested against several benchmark examples for second-order systems. In each case, it is observed that the *structured* surrogate models, computed intrusively or non-intrusively, outperform non-structure-preserving approaches.

Chapter 5

System-theoretic concepts for linear systems with quadratic outputs

5.1 Introduction

In this chapter, we turn our attention to a class of weakly nonlinear dynamical systems; those that are linear in the state equation and contain quadratic terms in the output equation. Dynamical systems with quadratic-output functions appear naturally in applications whenever one is interested in observing or simulating response quantities computed as the product of time- or frequency-components of the state. Relevant examples of quadratic outputs include the root mean squared displacement of the states [9, 188, 214], quadratic cost functions in design optimization problems [209, 241, 242]—e.g., in the optimal placement of tuned mass dampers used to minimize vibrations—and quantities pertaining to power or energy [109, 177]. Several selected examples of these quadratic-output systems from practical applications are presented in Section 5.2 to motivate our study. In state space, linear dynamical systems with quadratic-output functions, or so-called *linear quadratic-output* (LQO) systems, are formulated as

$$\mathcal{G}_{\text{lqo}} : \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{x}(0) = \mathbf{0}_n, \\ \mathbf{y}(t) = \underbrace{\mathbf{C}\mathbf{x}(t)}_{\stackrel{\text{def}}{=} \mathbf{y}_{\text{lo}}(t)} + \underbrace{\mathbf{M}(\mathbf{x}(t) \otimes \mathbf{x}(t))}_{\stackrel{\text{def}}{=} \mathbf{y}_{\text{qo}}(t)}, \end{cases} \quad (5.1)$$

where $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$ describe the evolution of the internal states $\mathbf{x} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^n$ under the influence of external control variables $\mathbf{u} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^m$ and the outputs $\mathbf{y} : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^p$ are modeled by the matrices $\mathbf{C} \in \mathbb{R}^{p \times n}$ and $\mathbf{M} \in \mathbb{R}^{p \times n^2}$. As already discussed for models of linear time-invariant systems in Section 2.4, in most practical applications the state dimension n of (5.1) is prohibitively large so that repeated evaluations of the full-order model are computationally infeasible, and system approximation is necessary. Systems of the form (5.1) have received increased attention in the model reduction literature in the past decade. Our interest in the approximation of linear quadratic-output systems stems largely from the applications highlighted at the onset of this section, although these systems also hold our interest from a theoretical point of view. Because the nonlinearity in (5.1) is restricted to the output equation, the system's input-to-state map is still fully linear; the

systems in (5.1) are thus attractive to study as a weakly nonlinear extension of classical linear-output systems (2.25).

In principle, one can emulate each of the quadratic outputs in (5.1) using an equivalent multiple linear-output system such as (2.25) under some mild assumptions. For instance, in the case of a single quadratic output, one can write $\mathbf{y}(t) = \mathbf{M}(\mathbf{x}(t) \otimes \mathbf{x}(t)) = \mathbf{x}(t)^\top \mathbf{M}_1 \mathbf{x}(t)$ where $\mathbf{M} = \text{vec}(\mathbf{M}_1)$ and $\mathbf{M}_1 \in \mathbb{R}^{n \times n}$. If \mathbf{M}_1 admits a square-root factorization $\mathbf{M}_1 = \mathbf{Z}^\top \mathbf{Z}$ for $\mathbf{Z} \in \mathbb{R}^{k \times n}$, $k \leq n$, then the single quadratic output can be recovered from a linear-output system (2.25) with same dynamics as (5.1) and the output equation $\mathbf{y}_{\text{lo}}(t) = \mathbf{Z}\mathbf{x}(t)$. Specifically, \mathbf{y} is given by the 2-norm of \mathbf{y}_{lo} :

$$\mathbf{y}(t) = \|\mathbf{y}_{\text{lo}}(t)\|_2^2 = (\mathbf{Z}\mathbf{x}(t))^\top (\mathbf{Z}\mathbf{x}(t)) = \mathbf{x}(t)^\top \mathbf{M}_1 \mathbf{x}(t).$$

Classical model-order reduction approaches for systems of the form (5.1) reduce the model order by first computing such a linearization, and subsequently applying one of the many well-established techniques from linear model reduction; see, e.g. [9, 214, 215]. However, this linearization often results in an intermediate model (2.25) with a very large number of outputs $p \sim n$, and traditional linear system approximation techniques tend to require more computational effort and produce lower-quality surrogates in this regime; see [7, 9]. In Section 5.2, we provide an example of a finite-element model from vibro-acoustics where the quadratic output is the root mean squared error of the spatial modes in the z -axis; emulating this quadratic output using an equivalent linear-output system (2.25) requires $p = 27\,298$ *individual linear outputs*.

Instead, we are interested here in the computation of structured surrogate models that:

- (i) Preserve the quadratic nonlinearity in the output equation (5.1), and;
- (ii) utilize the quadratic state-to-output map of (5.1) as is—that is, without any intermediate lifting or linearization—to determine suitable approximation subspaces for order reduction.

More precisely, given a linear quadratic-output system as in (5.1), our goal is the construction of another comparatively lower-order, linear quadratic-output system of the form

$$\tilde{\mathcal{G}}_{\text{lqo}} : \begin{cases} \tilde{\mathbf{E}}\dot{\tilde{\mathbf{x}}}(t) = \tilde{\mathbf{A}}\tilde{\mathbf{x}}(t) + \tilde{\mathbf{B}}\mathbf{u}(t), & \tilde{\mathbf{x}}(0) = \mathbf{0}_r, \\ \tilde{\mathbf{y}}(t) = \underbrace{\tilde{\mathbf{C}}\tilde{\mathbf{x}}(t)}_{\stackrel{\text{def}}{=} \tilde{\mathbf{y}}_{\text{lo}}(t)} + \underbrace{\tilde{\mathbf{M}}(\tilde{\mathbf{x}}(t) \otimes \tilde{\mathbf{x}}(t))}_{\stackrel{\text{def}}{=} \tilde{\mathbf{y}}_{\text{qo}}(t)}, \end{cases} \quad (5.2)$$

where $\tilde{\mathbf{E}}, \tilde{\mathbf{A}} \in \mathbb{R}^{r \times r}$, $\tilde{\mathbf{B}} \in \mathbb{R}^{r \times m}$, $\tilde{\mathbf{C}} \in \mathbb{R}^{p \times r}$ and $\tilde{\mathbf{M}} \in \mathbb{R}^{p \times r^2}$ for $r \ll n$. As with the generic linear model reduction problem, to be an effective surrogate, the reduced model (5.2) should accurately reproduce the input-to-output response of the original system (5.1) for all admissible inputs. Several successful attempts have already been made in this realm; for

example, there are generalizations of balanced truncation model reduction [16, 28, 176, 177, 178, 205, 214], methods based on the rational interpolation of the linear- and quadratic-output subsystem transfer functions of (5.1) or matching moments [49, 62, 87, 188, 215], and an extension of the popular adaptive Antoulas-Anderson algorithm [153] for the construction of surrogate models (5.2) from frequency-response data [88]. Two of the above works deserve special mention: Benner et al. [28] introduce a novel algebraic Gramian and linear quadratic-output system \mathcal{H}_2 norm based on the Volterra kernels of (5.1), and a related balanced truncation algorithm. Diaz et al. [62] introduce an overarching framework for tangential interpolation of dynamical systems with up to quadratic-bilinear dynamics and quadratic-bilinear outputs; this general model class includes (5.1) as a special case.

5.1.1 Chapter contents

The purpose of this chapter is to motivate and set up the generic model reduction problem for linear quadratic-output systems (5.1), as well as lay the theoretical groundwork for the results in Chapter 6 that consider the optimal- \mathcal{H}_2 approximation of (5.1). In Section 5.2, we present three examples of linear dynamical systems with quadratic-output functions from applications to motivate our study. Section 5.3 revises the system-theoretic concepts for linear quadratic-output systems that are required for the forthcoming results here and in Chapter 6. In Section 5.4 we derive new expressions for computing the linear quadratic-output system \mathcal{H}_2 norm and inner product. The first pair of formulae that we derive is expressed in terms of solutions to generalized Sylvester equations, while the second is expressed in terms of the poles and residues of the linear- and quadratic-output subsystem transfer functions of (5.1); see Theorems 5.10 and 5.11, respectively. Significantly, these will enable us to realize more computationally tractable parameterizations of the \mathcal{H}_2 model error that we will use as footholds to derive first-order optimality conditions for the best \mathcal{H}_2 approximation of linear quadratic-output systems in Chapter 6.

Portions of this introduction, as well as the contents of Sections 5.2—5.4 are available in the preprints [186, 188]. Theorem 6.9 is from a work in preparation [185].

5.2 Motivating examples

In this section, three examples of linear dynamical systems with quadratic output functions are presented to motivate our study and to illustrate the necessity for specially tailored surrogate models for such systems.

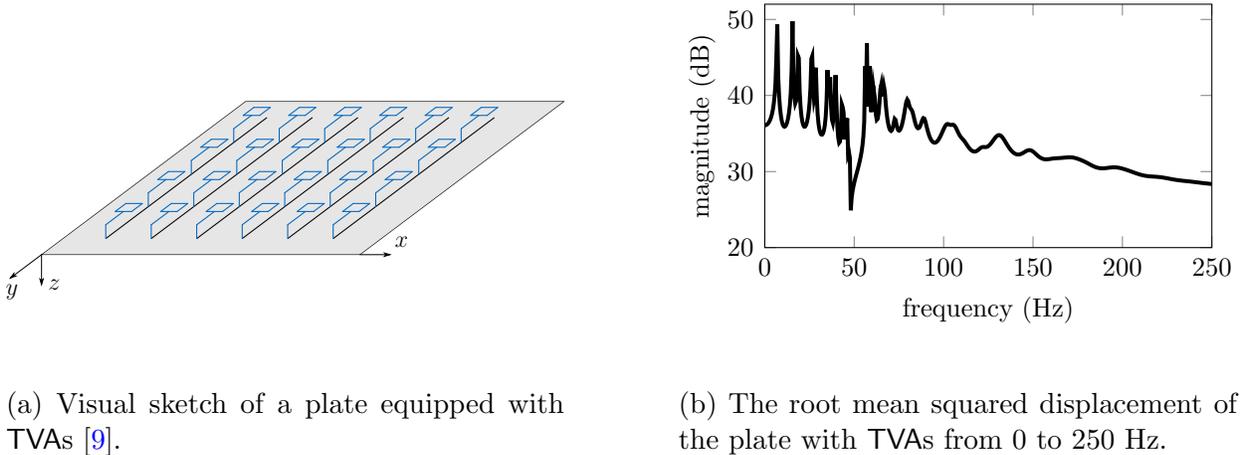


Figure 5.1: Plate equipped with TVAs from [9].

5.2.1 Vibration of a plate with tuned vibration absorbers

In the study of structural dynamics and vibro-acoustic systems, one is often interested in measuring the average spatial deformation or vibrational response of a given surface or structure in response to external excitations [9, 215, 241, 242]. These vibrational analyses are usually accomplished by frequency response analysis of a finite-element model using numerical simulations. For accurate predictions, high-order models (many elements) are required, and evaluating a single frequency sweep is excessively costly. As a specific example of these systems, we consider the vibrational response of a simply strutted plate with tuned vibration absorbers (TVAs) from [9]. A slightly different version of this model was also considered in Section 4.5.4. The plate has the dimensions 0.8×0.8 m with a thickness of 1 mm; it consists of aluminum, with the material parameters $E = 69$ GPa (Young’s Modulus), $\rho = 2650$ kg m⁻³ (density) and $\nu = 0.22$ (shear modulus). The damping is assumed to be proportional (Rayleigh) damping (4.78) with the scaling parameters $\alpha = 0.01$ and $\beta = 1 \cdot 10^{-4}$. The system matrices are real-valued and symmetric. The TVAs connected to the plate are used to reduce the vibrational response in the frequency region about 48 Hz, and are modeled as discrete mass-spring-damper systems. The discretized plate model is given in linear *second-order* form (4.72) with $n_{\text{so}} = 201\,900$ spatial modes. In this section, we consider its (equivalent) first-order formulation (4.75) with $n = 2n_{\text{so}} = 403\,800$ states. Acoustical engineers are especially interested in the root mean squared displacement of the plate points in the z -direction in response to point load excitation; see Figure 5.1a for a visual sketch. This response quantity for the full-order model is evaluated over the frequency range from 0 to 250 Hz and plotted in Figure 5.1b. Explicitly, the root mean squared displacement of the state (in frequency coordinates) is written down in general form as

$$y_{\text{rms}}(s) \stackrel{\text{def}}{=} \sqrt{\sum_{k=1}^n |m_k|^2 |X_k(s) - \check{x}_k|^2}, \quad (5.3)$$

where $X_k(s)$ is the k -th component of the state vector \mathbf{X} in (2.29), $\tilde{\mathbf{x}} \in \mathbb{R}^n$ is a point of reference, and $m_k \in \mathbb{R}$ are weights. In its general formulation, (5.3) can be simulated via the *transfer functions* of an LQO system (5.1), i.e.,

$$\begin{aligned} y(s) = y_{\text{rms}}(s)^2 &= \mathbf{M} (\mathbf{X}(s) \otimes \mathbf{X}(-s)) - 2\tilde{\mathbf{x}}^\top \mathbf{M}_1 \mathbf{X}(s) + \tilde{\mathbf{x}}^\top \mathbf{M}_1 \tilde{\mathbf{x}} \\ &= \mathbf{M} ((s\mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \otimes (-s\mathbf{E} - \mathbf{A})^{-1} \mathbf{B}) + \mathbf{C} (s\mathbf{E} - \mathbf{A})^{-1} \mathbf{B} + \tilde{\mathbf{x}}^\top \mathbf{M}_1 \tilde{\mathbf{x}} \end{aligned}$$

for $\mathbf{C} = -2\tilde{\mathbf{x}}^\top \mathbf{M}_1$ and $\mathbf{M} = \text{vec}(\mathbf{M}_1)$ with $\mathbf{M}_1 \stackrel{\text{def}}{=} \text{diag}(|m_1|^2, \dots, |m_n|^2) \in \mathbb{R}^{n \times n}$. See (5.12) in Section 5.3.2 for the formulation of the transfer functions. One may also take the reference state $\tilde{\mathbf{x}}$ to be zero without loss of generality by replacement of \mathbf{x} with $\mathbf{x} - \tilde{\mathbf{x}}$; in this case, (5.3) is purely quadratic with no linear component.

Model-order reduction is used to reduce the order of the plate model and thereby facilitate fast evaluation of the root mean squared displacement (5.3). In [9], the root mean squared displacement of the plate components is emulated using a $p \times n$ linear-output matrix to recover the displacement of the relevant spatial coordinates and then compute (5.3); this approach results in a linear-output system with $p = 27\,278$ linear outputs. Instead, we choose to model the root mean squared displacement (5.3) of the plate model directly as a single quadratic-output function, and develop surrogate models of the form (5.2). We investigate this example further using numerical experiments in Section 6.6.

5.2.2 The Hamiltonian energy functional

One important class of systems in the modeling community is that of port-Hamiltonian systems [141, 216]. In state-space, we say a linear time-invariant system has a port-Hamiltonian (pH) realization if it can be written in the form

$$\begin{aligned} \dot{\mathbf{x}}(t) &= (\mathbf{J} - \mathbf{R}) \mathbf{Q} \mathbf{x}(t) + \mathbf{B} \mathbf{u}(t), \\ \mathbf{y}_{\text{lo}}(t) &= \mathbf{B}^\top \mathbf{Q} \mathbf{x}(t), \end{aligned} \tag{5.4}$$

where $\mathbf{B} \in \mathbb{R}^{n \times m}$, $\mathbf{J}, \mathbf{R}, \mathbf{Q} \in \mathbb{R}^{n \times n}$ with $\mathbf{J} = -\mathbf{J}^\top$, \mathbf{R}, \mathbf{Q} symmetric positive semi-definite. We mention that (5.4) is one very specific formulation of a pH system, and more complicated forms exist; see [141]. The energy functional

$$\mathcal{H}: \mathbb{R}^n \rightarrow \mathbb{R} \quad \text{where} \quad \mathcal{H}(\mathbf{x}(t)) \stackrel{\text{def}}{=} \mathbf{x}(t)^\top \mathbf{Q} \mathbf{x}(t) \tag{5.5}$$

is called the *Hamiltonian* of the system (5.4). The Hamiltonian represents the total energy stored in the system at time $t \geq 0$, and obeys the energy-balanced equation

$$\mathcal{H}(\mathbf{x}(t_2)) \leq \mathcal{H}(\mathbf{x}(t_1)) + \int_{t_1}^{t_2} \mathbf{u}(t)^\top \mathbf{y}_{\text{lo}}(t) dt$$

for all $t_1 \leq t_2$ and solution trajectories \mathbf{x} that satisfy (5.4) with the input-output pair $\mathbf{u}, \mathbf{y}_{\text{lo}}$. By design, port-Hamiltonian systems are stable and passive, and any stable and

passive linear time-invariant system admits a pH representation (5.4); see [109, Sec. 2.2] for a constructive illustration of this equivalence. In the context of the MOR of pH systems, not only is preserving the traditional input-to-output map of interest, but also preserving the Hamiltonian (5.5). It is proposed in [109] that the Hamiltonian be included in (5.1) as an additional quadratic output; this yields a linear quadratic-output system (5.1) with the particular choice of output matrices

$$C = \begin{bmatrix} B^\top Q \\ \mathbf{0}_{1 \times n} \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} \mathbf{0}_{m \times n^2} \\ \text{vec}(Q)^\top \end{bmatrix} \quad \text{so that} \quad \mathbf{y}(t) = \begin{bmatrix} B^\top Q \mathbf{x}(t) \\ \mathbf{x}(t)^\top Q \mathbf{x}(t) \end{bmatrix}.$$

We refer to [109] for further details.

Other examples of energy-based quantities of interest appear in constrained optimization problems where the objective function is quadratic, e.g., the square of the 2-norm of the output \mathbf{y} in damping optimization; we refer to the examples in [241, Sec. 2] for further details.

5.2.3 1D advection-diffusion equation with a quadratic cost

As a final example, we consider the 1D advection-diffusion equation with a quadratic cost function that has been used as a benchmark [62, 186]. This example illustrates how an LQO system (5.1) can arise from the discretization of a PDE. The governing equations for this PDE are

$$\begin{aligned} \frac{\partial}{\partial t} v(t, x) - \alpha \frac{\partial^2}{\partial x^2} v(t, x) + \beta \frac{\partial}{\partial x} v(t, x) &= 0, \\ v(t, 0) = u_0(t), \quad \alpha \frac{\partial}{\partial x} v(t, 1) = u_1(t), \quad v(0, x) &= 0, \end{aligned} \tag{5.6}$$

for $x \in (0, 1)$ and $t \in (0, T)$ and inputs $u_0, u_1 \in \mathcal{L}_2(0, T)$; the diffusion and advection coefficients are $\alpha > 0$ and $\beta \geq 0$, respectively. The output that we consider is

$$C(x, t) = \frac{1}{2} \int_0^1 |v(t, x) - 1|^2 dx. \tag{5.7}$$

Such an observable may arise from, e.g., the objective cost function in an optimal control problem. Discretizing the equations in (5.6) using $n + 1$ equidistant spatial points yields an order- n state-space model of the form (5.1) with $m = 2$ inputs and $p = 1$ output. Let $\mathbf{x}(t) \in \mathbb{R}^n$ denote the spatial discretization of $v(t, x)$, $h = 1/n$, and $\mathbf{1}_s \in \mathbb{R}^s$ the vector consisting of all ones. Then, the discretization provides an approximation to the quadratic cost function $C(x, t)$ in (5.7):

$$C(x, t) \approx \frac{h}{2} \|\mathbf{x}(t) - \mathbf{1}\|_2^2 = -h \mathbf{1}_n^\top \mathbf{x}(t) + \frac{h}{2} \text{vec}(\mathbf{I}_n)^\top (\mathbf{x}(t) \otimes \mathbf{x}(t)) + \frac{h}{2} \|\mathbf{1}_n\|_2^2 = y(t) + \frac{h}{2} \|\mathbf{1}_n\|_2^2.$$

To fit (5.1), we take $\mathbf{C} = -h\mathbf{1}_n^\top \in \mathbb{R}^{1 \times n}$ and $\mathbf{M} = \frac{h}{2} \text{vec}(\mathbf{I}_n)^\top \in \mathbb{R}^{1 \times n^2}$, where \mathbf{I}_n is the $n \times n$ identity matrix. The approximation to the cost at time $t \geq 0$ is recovered from a single output $y(t)$ in (5.1) via $\frac{h}{2}\|\mathbf{x}(t) - \mathbf{1}_n\|_2^2 = y(t) + \frac{h}{2}\|\mathbf{1}_n\|_2^2$. We consider this specific benchmark in the numerical results that are presented in Section 6.6.

5.3 System-theoretic concepts

Before presenting the major theoretical results of this chapter, we revisit the relevant system-theoretic concepts for LQO systems (5.1). Many of these build directly from those presented for linear-output systems (2.25) in Section 2.3. We also review selected existing MOR algorithms for systems of the form (5.1).

5.3.1 Basic concepts and definitions

As already highlighted, the nonlinearity in (5.1) is restricted to the output equations; the input-to-state map of (5.1) is identical to that of a classical linear-output system (2.25) having the same \mathbf{E} , \mathbf{A} and \mathbf{B} matrices. Thus, much of the previous discussion for linear-time invariant systems presented in Section 2.3 passes directly to this setting. Specifically:

- The *order* n of the system in (5.1) is the number of differential equations in (5.1) (Definition 2.20).
- The system in (5.1) is *asymptotically stable* if all the eigenvalues of the matrix pencil $s\mathbf{E} - \mathbf{A}$ have negative real part (Definition 2.21).
- The system (5.1) is *reachable* if for any state $\tilde{\mathbf{x}} \in \mathbb{R}^n$ there exists a finite time $t_f > 0$ and finite-energy input \mathbf{u} such that the corresponding trajectory $\mathbf{x}(t)$ that solves the dynamics in (5.1) satisfies $\mathbf{x}(0) = \mathbf{0}_n$ and $\mathbf{x}(t_f) = \tilde{\mathbf{x}}$ (Definition 2.27).

Remark 5.1 (Quadratic-output terms in (5.1)). Using (2.7), one can arrive at an alternative expression for the quadratic terms \mathbf{y}_{qo} of the system outputs \mathbf{y} in (5.1), namely

$$\mathbf{y}_{\text{qo}}(t) = \mathbf{M}(\mathbf{x}(t) \otimes \mathbf{x}(t)) = \begin{bmatrix} \mathbf{x}(t)^\top \mathbf{M}_1 \mathbf{x}(t) \\ \mathbf{x}(t)^\top \mathbf{M}_2 \mathbf{x}(t) \\ \vdots \\ \mathbf{x}(t)^\top \mathbf{M}_p \mathbf{x}(t) \end{bmatrix} \quad \text{where } \mathbf{M} \stackrel{\text{def}}{=} \begin{bmatrix} \text{vec}(\mathbf{M}_1)^\top \\ \text{vec}(\mathbf{M}_2)^\top \\ \vdots \\ \text{vec}(\mathbf{M}_p)^\top \end{bmatrix}. \quad (5.8)$$

The matrix $\mathbf{M}_k \in \mathbb{R}^{n \times n}$ encodes the quadratic term of the k -th output. We will switch between these two (equivalent) formulations of \mathbf{y}_{qo} as needed for ease of exposition and

theoretical development. In the representation (5.8), one can always replace \mathbf{M}_k by its symmetric part by noting that

$$\mathbf{x}(t)^\top \mathbf{M}_k \mathbf{x}(t) = \frac{1}{2} \mathbf{x}(t)^\top (\mathbf{M}_k + \mathbf{M}_k^\top) \mathbf{x}(t)$$

since $\mathbf{x}(t)^\top \mathbf{M}_k \mathbf{x}(t)$ is a scalar quantity and \mathbf{M}_k is real valued. Thus, we henceforth assume that each \mathbf{M}_k is symmetric without loss of generality. This will allow us to derive useful symmetry properties of the system transfer functions of (5.1) later on. \diamond

5.3.2 Subsystem Volterra kernels and transfer functions

Multiple classes of weakly nonlinear dynamical systems can be understood via infinite series of *Volterra kernels*; see [194] for details. Because the nonlinearity in (5.1) is restricted to the output equation, only *two* kernels are required to fully describe the system's input-to-output mapping [28]. The solution $\mathbf{x}(t)$ to (5.1) at time $t \geq 0$ is precisely that in (2.26) of the linear time-invariant system (2.25). Substituting \mathbf{x} into the equation for \mathbf{y} reveals the input-to-output relationship

$$\mathbf{y}(t) = \underbrace{\int_0^t \mathbf{g}_{\text{lo}}(\tau) \mathbf{u}(t - \tau) d\tau}_{=\mathbf{y}_{\text{lo}}(t)} + \underbrace{\int_0^t \int_0^t \mathbf{g}_{\text{qo}}(\tau_1, \tau_2) (\mathbf{u}(t - \tau_1) \otimes \mathbf{u}(t - \tau_2)) d\tau_1 d\tau_2}_{=\mathbf{y}_{\text{qo}}(t)} \quad (5.9)$$

for any time $t \geq 0$. The univariate and multivariate Volterra kernels $\mathbf{g}_{\text{lo}}: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{p \times m}$ and $\mathbf{g}_{\text{qo}}: \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}^{p \times m^2}$ that appear in (5.9) are defined as

$$\mathbf{g}_{\text{lo}}(t) \stackrel{\text{def}}{=} \mathbf{C} e^{\mathbf{E}^{-1} \mathbf{A} t} \mathbf{E}^{-1} \mathbf{B}, \quad (5.10a)$$

$$\text{and } \mathbf{g}_{\text{qo}}(t_1, t_2) \stackrel{\text{def}}{=} \mathbf{M} \left(e^{\mathbf{E}^{-1} \mathbf{A} t_1} \mathbf{E}^{-1} \mathbf{B} \otimes e^{\mathbf{E}^{-1} \mathbf{A} t_2} \mathbf{E}^{-1} \mathbf{B} \right). \quad (5.10b)$$

For asymptotically stable systems (5.1), the Volterra kernels in (5.10) belong to the appropriate \mathcal{L}_2 spaces introduced in Section 2.2.3, i.e.,

$$\mathbf{g}_{\text{lo}} \in \mathcal{L}_2^{p \times m}(\mathbb{R}_{\geq 0}) \quad \text{and} \quad \mathbf{g}_{\text{qo}} \in \mathcal{L}_2^{p \times m^2}(\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}).$$

Remark 5.2 (Linear- and quadratic-output subsystems of (5.1)). The Volterra kernels in (5.10) provide a useful perspective on the system in (5.1); namely, the full LQO system (5.1) is comprised of two coupled subsystems. The first is a purely *linear*-output system having the same form as (2.25):

$$\mathcal{G}_{\text{lo}} : \begin{cases} \mathbf{E} \dot{\mathbf{x}}(t) = \mathbf{A} \mathbf{x}(t) + \mathbf{B} \mathbf{u}(t), & \mathbf{x}(0) = \mathbf{0}_n, \\ \mathbf{y}_{\text{lo}}(t) = \mathbf{C} \mathbf{x}(t), \end{cases}$$

that we denote by \mathcal{G}_{lo} . The (completely linear) input-to-output map of the subsystem \mathcal{G}_{lo} is specified by the convolution of the first (univariate) Volterra kernel with the input \mathbf{u} in the expansion (5.9). The second is the purely *quadratic*-output system defined as

$$\mathcal{G}_{\text{qo}} : \begin{cases} \mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{x}(0) = \mathbf{0}_n, \\ \mathbf{y}_{\text{qo}}(t) = \mathbf{M}(\mathbf{x}(t) \otimes \mathbf{x}(t)), \end{cases} \quad (5.11)$$

that we denote by \mathcal{G}_{qo} . The (completely quadratic) input-to-output map of the subsystem \mathcal{G}_{qo} is specified by the convolution of the second (multivariate) Volterra kernel with $\mathbf{u} \otimes \mathbf{u}$ in the expansion (5.9). If $\mathbf{M} = \mathbf{0}_{p \times n^2}$, the output $\mathbf{y} = \mathbf{y}_{\text{lo}}$ in (5.9) is purely linear and (5.1) is described entirely by its first linear-output subsystem \mathcal{G}_{lo} (which is just a classical linear time-invariant system (2.25)). If $\mathbf{C} = \mathbf{0}_{p \times n}$, the output $\mathbf{y} = \mathbf{y}_{\text{qo}}$ in (5.9) is purely quadratic and (5.1) is described entirely by its second quadratic-output subsystem in (5.11). This perspective of (5.1) marries well with the concept of coupled subsystem expansions for other classes of nonlinear systems; e.g., bilinear or quadratic-bilinear systems; see [194] for further details. The primary difference here is that \mathcal{G}_{lo} is fully specified by *two* subsystems as opposed to infinitely many. This subsystem perspective of (5.1) will also be useful for drawing comparisons with the analogous results and ideas from linear model reduction that were reviewed in Section 2.4, because LTI systems (2.25) can be viewed as a special case of LQO systems (5.1) with \mathbf{M} equal to zero. \diamond

In the linear setting of (2.25), we used the univariate Laplace transformation to derive a frequency-domain representation of the system (2.25). We will employ the same strategy here using the multivariate Laplace transform in Definition 2.18. By applying the univariate Laplace transform to \mathbf{g}_{lo} in (5.10a) and the bivariate Laplace transform to \mathbf{g}_{qo} in (5.10b), we obtain a frequency-domain representation of (5.1) in the form of *two* complex-matrix-valued functions $\mathbf{G}_{\text{lo}}: \mathbb{C} \rightarrow \mathbb{C}^{p \times m}$ and $\mathbf{G}_{\text{qo}}: \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{C}^{p \times m^2}$ defined as

$$\mathbf{G}_{\text{lo}}(s) \stackrel{\text{def}}{=} \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}, \quad (5.12a)$$

$$\text{and } \mathbf{G}_{\text{qo}}(s_1, s_2) \stackrel{\text{def}}{=} \mathbf{M}((s_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} \otimes (s_2\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}). \quad (5.12b)$$

See also [62, Section 3.1], [88, Lemma 2.1], [87]. Recalling Remark 5.2, \mathbf{G}_{lo} is the transfer function of the linear-output subsystem \mathcal{G}_{lo} of the LQO system \mathcal{G}_{lqo} defined according to (2.25), whereas \mathbf{G}_{qo} is the transfer function of the quadratic-output subsystem \mathcal{G}_{qo} of \mathcal{G}_{lqo} in (5.11). Note that \mathbf{G}_{qo} is a multivariate rational function of two complex variables s_1 and s_2 , and its poles are precisely the eigenvalues of $s\mathbf{E} - \mathbf{A}$. Thus, for asymptotically stable systems (5.1), the transfer functions belong to the appropriate Hardy spaces

$$\mathbf{G}_{\text{lo}} \in \mathcal{H}_2^{p \times m}(\mathbb{C}_{>0}) \quad \text{and} \quad \mathbf{G}_{\text{qo}} \in \mathcal{H}_2^{p \times m^2}(\mathbb{C}_{>0} \times \mathbb{C}_{>0})$$

introduced in Section 2.2.3. Notationally, we take $\overline{\mathbf{G}_{\text{lo}}}(s)$ or $\overline{\mathbf{G}_{\text{qo}}}(s)$ to mean that conjugation

is only applied to the matrices in the transfer function, and not the argument, i.e.,

$$\begin{aligned}\overline{\mathbf{G}}_{\text{lo}}(s) &= \overline{\mathbf{C}}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}, \\ \overline{\mathbf{G}}_{\text{lo}}(s) &= \overline{\mathbf{C}}(-s\overline{\mathbf{E}} - \overline{\mathbf{A}})^{-1}\overline{\mathbf{B}}, \\ \overline{\mathbf{G}}_{\text{qo}}(s_1, s_2) &= \overline{\mathbf{M}}\left((s_1\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} \otimes (s_2\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\right), \\ \overline{\mathbf{G}}_{\text{qo}}(s_1, s_2) &= \overline{\mathbf{M}}\left((s_1\overline{\mathbf{E}} - \overline{\mathbf{A}})^{-1}\overline{\mathbf{B}} \otimes (s_2\overline{\mathbf{E}} - \overline{\mathbf{A}})^{-1}\overline{\mathbf{B}}\right).\end{aligned}\tag{5.13}$$

For real-valued systems (5.1), we have that $\overline{\mathbf{G}}_{\text{lo}}(i\omega) = \overline{\mathbf{G}}_{\text{lo}}(-i\omega) = \overline{\mathbf{C}}(-i\omega\overline{\mathbf{E}} - \overline{\mathbf{A}})^{-1}\overline{\mathbf{B}} = \mathbf{G}_{\text{lo}}(-i\omega)$, and $\overline{\mathbf{G}}_{\text{qo}}(i\omega_1, i\omega_2) = \overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -i\omega_2) = \mathbf{G}_{\text{qo}}(-i\omega_1, -i\omega_2)$ for $\omega, \omega_1, \omega_2 \in \mathbb{R}$.

As a direct consequence of the symmetry underlying \mathbf{M} that we described in Remark 5.1 and properties of the Kronecker product presented in Propositions 2.5 and 2.6, the quadratic-output transfer function \mathbf{G}_{qo} in (5.12b) exhibits some useful symmetries. Specifically, \mathbf{G}_{qo} and its first partial derivatives are symmetric with respect to the interchange of the arguments s_1, s_2 and matrix-vector products. These symmetry conditions will be used to simplify the interpolation-based optimality conditions that we derive in Chapter 6. For the subsequent lemma, we introduce the notation

$$\frac{\partial}{\partial s_1}\mathbf{G}_{\text{qo}}(s, z) = \frac{\partial}{\partial s_1}\mathbf{G}_{\text{qo}}(s_1, s_2)|_{(s_1, s_2)=(s, z)} \quad \text{and} \quad \frac{\partial}{\partial s_2}\mathbf{G}_{\text{qo}}(s, z) = \frac{\partial}{\partial s_2}\mathbf{G}_{\text{qo}}(s_1, s_2)|_{(s_1, s_2)=(s, z)}.$$

Lemma 5.3 (Symmetry properties of \mathbf{G}_{qo}). Let $\mathbf{G}_{\text{qo}}: \mathbb{C} \times \mathbb{C} \rightarrow \mathbb{C}^{p \times m^2}$ be the quadratic-output transfer function of the system (5.1) as defined in (5.12b). Then for any $\mathbf{U} \in \mathbb{C}^{m \times \ell}$ and $\mathbf{v} \in \mathbb{C}^m$:

$$\mathbf{G}_{\text{qo}}(s, z)(\mathbf{U} \otimes \mathbf{v}) = \mathbf{G}_{\text{qo}}(z, s)(\mathbf{v} \otimes \mathbf{U}),\tag{5.14}$$

$$\frac{\partial}{\partial s_1}\mathbf{G}_{\text{qo}}(s, z)(\mathbf{U} \otimes \mathbf{v}) = \frac{\partial}{\partial s_2}\mathbf{G}_{\text{qo}}(z, s)(\mathbf{v} \otimes \mathbf{U}).\tag{5.15}$$

◇

Proof of Lemma 5.3. We first prove a more general identity that involves only the quadratic-output matrix \mathbf{M} . For any $\mathbf{X} \in \mathbb{C}^{n \times n}$ and $\mathbf{z} \in \mathbb{C}^\ell$, we have, by (2.14), that

$$\mathbf{M}(\mathbf{X} \otimes \mathbf{z}) = \mathbf{M}\mathbf{K}_{nn}(\mathbf{z} \otimes \mathbf{X}) = \begin{bmatrix} \text{vec}(\mathbf{M}_1)^\top \mathbf{K}_{nn} \\ \text{vec}(\mathbf{M}_2)^\top \mathbf{K}_{nn} \\ \vdots \\ \text{vec}(\mathbf{M}_p)^\top \mathbf{K}_{nn} \end{bmatrix} (\mathbf{z} \otimes \mathbf{X}) = \begin{bmatrix} \text{vec}(\mathbf{M}_1)^\top \\ \text{vec}(\mathbf{M}_2)^\top \\ \vdots \\ \text{vec}(\mathbf{M}_p)^\top \end{bmatrix} (\mathbf{z} \otimes \mathbf{X}),$$

where \mathbf{K}_{nn} is the perfect-shuffle matrix defined in (2.11), and the last equality follows from (2.15) along with the previous assumption that $\mathbf{M}_k = \mathbf{M}_k^\top$ for all k . In aggregate, this yields

$$\mathbf{M}(\mathbf{X} \otimes \mathbf{z}) = \mathbf{M}(\mathbf{z} \otimes \mathbf{X}).\tag{5.16}$$

Then, by (5.16) and the mixed product property (2.9), we have for any $\mathbf{U} \in \mathbb{C}^{m \times \ell}$ and $\mathbf{v} \in \mathbb{C}^m$ that

$$\begin{aligned} \mathbf{G}_{\text{qo}}(s, z)(\mathbf{U} \otimes \mathbf{v}) &= \mathbf{M}((s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} \otimes (z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B})(\mathbf{U} \otimes \mathbf{v}) \\ &= \mathbf{M}((s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U} \otimes (z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{v}) \\ &= \mathbf{M}((z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{v} \otimes (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}) \\ &= \mathbf{M}((z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} \otimes (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B})(\mathbf{v} \otimes \mathbf{U}) = \mathbf{G}_{\text{qo}}(z, s)(\mathbf{v} \otimes \mathbf{U}), \end{aligned}$$

proving (5.14). The second identity (5.15) follows similarly: we have

$$\begin{aligned} \frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(s, z)(\mathbf{U} \otimes \mathbf{v}) &= \mathbf{M}(-(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} \otimes (z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B})(\mathbf{U} \otimes \mathbf{v}) \\ &= \mathbf{M}(-(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U} \otimes (z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\mathbf{v}) \\ &= \mathbf{M}((z\mathbf{E} - \mathbf{A})^{-1}\mathbf{B} \otimes -(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{E}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B})(\mathbf{v} \otimes \mathbf{U}) \\ &= \frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(z, s)(\mathbf{v} \otimes \mathbf{U}), \end{aligned}$$

proving (5.15). □

We observe that in the single-input, single-output setting, Lemma 5.3 implies that the scalar-valued \mathbf{G}_{qo} is symmetric with respect to the interchange of its arguments, i.e., if $m = p = 1$:

$$\mathbf{G}_{\text{qo}}(s, z) = \mathbf{G}_{\text{qo}}(z, s) \text{ for all } s, z \in \mathbb{C}.$$

5.3.3 Gramians and the observability energy functional

Here, we introduce the notion of observability and the algebraic system Gramians of an LQO system (5.1) following the discussion in [16, 28]. As already pointed out in Section 5.3.1, the concept of reachability from linear systems theory is inherited by linear quadratic-output systems since the state-to-input map of (5.1) is wholly linear. Thus, the *infinite reachability Gramian* of a linear quadratic-output system (5.1) is precisely the same as that of the linear system defined in (2.40). In the interest of self-containment, we recall that the infinite-time reachability Gramian is the SPSD matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ defined as

$$\mathbf{P} = \int_0^\infty e^{\mathbf{E}^{-1}\mathbf{A}\tau} \mathbf{E}^{-1}\mathbf{B} \left(e^{\mathbf{E}^{-1}\mathbf{A}\tau} \mathbf{E}^{-1}\mathbf{B} \right)^\top d\tau.$$

The reachability energy $\ell_r(\check{\mathbf{x}})$ defined according to (2.70) of a state $\check{\mathbf{x}} \in \mathbb{R}^n$ is thus also unchanged. Moreover, recall from Section 2.3 that if the triple $(\mathbf{E}, \mathbf{A}, \mathbf{B})$ in (5.1) is controllable, then \mathbf{P} in (2.39) is in fact SPD, and uniquely satisfies the generalized Lyapunov equation in (2.43), i.e.,

$$\mathbf{A}\mathbf{P}\mathbf{E}^\top + \mathbf{E}\mathbf{P}\mathbf{A}^\top + \mathbf{B}\mathbf{B}^\top = \mathbf{0}_{n \times n}.$$

By contrast, observability is dependent on the state-to-output map of a system and is thus different for systems such as (5.1) compared to our discussion in Section 2.3. We recall that the energy associated with observing the state of a generic nonlinear system under zero external forcing with the initial condition $\tilde{\mathbf{x}} \in \mathbb{R}^n$ over an infinite time horizon is given by

$$\ell_o(\tilde{\mathbf{x}}) = \frac{1}{2} \int_0^\infty \|\mathbf{y}(\tau)\|_2^2 d\tau, \quad \mathbf{x}(0) = \tilde{\mathbf{x}}, \quad \mathbf{u}(t) = \mathbf{0}_m, \quad (5.17)$$

see [198, Definition 3.1], [199]. The notion of observability for generic nonlinear systems is more nuanced than the linear case; we refer the reader to [154, Ch. 3.2] for a detailed treatment. Fortunately, for the setting of linear quadratic-output systems (5.1), the usual notion will suffice. A state $\tilde{\mathbf{x}} \in \mathbb{R}^{n \times n}$ is *unobservable* if ℓ_o is zero, and the linear quadratic-output system (5.1) is *completely observable* if the set of unobservable states consists of only the zero vector. For a review of reachability and observability energy functionals in the context of nonlinear balancing, see [198, 199]. For linear systems, the reachability and observability energy functionals (2.70) are quadratic polynomials of the state vector \mathbf{x} ; this is not the case for generic nonlinear systems (though, these energy functionals can still be analytically characterized in special cases). In the recent work [16], Balicki and Gugercin write down the observability energy functionals of linear dynamical systems with *polynomial* output functions [16, Theorem 1]. Obviously, the LQO systems (5.1) are a special instance of this. Since balancing-based methods are not the primary focus of this chapter, we refer the reader to [16] for further discussion.

Alternative (albeit, partial) characterizations of the observability subspace and energy of the system (5.1) were developed in [28]. Therein, the authors derive an algebraic Gramian, the so-called *quadratic-output (QO) observability Gramian*, that aims to generalize the classical observability Gramian $\mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} \in \mathbb{R}^{n \times n}$ of the linear system (2.25), as defined in (2.44). The QO observability Gramian is defined as follows: Consider the LQO system \mathcal{G}_{lqo} in (5.1) and recall the *infinite-time observability Gramian* $\mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} \in \mathbb{R}^{n \times n}$ of its linear-output subsystem (2.25) introduced in Definition 2.34 and defined according to (2.40):

$$\mathbf{Q}_{\text{lo}} = \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{C}^\top \left(\mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{C}^\top \right)^\top d\tau.$$

Recall further that \mathbf{Q}_{lo} is the unique SPSD solution to the generalized Lyapunov equation (2.44), i.e.,

$$\mathbf{A}^\top \mathbf{Q}_{\text{lo}} \mathbf{E} + \mathbf{E}^\top \mathbf{Q}_{\text{lo}} \mathbf{A} + \mathbf{C}^\top \mathbf{C} = \mathbf{0}_{n \times n}.$$

Define also for each $k = 1, 2, \dots, p$ the intermediate matrices $\mathbf{Q}_{\text{qo},k} \in \mathbb{R}^{n \times n}$ by

$$\mathbf{Q}_{\text{qo},k} \stackrel{\text{def}}{=} \int_0^\infty \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau_1} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau_2} \mathbf{E}^{-1} \mathbf{B} \left(\mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau_1} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau_2} \mathbf{E}^{-1} \mathbf{B} \right)^\top d\tau_1 d\tau_2. \quad (5.18)$$

Then, the QO observability Gramian of \mathcal{G}_{lqo} in (5.1) is the matrix $\mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E} \in \mathbb{R}^{n \times n}$ defined via

$$\mathbf{Q}_{\text{lqo}} \stackrel{\text{def}}{=} \mathbf{Q}_{\text{lo}} + \mathbf{Q}_{\text{qo}} \quad \text{and} \quad \mathbf{Q}_{\text{qo}} \stackrel{\text{def}}{=} \sum_{k=1}^p \mathbf{Q}_{\text{qo},k} \in \mathbb{R}^{n \times n}. \quad (5.19)$$

For an asymptotically stable system (5.1), the matrix $\mathbf{Q}_{\text{lqo}} \in \mathbb{R}^{n \times n}$ exists, is SPSD, and uniquely satisfies a type of generalized Lyapunov equation.

Proposition 5.4 (Generalized Lyapunov equation for \mathbf{Q}_{lqo} [28, 176]). Consider an asymptotically stable linear quadratic-output system \mathcal{G}_{lqo} as in (5.1), and let $\mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{Q}_{\text{lqo}} \in \mathbb{R}^{n \times n}$ be the reachability and quadratic-output observability Gramians of \mathcal{G}_{lqo} defined according to (2.39) and (5.19). Then, $\mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E}$ is SPSD, and \mathbf{Q}_{lqo} is the unique solution of the generalized Lyapunov equation

$$\mathbf{A}^\top \mathbf{Q}_{\text{qo}} \mathbf{E} + \mathbf{E}^\top \mathbf{Q}_{\text{qo}} \mathbf{A} + \mathbf{C}^\top \mathbf{C} + \sum_{k=1}^p \mathbf{M}_k \mathbf{P} \mathbf{M}_k = \mathbf{0}_{n \times n}. \quad (5.20)$$

◇

This result was proven rigorously in [28, Lemma 2.1] for the case of (5.1) with $\mathbf{E} = \mathbf{I}_n$ and a single quadratic output, i.e., $p = 1$ with $\mathbf{C} = \mathbf{0}_{p \times n}$, whereas [176, Theorem 3.3] proves this for a general (possibly singular) \mathbf{E} . Here, we provide a simplified proof for the special case of nonsingular \mathbf{E} .

Proof of Proposition 5.4. The fact that $\mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E}$ is SPSD follows from the definition of its constituent parts in (2.40) and (5.18). Recall first that $\mathbf{Q}_{\text{lo}} \in \mathbb{R}^{n \times n}$ in (2.40) is the unique SPSD solution of the generalized Lyapunov equation (2.44). For each $k = 1, \dots, p$, $\mathbf{Q}_{\text{qo},k} \in \mathbb{R}^{n \times n}$ in (5.18) can be rewritten as

$$\mathbf{Q}_{\text{qo},k} = \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{M}_k \mathbf{P} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau} \mathbf{E}^{-1} d\tau. \quad (5.21)$$

We claim then that $\mathbf{Q}_{\text{qo},k}$ solves the generalized Lyapunov equation

$$\mathbf{A}^\top \mathbf{Q}_{\text{qo},k} \mathbf{E} + \mathbf{E}^\top \mathbf{Q}_{\text{qo},k} \mathbf{A} + \mathbf{M}_k \mathbf{P} \mathbf{M}_k = \mathbf{0}_{n \times n}. \quad (5.22)$$

This can be deduced as follows:

$$\begin{aligned} \mathbf{A}^\top \mathbf{Q}_{\text{qo},k} \mathbf{E} + \mathbf{E}^\top \mathbf{Q}_{\text{qo},k} \mathbf{A} &= \int_0^\infty \left(\mathbf{A}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{M}_k \mathbf{P} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau} \right. \\ &\quad \left. + e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{M}_k \mathbf{P} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau} \mathbf{E}^{-1} \mathbf{A} \right) d\tau \\ &= \int_0^\infty \frac{d}{d\tau} \left(e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{M}_k \mathbf{P} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau} \right) d\tau \\ &= \lim_{t \rightarrow \infty} \left(e^{\mathbf{A}^\top \mathbf{E}^{-\top} t} \mathbf{M}_k \mathbf{P} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} t} \right) \Big|_{t=0}^{t=\tau} \\ &= -\mathbf{M}_k \mathbf{P} \mathbf{M}_k. \end{aligned}$$

Moreover, the solution $\mathbf{Q}_{\text{qo},k}$ to (5.22) is unique by [4, Corollary 6.3] and the asymptotic stability assumption. Summing up the equations (2.44) and (5.22) over all k results in (5.20). Obviously, $\mathbf{Q}_{\text{lqo}} = \mathbf{Q}_{\text{lo}} + \mathbf{Q}_{\text{qo}}$ as defined in (5.19) is a solution to (5.20) by this logic. By the asymptotic stability of (5.1), it follows that any solution to (5.20) is unique [4, Corollary 6.3], and so \mathbf{Q}_{lqo} uniquely solves (5.20) as claimed. \square

For the ultimate goal of model reduction, the Gramian $\mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E}$ is informative about the observability subspace and energies of an LQO system (5.1) in the following sense.

Proposition 5.5 (Observability of a system (5.1) according to (5.19) [28]). Consider an LQO system \mathcal{G}_{lqo} as in (5.1). Let $\mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{Q}_{\text{lqo}} \in \mathbb{R}^{n \times n}$ be the reachability and QO observability Gramians of \mathcal{G}_{lqo} defined according to (2.39) and (5.19), and suppose additionally that $\mathbf{P} \succ 0$. For $\mathbf{u} = \mathbf{0}_m$ and all states $\check{\mathbf{x}} \in \mathbb{R}^{n \times n}$

$$\ell_o(\check{\mathbf{x}}) \leq \check{\mathbf{x}}^\top \mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E} \check{\mathbf{x}} (1 + \check{\mathbf{x}} \mathbf{P}^{-1} \check{\mathbf{x}}) = \tilde{\ell}_o(\check{\mathbf{x}}) \check{\mathbf{x}} (1 + \ell_r(\check{\mathbf{x}})), \quad (5.23)$$

where ℓ_r and ℓ_o are the reachability and observability energy functionals defined according to (2.70) and (5.17), and $\tilde{\ell}_o: \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by $\tilde{\ell}_o(\check{\mathbf{x}}) \stackrel{\text{def}}{=} \check{\mathbf{x}}^\top \mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E} \check{\mathbf{x}}$. \diamond

5.3.4 Generic model reduction of linear quadratic-output systems

We now consider specific details regarding the construction of LQO reduced models (5.2). As in the setting of Section 2.4, we consider reduced models determined by Petrov-Galerkin projection. The fundamental idea is the same; we choose left and right approximation subspaces $\text{Range}(\mathbf{W})$ and $\text{Range}(\mathbf{V})$ spanned by the model reduction bases stored in the columns of $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{n \times r}$, so that $\mathbf{x} \approx \mathbf{V} \tilde{\mathbf{x}}$, and the Petrov-Galerkin orthogonality condition (2.57) is satisfied. The only deviation from the linear setting is how the approximation affects the output equation:

$$\mathbf{y}(t) = \mathbf{C} \mathbf{x}(t) + \mathbf{M}(\mathbf{x}(t) \otimes \mathbf{x}(t)) \approx \mathbf{C} \mathbf{V} \tilde{\mathbf{x}}(t) + \mathbf{M}(\mathbf{V} \otimes \mathbf{V})(\tilde{\mathbf{x}}(t) \otimes \tilde{\mathbf{x}}(t)).$$

The reduced-order model (5.2) by projection is thereby given as

$$\tilde{\mathbf{E}} = \mathbf{W}^\top \mathbf{E} \mathbf{V}, \quad \tilde{\mathbf{A}} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \tilde{\mathbf{B}} = \mathbf{W}^\top \mathbf{B}, \quad \tilde{\mathbf{C}} = \mathbf{C} \mathbf{V}, \quad \text{and} \quad \tilde{\mathbf{M}} = \mathbf{M}(\mathbf{V} \otimes \mathbf{V}). \quad (5.24)$$

Remark 5.6. Storing $\mathbf{V} \otimes \mathbf{V}$ can be infeasible for large n . Thus, in any practical implementation, it is usually preferred to compute $\tilde{\mathbf{M}}$ by first unpacking its alternative representation in (5.8), and then projecting the individual quadratic-output matrices. The projection is then specified by

$$\tilde{\mathbf{M}}_k = \mathbf{V}^\top \mathbf{M}_k \mathbf{V}, \quad k = 1, \dots, p. \quad (5.25)$$

Henceforth, when computing $\tilde{\mathbf{M}} = \mathbf{M}(\mathbf{V} \otimes \mathbf{V})$ by projection we assume that it is done according to (5.25). \diamond

Algorithm 5.3.1: Linear quadratic-output Balanced Truncation (LQO-BT) [28].

Input: E, A, B, C, M from (5.1), order $1 \leq r < n$.

Output: $\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{M}$ —state-space matrices of (5.2).

- 1 Compute Cholesky factors $R \in \mathbb{R}^{n \times n}$ and $L_{\text{lqo}} \in \mathbb{R}^{n \times n}$ of $P \in \mathbb{R}^{n \times n}$ in (2.39) and $Q_{\text{lqo}} \in \mathbb{R}^{n \times n}$ in (5.19) from the generalized Lyapunov equations (2.43) and (5.20).
- 2 Compute the singular value decomposition of $L_{\text{lqo}}^T E R$ partitioned according to

$$L_{\text{lqo}}^T E R = U \Sigma Y^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix}$$

for $\Sigma_1 \in \mathbb{R}^{r \times r}$, $U_1, Y_1 \in \mathbb{R}^{n \times r}$ and $\Sigma_2 \in \mathbb{R}^{(n-r) \times (n-r)}$, $U_2, Y_2 \in \mathbb{R}^{n \times (n-r)}$.

- 3 Compute the model reduction basis matrices $W, V \in \mathbb{R}^{n \times r}$ as

$$W = L_{\text{lqo}} Y_1 \Sigma_1^{-1/2}, \quad V = R U_1 \Sigma_1^{-1/2}.$$

- 4 Compute the reduced model $\tilde{\mathcal{G}}_{\text{lqo}}$ by projection (5.24) using W and V .
-

As we recall from Section 2.4, two well-established classes of methods for system-theoretic model reduction are those based on the balancing of energy functionals and truncation of states, and the rational interpolation of system transfer functions. For LQO systems (5.1), we review the generalization of balanced truncation called linear quadratic-output BT (LQO-BT) from [28], and the tangential interpolation framework of [62]. We review these specifically because the results presented later in this chapter build upon the work of [28, 62]; we also use these as benchmark methods for the algorithms we propose in Chapter 6.

Linear quadratic-output balanced truncation.

A comprehensive theory for the balancing of nonlinear dynamical systems was developed in [198, 199]. While theoretically sound, computing the required reachability and observability energy functionals for generic nonlinear systems is a challenging task. Recent progress has been made in this realm; specifically, [16] proposes a computationally scalable implementation for the nonlinear balancing of linear systems with polynomial output functions. We refer to [16] and more generally [120] for further details.

A different attempt to generalize the classical BT of [151, 152] to systems of the form (5.1), called LQO-BT, was made in [28]. The fundamental idea is exactly that of the Lyapunov BT approach discussed in Section 2.4.3, with the QO observability Gramian $E^T Q_{\text{lqo}} E$ defined in (5.19) playing the role of the classical infinite observability Gramian $E^T Q_{\text{lo}} E$ in Algorithm 2.4.3. An implementation of LQO-BT is presented in Algorithm 5.3.1. Under

the hood, Algorithm 5.3.1 is balancing the quadratic cost functionals $\ell_r(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^\top \mathbf{P}^{-1} \tilde{\mathbf{x}}$ and $\tilde{\ell}_o(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^\top \mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E} \tilde{\mathbf{x}}$; the states corresponding to the smallest singular values of $\mathbf{L}_{\text{lqo}}^\top \mathbf{E} \mathbf{R} \in \mathbb{R}^{n \times n}$, which are precisely those states $\tilde{\mathbf{x}}$ that are difficult to control (observe) according to $\ell_r(\tilde{\mathbf{x}})$ ($\tilde{\ell}_o(\tilde{\mathbf{x}})$), are thereby truncated. Note that the singular values of $\mathbf{L}_{\text{lqo}}^\top \mathbf{E} \mathbf{R}$ in Algorithm 5.3.1 take the place of the usual Hankel singular values in Algorithm 2.4.3. This replacement of $\mathbf{E}^\top \mathbf{Q}_{\text{lqo}} \mathbf{E}$ with $\mathbf{E}^{\text{trans}} \mathbf{Q}_{\text{lqo}} \mathbf{E}$ is justified by Proposition 5.5; the states $\tilde{\mathbf{x}}$ that produce small energies $\tilde{\ell}_o(\tilde{\mathbf{x}})$ will necessarily produce small $\ell_o(\tilde{\mathbf{x}})$, where ℓ_o is the true nonlinear observability energy functional defined in (5.17).

Algorithm 5.3.1 preserves the asymptotic stability of the original system (see [28, Theorem 3.1]) although the resulting reduced model is not necessarily guaranteed to be minimal [28, Remark 3.2] or balanced [28, Remark 2.4], as is the case of linear BT. Several extensions of this work have also been proposed, for instance, to systems of differential algebraic equations (DAEs) [176]. As with Algorithm 2.4.3, the dominant cost of Algorithm 5.3.1 is in solving the large-scale generalized Lyapunov equations (2.43) and (5.20) for the factors \mathbf{R} and \mathbf{L}_{lqo} . Fortunately, there exist efficient numerical algorithms for computing (inexact) low-rank Cholesky factors; we refer to [34, 121, 203] for details.

Rational interpolation of \mathbf{G}_{lqo} and \mathbf{G}_{qo} .

The tangential interpolation theory for the model reduction of linear systems (2.25) reviewed in Section 2.4.1 can be generalized to nonlinear systems with the concept of *subsystem interpolation*. As the name suggests, this amounts to the rational interpolation of the subsystem transfer functions \mathbf{G}_{lqo} and \mathbf{G}_{qo} in the direction of specified tangent vectors. Diaz et al. [62] introduce an overarching framework for the subsystem interpolation of dynamical systems with up to quadratic-bilinear dynamics and quadratic-bilinear outputs; this general model class includes (5.1) as a special case [62, Corollary 1]. We focus on this idea here; other similar approaches for the interpolatory model reduction of linear quadratic-output systems (5.1) have been proposed in [49, 62, 87, 188, 215].

Formally, the Lagrange version of the tangential interpolation problem from [62] can be stated as follows: Given a collection of complex shifts $\sigma_1, \dots, \sigma_{2k} \in \mathbb{C}$ in conjunction with left and right tangential direction vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{2k} \in \mathbb{C}^p$ and $\mathbf{r}_1, \dots, \mathbf{r}_{2k} \in \mathbb{C}^m$, construct $\tilde{\mathbf{G}}_{\text{lqo}}$ in (5.2) by an appropriate choice of \mathbf{W} and \mathbf{V} so that the transfer functions $\tilde{\mathbf{G}}_{\text{lqo}}$ and

$\tilde{\mathbf{G}}_{\text{qo}}$ of $\tilde{\mathcal{G}}_{\text{lqo}}$ satisfy the tangential interpolation conditions

$$\begin{aligned}
\mathbf{G}_{\text{lo}}(\sigma_j)\mathbf{r}_j &= \tilde{\mathbf{G}}_{\text{lo}}(\sigma_j)\mathbf{r}_j, \\
\mathbf{G}_{\text{lo}}(\sigma_{2j})\mathbf{r}_j &= \tilde{\mathbf{G}}_{\text{lo}}(\sigma_{2j})\mathbf{r}_j, \\
\boldsymbol{\ell}_j^\top \mathbf{G}_{\text{lo}}(\sigma_{2j}) &= \boldsymbol{\ell}_j^\top \tilde{\mathbf{G}}_{\text{lo}}(\sigma_{2j}), \\
\mathbf{G}_{\text{qo}}(\sigma_j, \sigma_j)(\mathbf{r}_j \otimes \mathbf{r}_j) &= \tilde{\mathbf{G}}_{\text{qo}}(\sigma_{2j}, \sigma_{2j})(\mathbf{r}_{2j} \otimes \mathbf{r}_{2j}), \\
\mathbf{G}_{\text{qo}}(\sigma_{2j}, \sigma_{2j})(\mathbf{r}_{2j} \otimes \mathbf{r}_{2j}) &= \tilde{\mathbf{G}}_{\text{qo}}(\sigma_{2j}, \sigma_{2j})(\mathbf{r}_{2j} \otimes \mathbf{r}_{2j}), \\
\text{and } \boldsymbol{\ell}_j^\top \mathbf{G}_{\text{qo}}(\sigma_j, \sigma_j)(\mathbf{I}_m \otimes \mathbf{r}_j) &= \boldsymbol{\ell}_j^\top \tilde{\mathbf{G}}_{\text{qo}}(\sigma_j, \sigma_j)(\mathbf{I}_m \otimes \mathbf{r}_j),
\end{aligned} \tag{5.26}$$

for all $j = 1, \dots, k$. For completeness, we restate the result of [62, Corollary 1] here.

Theorem 5.7 (Lagrange interpolation of systems (5.1) [62]). Suppose that \mathcal{G}_{lqo} is an LQO system as in (5.1), and that $\tilde{\mathcal{G}}_{\text{lqo}}$ is a reduced model as in (5.2) constructed by projection with \mathbf{V} and \mathbf{W} . Consider the interpolation points $\sigma_1, \dots, \sigma_{2k} \in \mathbb{C}$ such that $\sigma_i \mathbf{E} - \mathbf{A}$ and $\sigma_i \tilde{\mathbf{E}} - \tilde{\mathbf{A}}$ are nonsingular for all $i = 1, \dots, 2k$, and the left and right tangential direction vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{2k} \in \mathbb{C}^p$ and $\mathbf{r}_1, \dots, \mathbf{r}_{2k} \in \mathbb{C}^m$. Let $\mathcal{K}: \mathbb{C} \rightarrow \mathbb{C}^{n \times n}$ be the matrix resolvent $\mathcal{K}(s) = s\mathbf{E} - \mathbf{A}$. Suppose for $r = 2k$ that $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{n \times r}$ have full rank and satisfy

$$\begin{aligned}
&\text{span} [\mathcal{K}(\sigma_j)^{-1} \mathbf{B} \mathbf{r}_j, \mathcal{K}(\sigma_{2j})^{-1} \mathbf{B} \mathbf{r}_j] \subset \text{Range}(\mathbf{V}), \\
&\text{and } \text{span} \left[(\boldsymbol{\ell}_j^\top \mathbf{C} \mathcal{K}(\sigma_j)^{-1})^\top, (\boldsymbol{\ell}_j^\top \mathbf{M} (\mathcal{K}(\sigma_j)^{-1} \mathbf{B} \otimes \mathcal{K}(\sigma_j)^{-1} \mathbf{B} \mathbf{r}_j))^\top \right] \subset \text{Range}(\mathbf{W}).
\end{aligned} \tag{5.27}$$

Then, $\tilde{\mathcal{G}}_{\text{lqo}}$ will satisfy the tangential interpolation conditions in (5.26) for all $j = 1, \dots, k$. \diamond

While Theorem 5.7 shows how to construct tangential interpolants of the form (5.2), there are still many open questions with regard to interpolatory model reduction of LQO systems. For instance, Theorem 5.7 does not consider any higher-order (Hermite) interpolation conditions; moreover, the placement of interpolation points, selection of tangent directions, and type of interpolation one should enforce (e.g., Lagrange or Hermite conditions) to guarantee quality surrogates is unknown for systems of the form (5.1). It is important to recall that, in the linear setting, \mathcal{H}_2 -optimal reduced models are necessarily bi-tangential *Hermite* interpolants of the full-order system, and the optimal interpolation points are the mirror images of the reduced model poles; see Theorem 2.45. Similar results hold for other classes of weakly nonlinear systems; e.g., in the \mathcal{H}_2 model reduction of *bilinear* dynamical systems, optimal approximants satisfy so-called multipoint rational interpolation conditions that respect the underlying Volterra series representation of the full-order model [24, 77, 78], *not* subsystem interpolation conditions like those in Theorem 5.7. This is also true for quadratic-bilinear systems; see [50]. It is thus natural to question whether there exist connections between \mathcal{H}_2 -optimality and transfer function interpolation for systems of the form (5.1). As one of the major theoretical contributions of this dissertation, we investigate and establish this connection in Chapter 6.

5.4 Two computationally tractable expressions for the system \mathcal{H}_2 norm and inner product

In order to assess the performance quality of a surrogate model (5.2), we require an appropriate measure for quantifying the approximation error induced by replacing the full-order system (5.1) with the lower-order surrogate (5.2). Here, we use the \mathcal{H}_2 norm of an LQO system; this can be formulated in the time domain (that is, in terms of the Volterra kernels in (5.10)) or in the frequency domain (in terms of the transfer functions in (5.12)). In either case, the \mathcal{H}_2 measure is derived from an underlying Hilbert space structure that we will leverage to derive optimality conditions later on. At the end of this section, we present two new formulas for computing the system \mathcal{H}_2 norm and inner product for systems of the form (5.1). We begin with the time-domain characterization.

Definition 5.8 (Time-domain formulation of the \mathcal{H}_2 system norm [186, Definition 2.1], [28, Definition 3.1]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2) with the Volterra kernels \mathbf{g}_{lo} , \mathbf{g}_{qo} and $\tilde{\mathbf{g}}_{\text{lo}}$, $\tilde{\mathbf{g}}_{\text{qo}}$ defined according to (5.10). The \mathcal{H}_2 inner product of \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ is defined as

$$\begin{aligned} \left\langle \mathcal{G}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \right\rangle_{\mathcal{H}_2} &\stackrel{\text{def}}{=} \int_0^\infty \text{tr} \left(\mathbf{g}_{\text{lo}}(\tau) \tilde{\mathbf{g}}_{\text{lo}}(\tau)^\top \right) d\tau \\ &\quad + \int_0^\infty \int_0^\infty \text{tr} \left(\mathbf{g}_{\text{qo}}(\tau_1, \tau_2) \tilde{\mathbf{g}}_{\text{qo}}(\tau_1, \tau_2)^\top \right) d\tau_1 d\tau_2. \end{aligned} \quad (5.28)$$

Likewise, the \mathcal{H}_2 norm of \mathcal{G}_{lqo} is defined as

$$\|\mathcal{G}_{\text{lqo}}\|_{\mathcal{H}_2} \stackrel{\text{def}}{=} \left(\int_0^\infty \|\mathbf{g}_{\text{lo}}(\tau)\|_{\mathbb{F}}^2 d\tau + \int_0^\infty \int_0^\infty \|\mathbf{g}_{\text{qo}}(\tau_1, \tau_2)\|_{\mathbb{F}}^2 d\tau_1 d\tau_2 \right)^{\frac{1}{2}}. \quad (5.29)$$

◇

Next, we present the frequency-domain formulation.

Definition 5.9 (Frequency-domain formulation of the \mathcal{H}_2 system norm [87]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2) with the transfer functions \mathbf{G}_{lo} , \mathbf{G}_{qo} and $\tilde{\mathbf{G}}_{\text{lo}}$, $\tilde{\mathbf{G}}_{\text{qo}}$ defined according to (5.12). The \mathcal{H}_2 inner product of \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ is defined as

$$\begin{aligned} \left\langle \mathcal{G}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \right\rangle_{\mathcal{H}_2} &\stackrel{\text{def}}{=} \frac{1}{2\pi} \int_{-\infty}^\infty \text{tr} \left(\overline{\mathbf{G}}_{\text{lo}}(-i\omega) \tilde{\mathbf{G}}_{\text{lo}}(i\omega)^\top \right) d\omega \\ &\quad + \frac{1}{(2\pi)^2} \int_{-\infty}^\infty \int_{-\infty}^\infty \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -i\omega_2) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, i\omega_2)^\top \right) d\omega_1 d\omega_2. \end{aligned} \quad (5.30)$$

Likewise, the \mathcal{H}_2 norm of \mathcal{G}_{lqo} is defined as

$$\|\mathcal{G}_{\text{lqo}}\|_{\mathcal{H}_2} \stackrel{\text{def}}{=} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \|\mathbf{G}_{\text{lo}}(i\omega)\|_{\text{F}}^2 d\omega + \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \|\mathbf{G}_{\text{qo}}(i\omega_1, i\omega_2)\|_{\text{F}}^2 d\omega_1 d\omega_2 \right)^{\frac{1}{2}}. \quad (5.31)$$

◇

Although the dynamics of the systems in (5.1) and (5.2) as written are real-valued, Definition 5.9 is valid for systems with complex-valued dynamics as well. Moreover, the inner product (5.30) is real-valued for real-valued dynamical systems.

It is a direct consequence of Plancherel's relation in one- and two-variables [43] that Definitions 5.8 and 5.9 are in fact equivalent, thereby justifying our abuse of notation. The time-domain characterization in Definition 5.8 is derived from the underlying \mathcal{L}_2 Hilbert space structure of the Volterra kernels (5.10); the frequency-domain characterization is derived from the underlying \mathcal{H}_2 Hardy space structure of the transfer functions (5.12). Indeed, the inner products (and thus the induced norms) in each of (5.28) and (5.30) are respectively the sums of the relevant \mathcal{L}_2 and \mathcal{H}_2 inner products of the full-order kernels and transfer functions with the reduced-order ones.

Next, we derive a pair of expressions for the system \mathcal{H}_2 norm and inner product that are more computationally tractable compared to those in Definitions 5.8 and 5.9. Each of these generalize the analogous expressions for calculating the \mathcal{H}_2 norm and inner product for linear dynamical systems from Theorems 2.41 and 2.42.

5.4.1 Sylvester-equation based formulation

First, we show that the system \mathcal{H}_2 norm and inner product can be computed in terms of the Gramians $\mathbf{P}, \mathbf{E}^T \mathbf{Q}_{\text{lqo}} \mathbf{E} \in \mathbb{R}^{n \times n}$ and solutions to generalized Sylvester equations. This was proved in [28, Proposition 3.3] for the special case of $\mathbf{C} = \mathbf{0}_{p \times n}$ and $\mathbf{E} = \mathbf{I}_n$. A similar formula was derived independently for systems of DAEs with quadratic-output functions [176]. The stated version of Theorem 5.10 was presented in the author's previous work [186], but for $\mathbf{E} = \mathbf{I}_n$. Here, we prove the formulae for arbitrary nonsingular \mathbf{E} .

Theorem 5.10 (\mathcal{H}_2 norm and inner product via generalized Sylvester equations [176, Lemma 5.2], [186, Theorem 2.1]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2). Let $\mathbf{X} \in \mathbb{R}^{n \times r}$ and $\mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ be solutions of the generalized Sylvester equations

$$\mathbf{A} \mathbf{X} \tilde{\mathbf{E}}^T + \mathbf{E} \mathbf{X} \tilde{\mathbf{A}}^T + \mathbf{B} \tilde{\mathbf{B}}^T = \mathbf{0}_{n \times r}, \quad (5.32)$$

$$\text{and } \mathbf{A}^T \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{E}} + \mathbf{E}^T \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{A}} - \sum_{k=1}^p \mathbf{M}_k \mathbf{X} \tilde{\mathbf{M}}_k - \mathbf{C}^T \tilde{\mathbf{C}} = \mathbf{0}_{n \times r}. \quad (5.33)$$

Then, \mathbf{X} and \mathbf{Z}_{lqo} are unique, and the \mathcal{H}_2 inner product of \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ is given by

$$\left\langle \mathcal{G}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \right\rangle_{\mathcal{H}_2} = -\text{tr} \left(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}} \right) \quad (5.34)$$

$$= \text{tr} \left(\mathbf{C} \mathbf{X} \tilde{\mathbf{C}}^\top \right) + \sum_{k=1}^p \text{tr} \left(\mathbf{X}^\top \mathbf{M}_k \mathbf{X} \tilde{\mathbf{M}}_k \right). \quad (5.35)$$

If $\mathcal{G}_{\text{lqo}} = \tilde{\mathcal{G}}_{\text{lqo}}$, then $\mathbf{X} = \mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{Z}_{\text{lqo}} = -\mathbf{Q}_{\text{lqo}} \in \mathbb{R}^{n \times n}$ according to (2.39) and (5.19); thus, the \mathcal{H}_2 norm of \mathcal{G}_{lqo} is given by

$$\|\mathcal{G}_{\text{lqo}}\|_{\mathcal{H}_2}^2 = \text{tr} \left(\mathbf{B}^\top \mathbf{Q}_{\text{lo}} \mathbf{B} \right) + \text{tr} \left(\mathbf{B}^\top \mathbf{Q}_{\text{qo}} \mathbf{B} \right) = \text{tr} \left(\mathbf{B}^\top \mathbf{Q}_{\text{lqo}} \mathbf{B} \right) \quad (5.36)$$

$$= \text{tr} \left(\mathbf{C} \mathbf{P} \mathbf{C}^\top \right) + \sum_{k=1}^p \text{tr} \left(\mathbf{P} \mathbf{M}_k \mathbf{P} \mathbf{M}_k \right). \quad (5.37)$$

◇

Proof of Theorem 5.10. Throughout, let $\mathbf{g}_{\text{lo}}, \mathbf{g}_{\text{qo}}$ and $\tilde{\mathbf{g}}_{\text{lo}}, \tilde{\mathbf{g}}_{\text{qo}}$ denote the Volterra kernels of \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ defined according to (5.10). Consider the solutions $\mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ and $\mathbf{Z}_{\text{qo},k} \in \mathbb{R}^{n \times r}$ to the generalized Sylvester equations

$$\mathbf{A}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{A}} - \mathbf{C}^\top \tilde{\mathbf{C}} = \mathbf{0}_{n \times r}, \quad (5.38)$$

$$\mathbf{A}^\top \mathbf{Z}_{\text{qo},k} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{Z}_{\text{qo},k} \tilde{\mathbf{A}} - \mathbf{M}_k \mathbf{X} \tilde{\mathbf{M}}_k = \mathbf{0}_{n \times r}. \quad (5.39)$$

These solutions exist and are unique due to the spectra of $\mathbf{E}^{-1} \mathbf{A}$ and $-\tilde{\mathbf{E}}^{-1} \tilde{\mathbf{A}}$ being disjoint by the asymptotic stability assumption; the solutions $\mathbf{X} \in \mathbb{R}^{n \times r}$ and $\mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ to (5.32) and (5.33) are in fact unique by the same logic. Thus, we can write down the solutions to (5.32), (5.38) and (5.39) analytically as

$$\begin{aligned} \mathbf{X} &= \int_0^\infty e^{\mathbf{E}^{-1} \mathbf{A} \tau} \mathbf{E}^{-1} \mathbf{B} \left(e^{\tilde{\mathbf{E}}^{-1} \tilde{\mathbf{A}} \tau} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}} \right)^\top d\tau, \\ \mathbf{Z}_{\text{lo}} &= - \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{C}^\top \left(\tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} \tau} \tilde{\mathbf{C}}^\top \right)^\top d\tau \\ \text{and } \mathbf{Z}_{\text{qo},k} &= - \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{E}^{-\top} \mathbf{A}^\top \tau} \mathbf{M}_k \mathbf{X} \left(\tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{E}}^{-\top} \tilde{\mathbf{A}}^\top \tau} \tilde{\mathbf{M}}_k \right)^\top d\tau. \end{aligned} \quad (5.40)$$

Summing up equations (5.38) with (5.39) over all k produces the equation in (5.33). Thus, $\mathbf{Z}_{\text{lo}} + \sum_{k=1}^p \mathbf{Z}_{\text{qo},k}$ is a solution to (5.33), and by uniqueness it holds that

$$\begin{aligned} \mathbf{Z}_{\text{lqo}} &= \mathbf{Z}_{\text{lo}} + \sum_{k=1}^p \mathbf{Z}_{\text{qo},k} = - \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau} \mathbf{C}^\top \left(\tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} \tau} \tilde{\mathbf{C}}^\top \right)^\top d\tau \\ &\quad - \sum_{k=1}^p \int_0^\infty \int_0^\infty \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau_1} \mathbf{M}_k e^{\mathbf{E}^{-1} \mathbf{A} \tau_2} \mathbf{E}^{-1} \mathbf{B} \left(\tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} \tau_1} \tilde{\mathbf{M}}_k e^{\tilde{\mathbf{E}}^{-1} \tilde{\mathbf{A}} \tau_2} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}} \right)^\top d\tau_1 d\tau_2. \end{aligned}$$

Note that we have substituted into each $\mathbf{Z}_{\text{qo},k}$ the analytic expression of \mathbf{X} in (5.40). Left and right multiplication of \mathbf{Z}_{lqo} with \mathbf{B}^\top and $\tilde{\mathbf{B}}$, respectively, as well as subsequently taking the trace of both sides, reveal

$$\begin{aligned} -\text{tr}\left(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}}\right) &= \int_0^\infty \text{tr}\left(\mathbf{C} e^{\mathbf{E}^{-1}\mathbf{A}\tau} \mathbf{E}^{-1} \mathbf{B} \left(\tilde{\mathbf{C}} e^{\tilde{\mathbf{E}}^{-1}\tilde{\mathbf{A}}\tau} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}}\right)^\top\right) d\tau \\ &+ \sum_{k=1}^p \int_0^\infty \int_0^\infty \text{tr}\left(\mathbf{B}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau_1} \mathbf{M}_k e^{\mathbf{E}^{-1}\mathbf{A}\tau_2} \mathbf{E}^{-1} \mathbf{B} \right. \\ &\quad \left. \times \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} \tau_1} \tilde{\mathbf{M}}_k e^{\tilde{\mathbf{E}}^{-1}\tilde{\mathbf{A}}\tau_2} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}}\right)^\top\right) d\tau_1 d\tau_2, \end{aligned}$$

where we have used the fact that the trace is invariant under cyclic permutation and transposition of matrices to simplify the expression. By definition of the kernels \mathbf{g}_{lo} and $\tilde{\mathbf{g}}_{\text{lo}}$ in (5.10), it follows already that first term in the above expression for $-\text{tr}\left(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}}\right)$ is precisely the first term in the time-domain expression of the \mathcal{H}_2 inner product (5.28). For the second term, using (2.7) we are able to express the kernel \mathbf{g}_{qo} in (5.10b) as

$$\mathbf{g}_{\text{qo}}(t_1, t_2) = \begin{bmatrix} \text{vec}\left(\mathbf{B}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} t_1} \mathbf{M}_1 e^{\mathbf{E}^{-1}\mathbf{A}t_2} \mathbf{E}^{-1} \mathbf{B}\right)^\top \\ \vdots \\ \text{vec}\left(\mathbf{B}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} t_1} \mathbf{M}_p e^{\mathbf{E}^{-1}\mathbf{A}t_2} \mathbf{E}^{-1} \mathbf{B}\right)^\top \end{bmatrix},$$

and likewise for $\tilde{\mathbf{g}}_{\text{qo}}$. Then, by (2.8) in Proposition 2.5 we have, for all $t_1, t_2 \geq 0$, that

$$\begin{aligned} \text{tr}\left(\mathbf{g}_{\text{qo}}(t_1, t_2) \tilde{\mathbf{g}}_{\text{qo}}(t_1, t_2)^\top\right) &= \sum_{k=1}^p \text{vec}\left(\mathbf{B}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} t_1} \mathbf{M}_k e^{\mathbf{E}^{-1}\mathbf{A}t_2} \mathbf{E}^{-1} \mathbf{B}\right)^\top \\ &\quad \times \text{vec}\left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} t_1} \tilde{\mathbf{M}}_k e^{\tilde{\mathbf{E}}^{-1}\tilde{\mathbf{A}}t_2} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}}\right) \\ &= \sum_{k=1}^p \text{tr}\left(\mathbf{B}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} t_1} \mathbf{M}_k e^{\mathbf{E}^{-1}\mathbf{A}t_2} \mathbf{E}^{-1} \mathbf{B} \right. \\ &\quad \left. \times \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} t_1} \tilde{\mathbf{M}}_k e^{\tilde{\mathbf{E}}^{-1}\tilde{\mathbf{A}}t_2} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}}\right)^\top\right). \end{aligned}$$

Integrating both sides over t_1 and t_2 from zero to ∞ yields

$$\begin{aligned} \int_0^\infty \int_0^\infty \text{tr}\left(\mathbf{g}_{\text{qo}}(\tau_1, \tau_2) \tilde{\mathbf{g}}_{\text{qo}}(\tau_1, \tau_2)^\top\right) d\tau_1 d\tau_2 &= \\ &\sum_{k=1}^p \int_0^\infty \int_0^\infty \text{tr}\left(\left(\mathbf{B}^\top \mathbf{E}^{-\top} e^{\mathbf{A}^\top \mathbf{E}^{-\top} \tau_1} \mathbf{M}_k e^{\mathbf{E}^{-1}\mathbf{A}\tau_2} \mathbf{E}^{-1} \mathbf{B}\right) \right. \\ &\quad \left. \times \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{E}}^{-\top} e^{\tilde{\mathbf{A}}^\top \tilde{\mathbf{E}}^{-\top} \tau_1} \tilde{\mathbf{M}}_k e^{\tilde{\mathbf{E}}^{-1}\tilde{\mathbf{A}}\tau_2} \tilde{\mathbf{E}}^{-1} \tilde{\mathbf{B}}\right)^\top\right) d\tau_1 d\tau_2. \end{aligned}$$

Hence, the second term in the previously derived expression for $-\operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}})$ is precisely the second term in the time-domain expression of the \mathcal{H}_2 inner product (5.28), and so $\langle \mathcal{G}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \rangle_{\mathcal{H}_2} = -\operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}})$ as claimed in (5.34). Applying this result for $\mathcal{G}_{\text{lqo}} = \tilde{\mathcal{G}}_{\text{lqo}}$ yields the formula for the \mathcal{H}_2 norm in (5.36).

The expression in (5.35) now follows straightforwardly from (5.34). Begin by vectorizing the generalized Sylvester equations in (5.32) and (5.33) to yield the equivalent linear systems of equations

$$\begin{aligned} (\tilde{\mathbf{E}} \otimes \mathbf{A} + \tilde{\mathbf{A}} \otimes \mathbf{E}) \operatorname{vec}(\mathbf{X}) &= -\operatorname{vec}(\mathbf{B} \tilde{\mathbf{B}}^\top), \\ (\tilde{\mathbf{E}}^\top \otimes \mathbf{A}^\top + \tilde{\mathbf{A}}^\top \otimes \mathbf{E}^\top) \operatorname{vec}(\mathbf{Z}_{\text{lqo}}) &= \operatorname{vec}(\mathbf{C}^\top \tilde{\mathbf{C}}) + \sum_{k=1}^p (\mathbf{M} \otimes \tilde{\mathbf{M}}) \operatorname{vec}(\mathbf{X}). \end{aligned}$$

Using the just-proven expression for the \mathcal{H}_2 inner product in (5.34), it follows that

$$\begin{aligned} \langle \mathcal{G}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \rangle_{\mathcal{H}_2} &= -\operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}}) = -\operatorname{vec}(\mathbf{B} \tilde{\mathbf{B}}^\top)^\top \operatorname{vec}(\mathbf{Z}_{\text{lqo}}) \quad \text{by (2.8)} \\ &= \operatorname{vec}(\mathbf{X})^\top (\tilde{\mathbf{E}}^\top \otimes \mathbf{A}^\top + \tilde{\mathbf{A}}^\top \otimes \mathbf{E}^\top) \operatorname{vec}(\mathbf{Z}_{\text{lqo}}) \\ &= \operatorname{vec}(\mathbf{X})^\top \left(\operatorname{vec}(\mathbf{C}^\top \tilde{\mathbf{C}}) + \sum_{k=1}^p (\mathbf{M} \otimes \tilde{\mathbf{M}}) \operatorname{vec}(\mathbf{X}) \right) \\ &= \operatorname{tr}(\mathbf{C} \mathbf{X} \tilde{\mathbf{C}}^\top) + \sum_{k=1}^p \operatorname{tr}(\mathbf{X}^\top \mathbf{M}_k \mathbf{X} \tilde{\mathbf{M}}_k) \quad \text{by (2.8) and (2.7)}. \end{aligned}$$

This proves (5.35); as before, the claim in (5.37) follows from (5.35) for $\tilde{\mathcal{G}}_{\text{lqo}} = \mathcal{G}_{\text{lqo}}$. \square

Note that Theorem 5.10 makes it obvious that the \mathcal{H}_2 inner product of two real-valued systems \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ —that is, systems \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ for which the state-space matrices in (5.1) and (5.2) are all real valued—is real valued as well.

5.4.2 Pole residue-based formulation

Next, we show that the system \mathcal{H}_2 norm and inner product can be computed in terms of the poles and residues of the subsystem transfer functions \mathbf{G}_{lqo} and \mathbf{G}_{qo} in (5.12). To this end, we first derive pole residue formulations of \mathbf{G}_{lqo} and \mathbf{G}_{qo} when the underlying system has simple poles: Consider an LQO system $\tilde{\mathcal{G}}_{\text{lqo}}$ as in (5.2), suppose that $\tilde{\mathcal{G}}_{\text{lqo}}$ is asymptotically stable with simple poles $\lambda_1, \dots, \lambda_r$. Let $\tilde{\mathbf{G}}_{\text{lqo}}$ and $\tilde{\mathbf{G}}_{\text{qo}}$ be the linear- and quadratic-output transfer functions of $\tilde{\mathcal{G}}_{\text{lqo}}$ defined according to (5.12). Because the poles of $\tilde{\mathcal{G}}_{\text{lqo}}$ are simple, the pair $\tilde{\mathbf{A}}, \tilde{\mathbf{E}}$ is diagonalizable and can be written as

$$\tilde{\mathbf{A}} \mathbf{S} = \tilde{\mathbf{E}} \mathbf{S} \mathbf{\Lambda} \quad \text{and} \quad \mathbf{T}^\top \tilde{\mathbf{A}} = \mathbf{\Lambda} \mathbf{T}^\top \tilde{\mathbf{E}},$$

or equivalently

$$\mathbf{T}^\top \tilde{\mathbf{A}} \mathbf{S} = \mathbf{\Lambda} \quad \text{and} \quad \mathbf{T}^\top \tilde{\mathbf{E}} \mathbf{S} = \mathbf{I}_r, \quad (5.41)$$

where $\mathbf{S}, \mathbf{T} \in \mathbb{C}^{r \times r}$ contain the right and left generalized eigenvectors of $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{A}}$, respectively, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_r)$ carries the generalized eigenvalues. Thus, under the equivalence transformation with $\mathbf{z} = \mathbf{S}\mathbf{x}$ and \mathbf{T}^\top , $\tilde{\mathbf{G}}_{\text{lo}}$ and $\tilde{\mathbf{G}}_{\text{qo}}$ can be expanded into *pole-residue* form as

$$\tilde{\mathbf{G}}_{\text{lo}}(s) = \sum_{i=1}^r \frac{\mathbf{c}_i \mathbf{b}_i^\top}{s - \lambda_i} \quad \text{and} \quad \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) = \sum_{j=1}^r \sum_{k=1}^r \frac{\mathbf{m}_{j,k} (\mathbf{b}_j \otimes \mathbf{b}_k)^\top}{(s_1 - \lambda_j)(s_2 - \lambda_k)}, \quad (5.42)$$

where

$$\mathbf{b}_i^\top \stackrel{\text{def}}{=} \mathbf{t}_i^\top \tilde{\mathbf{B}} \in \mathbb{C}^{1 \times m}, \quad \mathbf{c}_i \stackrel{\text{def}}{=} \tilde{\mathbf{C}} \mathbf{s}_i \in \mathbb{C}^p, \quad \text{and} \quad \mathbf{m}_{j,k} \stackrel{\text{def}}{=} \tilde{\mathbf{M}} (\mathbf{s}_j \otimes \mathbf{s}_k) \in \mathbb{C}^p, \quad (5.43)$$

for all $i, j, k = 1, \dots, r$ and the vectors $\mathbf{s}_i, \mathbf{t}_i \in \mathbb{C}^r$ denote the i -th columns of \mathbf{S} and \mathbf{T} . To derive this expansion for $\tilde{\mathbf{G}}_{\text{qo}}$, observe:

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) &= \tilde{\mathbf{M}} (\mathbf{S} \otimes \mathbf{S}) \left((s_1 \mathbf{T}^\top \tilde{\mathbf{E}} \mathbf{S} - \mathbf{T}^\top \tilde{\mathbf{A}} \mathbf{S})^{-1} \mathbf{T}^\top \mathbf{B} \otimes (s_2 \mathbf{T}^\top \tilde{\mathbf{E}} \mathbf{S} - \mathbf{T}^\top \tilde{\mathbf{A}} \mathbf{S})^{-1} \mathbf{T}^\top \mathbf{B} \right) \\ &= \tilde{\mathbf{M}} (\mathbf{S} (s_1 \mathbf{I}_r - \mathbf{\Lambda})^{-1} \mathbf{T}^\top \mathbf{B} \otimes \mathbf{S} (s_2 \mathbf{I}_r - \mathbf{\Lambda})^{-1} \mathbf{T}^\top \mathbf{B}) \\ &= \tilde{\mathbf{M}} \left(\sum_{j=1}^r \frac{\mathbf{s}_j \mathbf{t}_j^\top \tilde{\mathbf{B}}}{s_1 - \lambda_j} \otimes \sum_{k=1}^r \frac{\mathbf{s}_k \mathbf{t}_k^\top \tilde{\mathbf{B}}}{s_2 - \lambda_k} \right) = \sum_{j=1}^r \sum_{k=1}^r \frac{\tilde{\mathbf{M}} (\mathbf{s}_j \otimes \mathbf{s}_k) \mathbf{t}_j^\top \tilde{\mathbf{B}}}{(s_1 - \lambda_j)(s_2 - \lambda_k)} \\ &= \sum_{j=1}^r \sum_{k=1}^r \frac{\mathbf{m}_{j,k} (\mathbf{b}_j \otimes \mathbf{b}_k)^\top}{(s_1 - \lambda_j)(s_2 - \lambda_k)}, \end{aligned}$$

by (2.9). The expression for $\tilde{\mathbf{G}}_{\text{lo}}$ in (5.42) can be derived using similar manipulations. The rank-1 matrices $\mathbf{c}_i \mathbf{b}_i^\top \in \mathbb{C}^{p \times m}$ and $\mathbf{m}_{j,k} (\mathbf{b}_j \otimes \mathbf{b}_k)^\top \in \mathbb{C}^{p \times m^2}$ are defined as the *residues* of $\tilde{\mathbf{G}}_{\text{lo}}$ and $\tilde{\mathbf{G}}_{\text{qo}}$ corresponding to the poles $s = \lambda_i$ and $(s_1, s_2) = (\lambda_j, \lambda_k)$. We call the vectors $\mathbf{b}_i \in \mathbb{C}^m$, $\mathbf{c}_i \in \mathbb{C}^p$, and $\mathbf{m}_{j,k} \in \mathbb{C}^p$ in (5.43) the *residue directions*. Under the assumption that $\mathbf{M}_i = \mathbf{M}_i^\top$ for all i , the left residue directions $\mathbf{m}_{j,k}$ obey the symmetry condition

$$\mathbf{m}_{j,k} = \tilde{\mathbf{M}} (\mathbf{s}_j \otimes \mathbf{s}_k) = \begin{bmatrix} \mathbf{s}_k^\top \mathbf{M}_1 \mathbf{s}_j \\ \vdots \\ \mathbf{s}_k^\top \mathbf{M}_p \mathbf{s}_j \end{bmatrix} = \begin{bmatrix} \mathbf{s}_j^\top \mathbf{M}_1 \mathbf{s}_k \\ \vdots \\ \mathbf{s}_j^\top \mathbf{M}_p \mathbf{s}_k \end{bmatrix} = \tilde{\mathbf{M}} (\mathbf{s}_k \otimes \mathbf{s}_j) = \mathbf{m}_{k,j}, \quad (5.44)$$

for each $j, k = 1, \dots, r$. Similar pole-residue expansions to (5.42) can be derived in the case of multiple and higher-order poles; see [218] for the linear case. The pole-residue expansions in (5.42) enable us to derive the computational formula from Definition 5.9.

Theorem 5.11 (\mathcal{H}_2 norm and inner product via transfer function poles and residues). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2) with the transfer functions \mathbf{G}_{lo} , \mathbf{G}_{qo} and $\tilde{\mathbf{G}}_{\text{lo}}$, $\tilde{\mathbf{G}}_{\text{qo}}$ defined according to (5.12). Suppose

additionally that $\tilde{\mathcal{G}}_{\text{lqo}}$ has simple poles $\lambda_1, \dots, \lambda_r$. Then, the \mathcal{H}_2 inner product (5.30) of \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ is given by

$$\left\langle \mathcal{G}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \right\rangle_{\mathcal{H}_2} = \underbrace{\sum_{i=1}^r \mathbf{c}_i^\top \bar{\mathbf{G}}_{\text{lqo}}(-\lambda_i) \mathbf{b}_i}_{=\langle \mathbf{G}_{\text{lqo}}, \tilde{\mathbf{G}}_{\text{lqo}} \rangle_{\mathcal{H}_2^{p \times m}}} + \underbrace{\sum_{j=1}^r \sum_{k=1}^r \mathbf{m}_{j,k}^\top \bar{\mathbf{G}}_{\text{qo}}(-\lambda_j, -\lambda_k) (\mathbf{b}_j \otimes \mathbf{b}_k)}_{=\langle \mathbf{G}_{\text{qo}}, \tilde{\mathbf{G}}_{\text{qo}} \rangle_{\mathcal{H}_2^{p \times m^2}}}, \quad (5.45)$$

where $\bar{\mathbf{G}}_{\text{lqo}}(s) = \bar{\mathbf{C}}(s\bar{\mathbf{E}} - \bar{\mathbf{A}})^{-1} \bar{\mathbf{B}}$ and $\bar{\mathbf{G}}_{\text{qo}}(s_1, s_2) = \bar{\mathbf{M}} \left((s_1\bar{\mathbf{E}} - \bar{\mathbf{A}})^{-1} \bar{\mathbf{B}} \otimes (s_2\bar{\mathbf{E}} - \bar{\mathbf{A}})^{-1} \bar{\mathbf{B}} \right)$. The \mathcal{H}_2 norm (5.31) of $\tilde{\mathcal{G}}_{\text{lqo}}$ is given by

$$\|\tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 = \underbrace{\sum_{i=1}^r \mathbf{c}_i^\top \tilde{\bar{\mathbf{G}}}_{\text{lqo}}(-\lambda_i) \mathbf{b}_i}_{=\|\tilde{\mathbf{G}}_{\text{lqo}}\|_{\mathcal{H}_2^{p \times m}}^2} + \underbrace{\sum_{j=1}^r \sum_{k=1}^r \mathbf{m}_{j,k}^\top \tilde{\bar{\mathbf{G}}}_{\text{qo}}(-\lambda_j, -\lambda_k) (\mathbf{b}_j \otimes \mathbf{b}_k)}_{=\|\tilde{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2^{p \times m^2}}^2}. \quad (5.46)$$

◇

Proof of Theorem 5.11. First note that the first terms in (5.30) and (5.45) are equal:

$$\int_{-\infty}^{\infty} \text{tr} \left(\tilde{\bar{\mathbf{G}}}_{\text{lqo}}(-i\omega) \tilde{\mathbf{G}}_{\text{lqo}}(i\omega)^\top \right) d\omega = \sum_{i=1}^r \mathbf{c}_i^\top \tilde{\bar{\mathbf{G}}}_{\text{lqo}}(-\lambda_i) \mathbf{b}_i.$$

This follows from classical results for calculating the Hardy \mathcal{H}_2 inner product of two linear time-invariant systems (2.25) via their transfer functions; see, e.g. [97, Lemma 3.5], [5, Lemma 2.1.4]. Then, to prove (5.45) it suffices to prove the remaining equality

$$\begin{aligned} & \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{tr} \left(\bar{\mathbf{G}}_{\text{qo}}(-i\omega_1, -i\omega_2) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, i\omega_2)^\top \right) d\omega_1 d\omega_2 \\ &= \sum_{j=1}^r \sum_{k=1}^r \mathbf{m}_{j,k}^\top \bar{\mathbf{G}}_{\text{qo}}(-\lambda_j, -\lambda_k) (\mathbf{b}_j \otimes \mathbf{b}_k). \end{aligned} \quad (5.47)$$

For fixed but arbitrary constants $R_1, R_2 > 0$, define the contours $\Gamma_{R_i} \subset \mathbb{C}$ as

$$\Gamma_{R_i} \stackrel{\text{def}}{=} [-iR_i, iR_i] \cup \{z = R_i e^{i\theta} \mid \pi/2 \leq \theta \leq 3\pi/2\}, \quad i = 1, 2.$$

Choose $R_1, R_2 > 0$ to be sufficiently large such that each contour Γ_{R_1} and Γ_{R_2} encircles the poles of the reduced model. In other words, $\lambda_j \in \text{int}(\Gamma_{R_1}), \text{int}(\Gamma_{R_2})$ for each $j = 1, \dots, r$ where $\text{int}(\cdot)$ denotes the set of interior points. Let $z \in i\mathbb{R}$ be arbitrarily fixed, and consider:

$$\begin{aligned} \int_{\Gamma_{R_1}} \text{tr} \left(\bar{\mathbf{G}}_{\text{qo}}(-\zeta_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(\zeta_1, z)^\top \right) d\zeta_1 &= \int_{-R_1}^{R_1} \text{tr} \left(\bar{\mathbf{G}}_{\text{qo}}(-i\omega_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, z)^\top \right) d\omega_1 \\ &+ \int_{\pi/2}^{3\pi/2} \text{tr} \left(\bar{\mathbf{G}}_{\text{qo}}(-R_1 e^{i\theta}, -z) \tilde{\mathbf{G}}_{\text{qo}}(R_1 e^{i\theta}, z)^\top \right) R_1 e^{i\theta} d\theta. \end{aligned}$$

Because $\mathbf{G}_{\text{qo}}(-s_1, -z)$ and $\tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top$ are strictly proper rational functions and $R_1 > 0$ is arbitrarily specified, for any $\varepsilon > 0$, we may choose R_1 to be large enough so that $\|\mathbf{G}_{\text{qo}}(-R_1 e^{i\theta}, -z)\|_{\text{F}}$ and $\|\tilde{\mathbf{G}}_{\text{qo}}(R_1 e^{i\theta}, z)^\top\|_{\text{F}}$ are smaller than or equal to ε . In turn, this implies

$$\left| \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-R_1 e^{i\theta}, -z) \tilde{\mathbf{G}}_{\text{qo}}(R_1 e^{i\theta}, z)^\top \right) \right| \leq \|\mathbf{G}_{\text{qo}}(-R_1 e^{i\theta}, -z)\|_{\text{F}} \|\tilde{\mathbf{G}}_{\text{qo}}(R_1 e^{i\theta}, z)^\top\|_{\text{F}} \leq \varepsilon^2.$$

Note that this choice of R_1 still guarantees that Γ_{R_1} encircles the poles of the reduced model. Using standard ML-estimates from Theorem 2.15, we obtain

$$\left| \int_{\pi/2}^{3\pi/2} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-R_1 e^{i\theta}, -z) \tilde{\mathbf{G}}_{\text{qo}}(R_1 e^{i\theta}, z)^\top \right) R_1 e^{i\theta} d\theta \right| \leq \frac{\pi}{2} \varepsilon^2 R_1 \longrightarrow 0 \text{ as } R_1 \rightarrow \infty.$$

Thus, taking the limit as $R_1 \rightarrow \infty$ yields

$$\begin{aligned} \lim_{R_1 \rightarrow \infty} \int_{\Gamma_{R_1}} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-\zeta_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(\zeta_1, z)^\top \right) d\zeta_1 &= \lim_{R_1 \rightarrow \infty} \int_{-R_1}^{R_1} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, z)^\top \right) d\omega_1 \\ &= \int_{-\infty}^{\infty} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, z)^\top \right) d\omega_1. \end{aligned} \tag{5.48}$$

Next, note that $\text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-s_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right)$ is a complex-valued scalar function of the variable s_1 with poles at $-\mu_1, -\mu_2, \dots, -\mu_n \in \mathbb{C}_{>0}$ and *simple poles* $\lambda_1, \lambda_2, \dots, \lambda_r \in \mathbb{C}_{<0}$, where μ_i denotes the i -th eigenvalue of $\mathbf{E}^{-1}\mathbf{A}$. By Theorem 2.14 (the Residue Theorem) and (5.48), we have that

$$\begin{aligned} &\frac{1}{2\pi} \int_{-\infty}^{\infty} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, z)^\top \right) d\omega_1 \\ &= \lim_{R_1 \rightarrow \infty} \frac{1}{2\pi i} \int_{\Gamma_{R_1}} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-\zeta_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(\zeta_1, z)^\top \right) d\zeta_1 \\ &= \sum_{j=1}^r \text{Res} \left[\text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-s_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right), s_1 = \lambda_j \right]. \end{aligned}$$

Under the assumption that the poles λ_j are simple, for any fixed $z \in i\mathbb{R}$ we can compute the residue of $\text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-s_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right)$ at λ_j as

$$\begin{aligned} &\text{Res} \left[\text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-s_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right), s_1 = \lambda_j \right] \\ &= \lim_{s_1 \rightarrow \lambda_j} (s_1 - \lambda_j) \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-s_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right) \\ &= \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -z) \lim_{s_1 \rightarrow \lambda_j} (s_1 - \lambda_j) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right). \end{aligned}$$

Because the poles of $\tilde{\mathbf{G}}_{\text{qo}}$ are assumed simple, $\tilde{\mathbf{G}}_{\text{qo}}$ permits the pole-residue expansion in (5.42). Substituting in directly for (5.42) yields

$$\lim_{s_1 \rightarrow \lambda_j} (s_1 - \lambda_j) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top = \lim_{s_1 \rightarrow \lambda_j} (s_1 - \lambda_j) \sum_{i=1}^r \sum_{k=1}^r \frac{(\mathbf{b}_i \otimes \mathbf{b}_k) \mathbf{m}_{i,k}^\top}{(s_1 - \lambda_i)(z - \lambda_k)} = \sum_{k=1}^r \frac{(\mathbf{b}_j \otimes \mathbf{b}_k) \mathbf{m}_{j,k}^\top}{z - \lambda_k},$$

and so

$$\text{Res} \left[\text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-s_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(s_1, z)^\top \right), s_1 = \lambda_j \right] = \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -z) \sum_{k=1}^r \frac{(\mathbf{b}_j \otimes \mathbf{b}_k) \mathbf{m}_{j,k}^\top}{z - \lambda_k} \right)$$

for each $j = 1, \dots, r$. Substituting this into the previously computed contour integral, at last we have that

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -z) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, z)^\top \right) d\omega_1 &= \sum_{j=1}^r \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -z) \sum_{k=1}^r \frac{(\mathbf{b}_j \otimes \mathbf{b}_k) \mathbf{m}_{j,k}^\top}{z - \lambda_k} \right) \\ &= \sum_{j=1}^r \sum_{k=1}^r \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -z) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{z - \lambda_k}, \end{aligned}$$

where the ultimate equality follows from the fact that the trace operator $\text{tr}(\cdot)$ is invariant under cyclic permutations, and that the trace of a scalar is just said scalar. Returning to the desired equality in (5.47) our calculations up to this point yield

$$\begin{aligned} \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{tr} \left(\overline{\mathbf{G}}_{\text{qo}}(-i\omega_1, -i\omega_2) \tilde{\mathbf{G}}_{\text{qo}}(i\omega_1, i\omega_2)^\top \right) d\omega_1 d\omega_2 \\ = \sum_{j=1}^r \sum_{k=1}^r \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -i\omega_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{i\omega_2 - \lambda_k} d\omega_2. \end{aligned}$$

What remains is to evaluate the integral in the expression above. Recalling the definition of Γ_{R_2} , consider the contour integral

$$\begin{aligned} \int_{\Gamma_{R_2}} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -\zeta_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{\zeta_2 - \lambda_k} d\zeta_2 &= \int_{-R_2}^{R_2} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -i\omega_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{i\omega_2 - \lambda_k} d\omega_2 \\ &\quad + \int_{\pi/2}^{3\pi/2} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -R_2 e^{i\theta}) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{R_2 e^{i\theta} - \lambda_k} R_2 e^{i\theta} d\theta. \end{aligned}$$

Because the constant $R_2 > 0$ is arbitrarily specified, we may take it to be large enough such that $|\mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -R_2 e^{i\theta}) (\mathbf{b}_j \otimes \mathbf{b}_k) / (R_2 e^{i\theta} - \lambda_k)| \leq \varepsilon^2$ for all $\pi/2 \leq \theta \leq 3\pi/2$ and any desired $\varepsilon > 0$. Thus

$$\left| \int_{\pi/2}^{3\pi/2} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\text{qo}}(-\lambda_j, -R_2 e^{i\theta}) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{R_2 e^{i\theta} - \lambda_k} R_2 e^{i\theta} d\theta \right| \leq \frac{\pi}{2} \varepsilon^2 R_2 \rightarrow 0$$

as $R_2 \rightarrow \infty$. In the limit as $R_2 \rightarrow \infty$, we see that

$$\begin{aligned} & \lim_{R_2 \rightarrow \infty} \int_{\Gamma_{R_2}} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -\zeta_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{\zeta_2 - \lambda_k} d\zeta_2 \\ &= \int_{-\infty}^{\infty} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -z) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{i\omega_2 - \lambda_k} d\omega_2. \end{aligned} \quad (5.49)$$

At this point, each integral appearing within the nested sum in the simplified expression for (5.47) can be evaluated by a straightforward application of Theorem 2.14. For each $j, k = 1, \dots, r$, the integrand $\mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -z) (\mathbf{b}_j \otimes \mathbf{b}_k) / (z - \lambda_k)$ is a complex-valued scalar function with poles at $-\mu_1, \dots, -\mu_n \in \mathbb{C}_{>0}$, i.e., the eigenvalues of $-\mathbf{E}^{-1}\mathbf{A}$, and $\lambda_k \in \mathbb{C}_{<0}$. Thus, for each j, k we have

$$\begin{aligned} & \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -i\omega_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{i\omega_2 - \lambda_k} d\omega_2 \\ &= \lim_{R_2 \rightarrow \infty} \frac{1}{2\pi i} \int_{\Gamma_{R_2}} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -\zeta_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{\zeta_2 - \lambda_k} d\zeta_2 \quad (\text{by (5.49)}) \\ &= \text{Res} \left[\mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -s_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{s_2 - \lambda_k}, s_2 = \lambda_k \right] \\ &= \lim_{s_2 \rightarrow \lambda_k} (s_2 - \lambda_k) \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -s_2) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{s_2 - \lambda_k} \\ &= \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -\lambda_k) (\mathbf{b}_j \otimes \mathbf{b}_k). \end{aligned}$$

Finally, substituting this into the two-dimensional integral (5.47) yields

$$\begin{aligned} & \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{tr} \left(\overline{\mathbf{G}}_{\mathbf{qo}}(-i\omega_1, -i\omega_2) \tilde{\mathbf{G}}_{\mathbf{qo}}(i\omega_1, i\omega_2)^\top \right) d\omega_1 d\omega_2 \\ &= \sum_{j=1}^r \sum_{k=1}^r \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -z) (\mathbf{b}_j \otimes \mathbf{b}_k) \frac{1}{i\omega_2 - \lambda_k} d\omega_2 \\ &= \sum_{j=1}^r \sum_{k=1}^r \mathbf{m}_{j,k}^\top \overline{\mathbf{G}}_{\mathbf{qo}}(-\lambda_j, -\lambda_k) (\mathbf{b}_j \otimes \mathbf{b}_k), \end{aligned}$$

which proves the formula (5.47), and thus the inner product formula (5.45). The formula for the \mathcal{H}_2 norm in (5.46) then follows directly by applying (5.45) for $\mathcal{G}_{\mathbf{lqo}} = \tilde{\mathcal{G}}_{\mathbf{lqo}}$. \square

To summarize, we have presented two new computational formulations of the system \mathcal{H}_2 norm and inner product (Definitions 5.8 and 5.9) in Theorems 5.10 and 5.11. The pole- and residue-based norm formula (5.46) of Theorem 5.11 is more limited insofar as it only applies to systems with simple poles. The formulae in Theorem 5.11 may also involve complex arithmetic, whereas those in Theorem 5.10 can be implemented using only real arithmetic. When applied to systems (5.1) and (5.2) with the particular choice $\mathbf{M} = \mathbf{0}_{p \times n^2}$ and $\tilde{\mathbf{M}} =$

$\mathbf{0}_{p \times r^2}$, both Theorems 5.10 and 5.11 return the usual Sylvester-equation and pole-residue based expressions of the \mathcal{H}_2 norm and inner product for linear time-invariant systems; see Theorems 2.41 and (2.42) as well as the original references [5, Lemma 2.1.4] [97, Lemma 2.3, Lemma 3.5], [4, Ch. 5.5]. We mention that similar expressions exist for the \mathcal{H}_2 norm of other nonlinear systems, e.g., a bilinear or a quadratic-bilinear system; see [77, Theorem 2.2], [24, Proposition 3.3] and [50, Theorem 2], respectively.

Chapter 6

Optimal- \mathcal{H}_2 approximation of linear systems with quadratic outputs

6.1 Introduction

In this section, we formally consider the \mathcal{H}_2 -optimal model reduction problem for the linear quadratic-output (LQO) systems introduced in Chapter 5. Our interest in this best approximation problem is principally motivated by an upper bound on the \mathcal{L}_∞ output error in terms of the \mathcal{H}_2 system error; see Theorem 6.1 in Section 6.2. This motivates minimizing the \mathcal{H}_2 system error induced by (5.2) to achieve a uniformly small output error. Given an order- n asymptotically stable LQO system as in (5.1), we seek an asymptotically stable reduced model $\tilde{\mathcal{G}}_{\text{lqo}}$ as in (5.2) of a fixed approximation order $1 \leq r < n$ such that the \mathcal{H}_2 error in approximating (5.1) is minimized, i.e., $\tilde{\mathcal{G}}_{\text{lqo}}$ solves

$$\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 = \min_{\dim(\tilde{\mathcal{G}}_{\text{lqo}})=r} \|\mathcal{G}_{\text{lqo}} - \check{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 \quad \text{subj. to } \check{\mathcal{G}}_{\text{lqo}} \text{ is asymptotically stable.} \quad (6.1)$$

The squared \mathcal{H}_2 error is only used for the ease of deriving first-order optimality conditions later on. As was the case for the \mathcal{H}_2 -optimal model reduction of *linear* systems described in Section 2.4.2, the \mathcal{H}_2 -minimization problem (6.1) is in general nonconvex, and global minimizers are hard to characterize. Thus, we adopt the more modest goal of identifying reduced-order models (5.2) that satisfy some first-order necessary conditions for local optimality, and subsequently develop iterative algorithms for constructing approximants that satisfy these optimality conditions.

6.1.1 Chapter contents

In this chapter, we present a pair of solutions to (6.1) in the form of two independent \mathcal{H}_2 -optimality frameworks. The first is based on the solutions to generalized Sylvester equations and the LQO system Gramians (2.39) and (5.19); the second is based on the tangential (rational) interpolation of the linear- and quadratic-output subsystem transfer functions (5.12). These are presented in Sections 6.3 and 6.4, respectively. In either case, the blueprint we follow to develop the aforementioned optimality frameworks is the same.

1. We use Theorems 5.10 and 5.11 to derive more computationally amenable parameterizations of the \mathcal{H}_2 model error in (6.1). For the Sylvester equation-based framework, this is done explicitly in terms of the reduced model matrices $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ and $\tilde{\mathbf{M}}$ using Theorem 5.10; for the interpolatory framework, Theorem 5.11 is used implicitly to recast (6.1) as a multivariate rational approximation problem, where the degrees of freedom are the poles and residue directions $\lambda_i, \mathbf{b}_i, \mathbf{c}_i, \mathbf{m}_{j,k}$ of the reduced model transfer functions.
2. We then derive first-order optimality conditions with respect to these two parameterizations. For the Sylvester- or Gramian-based conditions, this is accomplished by taking gradients of the squared \mathcal{H}_2 error with respect to the reduced model matrices; for the interpolatory conditions, these are derived by choosing particular ε -perturbations of the reduced model transfer functions, and taking limits. The Sylvester- and interpolation-based optimality conditions are presented in Theorems 6.5 and 6.9, respectively.
3. We prove how to enforce each set of optimality conditions in a Petrov-Galerkin projection framework with appropriately chosen \mathbf{V} and \mathbf{W} .
4. Based on the theoretical Sylvester- and interpolation-based optimality frameworks, we propose a pair of numerically efficient algorithms for \mathcal{H}_2 -optimal model reduction of LQO systems. Each algorithm performs iteratively-corrected projection to construct a reduced model (5.2) that satisfies the relevant set of necessary \mathcal{H}_2 -optimality conditions and locally minimizes the \mathcal{H}_2 error. These generalize the two-sided iteration algorithm (TSIA) from [237] and the iterative rational Krylov algorithm (IRKA) from [97], and are presented at the end of Sections 6.3 and 6.4.

In effect, these results establish the Sylvester equation-based (or Wilson) [217, 233] and the interpolation-based (or Meier-Luenberger) [97, 142] optimality frameworks discussed in Section 2.4.2 for the best- \mathcal{H}_2 approximation of LQO systems (5.1). In Section 6.5, we show that the Sylvester-based conditions imply the interpolatory conditions when the optimal reduced model has simple poles, thereby justifying the use of either computational algorithm. Section 6.6 contains numerical results and closes our discussion.

Portions of this introduction, as well as the contents of Sections 6.2–6.6 are available in the preprints [185, 186] and the published work [188].

- [185] Reiter, S., Gosea, I. V., Pontes Duff, I., and Gugercin, S. (2025). \mathcal{H}_2 -optimal model reduction of linear quadratic-output systems by multivariate rational interpolation. arXiv, 2505.03057.
- [186] Reiter, S., Pontes Duff, I., Gosea, I. V., and Gugercin, S. (2024a). \mathcal{H}_2 -optimal model reduction of linear systems with multiple quadratic outputs. arXiv, 2405.05951. (Under review.)

[188] Reiter, S. and Werner, S. W. R. (2025b). [Interpolatory model reduction of dynamical systems with root mean squared error](#). *IFAC-PapersOnLine*, 59(1):385–390.

6.2 Motivating the optimal- \mathcal{H}_2 approximation problem

Our principal rationale for using the \mathcal{H}_2 model error as a minimization objective in computing (5.2) is based on an error bound from [28]. Put succinctly, the \mathcal{H}_2 error induced by replacing \mathcal{G}_{lqo} in (5.1) with $\tilde{\mathcal{G}}_{\text{lqo}}$ in (5.2) controls the associated \mathcal{L}_∞ output error $\mathbf{y} - \tilde{\mathbf{y}}$. Thus, a small \mathcal{H}_2 approximation in turn ensures a uniformly good approximation in the output. We prove this bound for the general case of nonsingular \mathbf{E} and mixed linear- and quadratic-output terms.

Theorem 6.1 (\mathcal{H}_2 upper bound on the \mathcal{L}_∞ output error [28, Theorem 3.4]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems in (5.1) and (5.2), and let \mathbf{u} be such that $\mathbf{u} \in \mathcal{L}_2^m$ and $\mathbf{u} \otimes \mathbf{u} \in \mathcal{L}_2^{m^2}$. Then, the output error $\mathbf{y} - \tilde{\mathbf{y}}$ satisfies

$$\|\mathbf{y} - \tilde{\mathbf{y}}\|_{\mathcal{L}_\infty^p}^2 \leq \|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 \left(\|\mathbf{u}\|_{\mathcal{L}_2^m}^2 + \|\mathbf{u} \otimes \mathbf{u}\|_{\mathcal{L}_2^{m^2}}^2 \right), \quad (6.2)$$

where the signal norms are defined according to Definition 2.17. \diamond

Proof of Theorem 6.1. Following (5.9), the output error $\mathbf{y}(t) - \tilde{\mathbf{y}}(t)$ at any time $t \geq 0$ can be expressed as

$$\begin{aligned} \mathbf{y}(t) - \tilde{\mathbf{y}}(t) &= \int_0^t (\mathbf{g}_{\text{lo}}(\tau) - \tilde{\mathbf{g}}_{\text{lo}}(\tau)) \mathbf{u}(t - \tau) d\tau \\ &\quad + \int_0^t \int_0^t (\mathbf{g}_{\text{qo}}(\tau_1, \tau_2) - \tilde{\mathbf{g}}_{\text{qo}}(\tau_1, \tau_2)) (\mathbf{u}(t - \tau_1) \otimes \mathbf{u}(t - \tau_2)) d\tau_1 d\tau_2, \end{aligned}$$

where $\mathbf{g}_{\text{lo}}, \mathbf{g}_{\text{qo}}$ and $\tilde{\mathbf{g}}_{\text{lo}}, \tilde{\mathbf{g}}_{\text{qo}}$ are the Volterra kernels of \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ defined according to (5.10). Applying the vector ∞ -norm to the output error, we obtain

$$\begin{aligned} \|\mathbf{y}(t) - \tilde{\mathbf{y}}(t)\|_\infty &\leq \int_0^t \|(\mathbf{g}_{\text{lo}}(\tau) - \tilde{\mathbf{g}}_{\text{lo}}(\tau)) \mathbf{u}(t - \tau)\|_\infty d\tau \\ &\quad + \int_0^t \int_0^t \|\mathbf{g}_{\text{qo}}(\tau_1, \tau_2) - \tilde{\mathbf{g}}_{\text{qo}}(\tau_1, \tau_2) (\mathbf{u}(t - \tau_1) \otimes \mathbf{u}(t - \tau_2))\|_\infty d\tau_1 d\tau_2 \\ &\leq \int_0^t \|\mathbf{g}_{\text{lo}}(\tau) - \tilde{\mathbf{g}}_{\text{lo}}(\tau)\|_{\text{F}} \|\mathbf{u}(t - \tau)\|_2 d\tau \\ &\quad + \int_0^t \int_0^t \|\mathbf{g}_{\text{qo}}(\tau_1, \tau_2) - \tilde{\mathbf{g}}_{\text{qo}}(\tau_1, \tau_2)\|_{\text{F}} \|\mathbf{u}(t - \tau_1) \otimes \mathbf{u}(t - \tau_2)\|_2 d\tau_1 d\tau_2, \end{aligned}$$

where the first inequality follows from the integral triangle inequality, and the second follows from the fact that $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2$ and $\|\mathbf{H}\mathbf{v}\|_2 \leq \|\mathbf{H}\|_2 \|\mathbf{v}\|_2 \leq \|\mathbf{H}\|_{\text{F}} \|\mathbf{v}\|_2$ for any $\mathbf{H} \in \mathbb{C}^{n_1 \times n_2}$

and $\mathbf{v} \in \mathbb{C}^{n_2}$. Because each integrand above is strictly nonnegative, we may take the upper integral limits $t \rightarrow \infty$. Under the assumption that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable, their linear- and quadratic-output Volterra kernels belong to $\mathcal{L}_2^{p \times m}$ and $\mathcal{L}_2^{p \times m^2}$, respectively, so that $\|\mathbf{g}_{\text{lo}} - \tilde{\mathbf{g}}_{\text{lo}}\|_{\mathcal{L}_2^{p \times m}}$ and $\|\mathbf{g}_{\text{qo}} - \tilde{\mathbf{g}}_{\text{qo}}\|_{\mathcal{L}_2^{p \times m^2}} \leq \infty$. By assumption, $\mathbf{u} \in \mathcal{L}_2^m$ and $\mathbf{u} \otimes \mathbf{u} \in \mathcal{L}_2^{m^2}$ so that $\|\mathbf{u}\|_{\mathcal{L}_2^m}$ and $\|\mathbf{u} \otimes \mathbf{u}\|_{\mathcal{L}_2^{m^2}} \leq \infty$. Under these assumptions, we may apply the Cauchy-Schwarz inequality for scalar-valued square integrable functions, i.e., $\mathcal{L}_2(0, t)$ and $\mathcal{L}_2(0, t) \times (0, t)$, to simplify the bound further:

$$\begin{aligned} \|\mathbf{y}(t) - \tilde{\mathbf{y}}(t)\|_\infty &\leq \left(\int_0^t \|\mathbf{g}_{\text{lo}}(\tau) - \tilde{\mathbf{g}}_{\text{lo}}(\tau)\|_{\text{F}} d\tau \right)^{\frac{1}{2}} \left(\int_0^t \|\mathbf{u}(t - \tau)\|_2^2 d\tau \right)^{\frac{1}{2}} \\ &+ \left(\int_0^t \int_0^t \|\mathbf{g}_{\text{qo}}(\tau_1, \tau_2) - \tilde{\mathbf{g}}_{\text{qo}}(\tau_1, \tau_2)\|_{\text{F}} d\tau_1 d\tau_2 \right)^{\frac{1}{2}} \left(\int_0^\infty \int_0^\infty \|\mathbf{u}(t - \tau_1) \otimes \mathbf{u}(t - \tau_2)\|_2^2 d\tau_1 d\tau_2 \right)^{\frac{1}{2}}. \\ &= \|\mathbf{g}_{\text{lo}} - \tilde{\mathbf{g}}_{\text{lo}}\|_{\mathcal{L}_2^{p \times m}} \|\mathbf{u}\|_{\mathcal{L}_2^m} + \|\mathbf{g}_{\text{qo}} - \tilde{\mathbf{g}}_{\text{qo}}\|_{\mathcal{L}_2^{p \times m^2}} \|\mathbf{u} \otimes \mathbf{u}\|_{\mathcal{L}_2^{m^2}}. \end{aligned}$$

Applying the Cauchy-Schwarz inequality again, this time to the 2-vectors containing the individual kernel errors, and squaring both sides, we retrieve a bound involving the \mathcal{H}_2 system error:

$$\|\mathbf{y}(t) - \tilde{\mathbf{y}}(t)\|_\infty^2 \leq \underbrace{\left(\|\mathbf{g}_{\text{lo}} - \tilde{\mathbf{g}}_{\text{lo}}\|_{\mathcal{L}_2^{p \times m}}^2 + \|\mathbf{g}_{\text{qo}} - \tilde{\mathbf{g}}_{\text{qo}}\|_{\mathcal{L}_2^{p \times m^2}}^2 \right)}_{=\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2} \left(\|\mathbf{u}\|_{\mathcal{L}_2^m}^2 + \|\mathbf{u} \otimes \mathbf{u}\|_{\mathcal{L}_2^{m^2}}^2 \right).$$

Because t is arbitrarily specified, we may take the supremum over $t > 0$ on both sides to yield the desired result in (6.2). \square

We emphasize that the upper bound holds *uniformly in time*, meaning that a small \mathcal{H}_2 error guarantees that the output error will be controlled at any time $t \geq 0$. While the result of Theorem 6.1 is used as a motivator for the \mathcal{H}_2 -optimal model reduction problem in (6.1), the bound is valid for any asymptotically stable reduced model. Thus, as long as $\tilde{\mathcal{G}}_{\text{lqo}}$ is asymptotically stable, one can use the \mathcal{H}_2 error to determine whether the corresponding output error will be small, regardless of the strategy or algorithm used to determine the reduced model.

6.3 Sylvester equation-based \mathcal{H}_2 optimality framework

6.3.1 The theoretical optimality framework

In this section, we establish the Sylvester equation-based (Wilson) optimality framework for the \mathcal{H}_2 model reduction of LQO systems (5.1). As our starting point, we apply the result

of Theorem 5.10 to derive an expression for the \mathcal{H}_2 model reduction error. To this end, we introduce the error system $\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}$; this is most easily represented using the alternative expression for the quadratic outputs described in Remark 5.1 and written in (5.8). Given \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$, the error system $\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}$ is a $(n+r)$ -dimensional system of the form (5.1) having the realization

$$\begin{aligned} \mathbf{E}_e &= \begin{bmatrix} \mathbf{E} & \\ & \tilde{\mathbf{E}} \end{bmatrix}, \quad \mathbf{A}_e = \begin{bmatrix} \mathbf{A} & \\ & \tilde{\mathbf{A}} \end{bmatrix}, \quad \mathbf{B}_e = \begin{bmatrix} \mathbf{B} \\ \tilde{\mathbf{B}} \end{bmatrix}, \quad \mathbf{C}_e = \begin{bmatrix} \mathbf{C} & -\tilde{\mathbf{C}} \end{bmatrix}, \\ \text{and } \mathbf{M}_{k,e} &= \begin{bmatrix} \mathbf{M}_k & \\ & -\tilde{\mathbf{M}}_k \end{bmatrix}, \quad \text{for all } k = 1, 2, \dots, p. \end{aligned} \quad (6.3)$$

Note that the realization (6.3) makes it rather obvious that $\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}$ is asymptotically stable whenever \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are. The Gramians of the error system, which we denote by $\mathbf{P}_e, \mathbf{E}_e^\top \mathbf{Q}_{\text{lqo},e} \mathbf{E}_e \in \mathbb{R}^{(n+r) \times (n+r)}$, uniquely solve the corresponding generalized Lyapunov equations (2.43) and (5.20) for the realization in (6.3), i.e.,

$$\begin{aligned} \mathbf{A}_e \mathbf{P}_e \mathbf{E}_e^\top + \mathbf{E}_e \mathbf{P}_e \mathbf{A}_e^\top + \mathbf{B}_e \mathbf{B}_e^\top &= \mathbf{0}_{(n+r) \times (n+r)}, \\ \mathbf{A}_e^\top \mathbf{Q}_{\text{lqo},e} \mathbf{E}_e + \mathbf{E}_e^\top \mathbf{Q}_{\text{lqo},e} \mathbf{A}_e + \mathbf{C}_e^\top \mathbf{C}_e + \sum_{k=1}^p \mathbf{M}_{k,e} \mathbf{P}_e \mathbf{M}_{k,e} &= \mathbf{0}_{(n+r) \times (n+r)}. \end{aligned} \quad (6.4)$$

Unpacking the matrix equations in (6.4) using the 2×2 block structure of the realization (6.3) reveals that the so-called error Gramians \mathbf{P}_e and $\mathbf{E}_e^\top \mathbf{Q}_{\text{lqo},e} \mathbf{E}_e$ can be written as

$$\mathbf{P}_e = \begin{bmatrix} \mathbf{P} & \mathbf{X} \\ \mathbf{X}^\top & \tilde{\mathbf{P}} \end{bmatrix} \quad \text{and} \quad \mathbf{E}_e^\top \mathbf{Q}_{\text{lqo},e} \mathbf{E}_e = \begin{bmatrix} \mathbf{E}_e^\top \mathbf{Q}_{\text{lqo}} \mathbf{E}_e & \mathbf{E}_e^\top \mathbf{Z}_{\text{lqo}} \\ \mathbf{Z}_{\text{lqo}}^\top \mathbf{E}_e & \mathbf{E}_e^\top \tilde{\mathbf{Q}}_{\text{lqo}} \mathbf{E}_e \end{bmatrix}, \quad (6.5)$$

where $\mathbf{P}, \mathbf{E}_e^\top \mathbf{Q}_{\text{lqo}} \mathbf{E}_e \in \mathbb{R}^{n \times n}$ and $\tilde{\mathbf{P}}, \tilde{\mathbf{E}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{E}} \in \mathbb{R}^{r \times r}$ are the full- and reduced-order system Gramians defined according to (2.39) and (5.19), while the off-diagonal submatrices $\mathbf{X}, \mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ are those appearing in Theorem 5.10. The matrices $\tilde{\mathbf{P}}, \tilde{\mathbf{E}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{E}} \in \mathbb{R}^{r \times r}$ and $\mathbf{X}, \mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ satisfy the matrix equations

$$\tilde{\mathbf{A}} \tilde{\mathbf{P}} \tilde{\mathbf{E}}^\top + \tilde{\mathbf{E}} \tilde{\mathbf{P}} \tilde{\mathbf{A}}^\top + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top = \mathbf{0}_{r \times r}, \quad (6.6a)$$

$$\tilde{\mathbf{A}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{E}} + \tilde{\mathbf{E}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{A}} + \sum_{k=1}^p \tilde{\mathbf{M}}_k \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k + \tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} = \mathbf{0}_{r \times r}, \quad (6.6b)$$

$$\mathbf{A} \mathbf{X} \tilde{\mathbf{E}}^\top + \mathbf{E} \mathbf{X} \tilde{\mathbf{A}}^\top + \mathbf{B} \tilde{\mathbf{B}}^\top = \mathbf{0}_{n \times r}, \quad (6.6c)$$

$$\mathbf{A}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{A}} - \sum_{k=1}^p \mathbf{M}_k \mathbf{X} \tilde{\mathbf{M}}_k - \mathbf{C}^\top \tilde{\mathbf{C}} = \mathbf{0}_{n \times r}. \quad (6.6d)$$

Recall that the matrices $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}_{\text{lqo}}$ are SPSD since they are the Gramians of the reduced model $\tilde{\mathcal{G}}_{\text{lqo}}$. At this point, applying Theorem 5.10 to the error system yields the following corollary.

Corollary 6.2 (Sylvester equation-based \mathcal{H}_2 model reduction error [186]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2). Then, the squared \mathcal{H}_2 model reduction error $\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2$ is given by

$$\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 = \text{tr} \left(\mathbf{B}^\top \mathbf{Q}_{\text{lqo}} \mathbf{B} + 2\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}} + \tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{B}} \right) \quad (6.7)$$

$$\begin{aligned} &= \text{tr} \left(\mathbf{C} \mathbf{P} \tilde{\mathbf{C}}^\top - 2\mathbf{C} \mathbf{X} \tilde{\mathbf{C}}^\top + \tilde{\mathbf{C}} \tilde{\mathbf{P}} \tilde{\mathbf{C}}^\top \right) \\ &\quad + \sum_{k=1}^p \text{tr} (\mathbf{P} \mathbf{M}_k \mathbf{P} \mathbf{M}_k) - 2 \text{tr} \left(\mathbf{X}^\top \mathbf{M}_k \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k + \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \right) \end{aligned} \quad (6.8)$$

where $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}_{\text{lqo}} \in \mathbb{R}^{r \times r}$ and $\mathbf{X}, \mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ satisfy the generalized matrix equations in (6.6). \diamond

Proof of Corollary 6.2. This follows directly by applying the Gramian-based expressions for the system \mathcal{H}_2 norm in (5.37) and (5.36) to the error system in (6.3), and resolving the block-matrix multiplication. \square

We now return our attention to the \mathcal{H}_2 -minimization problem (6.1): Consider the objective cost function $\mathcal{J}: \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times r} \times \mathbb{R}^{r \times m} \times \mathbb{R}^{p \times r} \times \mathbb{R}^{r \times r} \times \dots \times \mathbb{R}^{r \times r} \rightarrow \mathbb{R}$ defined as

$$\begin{aligned} \mathcal{J} \left(\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{M}}_1, \dots, \tilde{\mathbf{M}}_p \right) &\stackrel{\text{def}}{=} \|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 \\ &= \text{tr} \left(\mathbf{B}^\top \mathbf{Q}_{\text{lqo}} \mathbf{B} + 2\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \tilde{\mathbf{B}} + \tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{B}} \right) \\ &= \text{tr} \left(\mathbf{C} \mathbf{P} \tilde{\mathbf{C}}^\top - 2\mathbf{C} \mathbf{X} \tilde{\mathbf{C}}^\top + \tilde{\mathbf{C}} \tilde{\mathbf{P}} \tilde{\mathbf{C}}^\top \right) \\ &\quad + \sum_{k=1}^p \text{tr} (\mathbf{P} \mathbf{M}_k \mathbf{P} \mathbf{M}_k) \\ &\quad - 2 \text{tr} \left(\mathbf{X}^\top \mathbf{M}_k \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k + \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \right) \end{aligned} \quad (6.9)$$

subject to the constraints that $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}_{\text{lqo}} \in \mathbb{R}^{r \times r}$ and $\mathbf{X}, \mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ satisfy (6.6). In other words, we treat the squared \mathcal{H}_2 model reduction error \mathcal{J} in (6.9) as taking the reduced-order matrices that determine $\tilde{\mathcal{G}}_{\text{lqo}}$ as arguments; \mathcal{J} is thus a multivariate function defined over real-valued Hilbert spaces of matrices in each variable. Per Proposition 2.11, if an approximation $\tilde{\mathcal{G}}_{\text{lqo}}$ to \mathcal{G}_{lqo} minimizes the \mathcal{H}_2 model reduction error, then the partial gradients of \mathcal{J} with respect to the reduced model matrices $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}$ and $\tilde{\mathbf{M}}_k$ for each $k = 1, \dots, p$ are necessarily zero. This paves the way for us to derive first-order Sylvester equation-based necessary conditions for \mathcal{H}_2 optimality. In addition to the constraints in (6.6), we recall from Section 2.4.2 the matrices $\mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ and $\tilde{\mathbf{Q}}_{\text{lo}} \in \mathbb{R}^{r \times r}$, which satisfy

$$\tilde{\mathbf{A}}^\top \mathbf{Q}_{\text{lo}} \tilde{\mathbf{E}} + \tilde{\mathbf{E}}^\top \mathbf{Q}_{\text{lo}} \tilde{\mathbf{A}} + \tilde{\mathbf{C}}^\top \tilde{\mathbf{C}} = \mathbf{0}_{r \times r}, \quad (6.10a)$$

$$\mathbf{A}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{E}} + \tilde{\mathbf{E}}^\top \mathbf{Z}_{\text{lo}} \tilde{\mathbf{A}} - \mathbf{C}^\top \tilde{\mathbf{C}} = \mathbf{0}_{n \times r}. \quad (6.10b)$$

We emphasize however that the matrix equations in (6.6b) and (6.6d) are *distinct* from (6.10a) and (6.10b) when $\mathbf{M}_k \neq \mathbf{0}_{n \times n}$, i.e., the solutions are such that $\tilde{\mathbf{Q}}_{\text{lqo}} \neq \tilde{\mathbf{Q}}_{\text{lo}}$ and $\mathbf{Z}_{\text{lqo}} \neq \mathbf{Z}_{\text{lo}}$.

We are now prepared to state and prove the first major theoretical result of this section, establishing the gradients of the squared \mathcal{H}_2 system error with respect to matrices in (5.2). This was presented in the author's previous work [186] for the special case of $\mathbf{E} = \mathbf{I}_n$; as with the earlier results in this section, we derive gradients of the squared \mathcal{H}_2 error for arbitrary nonsingular \mathbf{E} . When it helps simplify notation, we drop the dependence of \mathcal{J} on the reduced model $\tilde{\mathcal{G}}_{\text{lqo}}$ in the proof of the subsequent result.

Theorem 6.3 (Gradients of the squared \mathcal{H}_2 system error [186, Theorem 3.1]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2). The gradients of the squared \mathcal{H}_2 error in (6.1) with respect to the reduced-order matrices in (5.2) are given as

$$\nabla_{\tilde{\mathbf{E}}} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) = 2 \left(\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{A}} \tilde{\mathbf{P}} + \left(2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top \right) \mathbf{A} \mathbf{X} \right), \quad (6.11a)$$

$$\nabla_{\tilde{\mathbf{A}}} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) = 2 \left(\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{E}} \tilde{\mathbf{P}} + \left(2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top \right) \mathbf{E} \mathbf{X} \right), \quad (6.11b)$$

$$\nabla_{\tilde{\mathbf{B}}} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) = 2 \left(\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{B}} + \left(2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top \right) \mathbf{B} \right), \quad (6.11c)$$

$$\nabla_{\tilde{\mathbf{C}}} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) = 2 \left(\tilde{\mathbf{C}} \tilde{\mathbf{P}} - \mathbf{C} \mathbf{X} \right), \quad (6.11d)$$

$$\nabla_{\tilde{\mathbf{M}}} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) = 2 \left(\tilde{\mathbf{M}} \left(\tilde{\mathbf{P}} \otimes \tilde{\mathbf{P}} \right) - \mathbf{M} \left(\mathbf{X} \otimes \mathbf{X} \right) \right), \quad (6.11e)$$

where $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}_{\text{lqo}} \in \mathbb{R}^{r \times r}$, $\mathbf{X}, \mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ satisfy the generalized matrix equations (6.6), and $\tilde{\mathbf{Q}}_{\text{lo}} \in \mathbb{R}^{r \times r}$ and $\mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ satisfy (6.10). \diamond

Before proving Theorem 6.3, we prove a lemma that we will invoke repeatedly to simplify the subsequent arguments.

Lemma 6.4 ([238, Lemma A.1]). Suppose that $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n \times n}$, $\tilde{\mathbf{E}}, \tilde{\mathbf{A}} \in \mathbb{R}^{r \times r}$ and $\mathbf{D}, \mathbf{F} \in \mathbb{R}^{n \times r}$. If $\mathbf{Y}, \mathbf{W} \in \mathbb{R}^{n \times r}$ solve the generalized Sylvester equations

$$\mathbf{A} \mathbf{Y} \tilde{\mathbf{E}}^\top + \mathbf{E} \mathbf{Y} \tilde{\mathbf{A}}^\top + \mathbf{D} = \mathbf{0}_{n \times r} \quad \text{and} \quad \mathbf{A}^\top \mathbf{W} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{W} \tilde{\mathbf{A}} + \mathbf{F} = \mathbf{0}_{n \times r},$$

then $\text{tr}(\mathbf{D}^\top \mathbf{W}) = \text{tr}(\mathbf{F}^\top \mathbf{Y})$. \diamond

Proof of Lemma 6.4. Observe, by invariance of the trace under cyclic permutations and matrix transposition

$$\begin{aligned} \text{tr}(\mathbf{D}^\top \mathbf{W}) &= \text{tr} \left(- \left(\mathbf{A} \mathbf{Y} \tilde{\mathbf{E}}^\top + \mathbf{E} \mathbf{Y} \tilde{\mathbf{A}}^\top \right)^\top \mathbf{W} \right) = \text{tr} \left(- \left(\tilde{\mathbf{E}} \mathbf{Y}^\top \mathbf{A}^\top \mathbf{W} + \tilde{\mathbf{A}} \mathbf{Y}^\top \mathbf{E}^\top \mathbf{W} \right) \right) \\ &= \text{tr} \left(- \mathbf{Y}^\top \left(\mathbf{A}^\top \mathbf{W} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{W} \tilde{\mathbf{A}} \right) \right) = \text{tr}(\mathbf{F}^\top \mathbf{Y}). \end{aligned}$$

\square

From Lemma 6.4, we also have $\text{tr}(\mathbf{W}^\top \mathbf{D}) = \text{tr}(\mathbf{Y}^\top \mathbf{F})$ by properties of the trace. In the subsequent result, we take for granted any such identities that arise from cyclic permutations or transposes applied to the result of Lemma 6.4.

Proof of Theorem 6.3. Throughout, \mathcal{G}_{lqo} is arbitrarily specified but fixed. We begin by calculating the gradient with respect to $\widetilde{\mathbf{M}}$. We prove this by taking gradients with respect to each $\widetilde{\mathbf{M}}_k$ individually, and then casting these in the form of (6.11e). We claim

$$\nabla_{\widetilde{\mathbf{M}}_k} \mathcal{J} = 2 \left(\widetilde{\mathbf{P}} \widetilde{\mathbf{M}}_k \widetilde{\mathbf{P}} - \mathbf{X}^\top \mathbf{M} \mathbf{X} \right) \quad \text{for each } k = 1, \dots, p. \quad (6.12)$$

Choose any $k = 1, \dots, p$ and consider an arbitrary infinitesimal perturbation $\Delta_{\mathbf{M}_k} \in \mathbb{R}^{r \times r}$. We wish to show that $\nabla_{\widetilde{\mathbf{M}}_k} \mathcal{J}$ in (6.12) satisfies

$$\mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right) = \mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} \right) + \left\langle \nabla_{\widetilde{\mathbf{M}}_k} \mathcal{J}, \Delta_{\mathbf{M}_k} \right\rangle_{\mathbb{F}} + O \left(\|\Delta_{\mathbf{M}_k}\|_{\mathbb{F}}^2 \right),$$

where $\Delta_{\mathcal{G}}$ denotes the perturbation in the reduced model $\widetilde{\mathcal{G}}_{\text{lqo}}$ due to $\Delta_{\mathbf{M}_k}$. The first-order perturbation in $\widetilde{\mathbf{M}}_k$ induces additional perturbations $\Delta_{\mathbf{Z}} \in \mathbb{R}^{n \times r}$ and $\Delta_{\mathbf{Q}} \in \mathbb{R}^{r \times r}$ in the solutions to the matrix equations (6.6d) and (6.6b). Taking the resulting expansions of the perturbed matrix equations (6.6d) and (6.6b) reveals that the perturbations $\Delta_{\mathbf{Z}}$ and $\Delta_{\mathbf{Q}}$ themselves satisfy

$$\mathbf{A}^\top \Delta_{\mathbf{Z}} \widetilde{\mathbf{E}} + \mathbf{E}^\top \Delta_{\mathbf{Z}} \widetilde{\mathbf{A}} - \mathbf{M}_k \mathbf{X} \Delta_{\mathbf{M}_k}^\top = \mathbf{0}_{n \times r} \quad (6.13a)$$

$$\text{and } \widetilde{\mathbf{A}}^\top \Delta_{\mathbf{Q}} \widetilde{\mathbf{E}} + \mathbf{E}^\top \Delta_{\mathbf{Q}} \widetilde{\mathbf{A}} + \widetilde{\mathbf{M}}_k \widetilde{\mathbf{P}} \Delta_{\mathbf{M}_k}^\top + \Delta_{\mathbf{M}_k} \widetilde{\mathbf{P}} \widetilde{\mathbf{M}}_k + \Delta_{\mathbf{M}_k} \widetilde{\mathbf{P}} \Delta_{\mathbf{M}_k}^\top = \mathbf{0}_{n \times r}. \quad (6.13b)$$

Note that the term $\Delta_{\mathbf{M}_k} \widetilde{\mathbf{P}} \Delta_{\mathbf{M}_k}^\top$ is $O(\|\Delta_{\mathbf{M}_k}\|_{\mathbb{F}}^2)$ in the sense of Section 2.2. Using the form of the objective function in (6.7), the first-order expansion of the error $\mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right)$ is given as

$$\mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right) = \mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} \right) + 2 \text{tr} \left(\mathbf{B}^\top \Delta_{\mathbf{Z}} \widetilde{\mathbf{B}} \right) + \text{tr} \left(\widetilde{\mathbf{B}}^\top \Delta_{\mathbf{Q}} \widetilde{\mathbf{B}} \right). \quad (6.14)$$

By applying Lemma 6.4 to the Sylvester equations (6.6c) and (6.13a) and using the invariance of the trace under permutation and matrix transposition, the terms in the first-order expansion (6.14) can be rewritten as

$$\begin{aligned} 2 \text{tr} \left(\mathbf{B}^\top \Delta_{\mathbf{Z}} \widetilde{\mathbf{B}} \right) &= 2 \text{tr} \left(\widetilde{\mathbf{B}} \mathbf{B}^\top \Delta_{\mathbf{Z}} \right) = 2 \text{tr} \left(-\mathbf{M}_k \mathbf{X} \Delta_{\mathbf{M}_k}^\top \mathbf{X}^\top \right) \\ &= 2 \text{tr} \left(-\mathbf{X}^\top \mathbf{M}_k \mathbf{X} \Delta_{\mathbf{M}_k} \right). \end{aligned}$$

Lemma 6.4 can similarly be applied to equations (6.6a) and (6.13b) to show

$$\text{tr} \left(\widetilde{\mathbf{B}}^\top \Delta_{\mathbf{Q}} \widetilde{\mathbf{B}} \right) = \text{tr} \left(\widetilde{\mathbf{B}} \widetilde{\mathbf{B}}^\top \Delta_{\mathbf{Q}} \right) = 2 \text{tr} \left(\widetilde{\mathbf{P}} \widetilde{\mathbf{M}}_k \widetilde{\mathbf{P}} \Delta_{\mathbf{M}_k} \right) + O \left(\|\Delta_{\mathbf{M}_k}\|_{\mathbb{F}}^2 \right).$$

Substituting these expressions for $\text{tr}(\mathbf{B}^\top \Delta_Z \tilde{\mathbf{B}})$ and $\text{tr}(\tilde{\mathbf{B}}^\top \Delta_Q \tilde{\mathbf{B}})$ into the expansion (6.14), we get

$$\begin{aligned} \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}}) &= \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}}) + 2 \text{tr}(-\mathbf{X}^\top \mathbf{M}_k \mathbf{X} \Delta_{\mathbf{M}_k}) + 2 \text{tr}(\tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \tilde{\mathbf{P}} \Delta_{\mathbf{M}_k}) + O(\|\Delta_{\mathbf{M}_k}\|_{\mathbb{F}}^2) \\ &= \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}}) + \left\langle 2 \left(\tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \tilde{\mathbf{P}} - \mathbf{X}^\top \mathbf{M}_k \mathbf{X} \right), \Delta_{\mathbf{M}_k} \right\rangle_{\mathbb{F}} + O(\|\Delta_{\mathbf{M}_k}\|_{\mathbb{F}}^2). \end{aligned}$$

So, the gradients with respect to $\tilde{\mathbf{M}}_k$ satisfy $\nabla_{\tilde{\mathbf{M}}_k} \mathcal{J} = 2 \left(\tilde{\mathbf{P}} \tilde{\mathbf{M}}_k \tilde{\mathbf{P}} - \mathbf{X}^\top \mathbf{M}_k \mathbf{X} \right)$ for each k as claimed in (6.11e). Vectorizing these for each k and concatenating them vertically gives the gradient with respect to $\tilde{\mathbf{M}}$, which is

$$\begin{aligned} \nabla_{\tilde{\mathbf{M}}} \mathcal{J} &= \begin{bmatrix} \text{vec}(\nabla_{\tilde{\mathbf{M}}_1} \mathcal{J})^\top \\ \vdots \\ \text{vec}(\nabla_{\tilde{\mathbf{M}}_p} \mathcal{J})^\top \end{bmatrix} = 2 \begin{bmatrix} \text{vec}(\tilde{\mathbf{P}} \tilde{\mathbf{M}}_1 \tilde{\mathbf{P}} - \mathbf{X}^\top \mathbf{M}_1 \mathbf{X}) \\ \vdots \\ \text{vec}(\tilde{\mathbf{P}} \tilde{\mathbf{M}}_p \tilde{\mathbf{P}} - \mathbf{X}^\top \mathbf{M}_p \mathbf{X}) \end{bmatrix} \\ &= 2 \left(\tilde{\mathbf{M}} \left(\tilde{\mathbf{P}} \otimes \tilde{\mathbf{P}} \right) - \mathbf{M} \left(\mathbf{X} \otimes \mathbf{X} \right) \right) \end{aligned}$$

by (2.7). This proves (6.11e),

We next compute $\nabla_{\tilde{\mathbf{C}}} \mathcal{J}$. Consider an infinitesimal arbitrary perturbation $\Delta_{\mathbf{C}} \in \mathbb{R}^{p \times r}$ to $\tilde{\mathbf{C}}$ and let $\Delta_{\mathcal{G}}$ be the perturbation to $\tilde{\mathcal{G}}_{\text{lqo}}$ corresponding to $\Delta_{\mathbf{C}}$. Note that, in contrast to the previous argument, this perturbation in $\tilde{\mathbf{C}}$ does not induce any further perturbations in $\tilde{\mathbf{P}}$ and \mathbf{X} . From (6.8), the first-order expansion of $\mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}})$ is

$$\begin{aligned} \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}}) &= \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}}) - 2 \text{tr}(\mathbf{C} \mathbf{X} \Delta_{\mathbf{C}}^\top) + 2 \text{tr}(\tilde{\mathbf{C}} \tilde{\mathbf{P}} \Delta_{\mathbf{C}}^\top) + O(\|\Delta_{\mathbf{C}}\|_{\mathbb{F}}^2) \\ &= \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}}) + \left\langle 2 \left(\tilde{\mathbf{C}} \tilde{\mathbf{P}} - \mathbf{C} \mathbf{X} \right), \Delta_{\mathbf{C}} \right\rangle_{\mathbb{F}} + O(\|\Delta_{\mathbf{C}}\|_{\mathbb{F}}^2). \end{aligned}$$

So $\nabla_{\tilde{\mathbf{C}}} \mathcal{J} = 2 \left(\tilde{\mathbf{C}} \tilde{\mathbf{P}} - \mathbf{C} \mathbf{X} \right)$ as claimed in (6.11d).

We next compute the gradient $\nabla_{\tilde{\mathbf{B}}} \mathcal{J}$ in (6.11c). Consider an arbitrary infinitesimal perturbation $\Delta_{\mathbf{B}} \in \mathbb{R}^{r \times m}$ to $\tilde{\mathbf{B}}$. This induces perturbations $\Delta_{\mathbf{X}} \in \mathbb{R}^{n \times r}$ and $\Delta_{\mathbf{P}} \in \mathbb{R}^{n \times r}$ in the solutions of (6.6c) and (6.6a). The resulting perturbations satisfy

$$\mathbf{A} \Delta_{\mathbf{X}} \tilde{\mathbf{E}}^\top + \mathbf{E} \Delta_{\mathbf{X}} \tilde{\mathbf{A}}^\top + \mathbf{B} \Delta_{\mathbf{B}}^\top = \mathbf{0}_{n \times r}, \quad (6.15a)$$

$$\text{and } \tilde{\mathbf{A}} \Delta_{\mathbf{P}} \tilde{\mathbf{E}}^\top + \tilde{\mathbf{E}} \Delta_{\mathbf{P}} \tilde{\mathbf{A}}^\top + \Delta_{\mathbf{B}} \tilde{\mathbf{B}}^\top + \tilde{\mathbf{B}} \Delta_{\mathbf{B}}^\top + O(\|\Delta_{\mathbf{B}}\|_{\mathbb{F}}^2) = \mathbf{0}_{r \times r}. \quad (6.15b)$$

The solutions \mathbf{Z}_{lqo} and $\tilde{\mathbf{Q}}_{\text{lqo}}$ to (6.6b) and (6.6d) depend linearly upon \mathbf{X} and $\tilde{\mathbf{P}}$, so $\Delta_{\mathbf{X}}$ and $\Delta_{\mathbf{P}}$ induce further perturbations $\Delta_{\mathbf{Z}} \in \mathbb{R}^{n \times r}$ and $\Delta_{\mathbf{Q}} \in \mathbb{R}^{n \times r}$ in the solutions to (6.6d)

and (6.6b). These perturbations satisfy

$$\mathbf{A}^\top \Delta_Z \tilde{\mathbf{E}} + \mathbf{E} \Delta_Z \tilde{\mathbf{A}} - \sum_{k=1}^p \mathbf{M}_k \Delta_X \tilde{\mathbf{M}}_k = \mathbf{0}, \quad (6.16a)$$

$$\text{and } \tilde{\mathbf{A}}^\top \Delta_Q \tilde{\mathbf{E}} + \tilde{\mathbf{E}}^\top \Delta_Q \tilde{\mathbf{A}} + \sum_{k=1}^p \tilde{\mathbf{M}}_k \Delta_P \tilde{\mathbf{M}}_k = \mathbf{0}. \quad (6.16b)$$

From (6.7) we may expand $\mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_G)$ as

$$\begin{aligned} \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_G) &= \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}}) + 2 \operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z) \\ &\quad + \operatorname{tr}(2 \tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Q) + O(\|\Delta_B\|_{\mathbb{F}}^2). \end{aligned} \quad (6.17)$$

We handle the terms in (6.17) individually. First, $\operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z)$ splits into the individual terms $\operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B)$ and $\operatorname{tr}(\tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z)$. Using properties of the trace and applying Lemma 6.4 to equations (6.6c) and (6.16a), we see that $\operatorname{tr}(\tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z)$ can be written as

$$\begin{aligned} \operatorname{tr}(\tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z) &= \operatorname{tr}\left(-\left(\sum_{k=1}^p \tilde{\mathbf{M}}_k \Delta_X^\top \mathbf{M}_k\right) \mathbf{X}\right) \\ &= \operatorname{tr}\left(-\Delta_X \sum_{k=1}^p \tilde{\mathbf{M}}_k \mathbf{X}^\top \mathbf{M}_k\right) \\ &= \operatorname{tr}\left(-\left(\tilde{\mathbf{E}}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{A} + \tilde{\mathbf{A}}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{E}\right) \Delta_X\right) + \operatorname{tr}(\Delta_X \tilde{\mathbf{C}}^\top \mathbf{C}) \\ &= \operatorname{tr}\left(-\mathbf{Z}_{\text{lqo}}^\top \left(\mathbf{A} \Delta_X \tilde{\mathbf{E}}^\top + \mathbf{E} \Delta_X \tilde{\mathbf{A}}^\top\right)\right) + \operatorname{tr}(\Delta_X \tilde{\mathbf{C}}^\top \mathbf{C}) \\ &= \operatorname{tr}(\mathbf{Z}_{\text{lqo}}^\top \mathbf{B} \Delta_B^\top) + \operatorname{tr}(\Delta_X \tilde{\mathbf{C}}^\top \mathbf{C}) \quad \text{by (6.15a)} \\ &= \operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B) + \operatorname{tr}(\Delta_X \tilde{\mathbf{C}}^\top \mathbf{C}). \end{aligned}$$

Applying Lemma 6.4 to equations (6.10b) and (6.15a), it follows that

$$\operatorname{tr}(\Delta_X \tilde{\mathbf{C}} \mathbf{C}^\top) = \operatorname{tr}(\tilde{\mathbf{C}} \mathbf{C}^\top \Delta_X) = \operatorname{tr}(-\Delta_B \mathbf{B}^\top \mathbf{Z}_{\text{lo}}) = \operatorname{tr}(-\mathbf{B}^\top \mathbf{Z}_{\text{lo}} \Delta_B).$$

So the term $2 \operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z)$ in (6.17) becomes

$$2 \operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Z) = 4 \operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B) - 2 \operatorname{tr}(\mathbf{B}^\top \mathbf{Z}_{\text{lo}} \Delta_B).$$

The term $\text{tr} \left(2\tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Q \right)$ in (6.17) can be dealt with following similar calculations to show that

$$\text{tr} \left(2\tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_B + \tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Q \right) = 4 \text{tr} \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_B \right) - 2 \text{tr} \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lo}} \Delta_B \right) + O \left(\|\Delta_B\|_{\mathbb{F}}^2 \right).$$

The expansion in (6.17) thus becomes

$$\begin{aligned} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right) &= \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) + 4 \text{tr} \left(\mathbf{B}^\top \mathbf{Z}_{\text{lqo}} \Delta_B \right) - 2 \text{tr} \left(\mathbf{B}^\top \mathbf{Z}_{\text{lo}} \Delta_B \right) \\ &\quad + 4 \text{tr} \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_B \right) - 2 \text{tr} \left(\tilde{\mathbf{B}}^\top \tilde{\mathbf{Q}}_{\text{lo}} \Delta_B \right) + O \left(\|\Delta_B\|_{\mathbb{F}}^2 \right) \\ &= \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) + \left\langle 2 \left(\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{B}} + \left(2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top \right) \mathbf{B} \right), \Delta_B \right\rangle_{\mathbb{F}} \\ &\quad + O \left(\|\Delta_B\|_{\mathbb{F}}^2 \right). \end{aligned}$$

Therefore $\nabla_{\tilde{\mathbf{B}}} \mathcal{J} = 2 \left(\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{B}} + \left(2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top \right) \mathbf{B} \right)$ as claimed.

We next compute $\nabla_{\tilde{\mathbf{A}}} \mathcal{J}$. Consider an arbitrary infinitesimal perturbation $\Delta_{\mathbf{A}} \in \mathbb{R}^{r \times r}$ about zero to $\tilde{\mathbf{A}}$. As was the case for the gradient with respect to $\tilde{\mathbf{B}}$, this first-order perturbation induces perturbations $\Delta_{\mathbf{X}}, \Delta_{\mathbf{Z}} \in \mathbb{R}^{n \times r}$, and $\Delta_{\mathbf{P}}, \Delta_{\mathbf{Q}} \in \mathbb{R}^{r \times r}$ in the solutions to (6.6). These satisfy

$$\mathbf{A} \Delta_{\mathbf{X}} \tilde{\mathbf{E}}^\top + \mathbf{E} \Delta_{\mathbf{X}} \tilde{\mathbf{A}}^\top + \mathbf{E} \mathbf{X} \Delta_{\mathbf{A}}^\top + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right) = \mathbf{0}_{n \times r}, \quad (6.18a)$$

$$\tilde{\mathbf{A}} \Delta_{\mathbf{P}} \tilde{\mathbf{E}}^\top + \tilde{\mathbf{E}} \Delta_{\mathbf{P}} \tilde{\mathbf{A}}^\top + \Delta_{\mathbf{A}} \tilde{\mathbf{P}} \tilde{\mathbf{E}}^\top + \tilde{\mathbf{E}} \tilde{\mathbf{P}} \Delta_{\mathbf{A}}^\top + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right) = \mathbf{0}_{r \times r}, \quad (6.18b)$$

$$\mathbf{A}^\top \Delta_{\mathbf{Z}} \tilde{\mathbf{E}} + \mathbf{E}^\top \Delta_{\mathbf{Z}} \tilde{\mathbf{A}} + \mathbf{E}^\top \mathbf{Z}_{\text{lqo}} \Delta_{\mathbf{A}} - \sum_{k=1}^p \mathbf{M}_k \Delta_{\mathbf{X}} \tilde{\mathbf{M}}_k + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right) = \mathbf{0}_{n \times r}, \quad (6.18c)$$

$$\tilde{\mathbf{A}}^\top \Delta_{\mathbf{Q}} \tilde{\mathbf{E}} + \tilde{\mathbf{E}}^\top \Delta_{\mathbf{Q}} \tilde{\mathbf{A}} + \Delta_{\mathbf{A}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \tilde{\mathbf{E}} + \tilde{\mathbf{E}}^\top \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_{\mathbf{A}} + \sum_{k=1}^p \tilde{\mathbf{M}}_k \Delta_{\mathbf{P}} \tilde{\mathbf{M}}_k + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right) = \mathbf{0}_{r \times r}. \quad (6.18d)$$

By (6.7), the error $\mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right)$ may be expanded as

$$\mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right) = \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) + 2 \text{tr} \left(\tilde{\mathbf{B}} \mathbf{B}^\top \Delta_{\mathbf{Z}} \right) + \text{tr} \left(\tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_{\mathbf{Q}} \right). \quad (6.19)$$

We deal with the terms in (6.19) individually as follows. Applying Lemma 6.4 to equations (6.6c) and (6.18c), $\text{tr} \left(\tilde{\mathbf{B}} \mathbf{B}^\top \Delta_{\mathbf{Z}} \right)$ can be re-written as

$$\begin{aligned} \text{tr} \left(\tilde{\mathbf{B}} \mathbf{B}^\top \Delta_{\mathbf{Z}} \right) &= \text{tr} \left(- \left(\sum_{k=1}^p \tilde{\mathbf{M}}_k \Delta_{\mathbf{X}}^\top \mathbf{M}_k - \Delta_{\mathbf{A}}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{E} \right) \mathbf{X} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right) \\ &= \text{tr} \left(\mathbf{X}^\top \mathbf{E}^\top \mathbf{Z}_{\text{lqo}} \Delta_{\mathbf{A}} \right) - \text{tr} \left(\sum_{k=1}^p \tilde{\mathbf{M}}_k \mathbf{X}^\top \mathbf{M}_k \Delta_{\mathbf{X}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right), \quad (6.20a) \end{aligned}$$

where we have used the fact that \mathbf{M}_k and $\widetilde{\mathbf{M}}_k$ are symmetric. From (6.6d) and (6.18a), observe that

$$\begin{aligned} \operatorname{tr} \left(\sum_{k=1}^p \widetilde{\mathbf{M}}_k \mathbf{X}^\top \mathbf{M}_k \Delta_{\mathbf{X}} \right) &= \operatorname{tr} \left(\left(\widetilde{\mathbf{E}}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{A} + \widetilde{\mathbf{A}}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{E} \right) \Delta_{\mathbf{X}} \right) - \operatorname{tr} \left(\widetilde{\mathbf{C}}^\top \mathbf{C} \Delta_{\mathbf{X}} \right) \\ &= \operatorname{tr} \left(\mathbf{Z}_{\text{lqo}}^\top \left(\mathbf{A} \Delta_{\mathbf{X}} \widetilde{\mathbf{E}}^\top + \mathbf{E} \Delta_{\mathbf{X}} \widetilde{\mathbf{A}}^\top \right) \right) - \operatorname{tr} \left(\widetilde{\mathbf{C}}^\top \mathbf{C} \Delta_{\mathbf{X}} \right) \\ &= -\operatorname{tr} \left(\mathbf{Z}_{\text{lqo}}^\top \mathbf{E} \mathbf{X} \Delta_{\mathbf{A}}^\top \right) - \operatorname{tr} \left(\widetilde{\mathbf{C}}^\top \mathbf{C} \Delta_{\mathbf{X}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right). \end{aligned} \quad (6.20b)$$

Applying Lemma 6.4 to (6.10b) and (6.18a) allows us to simplify the $\operatorname{tr} \left(\widetilde{\mathbf{C}}^\top \mathbf{C} \Delta_{\mathbf{X}} \right)$ term further as

$$-\operatorname{tr} \left(\widetilde{\mathbf{C}}^\top \mathbf{C} \Delta_{\mathbf{X}} \right) = \operatorname{tr} \left(\Delta_{\mathbf{A}} \mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lo}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right).$$

So, the term (6.20b) appearing in (6.20a) ultimately becomes

$$-\operatorname{tr} \left(\sum_{k=1}^p \widetilde{\mathbf{M}}_k \mathbf{X}^\top \mathbf{M}_k \Delta_{\mathbf{X}} \right) = \operatorname{tr} \left(\mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lqo}} \Delta_{\mathbf{A}} \right) - \operatorname{tr} \left(\mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lo}} \Delta_{\mathbf{A}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right)$$

and (6.20a) in (6.19) is given by

$$\operatorname{tr} \left(\widetilde{\mathbf{B}} \mathbf{B}^\top \Delta_{\mathbf{Z}} \right) = 2 \operatorname{tr} \left(\mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lqo}} \Delta_{\mathbf{A}} \right) - \operatorname{tr} \left(\mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lo}} \Delta_{\mathbf{A}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right).$$

Following similar calculations involving (6.18b) and (6.18d), the term $\operatorname{tr} \left(\widetilde{\mathbf{B}} \widetilde{\mathbf{B}}^\top \Delta_{\mathbf{Q}} \right)$ in (6.19) can be expressed as

$$\operatorname{tr} \left(\widetilde{\mathbf{B}} \widetilde{\mathbf{B}}^\top \Delta_{\mathbf{Q}} \right) = 4 \operatorname{tr} \left(\widetilde{\mathbf{P}} \widetilde{\mathbf{E}} \widetilde{\mathbf{Q}}_{\text{lqo}} \Delta_{\mathbf{A}} \right) - 2 \operatorname{tr} \left(\widetilde{\mathbf{P}} \widetilde{\mathbf{E}} \widetilde{\mathbf{Q}}_{\text{lo}} \Delta_{\mathbf{A}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right).$$

So, the expansion (6.19) simplifies to

$$\begin{aligned} \mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right) &= \mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} \right) + 4 \operatorname{tr} \left(\mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lqo}} \Delta_{\mathbf{A}} \right) - 2 \operatorname{tr} \left(\mathbf{X}^\top \mathbf{E} \mathbf{Z}_{\text{lo}} \Delta_{\mathbf{A}} \right) \\ &\quad + 4 \operatorname{tr} \left(\widetilde{\mathbf{P}} \widetilde{\mathbf{E}} \widetilde{\mathbf{Q}}_{\text{lqo}} \Delta_{\mathbf{A}} \right) - 2 \operatorname{tr} \left(\widetilde{\mathbf{P}} \widetilde{\mathbf{E}} \widetilde{\mathbf{Q}}_{\text{lo}} \Delta_{\mathbf{A}} \right) + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right) \\ &= \mathcal{J} \left(\widetilde{\mathcal{G}}_{\text{lqo}} \right) + \left\langle 2 \left((2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{E} \mathbf{X} + (2\widetilde{\mathbf{Q}}_{\text{lqo}} - \widetilde{\mathbf{Q}}_{\text{lo}}) \widetilde{\mathbf{E}} \widetilde{\mathbf{P}} \right), \Delta_{\mathbf{A}} \right\rangle_{\mathbb{F}} \\ &\quad + O \left(\|\Delta_{\mathbf{A}}\|_{\mathbb{F}}^2 \right). \end{aligned}$$

Thus $\nabla_{\widetilde{\mathbf{A}}} \mathcal{J} = 2 \left((2\mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{E} \mathbf{X} + (2\widetilde{\mathbf{Q}}_{\text{lqo}} - \widetilde{\mathbf{Q}}_{\text{lo}}) \widetilde{\mathbf{E}} \widetilde{\mathbf{P}} \right)$ holds as claimed.

Finally, we compute the gradient $\nabla_{\widetilde{\mathbf{E}}} \mathcal{J}$. This calculation is nearly identical to the previous one for $\nabla_{\widetilde{\mathbf{A}}} \mathcal{J}$, but we include it here for completeness. Consider an arbitrary infinitesimal

perturbation $\Delta_E \in \mathbb{R}^{r \times r}$ to \tilde{E} . This first-order perturbation induces perturbations Δ_X , $\Delta_Z \in \mathbb{R}^{n \times r}$, and $\Delta_P, \Delta_Q \in \mathbb{R}^{r \times r}$ in the solutions to (6.6). These satisfy

$$\mathbf{A}\mathbf{X}\Delta_E^\top + \mathbf{A}\Delta_X\tilde{E}^\top + \mathbf{E}\Delta_X\tilde{A}^\top + O(\|\Delta_E\|_F^2) = \mathbf{0}_{n \times r}, \quad (6.21a)$$

$$\tilde{A}\tilde{P}\Delta_E^\top + \Delta_E\tilde{P}\tilde{A}^\top + \tilde{A}\Delta_P\tilde{E}^\top + \tilde{E}\Delta_P\tilde{A}^\top + O(\|\Delta_E\|_F^2) = \mathbf{0}_{r \times r}, \quad (6.21b)$$

$$\mathbf{A}^\top \mathbf{Z}_{\text{lqo}} \Delta_E + \mathbf{A}^\top \Delta_Z \tilde{E} + \mathbf{E}^\top \Delta_Z \tilde{A}^\top - \sum_{k=1}^p \mathbf{M}_k \Delta_X \tilde{M}_k + O(\|\Delta_E\|_F^2) = \mathbf{0}_{n \times r}, \quad (6.21c)$$

$$\tilde{A}^\top \Delta_Q \tilde{E} + \tilde{E}^\top \Delta_Q \tilde{A} + \tilde{A}^\top \mathbf{Q}_{\text{lqo}} \Delta_E + \Delta_E^\top \mathbf{Q}_{\text{lqo}} \tilde{A} + \sum_{k=1}^p \tilde{M}_k \Delta_P \tilde{M}_k + O(\|\Delta_E\|_F^2) = \mathbf{0}_{r \times r}. \quad (6.21d)$$

The expansion of the error $\mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_G)$ due to Δ_E is the same as (6.19); here, we again deal with the terms individually. Applying Lemma 6.4 to equations (6.6c) and (6.21c), $\text{tr}(\tilde{\mathbf{B}}\mathbf{B}^\top \Delta_Z)$ becomes

$$\begin{aligned} \text{tr}(\tilde{\mathbf{B}}\mathbf{B}^\top \Delta_Z) &= \text{tr}\left(-\left(\sum_{k=1}^p \tilde{M}_k \Delta_X \mathbf{M}_k - \Delta_E^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{A}\right) \mathbf{X}\right) + O(\|\Delta_E\|_F^2) \\ &= \text{tr}(\mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lqo}} \Delta_E) - \text{tr}\left(\sum_{k=1}^p \tilde{M}_k \mathbf{X}^\top \mathbf{M}_k \Delta_X\right) + O(\|\Delta_E\|_F^2). \end{aligned} \quad (6.22a)$$

From (6.6d) and (6.21a), observe that

$$\begin{aligned} \text{tr}\left(\sum_{k=1}^p \tilde{M}_k \mathbf{X}^\top \mathbf{M}_k \Delta_X\right) &= \text{tr}\left(\left(\tilde{E}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{A} + \tilde{A}^\top \mathbf{Z}_{\text{lqo}}^\top \mathbf{E}\right) \Delta_X\right) - \text{tr}(\tilde{\mathbf{C}}^\top \mathbf{C} \Delta_X) \\ &= \text{tr}\left(\mathbf{Z}_{\text{lqo}}^\top \left(\mathbf{A} \Delta_X \tilde{E}^\top + \mathbf{E} \Delta_X \tilde{A}^\top\right)\right) - \text{tr}(\tilde{\mathbf{C}}^\top \mathbf{C} \Delta_X) \\ &= -\text{tr}\left(\mathbf{Z}_{\text{lqo}}^\top \mathbf{A} \mathbf{X} \Delta_E^\top\right) - \text{tr}(\tilde{\mathbf{C}}^\top \mathbf{C} \Delta_X) + O(\|\Delta_E\|_F^2). \end{aligned} \quad (6.22b)$$

Applying Lemma 6.4 to (6.10b) and (6.21a) allows us to simplify the $\text{tr}(\tilde{\mathbf{C}}^\top \mathbf{C} \Delta_X)$ term further as

$$-\text{tr}(\tilde{\mathbf{C}}^\top \mathbf{C} \Delta_X) = \text{tr}(\Delta_E \mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lo}}) + O(\|\Delta_E\|_F^2).$$

So, the term (6.22b) in (6.22a) ultimately becomes

$$-\text{tr}\left(\sum_{k=1}^p \tilde{M}_k \mathbf{X}^\top \mathbf{M}_k \Delta_X\right) = \text{tr}(\mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lqo}} \Delta_E) - \text{tr}(\mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lo}} \Delta_E) + O(\|\Delta_E\|_F^2).$$

Thus, (6.22a) in (6.19) is given by

$$\text{tr}(\tilde{\mathbf{B}}\mathbf{B}^\top \Delta_Z) = 2 \text{tr}(\mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lqo}} \Delta_E) - \text{tr}(\mathbf{X}^\top \mathbf{Z}_{\text{lo}} \mathbf{A} \Delta_E) + O(\|\Delta_E\|_F^2).$$

Following similar calculations involving (6.21b) and (6.21d), the term $\text{tr} \left(\tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Q \right)$ in (6.19) can be expressed as

$$\text{tr} \left(\tilde{\mathbf{B}} \tilde{\mathbf{B}}^\top \Delta_Q \right) = \text{tr} \left(4 \tilde{\mathbf{P}} \tilde{\mathbf{A}} \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_E \right) - \text{tr} \left(2 \tilde{\mathbf{P}} \tilde{\mathbf{A}} \tilde{\mathbf{Q}}_{\text{lo}} \Delta_E \right) + O \left(\|\Delta_E\|_{\mathbb{F}}^2 \right).$$

So, the expansion (6.19) simplifies to

$$\begin{aligned} \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} + \Delta_{\mathcal{G}} \right) &= \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) + 4 \text{tr} \left(\mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lqo}} \Delta_E \right) - 2 \text{tr} \left(\mathbf{X}^\top \mathbf{A} \mathbf{Z}_{\text{lo}} \Delta_E \right) \\ &\quad + 4 \text{tr} \left(\tilde{\mathbf{P}} \tilde{\mathbf{A}} \tilde{\mathbf{Q}}_{\text{lqo}} \Delta_E \right) - 2 \text{tr} \left(\tilde{\mathbf{P}} \tilde{\mathbf{A}} \tilde{\mathbf{Q}}_{\text{lo}} \Delta_E \right) + O \left(\|\Delta_E\|_{\mathbb{F}}^2 \right) \\ &= \mathcal{J} \left(\tilde{\mathcal{G}}_{\text{lqo}} \right) + \left\langle 2 \left((2 \mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{A} \mathbf{X} + (2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}) \tilde{\mathbf{A}} \tilde{\mathbf{P}} \right), \Delta_E \right\rangle_{\mathbb{F}} \\ &\quad + O \left(\|\Delta_E\|_{\mathbb{F}}^2 \right). \end{aligned}$$

Thus $\nabla_{\tilde{\mathcal{E}}} \mathcal{J} = 2 \left((2 \mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{A} \mathbf{X} + (2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}) \tilde{\mathbf{A}} \tilde{\mathbf{P}} \right)$ holds as claimed. This completes the proof. \square

The stationary points of the gradients in (6.11) lead directly to first-order necessary conditions for \mathcal{H}_2 optimality, which is our second major theoretical contribution of this section. In contrast to that of the author's previous work [186, Theorem 3.2], Theorem 6.5 is stated for generic nonsingular \mathbf{E} .

Theorem 6.5 (Sylvester equation-based \mathcal{H}_2 -optimality condition [186, Theorem 3.2]). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2), and suppose additionally that $\tilde{\mathcal{G}}_{\text{lqo}}$ minimizes the squared \mathcal{H}_2 error in (6.1). Then

$$\mathbf{0} = \left(2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{A}} \tilde{\mathbf{P}} + (2 \mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{A} \mathbf{X}, \quad (6.23a)$$

$$\mathbf{0} = \left(2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{E}} \tilde{\mathbf{P}} + (2 \mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{E} \mathbf{X}, \quad (6.23b)$$

$$\mathbf{0} = \left(2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right) \tilde{\mathbf{B}} + (2 \mathbf{Z}_{\text{lqo}}^\top - \mathbf{Z}_{\text{lo}}^\top) \mathbf{B}, \quad (6.23c)$$

$$\mathbf{0} = \tilde{\mathbf{C}} \tilde{\mathbf{P}} - \mathbf{C} \mathbf{X}, \quad (6.23d)$$

$$\mathbf{0} = \tilde{\mathbf{M}} \left(\tilde{\mathbf{P}} \otimes \tilde{\mathbf{P}} \right) - \mathbf{M} \left(\mathbf{X} \otimes \mathbf{X} \right), \quad (6.23e)$$

where $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}}_{\text{lqo}} \in \mathbb{R}^{r \times r}$, $\mathbf{X}, \mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ satisfy the generalized matrix equations (6.6), and $\tilde{\mathbf{Q}}_{\text{lo}} \in \mathbb{R}^{r \times r}$ and $\mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ satisfy (6.10). Moreover, if $\tilde{\mathbf{P}}$ and $2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}$ are nonsingular, then the locally \mathcal{H}_2 -optimal reduced model $\tilde{\mathcal{G}}_{\text{lqo}}$ is defined by a Petrov-Galerkin projection (5.24) where the model reduction bases $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{n \times r}$ are given by

$$\mathbf{V} = \mathbf{X} \tilde{\mathbf{P}}^{-1} \quad \text{and} \quad \mathbf{W} = - (2 \mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}) \left(2 \tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right). \quad (6.24)$$

\diamond

Proof of Theorem 6.5. Under the assumption that $\tilde{\mathcal{G}}_{\text{lqo}}$ minimizes the squared \mathcal{H}_2 error, the gradients in (6.11) are necessarily stationary. The optimality conditions (6.23) follow directly from this. Then it remains to show that the reduced model is defined by Petrov-Galerkin projection using the matrices in (6.24). This follows directly from the choice of \mathbf{W} and \mathbf{V} along with the conditions in (6.23). The conditions corresponding to $\nabla_{\tilde{\mathbf{E}}}\mathcal{J}$ and $\nabla_{\tilde{\mathbf{A}}}\mathcal{J}$ show

$$\begin{aligned}\nabla_{\tilde{\mathbf{E}}}\mathcal{J} = \mathbf{0} &= \left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}\right) \tilde{\mathbf{A}}\tilde{\mathbf{P}} + \left(2\mathbf{Z}_{\text{lqo}}^{\top} - \mathbf{Z}_{\text{lo}}^{\top}\right) \mathbf{A}\mathbf{X} \\ \implies \tilde{\mathbf{A}} &= -\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}\right)^{-1} \left(2\mathbf{Z}_{\text{lqo}}^{\top} - \mathbf{Z}_{\text{lo}}^{\top}\right) \mathbf{A}\mathbf{X}\tilde{\mathbf{P}}^{-1} = \mathbf{W}^{\top}\mathbf{A}\mathbf{V} \\ \nabla_{\tilde{\mathbf{A}}}\mathcal{J} = \mathbf{0} &= \left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}\right) \tilde{\mathbf{E}}\tilde{\mathbf{P}} + \left(2\mathbf{Z}_{\text{lqo}}^{\top} - \mathbf{Z}_{\text{lo}}^{\top}\right) \mathbf{E}\mathbf{X} \\ \implies \tilde{\mathbf{E}} &= -\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}\right)^{-1} \left(2\mathbf{Z}_{\text{lqo}}^{\top} - \mathbf{Z}_{\text{lo}}^{\top}\right) \mathbf{E}\mathbf{X}\tilde{\mathbf{P}}^{-1} = \mathbf{W}^{\top}\mathbf{E}\mathbf{V}.\end{aligned}$$

Similarly, the conditions corresponding to $\nabla_{\tilde{\mathbf{B}}}\mathcal{J}$, $\nabla_{\tilde{\mathbf{C}}}\mathcal{J}$, and $\nabla_{\tilde{\mathbf{M}}}\mathcal{J}$ show

$$\begin{aligned}\nabla_{\tilde{\mathbf{B}}}\mathcal{J} = \mathbf{0} &= \left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}\right) \tilde{\mathbf{B}} + \left(2\mathbf{Z}_{\text{lqo}}^{\top} - \mathbf{Z}_{\text{lo}}^{\top}\right) \mathbf{B} \\ \implies \tilde{\mathbf{B}} &= -\left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}}\right)^{-1} \left(2\mathbf{Z}_{\text{lqo}}^{\top} - \mathbf{Z}_{\text{lo}}^{\top}\right) \mathbf{B} = \mathbf{W}^{\top}\mathbf{B} \\ \nabla_{\tilde{\mathbf{C}}}\mathcal{J} = \mathbf{0} &= \tilde{\mathbf{C}}\tilde{\mathbf{P}} - \mathbf{C}\mathbf{X} \\ \implies \tilde{\mathbf{C}} &= \mathbf{C}\mathbf{X}\tilde{\mathbf{P}}^{-1} = \mathbf{C}\mathbf{V} \\ \nabla_{\tilde{\mathbf{M}}}\mathcal{J} = \mathbf{0} &= \tilde{\mathbf{M}}\left(\tilde{\mathbf{P}} \otimes \tilde{\mathbf{P}}\right) - \mathbf{M}\left(\mathbf{X} \otimes \mathbf{X}\right) \\ \implies \tilde{\mathbf{M}} &= \mathbf{M}\left(\mathbf{X}\tilde{\mathbf{P}}^{-1} \otimes \mathbf{X}\tilde{\mathbf{P}}^{-1}\right) = \mathbf{M}\left(\mathbf{V} \otimes \mathbf{V}\right) \text{ by (2.9)}\end{aligned}$$

thus completing the proof. \square

The gradients of $\mathcal{J}\left(\tilde{\mathcal{G}}_{\text{lqo}}\right)$ in Theorem 6.3 and the Sylvester equation-based \mathcal{H}_2 optimality conditions in Theorem 6.5 generalize the analogous results in the LTI setting to that of (5.1), i.e., *they establish the Wilson framework for the optimal- \mathcal{H}_2 approximation of linear quadratic-output systems.* Indeed, in the particular instance where $\mathbf{M}_k = \mathbf{0}_{n \times n}$ for each output k and $\mathcal{G}_{\text{lqo}} = \mathcal{G}_{\text{lo}}$ is a linear system as in (2.25), the reduced-order quadratic-output observability Gramian $\tilde{\mathbf{Q}}_{\text{lqo}} \in \mathbb{R}^{r \times r}$ and solution $\mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ to (6.6d) reduce to $\tilde{\mathbf{Q}}_{\text{lqo}} = \tilde{\mathbf{Q}}_{\text{lo}}$ solving (2.40) and $\mathbf{Z}_{\text{lqo}} = \mathbf{Z}_{\text{lo}} \in \mathbb{R}^{n \times r}$ solving (6.10b), respectively. Applying the result of Theorem 6.3 in this case, the gradients with respect to $\tilde{\mathbf{E}}$, $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ then resolve to

$$\begin{aligned}\nabla_{\tilde{\mathbf{E}}}\mathcal{J} &= 2\left(\tilde{\mathbf{Q}}_{\text{lo}}\tilde{\mathbf{A}}\tilde{\mathbf{P}} + \mathbf{Z}_{\text{lo}}^{\top}\mathbf{A}\mathbf{X}\right), \\ \nabla_{\tilde{\mathbf{A}}}\mathcal{J} &= 2\left(\tilde{\mathbf{Q}}_{\text{lo}}\tilde{\mathbf{E}}\tilde{\mathbf{P}} + \mathbf{Z}_{\text{lo}}^{\top}\mathbf{E}\mathbf{X}\right), \\ \nabla_{\tilde{\mathbf{B}}}\mathcal{J} &= 2\left(\tilde{\mathbf{Q}}_{\text{lo}}\tilde{\mathbf{B}} + \mathbf{Z}_{\text{lo}}^{\top}\mathbf{B}\right),\end{aligned}$$

which are precisely those in (2.65); the gradient with respect to $\tilde{\mathbf{C}}$ is unchanged. The Sylvester equation-based optimality conditions in Theorem 6.5 reduce in an obviously similar way. Thus, we conclude that Theorems 6.3 and 6.5 contain Theorem 2.44 as a special case.

The Wilson optimality framework prescribed by Theorem 6.5 also bears a resemblance to analogous Sylvester-based optimality conditions that appear in the optimal- \mathcal{H}_2 approximation of other classes of weakly nonlinear dynamical systems; cf., [24, 44, 243] and [26, 91] for the settings of bilinear and quadratic-bilinear systems. These model classes are significant because a gamut of nonlinearities can be recast as bilinear or quadratic-bilinear systems (either approximately or exactly) using Carleman bilinearization or McCormick relaxation [140, 179, 194]. To the author's knowledge, the results of this section are the first to establish any sort of \mathcal{H}_2 -optimal approximation framework for systems with nonlinearities in the output equation, thus distinguishing this dissertation from the work of [24, 44, 243] and [26, 91].

Remark 6.6 (Conditions for the quadratic-output subsystem). As already discussed in Section 5.2, in some applications, the outputs are purely quadratic. Then $\mathbf{C} = \mathbf{0}_{p \times n}$, and so \mathcal{G}_{lqo} reduces to the quadratic-output subsystem \mathcal{G}_{qo} in (5.11). In these instances, $\mathbf{Q}_{\text{lo}} \in \mathbb{R}^{n \times n}$, $\tilde{\mathbf{Q}}_{\text{lo}} \in \mathbb{R}^{r \times r}$ and $\mathbf{Z}_{\text{lo}} \in \mathbb{R}^{r \times r}$ are all zero. The gradients of \mathcal{J} in Theorem 6.3 thus become

$$\begin{aligned}\nabla_{\tilde{\mathbf{E}}}\mathcal{J} &= 4 \left(\tilde{\mathbf{Q}}_{\text{lqo}}\tilde{\mathbf{A}}\tilde{\mathbf{P}} + \mathbf{Z}_{\text{lqo}}^{\top}\mathbf{A}\mathbf{X} \right), \\ \nabla_{\tilde{\mathbf{A}}}\mathcal{J} &= 4 \left(\tilde{\mathbf{Q}}_{\text{lqo}}\tilde{\mathbf{E}}\tilde{\mathbf{P}} + \mathbf{Z}_{\text{lqo}}^{\top}\mathbf{E}\mathbf{X} \right), \\ \nabla_{\tilde{\mathbf{B}}}\mathcal{J} &= 4 \left(\tilde{\mathbf{Q}}_{\text{lqo}}\tilde{\mathbf{B}} + \mathbf{Z}^{\top}\mathbf{B} \right), \\ \nabla_{\tilde{\mathbf{M}}}\mathcal{J} &= 2 \left(\tilde{\mathbf{M}} \left(\tilde{\mathbf{P}} \otimes \tilde{\mathbf{P}} \right) - \mathbf{M} \left(\mathbf{X} \otimes \mathbf{X} \right) \right).\end{aligned}$$

where $\tilde{\mathbf{Q}}_{\text{lqo}} \in \mathbb{R}^{r \times r}$, $\mathbf{Z}_{\text{lqo}} \in \mathbb{R}^{n \times r}$ solve (6.6b) and (6.6d) with \mathbf{C} and $\tilde{\mathbf{C}}$ equal to zero. \diamond

Theorem 6.5 states that any local minimizer of the \mathcal{H}_2 error $\tilde{\mathcal{G}}_{\text{lqo}}$ in (6.1) is necessarily defined by a Petrov-Galerkin projection where the optimal matrices are given by $\mathbf{V} = \mathbf{X}\tilde{\mathbf{P}}^{-1} \in \mathbb{R}^{n \times r}$ and $\mathbf{W} = (2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}) \left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right)^{-1} \in \mathbb{R}^{n \times r}$. Under the equivalence transformation $\mathbf{T} = \tilde{\mathbf{P}}$ and $\mathbf{S} = - \left(2\tilde{\mathbf{Q}}_{\text{lqo}} - \tilde{\mathbf{Q}}_{\text{lo}} \right)$, we see the \mathcal{H}_2 -optimal reduced model has an equivalent realization given by

$$\begin{aligned}\tilde{\mathbf{E}} &= (2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}})^{\top}\mathbf{E}\mathbf{X}, & \tilde{\mathbf{A}} &= (2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}})^{\top}\mathbf{A}\mathbf{X}, & \tilde{\mathbf{B}} &= (2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}})^{\top}\mathbf{B}, \\ \tilde{\mathbf{C}} &= \mathbf{C}\mathbf{X}, & \tilde{\mathbf{M}} &= \mathbf{M} \left(\mathbf{X} \otimes \mathbf{X} \right).\end{aligned}\tag{6.25}$$

The right optimal projection matrix $\mathbf{V} = \mathbf{X}$ satisfies the generalized Sylvester equation in (6.6c); because \mathbf{Z}_{lqo} and \mathbf{Z}_{lo} satisfy (6.6d) and (6.10b), the left optimal projection matrix

$\mathbf{W} = \mathbf{Z}_{2\text{lqo}-\text{qo}} \stackrel{\text{def}}{=} 2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}$ satisfies a linear combination of these equations, namely

$$\mathbf{A}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}} \tilde{\mathbf{E}} + \mathbf{E}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}} \tilde{\mathbf{A}} - 2 \sum_{k=1}^p \mathbf{M}_k \mathbf{V} \tilde{\mathbf{M}}_k - \mathbf{C}^\top \tilde{\mathbf{C}} = \mathbf{0} \quad (6.26)$$

where $\mathbf{V} = \mathbf{X} \in \mathbb{R}^{n \times r}$ satisfies $\mathbf{A}\mathbf{X}\tilde{\mathbf{E}}^\top + \mathbf{E}\mathbf{X}\tilde{\mathbf{A}}^\top + \mathbf{B}\tilde{\mathbf{B}}^\top = \mathbf{0}$.

Note that the generalized Sylvester equations in (6.26) depend explicitly upon the \mathcal{H}_2 -optimal reduced model via its realization in (6.25). In other words, if we wanted to obtain a reduced model that satisfies the first-order optimality conditions of Theorem 6.5 by projecting the full-order matrix operators in (5.1) using $\mathbf{V} = \mathbf{X}$ and $\mathbf{W} = \mathbf{Z}_{2\text{lqo}-\text{qo}}$, this would require *a priori* knowledge of a locally \mathcal{H}_2 -optimal reduced model in order to solve the matrix equations in (6.26), which is of course unavailable. This discussion highlights the major challenge faced by practitioners in (any area of) \mathcal{H}_2 -optimal model reduction; one needs to develop iterative algorithms for computing reduced models that satisfy first-order necessary conditions for optimality. In the next section, we propose the first of two such algorithms for the optimal \mathcal{H}_2 -approximation of the systems in (5.1) that generalizes the core ideas of [30, 237] to the problem (6.1). The theoretical and computational backbone of the iteration is based on the optimality conditions in (6.23) and the pair of generalized Sylvester equations in (6.26).

6.3.2 A two-sided iterative algorithm for optimal- \mathcal{H}_2 approximation of linear quadratic-output systems

Recall the pair of generalized Sylvester equations from (6.26) that determines the optimal projection matrices $\mathbf{V} = \mathbf{X}$ and $\mathbf{W} = \mathbf{Z}_{2\text{lqo}-\text{qo}} \stackrel{\text{def}}{=} 2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}$. These matrix equations motivate an iterative procedure where, given a reduced model $\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}$ at step i , the equations in (6.26) are solved for $\mathbf{V} = \mathbf{X}^{(i)}$ and $\mathbf{W} = \mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)}$ using the i -th model iterate; the subsequent model iterate $\tilde{\mathcal{G}}_{\text{lqo}}^{(i+1)}$ is then computed by projection:

$$\begin{aligned} \tilde{\mathbf{E}}^{(i+1)} &= \left(\mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)} \right)^\top \mathbf{E} \mathbf{X}^{(i)}, & \tilde{\mathbf{A}}^{(i+1)} &= \left(\mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)} \right)^\top \mathbf{A} \mathbf{X}^{(i)}, & \tilde{\mathbf{B}}^{(i+1)} &= \left(\mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)} \right)^\top \mathbf{B}, \\ \tilde{\mathbf{C}}^{(i+1)} &= \mathbf{C} \mathbf{X}^{(i)}, & \tilde{\mathbf{M}}^{(i+1)} &= \mathbf{M} \left(\mathbf{X}^{(i)} \otimes \mathbf{X}^{(i)} \right). \end{aligned}$$

In practice, $\mathbf{X}^{(k)}$ and $\mathbf{Z}_{2\text{lqo}-\text{qo}}^{(k)}$ are first orthonormalized, since this does not change the resulting system. The corresponding iteration is presented in Algorithm 6.3.1. If the \mathcal{H}_2 error between consecutive model iterates is stationary, then the first-order optimality conditions of Theorem 6.5 will be satisfied since the gradients in (6.11) are necessarily zero at this point. This is the fundamental idea behind the so-called two-sided iterative algorithm (TSIA) proposed by Xu and Zeng [237] for linear \mathcal{H}_2 -optimal model reduction; in the iteration described above, $\mathbf{Z}_{2\text{lqo}-\text{qo}}$ takes the place of \mathbf{Z}_{lo} that solves (2.40). Based on this connection,

Algorithm 6.3.1: Linear quadratic-output two-sided iterative algorithm (LQO-TSIA).

Input: $\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{M}$ from (5.1), order $1 \leq r < n$, tolerance $\tau > 0$, maximum number of iteration steps $M \geq 1$, initial reduced model (5.2) given by $\tilde{\mathbf{E}}^{(0)}, \tilde{\mathbf{A}}^{(0)}, \tilde{\mathbf{B}}^{(0)}, \tilde{\mathbf{C}}^{(0)}, \tilde{\mathbf{M}}^{(0)}$.

Output: $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{M}}$ —state-space matrices of the converged model (5.2).

1 Iteration count $i = 0$ and change in error $\mathcal{E}^{(0)} = \mathcal{J}(\tilde{\mathcal{G}}_{\text{lqo}}^{(0)})$.

2 **while** $\mathcal{E}^{(i)} > \tau$ and $i \leq M$ **do**

3 Solve generalized Sylvester equations (6.26) for $\mathbf{X}^{(i)}, \mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)} \in \mathbb{R}^{n \times r}$:

$$\begin{aligned} \mathbf{A}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)} \tilde{\mathbf{E}}^{(i)} + \mathbf{E}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)} \tilde{\mathbf{A}}^{(i)} - 2 \sum_{k=1}^p \mathbf{M}_k \mathbf{X}^{(i)} \tilde{\mathbf{M}}_k^{(i)} - \mathbf{C}^\top \tilde{\mathbf{C}}^{(i)} &= \mathbf{0} \\ \mathbf{A} \mathbf{X}^{(i)} \tilde{\mathbf{E}}^{(i)\top} + \mathbf{E} \mathbf{X}^{(i)} \tilde{\mathbf{A}}^{(i)\top} + \mathbf{B} \tilde{\mathbf{B}}^{(i)\top} &= \mathbf{0}. \end{aligned}$$

4 Orthonormalize $\mathbf{X}^{(i)}$ and $\mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)}$ to obtain \mathbf{V} and \mathbf{W} .

$$\mathbf{V} \leftarrow \text{orth}(\mathbf{X}^{(i)}), \quad \mathbf{W} \leftarrow \text{orth}(\mathbf{Z}_{2\text{lqo}-\text{qo}}^{(i)}).$$

5 Compute reduced-order matrices by Petrov-Galerkin projection using \mathbf{V} and \mathbf{W}

$$\begin{aligned} \tilde{\mathbf{E}}^{(i+1)} &= \mathbf{W}^\top \mathbf{E} \mathbf{V}, \quad \tilde{\mathbf{A}}^{(i+1)} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \tilde{\mathbf{B}}^{(i+1)} = \mathbf{W}^\top \mathbf{B}, \\ \tilde{\mathbf{C}}^{(i+1)} &= \mathbf{C} \mathbf{V}, \quad \tilde{\mathbf{M}}^{(i+1)} = \mathbf{M} (\mathbf{V} \otimes \mathbf{V}). \end{aligned}$$

6 Compute the normalized \mathcal{H}_2 distance between model iterates:

$$\mathcal{E}^{(i+1)} = \frac{\left| \mathcal{J}_{\text{rel}}(\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}) - \mathcal{J}_{\text{rel}}(\tilde{\mathcal{G}}_{\text{lqo}}^{(i+1)}) \right|}{\mathcal{J}_{\text{rel}}(\tilde{\mathcal{G}}_{\text{lqo}}^{(0)})}.$$

7 Set $i \leftarrow i + 1$.

8 **end**

we refer to the proposed computational procedure in Algorithm 6.3.1 as the *linear quadratic-output two-sided iterative algorithm* (LQO-TSIA).

We now discuss some implementation details specific to Algorithm 6.3.1. General comments that compare and contrast Algorithm 6.3.1 and Algorithm 6.4.1, which is presented next in Section 6.4.2, are provided in Section 6.5. We mention that, as an alternative to Algorithm 6.3.1, one could use the gradients of Theorem 6.3 in any off-the-shelf optimization

method to solve (6.1). We will not consider such an approach in this dissertation.

Solving the sparse-dense Sylvester equations.

The dominant computational cost at each iteration is the solution of the two generalized Sylvester equations in Step 2 of Algorithm 6.3.1. These are often called *sparse-dense* Sylvester equations because, in most applications, the large $n \times n$ matrices that appear in (6.26) have some inherent sparsity, while the small $r \times r$ matrices are dense, as an artifact of the Petrov-Galerkin projection. Because the solution matrices $\mathbf{X}, \mathbf{Z}_{2l_{qo}-qo} \in \mathbb{R}^{n \times r}$ in (6.26) are tall and skinny, they can be obtained efficiently and directly by computing a Schur decomposition of the reduced matrix $\tilde{\mathbf{A}}$, and subsequently solving for their columns via shifted linear systems. Direct (exact) methods for solving this type of Sylvester equation are provided in [30]. For completeness, we briefly describe how these methods can be applied to the equation for $\mathbf{Z}_{2l_{qo}-qo}$ in (6.26) when $\mathbf{E} = \mathbf{I}_n$ and $\tilde{\mathbf{E}} = \mathbf{I}_r$. The generalized case is treated in [30, Sec. 3.2].

Let $\tilde{\mathbf{A}}^T = \tilde{\mathbf{U}}^H \tilde{\mathbf{T}} \tilde{\mathbf{U}}$ be the Schur form of $\tilde{\mathbf{A}}^T$ so that $\tilde{\mathbf{U}} \in \mathbb{C}^{r \times r}$ is unitary, and $\tilde{\mathbf{T}} \in \mathbb{C}^{r \times r}$ is an upper triangular matrix. Replacing $\tilde{\mathbf{A}}^T$ with its Schur form in the equation for $\mathbf{Z}_{2l_{qo}-qo}$ in (6.26) and multiplying on the right by $\tilde{\mathbf{U}}^H$ yields a similar Sylvester equation

$$\mathbf{A}^T \left(\mathbf{Z}_{2l_{qo}-qo} \tilde{\mathbf{U}}^H \right) + \left(\mathbf{Z}_{2l_{qo}-qo} \tilde{\mathbf{U}}^H \right) \tilde{\mathbf{T}}^T - 2 \sum_{k=1}^p \mathbf{M}_k \mathbf{X} \tilde{\mathbf{M}}_k \tilde{\mathbf{U}}^H - \mathbf{C}^T \tilde{\mathbf{C}} \tilde{\mathbf{U}}^H = \mathbf{0}.$$

The columns of $\mathbf{Z}_{2l_{qo}-qo} \tilde{\mathbf{U}}^H$ are computed by backward substitution; the matrix \mathbf{X} can be obtained by a nearly identical procedure. Because $\tilde{\mathbf{A}}$ is a small $r \times r$ matrix, its Schur form is computable in $O(r^3)$ operations, where $r \ll n$. Thus, the dominant cost in obtaining \mathbf{X} and $\mathbf{Z}_{2l_{qo}-qo}$ lies in the $2r$ shifted linear system solves employed in solving the Sylvester equations. Usually, the large-scale coefficient matrix \mathbf{A} has some inherent sparsity that one can exploit. So, the complexity of the algorithm is roughly $2r$ times the complexity of the solver used. In the worst case, where \mathbf{A} is dense, the complexity of the solver is bounded above by $O(n^3)$. However, modern solvers are typically much more efficient, and so the complexity of solving (6.26) will likely be more favorable than $O(n^3)$; we refer the reader to [30, Sec. 3] and the references therein. Lastly, while the described procedure involves complex arithmetic, this can be avoided by using the real-valued block Schur form of $\tilde{\mathbf{A}}$ instead.

This discussion describes how the Sylvester equations in (6.26) can be solved *directly* using only *sparse* calculations involving the full-order coefficient matrix \mathbf{A} . As a point of comparison, the LQO-BT presented in Algorithm 5.3.1 requires the one-time solution of the two large-scale Lyapunov equations (2.43) and (5.20) to obtain the system Gramians. Solving these via direct methods, such as the Bartels-Stewart algorithm, requires the Schur decomposition of the large-scale \mathbf{A} matrix and has a complexity of $O(n^3)$.

Convergence monitoring and initialization strategies.

With regard to the convergence of Algorithm 6.3.1, the iteration repeats until either some preset number of steps M is reached, or the algorithm converges within the tolerance $\tau > 0$ based on some pre-determined stopping criterion. As is the case with any optimization problem, there exists a variety of possible choices for measuring convergence. Because we are seeking to minimize the squared \mathcal{H}_2 error \mathcal{J} in (6.1), for simplicity, we use the (relative) change in the relative squared \mathcal{H}_2 error between consecutive iterates to monitor convergence. From (6.7), the square of the relative error due to $\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}$ at step i of the iteration is given by

$$\mathcal{J}_{\text{rel}}\left(\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}\right) = \frac{\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}^{(i)}\|_{\mathcal{H}_2}^2}{\|\mathcal{G}_{\text{lqo}}\|_{\mathcal{H}_2}^2} = \frac{\|\mathcal{G}_{\text{lqo}}\|_{\mathcal{H}_2}^2 + \|\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}\|_{\mathcal{H}_2}^2 + 2 \operatorname{tr}\left(\mathbf{B}\mathbf{Z}_{\text{lqo}}^{(i)}\tilde{\mathbf{B}}^{(i)\top}\right)}{\|\mathcal{G}_{\text{lqo}}\|_{\mathcal{H}_2}^2},$$

where $\mathbf{Z}_{\text{lqo}}^{(i)} \in \mathbb{R}^{n \times r}$ satisfies (6.6d) for the current model iterate $\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}$. Then Algorithm 6.3.1 is deemed to have converged at step i if

$$\mathcal{E}^{(i)} = \frac{\left|\mathcal{J}_{\text{rel}}\left(\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}\right) - \mathcal{J}_{\text{rel}}\left(\tilde{\mathcal{G}}_{\text{lqo}}^{(i-1)}\right)\right|}{\mathcal{J}_{\text{rel}}\left(\tilde{\mathcal{G}}_{\text{lqo}}^{(0)}\right)} \leq \tau \text{ for } \tau \geq 0. \quad (6.27)$$

The \mathcal{H}_2 norm of the full-order model in (6.27) can be pre-computed at the start of the iteration. Another natural option is to monitor changes in the gradients $\nabla_{\tilde{\mathbf{A}}}\mathcal{J}$, $\nabla_{\tilde{\mathbf{B}}}\mathcal{J}$, $\nabla_{\tilde{\mathbf{C}}}\mathcal{J}$, and $\nabla_{\tilde{\mathbf{M}}_k}\mathcal{J}$ of the error function \mathcal{J} , and terminate when they are sufficiently small. The information required to compute these quantities is readily available from the iteration itself. However, a relative metric that uses scaled gradients (as is usually done in practice) would require computing the Hessian of \mathcal{J} , which is not directly available from already computed quantities. So, we do not consider this criterion further.

Remark 6.7. Computing the \mathcal{H}_2 norm of the full-order model using, e.g., the Gramian-based formulae in (5.37) and (5.36), requires the solution of a large-scale Lyapunov equation, which, as already discussed, may be infeasible for problems with truly large state-space dimension n . Thus, monitoring the convergence with (6.27) may not always be feasible. As an alternative, note that in the error formula (6.27), the only part that varies in each iteration is the “tail” of the error, i.e.,

$$\eta^{(i)} \stackrel{\text{def}}{=} \|\tilde{\mathcal{G}}_{\text{lqo}}^{(i)}\|_{\mathcal{H}_2}^2 + 2 \operatorname{tr}\left(\mathbf{B}\mathbf{Z}_{\text{lqo}}^{(i)}\tilde{\mathbf{B}}^{(i)\top}\right). \quad (6.28)$$

Therefore, if computing the true (or an approximate) \mathcal{H}_2 norm of the full-order model is not feasible, one may monitor the relative change in the tails (6.28) and terminate the algorithm when this quantity falls below the specified convergence tolerance, i.e.

$$|\eta^{(i)} - \eta^{(i-1)}|/|\eta^{(1)}| \leq \tau \text{ for } \tau > 0. \quad (6.29)$$

We refer to [186, Section IV] for a comparison of these two convergence strategies. \diamond

As is the case for the linear TSIA [237], convergence of Algorithm 6.3.1 is *not* guaranteed in general since it is a fixed-point iteration. The algorithm is also not guaranteed to produce an asymptotically stable reduced model. However, similar to TSIA and IRKA, in practice the algorithm performs well; in practice, we have never observed LQO-TSIA converge to an unstable reduced model given a stable initialization. We include a numerical study of the convergence of Algorithm 6.3.1 in Section 6.6. For guaranteed convergence, one may consider developing a descent-based algorithm based on the explicit gradient formulae (6.11) we derived, although we will not pursue this further.

As with any minimization problem, the initialization of Algorithm 6.3.1 will affect the quality of the final result. We leave these considerations to the discussion in Section 6.4.1, since the same comments regarding the initialization problem also apply to the general \mathcal{H}_2 minimization problem, and thus apply here.

6.4 Interpolation-based \mathcal{H}_2 optimality framework

The Wilson conditions of Theorem 6.5 provide a set of first-order necessary conditions that the Gramians and matrices of a reduced-order model (5.2) must satisfy to be \mathcal{H}_2 -optimal. In some sense, these conditions provide a natural characterization of all optimal approximants since they are derived from gradients of the \mathcal{H}_2 system error. We have already mentioned in the previous section that the Wilson framework, which originated in the context of linear system approximation [97, 111, 217, 233], has been generalized to other classes of weakly nonlinear dynamical systems in the last two decades; cf., [24, 44, 243] and [26, 91] for the case of bilinear and quadratic-bilinear systems.

Simultaneously, in each of the aforementioned settings, the Wilson framework has an alternative interpretation in terms of *rational function interpolation*. Indeed, for linear time-invariant [97, 142, 217, 218] as well as bilinear [77, 78] and quadratic-bilinear [50] dynamical systems, it has been shown that if the Wilson optimality conditions are satisfied, then the corresponding reduced model (or more specifically, its subsystem transfer functions) will be a *rational interpolant* of the full-order system in some sense. In the linear time-invariant setting of (2.25), we have already highlighted in Section 2.4.2 that optimal approximants are bi-tangential Hermite interpolants of the original system. For bilinear and quadratic-bilinear systems, it has been shown in [77, 78] and [50] that \mathcal{H}_2 -optimal reduced models satisfy so-called *multipoint Volterra series interpolation* conditions that, as the name suggests, respect the underlying Volterra series expansion of the original system. In *all* of these cases, the optimal interpolation points are the *mirror images of the reduced model poles*. These rich and ubiquitous connections between optimal- \mathcal{H}_2 approximations and rational function interpolation raise the question: Does there exist a similar characterization of \mathcal{H}_2 -optimal reduced models of the form (5.2) based on rational interpolation of \mathbf{G}_{lo} and \mathbf{G}_{qo} ? In this section, we provide an affirmative answer to this question, and develop an interpolatory framework for the best \mathcal{H}_2 -approximation of LQO systems (5.1). The results of this section are derived

independently from those of Section 6.3.

6.4.1 Theoretical optimality framework

Consider a pair of LQO systems \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ as in (5.1) and (5.2) with the transfer functions \mathbf{G}_{lo} , \mathbf{G}_{qo} and $\tilde{\mathbf{G}}_{\text{lo}}$, $\tilde{\mathbf{G}}_{\text{qo}}$ defined according to (5.12). Recall from (5.42) that, \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ have simple poles μ_1, \dots, μ_n and $\lambda_1, \dots, \lambda_r$, their transfer functions admit pole-residue expansions

$$\begin{aligned}\tilde{\mathbf{G}}_{\text{lo}}(s) &= \sum_{i=1}^r \frac{\mathbf{c}_i \mathbf{b}_i^\top}{s - \lambda_i}, & \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) &= \sum_{j=1}^r \sum_{k=1}^r \frac{\mathbf{m}_{j,k} (\mathbf{b}_j \otimes \mathbf{b}_k)^\top}{(s_1 - \lambda_j)(s_2 - \lambda_k)} \\ \mathbf{G}_{\text{lo}}(s) &= \sum_{i=1}^n \frac{\boldsymbol{\delta}_i \boldsymbol{\beta}_i^\top}{s - \mu_i}, & \mathbf{G}_{\text{qo}}(s_1, s_2) &= \sum_{j=1}^n \sum_{k=1}^n \frac{\boldsymbol{\pi}_{j,k} (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k)^\top}{(s_1 - \mu_j)(s_2 - \mu_k)},\end{aligned}\tag{6.30}$$

where $\mathbf{b}_i \in \mathbb{C}^m$, $\mathbf{c}_i \in \mathbb{C}^p$, $\mathbf{m}_{j,k} \in \mathbb{C}^p$ are the residue directions of $\tilde{\mathbf{G}}_{\text{lo}}$, $\tilde{\mathbf{G}}_{\text{qo}}$ corresponding to the poles λ_i , (λ_j, λ_k) , and $\boldsymbol{\beta}_i \in \mathbb{C}^m$, $\boldsymbol{\delta}_i \in \mathbb{C}^p$, $\boldsymbol{\pi}_{j,k} \in \mathbb{C}^p$ are the residue directions of \mathbf{G}_{lo} , \mathbf{G}_{qo} corresponding to the poles μ_i , (μ_j, μ_k) . Before formally presenting our interpolation-based optimality framework, we exploit Theorem 5.11 and the pole-residue expansions (6.30) to make an observation regarding the \mathcal{H}_2 model reduction error in terms of the mismatch of the corresponding transfer functions.

Corollary 6.8 (Pole-residue based \mathcal{H}_2 model reduction error). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2) with the transfer functions \mathbf{G}_{lo} , \mathbf{G}_{qo} and $\tilde{\mathbf{G}}_{\text{lo}}$, $\tilde{\mathbf{G}}_{\text{qo}}$ defined according to (5.12), and that both \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ have simple poles μ_1, \dots, μ_n and $\lambda_1, \dots, \lambda_r$. Then the squared \mathcal{H}_2 model reduction error $\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2$ is given by

$$\begin{aligned}\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 &= \sum_{i=1}^n \boldsymbol{\delta}_i^\top \left(\mathbf{G}_{\text{lo}}(-\mu_i) - \tilde{\mathbf{G}}_{\text{lo}}(-\mu_i) \right) \boldsymbol{\beta}_i - \sum_{i=1}^r \mathbf{c}_i^\top \left(\mathbf{G}_{\text{lo}}(-\lambda_i) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_i) \right) \mathbf{b}_i \\ &\quad + \sum_{j=1}^n \sum_{k=1}^n \boldsymbol{\pi}_{j,k}^\top \left(\mathbf{G}_{\text{qo}}(-\mu_j, -\mu_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\mu_j, -\mu_k) \right) (\boldsymbol{\beta}_j \otimes \boldsymbol{\beta}_k) \\ &\quad + \sum_{j=1}^r \sum_{k=1}^r \mathbf{m}_{j,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_j, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_j, -\lambda_k) \right) (\mathbf{b}_j \otimes \mathbf{b}_k),\end{aligned}\tag{6.31}$$

where $\mathbf{b}_i \in \mathbb{C}^m$, $\mathbf{c}_i \in \mathbb{C}^p$, $\mathbf{m}_{j,k} \in \mathbb{C}^p$, and $\boldsymbol{\beta}_i \in \mathbb{C}^m$, $\boldsymbol{\delta}_i \in \mathbb{C}^p$, $\boldsymbol{\pi}_{j,k} \in \mathbb{C}^p$ are defined as in (6.30). \diamond

Proof of Corollary 6.8. The results follow from expanding the \mathcal{H}_2 error as an inner product:

$$\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 = \left\langle \mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}, \mathcal{G}_{\text{lqo}} \right\rangle_{\mathcal{H}_2} - \left\langle \mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \right\rangle_{\mathcal{H}_2}$$

and applying Theorem 5.11 to $\langle \mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}, \mathcal{G}_{\text{lqo}} \rangle_{\mathcal{H}_2}$ and $\langle \mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}, \tilde{\mathcal{G}}_{\text{lqo}} \rangle_{\mathcal{H}_2}$. \square

Corollary 6.8 suggests the following: One can make the \mathcal{H}_2 model error small by eliminating the mismatch between a bi-tangential sum of the full- and reduced-order transfer functions evaluated at all combinations of $-\lambda_i$ and $-\mu_i$, the reduced- and full-order model poles. As it turns out, this sort of interpolation of the sum of \mathbf{G}_{lo} and \mathbf{G}_{qo} is in fact a *necessary condition* for \mathcal{H}_2 optimality, and interpolation at the reduced model poles is more important.

This brings us to the first major theoretical contribution of this section in Theorem 6.9, which derives first-order optimality conditions framed in terms of the rational interpolation of the linear- and quadratic-output transfer functions. As suggested by Corollary 6.8, a subset of the optimality conditions require interpolating a linear combination of the action of \mathbf{G}_{lo} and \mathbf{G}_{qo} evaluated at all possible combinations of the optimal interpolation points; the weights in the combination are tangential interpolation directions.

Theorem 6.9 (Interpolation-based \mathcal{H}_2 -optimality conditions). Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2) with the transfer functions \mathbf{G}_{lo} , \mathbf{G}_{qo} and $\tilde{\mathbf{G}}_{\text{lo}}$, $\tilde{\mathbf{G}}_{\text{qo}}$ defined according to (5.12), and that $\tilde{\mathcal{G}}_{\text{lqo}}$ has simple poles $\lambda_1, \dots, \lambda_r$. Let $\mathbf{b}_i \in \mathbb{C}^m$, $\mathbf{c}_i \in \mathbb{C}^p$, $\mathbf{m}_{j,k} \in \mathbb{C}^p$ be the corresponding residue directions defined in (5.43). If $\tilde{\mathcal{G}}_{\text{lqo}}$ minimizes the squared \mathcal{H}_2 error in (6.1), then $\tilde{\mathcal{G}}_{\text{lqo}}$ satisfies the following tangential interpolation conditions:

$$\mathbf{0}_p = \left(\mathbf{G}_{\text{lo}}(-\lambda_i) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_i) \right) \mathbf{b}_i, \quad (6.32a)$$

$$\mathbf{0}_p = \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j), \quad (6.32b)$$

$$\begin{aligned} \mathbf{0}_m &= \mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \\ &\quad + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{I}_m \otimes \mathbf{b}_\ell) \\ &\quad + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{I}_m), \end{aligned} \quad (6.32c)$$

$$\begin{aligned} 0 &= \mathbf{c}_k^\top \left(\frac{d}{ds} \mathbf{G}_{\text{lo}}(-\lambda_k) - \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \mathbf{b}_k \\ &\quad + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{b}_k \otimes \mathbf{b}_\ell) \\ &\quad + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{b}_k) \end{aligned} \quad (6.32d)$$

for all $i, j, k = 1, \dots, r$. \diamond

Proof of Theorem 6.9. Take $\check{\mathcal{G}}_{\text{lqo}}$ to be any order- r , asymptotically stable LQO system defined according to (5.2) that exists in a local neighborhood about $\tilde{\mathcal{G}}_{\text{lqo}}$ such that $\check{\mathcal{G}}_{\text{lqo}}$ is not a locally optimal \mathcal{H}_2 approximation of \mathcal{G}_{lqo} . Let $\check{\mathbf{G}}_{\text{lo}}$ and $\check{\mathbf{G}}_{\text{qo}}$ be the transfer functions of $\check{\mathcal{G}}_{\text{lqo}}$ according to (5.12). Then, by assumption, direct manipulations of the transfer function norms and inner products yield the inequality

$$\begin{aligned} \|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 &\leq \|\mathcal{G}_{\text{lqo}} - \check{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}^2 = \|\mathbf{G}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2^{p \times m}}^2 + \|\mathbf{G}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2^{p \times m^2}}^2 \\ \Rightarrow 0 &\leq 2 \operatorname{Re} \left\langle \mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}}, \tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}} \right\rangle_{\mathcal{H}_2^{p \times m}} + \|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2^{p \times m}}^2 \\ &\quad + 2 \operatorname{Re} \left\langle \mathbf{G}_{\text{qo}} - \tilde{\mathbf{G}}_{\text{qo}}, \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}} \right\rangle_{\mathcal{H}_2^{p \times m^2}} + \|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2^{p \times m^2}}^2. \end{aligned} \quad (6.33)$$

Henceforth, we drop the matrix dimensions when invoking the Hardy space norms and inner products of the transfer functions (5.12) since they will be clear from context. Take $\varepsilon > 0$ to be arbitrarily specified, and $\boldsymbol{\xi}$ to be an arbitrary unit vector in \mathbb{C}^p or \mathbb{C}^m , which we will specify depending on the context. We will prove each set of interpolation conditions in (6.32) by choosing $\check{\mathbf{G}}_{\text{lo}}$ and $\check{\mathbf{G}}_{\text{qo}}$ to differ from the \mathcal{H}_2 -optimal transfer functions $\tilde{\mathbf{G}}_{\text{lo}}$ and $\tilde{\mathbf{G}}_{\text{qo}}$ by carefully selected ε -perturbations of the optimal reduced model poles and residue directions. Because the state-space matrices in (5.1) and (5.2) are assumed real, we take for granted that $\overline{\mathbf{G}}_{\text{lo}}(s) = \mathbf{G}_{\text{lo}}(s)$ and $\overline{\mathbf{G}}_{\text{qo}}(s_1, s_2) = \mathbf{G}_{\text{qo}}(s_1, s_2)$ for any $s, s_1, s_2 \in \mathbb{C}$ (and likewise for the transfer functions of (5.2)) when invoking Theorem 5.11, where $\overline{\mathbf{G}}_{\text{lo}}(s)$ and $\overline{\mathbf{G}}_{\text{qo}}(s_1, s_2)$ are defined according to (5.13).

We first deal with the right-tangential interpolation conditions in (6.32a) and (6.32b). Since the conditions in (6.32a) relate to the purely linear output, their derivation follows similarly to that of [5, Thm. 5.1.1] for deriving the linear \mathcal{H}_2 -optimality conditions. For the sake of contradiction, assume that the (i, j) -th interpolation condition in (6.32b) does not hold, and take $\boldsymbol{\xi} \in \mathbb{C}^p$. Define $\check{\mathcal{G}}_{\text{lqo}}$ to be the system obtained by perturbing the (i, j) -th residue direction $\mathbf{m}_{i,j}$ of $\tilde{\mathcal{G}}_{\text{qo}}$ by $-\varepsilon e^{i\theta} \boldsymbol{\xi}$ for $\theta \in \mathbb{C}$ that is to be defined. In other words, $\check{\mathcal{G}}_{\text{qo}}$ is defined as

$$\check{\mathbf{G}}_{\text{qo}}(s_1, s_2) = \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) - \varepsilon e^{i\theta} \frac{\boldsymbol{\xi} (\mathbf{b}_i \otimes \mathbf{b}_j)^\top}{(s_1 - \lambda_i)(s_2 - \lambda_j)},$$

Thus, the transfer functions of $\check{\mathcal{G}}_{\text{lqo}}$ satisfy

$$\check{\mathbf{G}}_{\text{lo}}(s) = \tilde{\mathbf{G}}_{\text{lo}}(s) \quad \text{and} \quad \check{\mathbf{G}}_{\text{qo}}(s_1, s_2) - \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) = \varepsilon e^{i\theta} \frac{\boldsymbol{\xi} (\mathbf{b}_i \otimes \mathbf{b}_j)^\top}{(s_1 - \lambda_i)(s_2 - \lambda_j)},$$

where we define $\theta \in \mathbb{C}$ as

$$\theta \stackrel{\text{def}}{=} \pi - \arg \left(\underbrace{\boldsymbol{\xi}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j)}_{\stackrel{\text{def}}{=} z} \right) = \pi - \arg(z).$$

Note that θ is well defined under the assumption that the (i, j) -th condition (6.32b) is nonzero and $\boldsymbol{\xi} \in \mathbb{C}^p$ is nontrivial. Applying (5.45) and (5.46) to the quantities in (6.33) for $\check{\mathcal{G}}_{\text{lqo}}$ as

just defined as well as using the identity $z = |z|e^{i\arg(z)}$ yields

$$\begin{aligned} \left\langle \mathbf{G}_{\text{qo}} - \tilde{\mathbf{G}}_{\text{qo}}, \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}} \right\rangle_{\mathcal{H}_2} &= \varepsilon e^{i\pi} e^{-i\arg(z)} \boldsymbol{\xi}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j) \\ &= -\varepsilon \left| \boldsymbol{\xi}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j) \right| \neq 0, \\ \text{and } \|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 &= \varepsilon^2 |e^{i\theta}|^2 \boldsymbol{\xi}^\top \left(\overline{\tilde{\mathbf{G}}_{\text{qo}}}(-\lambda_i, -\lambda_j) - \overline{\check{\mathbf{G}}_{\text{qo}}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j) \\ &= \varepsilon^2 \frac{\boldsymbol{\xi}^\top \boldsymbol{\xi} (\mathbf{b}_i \otimes \mathbf{b}_j)^\top (\mathbf{b}_i \otimes \mathbf{b}_j)}{(\lambda_i - \bar{\lambda}_i)(\lambda_j - \bar{\lambda}_j)} \\ &= \varepsilon^2 \frac{\|\mathbf{b}_i \otimes \mathbf{b}_j\|_2^2}{4 \operatorname{Re}(\lambda_i) \operatorname{Re}(\lambda_j)} = O(\varepsilon^2). \end{aligned}$$

Moreover, $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 \geq 0$ since this holds for any norm. Because $\check{\mathbf{G}}_{\text{lo}} = \tilde{\mathbf{G}}_{\text{lo}}$, the inner products and norms involving the linear-output transfer functions in (6.33) are zero. Thus, substituting the above calculations into (6.33), we obtain

$$0 \leq -\varepsilon \left| \boldsymbol{\xi}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j) \right| + O(\varepsilon^2).$$

Since $\varepsilon > 0$ is arbitrarily specified, we may take it to be sufficiently small such that the negative $O(\varepsilon)$ term above is greater in magnitude than the positive $O(\varepsilon^2)$ term, yielding a contradiction. However, we assumed initially that the (i, j) -th interpolation condition in (6.32b) does not hold. Therefore, we must conclude by contradiction that it does. Repeating this argument for all i, j pairs yields

$$\left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_j) \right) (\mathbf{b}_i \otimes \mathbf{b}_j) = \mathbf{0}_p \text{ for each } i, j = 1, \dots, r,$$

which are precisely the right tangential conditions in (6.32b).

Next, assume that the k -th interpolation condition in (6.32c) does not hold. We obtain $\check{\mathbf{G}}_{\text{lo}}$ by applying the perturbation $-\varepsilon e^{i\theta} \boldsymbol{\xi}$ to the k -th residue direction \mathbf{b}_k in (5.42), where θ is to be re-defined (but using the same notation as before). Specifically, $\check{\mathbf{G}}_{\text{lo}}$'s transfer functions in this instance satisfy

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{lo}}(s) - \check{\mathbf{G}}_{\text{lo}}(s) &= \varepsilon e^{i\theta} \frac{\mathbf{c}_k \boldsymbol{\xi}^\top}{s - \lambda_k} \text{ and} \\ \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) - \check{\mathbf{G}}_{\text{qo}}(s_1, s_2) &= \varepsilon e^{i\theta} \left(\sum_{\ell=1}^r \frac{\mathbf{m}_{\ell,k} (\mathbf{b}_i \otimes \boldsymbol{\xi})^\top}{(s_1 - \lambda_\ell)(s_2 - \lambda_k)} + \sum_{\ell=1}^r \frac{\mathbf{m}_{k,\ell} (\boldsymbol{\xi} \otimes \mathbf{b}_\ell)^\top}{(s_1 - \lambda_k)(s_2 - \lambda_\ell)} \right) \\ &\quad - \varepsilon^2 e^{2i\theta} \frac{\mathbf{m}_{k,k} (\boldsymbol{\xi} \otimes \boldsymbol{\xi})^\top}{(s_1 - \lambda_k)(s_2 - \lambda_k)}. \end{aligned}$$

Implicitly, we have used the fact that the Kronecker product is bilinear [47] in simplifying the expression for $\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}$. We redefine $\theta \in \mathbb{C}$ as

$$\begin{aligned} \theta \stackrel{\text{def}}{=} \pi - \arg \left[\mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \boldsymbol{\xi} \right. \\ \left. + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{I}_m \otimes \mathbf{b}_\ell) \boldsymbol{\xi} \right. \\ \left. + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{I}_m) \boldsymbol{\xi} \right], \end{aligned} \quad (6.34a)$$

which is well-defined, since the quantity in the argument is nonzero. As before, we apply the formulae in Theorem 5.11 to compute the relevant terms in (6.33). First, by (5.45) the inner products are

$$\begin{aligned} \left\langle \mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}}, \tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}} \right\rangle_{\mathcal{H}_2} &= \varepsilon e^{i\theta} \mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \boldsymbol{\xi}, \\ \left\langle \mathbf{G}_{\text{qo}} - \tilde{\mathbf{G}}_{\text{qo}}, \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}} \right\rangle_{\mathcal{H}_2} &= \\ \varepsilon e^{i\theta} \left[\sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{I}_m) \boldsymbol{\xi} \right. \\ &+ \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{I}_m \otimes \mathbf{b}_\ell) \boldsymbol{\xi} \left. \right] \\ &- \varepsilon^2 e^{2i\theta} \mathbf{m}_{k,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) \right) (\boldsymbol{\xi} \otimes \boldsymbol{\xi}). \end{aligned} \quad (6.34b)$$

In the latter, we have used the fact that $(\mathbf{b}_i \otimes \boldsymbol{\xi}) = (\mathbf{b}_i \otimes \mathbf{I}_m) \boldsymbol{\xi}$ and $(\boldsymbol{\xi} \otimes \mathbf{b}_j) = (\mathbf{I}_m \otimes \mathbf{b}_j) \boldsymbol{\xi}$; this follows straightforwardly from the definition of the Kronecker product. By (5.46), the norm of $\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}$ is

$$\|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2 = \varepsilon^2 \frac{\|\mathbf{c}_k\|_2^2}{-2 \operatorname{Re}(\lambda_k)} = O(\varepsilon^2). \quad (6.34c)$$

At first pass, the norm of $\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}$ is

$$\begin{aligned} \|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 &= \varepsilon |e^{i\theta}| \left[\sum_{i=1}^r \mathbf{m}_{i,k}^\top \left(\overline{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) - \overline{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) \right) (\mathbf{b}_i \otimes \mathbf{I}_m) \boldsymbol{\xi} \right. \\ &+ \sum_{j=1}^r \mathbf{m}_{k,j}^\top \left(\overline{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_j) - \overline{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_j) \right) (\mathbf{I}_m \otimes \mathbf{b}_j) \boldsymbol{\xi} \left. \right] \\ &- \varepsilon^2 |e^{2i\theta}| \mathbf{m}_{k,k}^\top \left(\overline{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \overline{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) \right) (\boldsymbol{\xi} \otimes \boldsymbol{\xi}). \end{aligned}$$

Substituting directly for the pole residue form of the error function $\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}$ allows us to

realize its norm as an $O(\varepsilon^2)$ term, i.e.,

$$\begin{aligned} \|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 &= \varepsilon^2 \sum_{i=1}^r \mathbf{m}_{i,k}^\top \left[\sum_{\ell=1}^r \frac{\overline{\mathbf{m}}_{\ell,k} (\bar{\mathbf{b}}_\ell \otimes \boldsymbol{\xi})^\top}{(-\lambda_i - \bar{\lambda}_\ell)(-2\operatorname{Re}(\lambda_k))} \right. \\ &+ \sum_{\ell=1}^r \frac{\overline{\mathbf{m}}_{k,\ell} (\boldsymbol{\xi} \otimes \bar{\mathbf{b}}_\ell)^\top}{(-\lambda_i - \bar{\lambda}_k)(-\lambda_k - \bar{\lambda}_\ell)} \left. \right] (\bar{\mathbf{b}}_i \otimes \mathbf{I}_m) \boldsymbol{\xi} + \varepsilon^2 \sum_{j=1}^r \mathbf{m}_{k,j}^\top \left[\sum_{\ell=1}^r \frac{\overline{\mathbf{m}}_{\ell,k} (\bar{\mathbf{b}}_\ell \otimes \boldsymbol{\xi})^\top}{(-\lambda_k - \bar{\lambda}_\ell)(-\lambda_j - \bar{\lambda}_k)} \right. \\ &\left. + \sum_{\ell=1}^r \frac{\overline{\mathbf{m}}_{k,\ell} (\boldsymbol{\xi} \otimes \bar{\mathbf{b}}_\ell)^\top}{(-2\operatorname{Re}(\lambda_k))(-\lambda_k - \bar{\lambda}_\ell)} \right] (\mathbf{I}_m \otimes \mathbf{b}_j) \boldsymbol{\xi} + O(\varepsilon^4) = O(\varepsilon^2). \end{aligned} \quad (6.34d)$$

Then, by the definition of θ in (6.34a), substituting the calculations (6.34b), (6.34c), and (6.34d) into (6.33) yields

$$\begin{aligned} 0 \leq -\varepsilon \left| \mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \boldsymbol{\xi} \right. \\ \left. + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{I}_m \otimes \mathbf{b}_\ell) \boldsymbol{\xi} \right. \\ \left. + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{I}_m) \boldsymbol{\xi} \right| + O(\varepsilon^2). \end{aligned}$$

For sufficiently small $\varepsilon \geq 0$, this yields a contradiction. Because $\boldsymbol{\xi}$ is nontrivial, we must conclude

$$\begin{aligned} \mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{I}_m \otimes \mathbf{b}_\ell) \\ + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{I}_m) = \mathbf{0}_m \quad \text{for } k = 1, \dots, r, \end{aligned}$$

by repeating this argument for all k , thereby proving (6.32c).

Finally, we prove the bi-tangential Hermite condition in (6.32d). As before, we assume that the k -th condition in (6.32d) does not hold. Re-define $\theta \in \mathbb{C}$ as

$$\begin{aligned} \theta \stackrel{\text{def}}{=} -\arg \left[\mathbf{c}_k^\top \left(\frac{d}{ds} \mathbf{G}_{\text{lo}}(-\lambda_k) - \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \mathbf{b}_k \right. \\ \left. + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{b}_k \otimes \mathbf{b}_\ell) \right. \\ \left. + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{b}_k) \right]. \end{aligned} \quad (6.35a)$$

Take $\varepsilon > 0$ to be small enough so that $\eta_k \stackrel{\text{def}}{=} \lambda_k + \varepsilon e^{i\theta}$ does not coincide with any of the remaining poles of $\tilde{\mathbf{G}}_{\text{lo}}$ and $\text{Re}(\eta_k) < 0$. We obtain $\check{\mathbf{G}}_{\text{lo}}$ by replacing the k -th pole λ_k of $\tilde{\mathbf{G}}_{\text{lo}}$ with η_k defined above. Then, the transfer functions of $\check{\mathbf{G}}_{\text{lo}}$ are such that

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{lo}}(s) - \check{\mathbf{G}}_{\text{lo}}(s) &= \mathbf{c}_k \mathbf{b}_k^\top \left(\frac{1}{s - \lambda_k} - \frac{1}{s - \eta_k} \right) \\ \text{and } \tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2) - \check{\mathbf{G}}_{\text{qo}}(s_1, s_2) &= \sum_{\ell \neq k}^r \frac{\mathbf{m}_{\ell, k} (\mathbf{b}_\ell \otimes \mathbf{b}_k)^\top}{s_1 - \lambda_\ell} \left(\frac{1}{s_2 - \lambda_k} - \frac{1}{s_2 - \eta_k} \right) \\ &\quad + \sum_{\ell \neq k}^r \left(\frac{1}{s_1 - \lambda_k} - \frac{1}{s_1 - \eta_k} \right) \frac{\mathbf{m}_{k, \ell} (\mathbf{b}_k \otimes \mathbf{b}_\ell)^\top}{s_2 - \lambda_\ell} \\ &\quad + \mathbf{m}_{k, k} (\mathbf{b}_k \otimes \mathbf{b}_k)^\top \left(\frac{1}{(s_1 - \lambda_k)(s_2 - \lambda_k)} - \frac{1}{(s_1 - \eta_k)(s_2 - \eta_k)} \right). \end{aligned} \quad (6.35b)$$

From its pole-residue form, we observe that the difference function $\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}$ has two poles λ_k and η_k corresponding to the residues $\mathbf{c}_k \mathbf{b}_k^\top$ and $-\mathbf{c}_k \mathbf{b}_k^\top$. Thus, applying (5.45) yields

$$\begin{aligned} \left\langle \mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}}, \tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}} \right\rangle_{\mathcal{H}_2} &= \mathbf{c}_k^\top \underbrace{\left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right)}_{= \mathbf{0}_p \text{ by (6.32a)}} \mathbf{b}_k - \mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\eta_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\eta_k) \right) \mathbf{b}_k. \end{aligned}$$

To resolve this further, we recognize that $\mathbf{G}_{\text{lo}}(s)$ and $\tilde{\mathbf{G}}_{\text{lo}}(s)$ are both analytic at $s = -\lambda_k$, and thus admit power series representations about this point. Expanding both $\mathbf{G}_{\text{lo}}(s)$ and $\tilde{\mathbf{G}}_{\text{lo}}(s)$ about $-\lambda_k$, and evaluating at $s = -\eta_k$ gives

$$\begin{aligned} \left\langle \mathbf{G}_{\text{lo}} - \tilde{\mathbf{G}}_{\text{lo}}, \tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}} \right\rangle_{\mathcal{H}_2} &= -\mathbf{c}_k^\top \left(\mathbf{G}_{\text{lo}}(-\eta_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\eta_k) \right) \mathbf{b}_k \\ &= -\mathbf{c}_k^\top \left[\left(\mathbf{G}_{\text{lo}}(-\lambda_k) + \underbrace{(-\eta_k - \lambda_k)}_{=-\varepsilon e^{i\theta}} \frac{d}{ds} \mathbf{G}_{\text{lo}}(-\lambda_k) + O(\varepsilon^2) \right) \right. \\ &\quad \left. - \left(\tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) + \underbrace{(-\eta_k - \lambda_k)}_{=-\varepsilon e^{i\theta}} \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) + O(\varepsilon^2) \right) \right] \mathbf{b}_k \\ &= -\varepsilon e^{i\theta} \mathbf{c}_k^\top \left(\frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) - \frac{d}{ds} \mathbf{G}_{\text{lo}}(-\lambda_k) \right) \mathbf{b}_k + O(\varepsilon^2), \end{aligned} \quad (6.35c)$$

since $\left(\mathbf{G}_{\text{lo}}(-\lambda_k) - \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \mathbf{b}_k = \mathbf{0}$ by (6.32a). Accounting for all the pole-residue pairs

of $\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}$, applying (5.45) yields

$$\begin{aligned}
\left\langle \mathbf{G}_{\text{qo}} - \tilde{\mathbf{G}}_{\text{qo}}, \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}} \right\rangle_{\mathcal{H}_2} &= \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \underbrace{\left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) \right)}_{=0 \text{ by (6.32b)}} (\mathbf{b}_i \otimes \mathbf{b}_k) \\
&\quad - \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\eta_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\eta_k) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) \\
&\quad + \sum_{j \neq k}^r \mathbf{m}_{k,j}^\top \underbrace{\left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_j) \right)}_{=0 \text{ by (6.32b)}} (\mathbf{b}_k \otimes \mathbf{b}_j) \\
&\quad - \sum_{j \neq k}^r \mathbf{m}_{k,j}^\top \left(\mathbf{G}_{\text{qo}}(-\eta_k, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\lambda_j) \right) (\mathbf{b}_k \otimes \mathbf{b}_j) \\
&\quad + \mathbf{m}_{k,k}^\top \underbrace{\left(\mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) \right)}_{=0 \text{ by (6.32b)}} (\mathbf{b}_k \otimes \mathbf{b}_k) \\
&\quad - \mathbf{m}_{k,k}^\top \left(\mathbf{G}_{\text{qo}}(-\eta_k, -\eta_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\eta_k) \right) (\mathbf{b}_k \otimes \mathbf{b}_k).
\end{aligned} \tag{6.35d}$$

Both $\mathbf{G}_{\text{qo}}(s_1, s_2)$ and $\tilde{\mathbf{G}}_{\text{qo}}(s_1, s_2)$ are analytic at $s = -\lambda_k$ in each separate argument, and thus admit power series expansions about this point in each separate argument. Expanding, e.g., $\mathbf{G}_{\text{qo}}(-\lambda_i, s_2) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, s_2)$ in s_2 about $-\lambda_k$ and evaluating at $s_2 = \eta_k$ for each $i \neq k$ gives

$$\begin{aligned}
&\mathbf{m}_{i,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\eta_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\eta_k) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) \\
&= \mathbf{m}_{i,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_k) + \underbrace{(-\eta_k - \lambda_k)}_{=-\varepsilon e^{i\theta}} \frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_k) + O(\varepsilon^2) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) \\
&\quad - \mathbf{m}_{i,k}^\top \left(\tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) + \underbrace{(-\eta_k - \lambda_k)}_{=-\varepsilon e^{i\theta}} \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) + O(\varepsilon^2) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) \\
&= \varepsilon e^{i\theta} \mathbf{m}_{i,k}^\top \left(\frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) - \frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_k) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) + O(\varepsilon^2),
\end{aligned}$$

since $\left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\lambda_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\lambda_k) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) = \mathbf{0}$ by (6.32b). Similarly, expanding $\mathbf{G}_{\text{qo}}(s_1, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(s_1, -\lambda_j)$ in s_1 about $-\lambda_k$ and evaluating at $s_1 = \eta_k$ for each $j \neq k$

gives

$$\begin{aligned} & \mathbf{m}_{k,j}^\top \left(\mathbf{G}_{\text{qo}}(-\eta_k, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\lambda_j) \right) (\mathbf{b}_k \otimes \mathbf{b}_j) \\ &= \varepsilon e^{i\theta} \mathbf{m}_{k,j}^\top \left(\frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_j) - \frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_j) \right) (\mathbf{b}_k \otimes \mathbf{b}_j) + O(\varepsilon^2). \end{aligned}$$

To finish simplifying the inner product (6.35d), in the (k, k) -th term, expand $\mathbf{G}_{\text{qo}}(s_1, -\eta_k)$ in s_1 about $-\lambda_k$ and evaluate at $s_1 = -\eta_k$ to obtain

$$\mathbf{G}_{\text{qo}}(-\eta_k, -\eta_k) = \mathbf{G}_{\text{qo}}(-\lambda_k, -\eta_k) - \varepsilon e^{i\theta} \frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\eta_k) + O(\varepsilon^2).$$

Then express $\mathbf{G}_{\text{qo}}(-\lambda_k, -\eta_k)$ as a series expansion of $\mathbf{G}_{\text{qo}}(-\lambda_k, s_2)$ in s_2 about $-\lambda_k$, evaluated at $s_2 = -\eta_k$:

$$\mathbf{G}_{\text{qo}}(-\lambda_k, -\eta_k) = \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) - \varepsilon e^{i\theta} \frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) + O(\varepsilon^2).$$

Because $\mathbf{G}_{\text{qo}}(s_1, s_2)$ is analytic in each argument, it is in fact infinitely differentiable. So, its partial derivative $\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, s_2)$ is analytic in s_2 and may also be expressed as a power series about $-\lambda_k$. Expand about this point and evaluate at $s_2 = -\eta_k$:

$$\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\eta_k) = \frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) - \varepsilon e^{i\theta} \frac{\partial}{\partial s_2} \frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) + O(\varepsilon^2).$$

Putting this all together, we have

$$\begin{aligned} \mathbf{G}_{\text{qo}}(-\eta_k, -\eta_k) &= \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) - \varepsilon e^{i\theta} \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) + \frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) \right) \\ &\quad + O(\varepsilon^2). \end{aligned}$$

Applying the exact same logic to the $\tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\eta_k)$ term, we have

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\eta_k) &= \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \varepsilon e^{i\theta} \left(\frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) + \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) \right) \\ &\quad + O(\varepsilon^2). \end{aligned}$$

Putting all of these calculations together, the (k, k) -th term in (6.35d) simplifies to

$$\begin{aligned} & \mathbf{m}_{k,k}^\top \left(\mathbf{G}_{\text{qo}}(-\eta_k, -\eta_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\eta_k) \right) (\mathbf{b}_k \otimes \mathbf{b}_k) \\ &= \varepsilon e^{i\theta} \mathbf{m}_{k,k}^\top \left(\frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) \right) (\mathbf{b}_k \otimes \mathbf{b}_k) \\ &\quad + \varepsilon e^{i\theta} \mathbf{m}_{k,k}^\top \left(\frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_k) \right) (\mathbf{b}_k \otimes \mathbf{b}_k) + O(\varepsilon^2), \end{aligned}$$

and so the expression for the inner product (6.35d) ultimately simplifies to

$$\begin{aligned}
\left\langle \mathbf{G}_{\text{qo}} - \tilde{\mathbf{G}}_{\text{qo}}, \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}} \right\rangle_{\mathcal{H}_2} &= - \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \left(\mathbf{G}_{\text{qo}}(-\lambda_i, -\eta_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_i, -\eta_k) \right) (\mathbf{b}_i \otimes \mathbf{b}_k) \\
&\quad - \sum_{j \neq k}^r \mathbf{m}_{k,j}^\top \left(\mathbf{G}_{\text{qo}}(-\eta_k, -\lambda_j) - \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\lambda_j) \right) (\mathbf{b}_k \otimes \mathbf{b}_j) \\
&\quad - \mathbf{m}_{k,k}^\top \left(\mathbf{G}_{\text{qo}}(-\eta_k, -\eta_k) - \tilde{\mathbf{G}}_{\text{qo}}(-\eta_k, -\eta_k) \right) (\mathbf{b}_k \otimes \mathbf{b}_k) \\
&= \varepsilon e^{i\theta} \left[\sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{b}_k \otimes \mathbf{b}_\ell) \right. \\
&\quad \left. + \sum_{\ell=k}^r \mathbf{m}_{\ell,k}^\top \left(\frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{b}_k) \right] \\
&\quad + O(\varepsilon^2). \tag{6.35e}
\end{aligned}$$

Note that in passing from the first to the second equality, we have relabeled the sums over i and j to run over ℓ to agree with the claim (6.32d), and grouped the (k, k) -th terms into each of these sums. What remains is to deal with the norms in (6.33) for this case. Similar to the previous arguments, we show that $\|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2$ and $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2$ are $O(\varepsilon^2)$ by direct calculation. The calculations required to do so are straightforward, but technically involved, and so we present them in Appendix A.

Using the fact that $\|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2$, $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2$ are $O(\varepsilon^2)$, we substitute the computed inner products (6.35c) and (6.35e) into (6.33). Then, using $z = e^{i\theta}|z|$ with the definition of θ in (6.35a), we observe

$$\begin{aligned}
0 \leq & -\varepsilon \left| \mathbf{c}_k^\top \left(\frac{d}{ds} \mathbf{G}_{\text{lo}}(-\lambda_k) - \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) \right) \mathbf{b}_k \right. \\
& + \sum_{\ell=1}^r \mathbf{m}_{k,\ell}^\top \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(-\lambda_k, -\lambda_\ell) - \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_k, -\lambda_\ell) \right) (\mathbf{b}_k \otimes \mathbf{b}_\ell) \\
& \left. + \sum_{\ell=1}^r \mathbf{m}_{\ell,k}^\top \left(\frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(-\lambda_\ell, -\lambda_k) - \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(-\lambda_\ell, -\lambda_k) \right) (\mathbf{b}_\ell \otimes \mathbf{b}_k) \right| + O(\varepsilon^2).
\end{aligned}$$

By the same logic used to prove (6.32c), this inequality yields a contradiction for small values of $\varepsilon > 0$, and thus the interpolation conditions in (6.32d) must hold. \square

Theorem 6.9 explicitly ties the optimal- \mathcal{H}_2 approximation of LQO systems (5.1) with multivariate rational interpolation. More precisely, Theorem 6.9 states that any \mathcal{H}_2 -optimal reduced model is necessarily a *tangential interpolant* of the original system in the sense of (6.32). The interpolatory optimality conditions amount to:

- (i) The *right-tangential Lagrange interpolation* of the linear- and quadratic-output transfer functions \mathbf{G}_{lo} and \mathbf{G}_{qo} , individually;
- (ii) The *left-tangential Lagrange interpolation* of a sum of \mathbf{G}_{lo} and \mathbf{G}_{qo} evaluated at all possible combinations of the optimal interpolation points;
- (iii) The *bi-tangential Hermite interpolation* of a sum of \mathbf{G}_{lo} and \mathbf{G}_{qo} evaluated at all possible combinations of the optimal interpolation points.

Henceforth, we refer to the latter type of tangential interpolation conditions appearing in (6.32c) and (6.32d) as *mixed-multipoint tangential interpolation conditions*, since they correspond to interpolating a linear combination (or mix) of \mathbf{G}_{lo} and \mathbf{G}_{qo} at multiple (and in fact, all possible) combinations of the optimal interpolation points.

How does the \mathcal{H}_2 -optimality framework prescribed by Theorem 6.9 compare with other interpolation-based optimality frameworks? As is the case with the \mathcal{H}_2 -optimal model reduction of linear [97, 217], bilinear [77, 78], and quadratic-bilinear [50] systems, the optimal interpolation points in Theorem 6.9 are the *mirror images of the reduced model poles reflected across the imaginary axis*; the optimal tangential directions are the residue directions (5.43) associated with these poles. Moreover, the conditions in (6.32) provide a satisfying generalization of the interpolatory \mathcal{H}_2 -optimality conditions for the approximation of linear systems written in Theorem 2.44. Indeed, if $\mathbf{M} = \mathbf{0}_{p \times n^2}$ and $\widetilde{\mathbf{M}} = \mathbf{0}_{p \times r^2}$ in (5.1) and (5.2), then the quadratic-output transfer functions \mathbf{G}_{qo} and $\widetilde{\mathbf{G}}_{\text{qo}}$ vanish, and the conditions in (6.32) reduce to the familiar interpolation-based first-order optimality conditions (2.69) from linear model reduction. In other words: *Theorem 6.9 establishes the Meier-Luenberger framework for the optimal- \mathcal{H}_2 approximation of linear quadratic-output systems*, and includes Theorem 2.45 as a special case.

From a different point of view, the mixed-multipoint conditions in (6.32c) and (6.32d) can be interpreted as multipoint Volterra series interpolation conditions, since each of the transfer functions \mathbf{G}_{lo} and \mathbf{G}_{qo} correspond to a kernel in the (finite) Volterra series expansion of (5.1); recall Remark 5.2. As already highlighted, multipoint Volterra series interpolation is a necessary condition for the optimal- \mathcal{H}_2 approximation of bilinear [77, 78] and quadratic-bilinear [50] systems.

Remark 6.10 (\mathcal{H}_2 -optimal approximation of structured systems). While the hypotheses of Theorem 6.9 are stated to accommodate the approximation of LQO systems with first-order time derivatives as in (5.1), the proof of Theorem 6.9 that we provide does not make any explicit reference to this first-order requirement. Indeed, the proof only assumes:

1. The transfer functions of the \mathcal{H}_2 -optimal reduced model permit pole-residue expansions as in (5.42).
2. The full-order functions \mathbf{G}_{lo} and \mathbf{G}_{qo} are members of the relevant Hardy spaces.

This observation allows for the application of Theorem 6.9 to other classes of LQO systems with various internal structures. For instance, the state dynamics may have second-order differential [9, 226] or delay [173] structure. In any case, the output equation is still $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{M}(\mathbf{x}(t) \otimes \mathbf{x}(t))$. Then, Theorem 6.9 states that \mathcal{H}_2 -optimal approximants of the form (5.2) (those described by rational transfer functions with simple poles) satisfy the interpolation conditions in (6.32), the internal structure of the full-order model notwithstanding. Nonetheless, it is not clear how to construct linear, first-order quadratic-output approximations (5.2) to structured systems, and we leave this question to future work.

◇

Suppose that the optimal interpolation data (that is, the poles λ_i and residue directions $\mathbf{b}_i, \mathbf{c}_i, \mathbf{m}_{j,k}$ of an optimal- \mathcal{H}_2 approximation (5.2)) are given. Can the interpolation-based optimality conditions of Theorem 6.9 be enforced by Petrov-Galerkin projection? From Theorem 6.5, it follows that any optimal- \mathcal{H}_2 approximation of the form (5.2) is necessarily obtained via Petrov-Galerkin projection. As an immediate consequence, the interpolatory \mathcal{H}_2 -optimal reduced models characterized by Theorem 6.9 are necessarily projection-based, as well. However, it is not *a priori* clear how to enforce all of the $3r+r^2$ interpolation conditions in (6.32) simultaneously by an appropriate choice of model reduction bases \mathbf{V} and \mathbf{W} . This is in contrast to the Sylvester equation-based optimality framework of Theorem 6.5, which reveals the associated \mathbf{V} and \mathbf{W} rather plainly. Theorem 5.7 shows how to enforce the right-tangential Lagrange conditions (6.32a) and (6.32b), but not the newly derived mixed-multipoint conditions (6.32c) and (6.32d) that are necessary for optimality. In the subsequent result, we prove how to enforce *all* of the necessary interpolation conditions simultaneously by explicit construction of \mathbf{V} and \mathbf{W} in (5.24). Given $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{n \times r}$, we recall that a LQO-ROM computed via Petrov-Galerkin projection is given as

$$\tilde{\mathbf{E}} = \mathbf{W}^\top \mathbf{E} \mathbf{V}, \quad \tilde{\mathbf{A}} = \mathbf{W}^\top \mathbf{A} \mathbf{V}, \quad \tilde{\mathbf{B}} = \mathbf{W}^\top \mathbf{B}, \quad \tilde{\mathbf{C}} = \mathbf{C} \mathbf{V}, \quad \text{and} \quad \tilde{\mathbf{M}} = \mathbf{M}(\mathbf{V} \otimes \mathbf{V}).$$

Theorem 6.11. Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2) with the transfer functions $\mathbf{G}_{\text{lo}}, \mathbf{G}_{\text{qo}}$ and $\tilde{\mathbf{G}}_{\text{lo}}, \tilde{\mathbf{G}}_{\text{qo}}$ defined according to (5.12), where $\tilde{\mathcal{G}}_{\text{lqo}}$ is computed by projection using $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{n \times r}$. Consider interpolation points $\sigma_1, \dots, \sigma_r \in \mathbb{C}$ such that $\sigma_i \mathbf{E} - \mathbf{A}$ and $\sigma_i \tilde{\mathbf{E}} - \tilde{\mathbf{A}}$ are invertible for all $i = 1, \dots, r$, right-tangential directions $\mathbf{r}_1, \dots, \mathbf{r}_r \in \mathbb{C}^m$, and left-tangential directions $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_r \in \mathbb{C}^p$ and $\mathbf{q}_{1,1}, \dots, \mathbf{q}_{r,r} \in \mathbb{C}^p$ such that $\mathbf{q}_{j,k} = \mathbf{q}_{k,j}$ for all $j, k = 1, \dots, r$. Suppose that \mathbf{V} and \mathbf{W} have full rank and satisfy

$$\mathbf{v}_k \stackrel{\text{def}}{=} (\sigma_k \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{r}_k \in \text{Range}(\mathbf{V}), \quad (6.36)$$

$$\mathbf{w}_k \stackrel{\text{def}}{=} (\sigma_k \mathbf{E}^\top - \mathbf{A}^\top)^{-1} \left(2 \sum_{\ell=1}^r [\mathbf{M}_1 \mathbf{v}_\ell \quad \dots \quad \mathbf{M}_p \mathbf{v}_\ell] \mathbf{q}_{k,\ell} + \mathbf{C}^\top \boldsymbol{\ell}_k \right) \in \text{Range}(\mathbf{W}), \quad (6.37)$$

for all $k = 1, \dots, r$. Then, $\tilde{\mathcal{G}}_{\text{loqo}}$ satisfies the $3r + r^2$ tangential interpolation conditions:

$$\mathbf{0}_p = \left(\mathbf{G}_{\text{lo}}(\sigma_i) - \tilde{\mathbf{G}}_{\text{lo}}(\sigma_i) \right) \mathbf{r}_i, \quad (6.38a)$$

$$\mathbf{0}_p = \left(\mathbf{G}_{\text{qo}}(\sigma_i, \sigma_j) - \tilde{\mathbf{G}}_{\text{qo}}(\sigma_i, \sigma_j) \right) (\mathbf{r}_i \otimes \mathbf{r}_j), \quad (6.38b)$$

$$\begin{aligned} \mathbf{0}_m &= \boldsymbol{\ell}_k^\top \left(\mathbf{G}_{\text{lo}}(\sigma_k) - \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) \right) + \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(\sigma_k, \sigma_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \\ &\quad + \sum_{\ell=1}^r \mathbf{q}_{\ell,k}^\top \left(\mathbf{G}_{\text{qo}}(\sigma_\ell, \sigma_k) - \tilde{\mathbf{G}}_{\text{qo}}(\sigma_\ell, \sigma_k) \right) (\mathbf{r}_\ell \otimes \mathbf{I}_m), \end{aligned} \quad (6.38c)$$

$$\begin{aligned} 0 &= \boldsymbol{\ell}_k^\top \left(\frac{d}{ds} \mathbf{G}_{\text{lo}}(\sigma_k) - \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) \right) \mathbf{r}_k \\ &\quad + \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(\sigma_k, \sigma_\ell) - \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) \right) (\mathbf{r}_k \otimes \mathbf{r}_\ell) \\ &\quad + \sum_{\ell=1}^r \mathbf{q}_{\ell,k}^\top \left(\frac{\partial}{\partial s_2} \mathbf{G}_{\text{qo}}(\sigma_\ell, \sigma_k) - \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_\ell, \sigma_k) \right) (\mathbf{r}_\ell \otimes \mathbf{r}_k), \end{aligned} \quad (6.38d)$$

for all $i, j, k = 1, \dots, r$. ◇

Proof of Theorem 6.11. First, we derive two identities that are invoked repeatedly throughout the proof. Define $\mathcal{K}(s) \stackrel{\text{def}}{=} s\mathbf{E} - \mathbf{A}$ and $\tilde{\mathcal{K}}(s) \stackrel{\text{def}}{=} s\tilde{\mathbf{E}} - \tilde{\mathbf{A}}$. By the construction of $\mathbf{V} \in \mathbb{C}^{n \times r}$ in (6.36) and the assumption that \mathbf{V} is full rank, there exists $\check{\mathbf{v}}_k \in \mathbb{C}^r$ so that $\mathbf{V}\check{\mathbf{v}}_k = \mathbf{v}_k = \mathcal{K}(\sigma_k)^{-1}\mathbf{B}\mathbf{r}_k$ and

$$\begin{aligned} \tilde{\mathcal{K}}(\sigma_k)\check{\mathbf{v}}_k &= (\sigma_k \mathbf{W}^\top \mathbf{E} \mathbf{V} - \mathbf{W}^\top \mathbf{A} \mathbf{V}) \check{\mathbf{v}}_k = \mathbf{W}^\top (\sigma_k \mathbf{E} - \mathbf{A}) \mathbf{V} \check{\mathbf{v}}_k \\ &= \mathbf{W}^\top \mathcal{K}(\sigma_k) \mathcal{K}(\sigma_k)^{-1} \mathbf{B} \mathbf{r}_k \text{ by design of } \check{\mathbf{v}}_k, \\ &\text{which implies } \check{\mathbf{v}}_k = \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \mathbf{r}_k. \end{aligned} \quad (6.39)$$

Equation (6.39) is the first of the aforementioned identities. To prove the second, first note that by construction of \mathbf{V} and (6.39), we have for each $i = 1, \dots, r$ and $j = 1, \dots, r$

$$\mathbf{r}_i^\top \mathbf{B}^\top \mathcal{K}(\sigma_i)^{-\top} \mathbf{M}_j \mathbf{V} = \check{\mathbf{v}}_i^\top \mathbf{V}^\top \mathbf{M}_j \mathbf{V} = \check{\mathbf{v}}_i^\top \tilde{\mathbf{M}}_j = \mathbf{r}_i^\top \tilde{\mathbf{B}}^\top \tilde{\mathcal{K}}(\sigma_i)^{-\top} \tilde{\mathbf{M}}_j. \quad (6.40)$$

By the construction of $\mathbf{W} \in \mathbb{C}^{n \times r}$ in (6.37) and the assumption that \mathbf{W} is full rank, there exists $\check{\mathbf{w}}_k \in \mathbb{C}^r$ so that

$$\check{\mathbf{w}}_k^\top \mathbf{W}^\top = \boldsymbol{\ell}_k^\top \mathbf{C} \mathcal{K}(\sigma_k)^{-1} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \begin{bmatrix} \mathbf{r}_\ell^\top \mathbf{B}^\top \mathcal{K}(\sigma_\ell)^{-\top} \mathbf{M}_1 \mathcal{K}(\sigma_k)^{-1} \\ \vdots \\ \mathbf{r}_\ell^\top \mathbf{B}^\top \mathcal{K}(\sigma_\ell)^{-\top} \mathbf{M}_p \mathcal{K}(\sigma_k)^{-1} \end{bmatrix}.$$

By the above equality as well as (6.39), we have that, for each $k = 1, \dots, r$,

$$\begin{aligned} \check{\mathbf{w}}_k^\top \tilde{\mathcal{K}}(\sigma_k) &= \check{\mathbf{w}}_k^\top \mathbf{W}^\top \mathcal{K}(\sigma_k) \mathbf{V} = \ell_k^\top \underbrace{\mathbf{C}\mathbf{V}}_{=\tilde{\mathbf{C}}} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \begin{bmatrix} \mathbf{r}_\ell^\top \mathbf{B}^\top \mathcal{K}(\sigma_\ell)^{-\top} \mathbf{M}_1 \mathbf{V} \\ \vdots \\ \mathbf{r}_\ell^\top \mathbf{B}^\top \mathcal{K}(\sigma_\ell)^{-\top} \mathbf{M}_p \mathbf{V} \end{bmatrix}, \\ \text{which implies } \check{\mathbf{w}}_k^\top &= \ell_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \begin{bmatrix} \mathbf{r}_\ell^\top \tilde{\mathbf{B}}^\top \tilde{\mathcal{K}}(\sigma_\ell)^{-\top} \tilde{\mathbf{M}}_1 \tilde{\mathcal{K}}(\sigma_k)^{-1} \\ \vdots \\ \mathbf{r}_\ell^\top \tilde{\mathbf{B}}^\top \tilde{\mathcal{K}}(\sigma_\ell)^{-\top} \tilde{\mathbf{M}}_p \tilde{\mathcal{K}}(\sigma_k)^{-1} \end{bmatrix}, \end{aligned}$$

where the second line follows from right-inversion of $\tilde{\mathcal{K}}(\sigma_k)$ and (6.40). Then, by applying (2.7) to the above, we arrive at our second useful identity:

$$\check{\mathbf{w}}_k^\top = \ell_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_k)^{-1} \otimes \tilde{\mathcal{K}}(\sigma_\ell)^{-1} \tilde{\mathbf{B}} \mathbf{r}_\ell \right). \quad (6.41)$$

We are now prepared to prove that the Petrov-Galerkin reduced model $\tilde{\mathcal{G}}_{\text{lqo}}$ constructed using \mathbf{V} and \mathbf{W} in (6.36) and (6.37) satisfies the tangential interpolation conditions in (6.38). The construction of \mathbf{V} gives the conditions (6.38a) and (6.38b); observe that

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{lo}}(\sigma_i) \mathbf{r}_i &= \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_i)^{-1} \tilde{\mathbf{B}} \mathbf{r}_i = \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_i)^{-1} \tilde{\mathcal{K}}(\sigma_i)^{-1} \check{\mathbf{v}}_i \quad \text{by (6.39)} \\ &= \mathbf{C} \mathbf{V} \check{\mathbf{v}}_i \\ &= \mathbf{C} \mathcal{K}(\sigma_i)^{-1} \mathbf{B} \mathbf{r}_i = \mathbf{G}_{\text{lo}}(\sigma_i) \mathbf{r}_i. \end{aligned}$$

The last line follows from the previous choice of $\check{\mathbf{v}}_k$ and by the construction of \mathbf{V} in (6.36). This proves (6.38a) for all $i = 1, \dots, r$. Similarly, we have that

$$\begin{aligned} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_i, \sigma_j) (\mathbf{r}_i \otimes \mathbf{r}_j) &= \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_i)^{-1} \tilde{\mathbf{B}} \mathbf{r}_i \otimes \tilde{\mathcal{K}}(\sigma_j)^{-1} \tilde{\mathbf{B}} \mathbf{r}_j \right) \\ &= \mathbf{M} (\mathbf{V} \otimes \mathbf{V}) (\check{\mathbf{v}}_i \otimes \check{\mathbf{v}}_j) \quad \text{by (6.39)} \\ &= \mathbf{M} (\mathbf{V} \check{\mathbf{v}}_i \otimes \mathbf{V} \check{\mathbf{v}}_j) \quad \text{by (2.9)} \\ &= \mathbf{M} (\mathcal{K}(\sigma_i)^{-1} \mathbf{B} \mathbf{r}_i \otimes \mathcal{K}(\sigma_j)^{-1} \mathbf{B} \mathbf{r}_j) \\ &= \mathbf{G}_{\text{qo}}(\sigma_i, \sigma_j) (\mathbf{r}_i \otimes \mathbf{r}_j). \end{aligned}$$

This proves (6.38b) for all $i, j = 1, \dots, r$. For the left-tangential Lagrange conditions (6.38c), we observe that the reduced-order portion of the interpolation conditions is written as

$$\ell_k^\top \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) + \sum_{\ell=1}^r \left(\mathbf{q}_{k,\ell}^\top \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{I}_m \otimes \mathbf{r}_\ell) + \mathbf{q}_{\ell,k}^\top \tilde{\mathbf{G}}_{\text{qo}}(\sigma_\ell, \sigma_k) (\mathbf{r}_\ell \otimes \mathbf{I}_m) \right).$$

It follows directly from Lemma 5.3 and the assumption that $\mathbf{q}_{\ell,k} = \mathbf{q}_{k,\ell}$ for all $\ell, k = 1, \dots, r$ that $\mathbf{q}_{k,\ell}^\top \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{I}_m \otimes \mathbf{r}_\ell) = \mathbf{q}_{\ell,k}^\top \tilde{\mathbf{G}}_{\text{qo}}(\sigma_\ell, \sigma_k) (\mathbf{r}_\ell \otimes \mathbf{I}_m)$; a similar equality holds for \mathbf{G}_{qo} . Thus, to prove (6.38b) it instead suffices to show that

$$\mathbf{0}_m = \boldsymbol{\ell}_k^\top \left(\mathbf{G}_{\text{lo}}(\sigma_k) - \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) \right) + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \left(\mathbf{G}_{\text{qo}}(\sigma_k, \sigma_\ell) - \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell). \quad (6.42)$$

Noting that $(\mathbf{I}_m \otimes \mathbf{r}_\ell) \tilde{\mathbf{B}} = (\mathbf{I}_m \otimes \mathbf{r}_\ell) (\tilde{\mathbf{B}} \otimes \mathbf{1}) = (\tilde{\mathbf{B}} \otimes \mathbf{r}_\ell)$ for all $\ell = 1, \dots, r$, it follows that

$$\begin{aligned} & \boldsymbol{\ell}_k^\top \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \\ &= \boldsymbol{\ell}_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \otimes \tilde{\mathcal{K}}(\sigma_\ell)^{-1} \tilde{\mathbf{B}} \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \\ &= \boldsymbol{\ell}_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_k)^{-1} \otimes \tilde{\mathcal{K}}(\sigma_\ell)^{-1} \tilde{\mathbf{B}} \right) (\tilde{\mathbf{B}} \otimes \mathbf{r}_\ell) \quad (\text{by (2.9)}) \\ &= \left(\boldsymbol{\ell}_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_k)^{-1} \otimes \tilde{\mathcal{K}}(\sigma_\ell)^{-1} \tilde{\mathbf{B}} \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \right) \tilde{\mathbf{B}} \\ &= \check{\mathbf{w}}_k^\top \tilde{\mathbf{B}} \quad (\text{by (6.41)}). \end{aligned}$$

Finally, our initial choice of $\check{\mathbf{w}}_k$ yields

$$\begin{aligned} \check{\mathbf{w}}_k^\top \tilde{\mathbf{B}} &= \check{\mathbf{w}}_k^\top \mathbf{W}^\top \mathbf{B} = \boldsymbol{\ell}_k^\top \mathbf{C} \mathcal{K}(\sigma_k)^{-1} \mathbf{B} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \mathbf{M} \left(\mathcal{K}(\sigma_k)^{-1} \mathbf{B} \otimes \mathcal{K}(\sigma_\ell)^{-1} \mathbf{B} \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \\ &= \boldsymbol{\ell}_k^\top \mathbf{G}_{\text{lo}}(\sigma_k) + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \mathbf{G}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{I}_m \otimes \mathbf{r}_\ell). \end{aligned}$$

Chaining these equalities together proves the simplified claim in (6.42).

As with the zeroth-order conditions, the symmetry relation from Lemma 5.3 implies that $\mathbf{q}_{k,\ell}^\top \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{r}_k, \otimes \mathbf{r}_\ell) = \mathbf{q}_{\ell,k}^\top \frac{\partial}{\partial s_2} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_\ell, \sigma_k) (\mathbf{r}_\ell, \otimes \mathbf{r}_k)$, and likewise for \mathbf{G}_{qo} . So, it suffices to prove

$$\begin{aligned} 0 &= \boldsymbol{\ell}_k^\top \left(\frac{d}{ds} \mathbf{G}_{\text{lo}}(\sigma_k) - \frac{d}{ds} \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) \right) \mathbf{r}_k \\ &\quad + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \left(\frac{\partial}{\partial s_1} \mathbf{G}_{\text{qo}}(\sigma_k, \sigma_\ell) - \frac{\partial}{\partial s_1} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) \right) (\mathbf{r}_k \otimes \mathbf{r}_\ell). \end{aligned} \quad (6.43)$$

Observe that

$$\begin{aligned}
& \boldsymbol{\ell}_k^\top \frac{d}{dS} \tilde{\mathbf{G}}_{\text{lo}}(\sigma_k) \mathbf{r}_k + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \frac{\partial}{\partial S_1} \tilde{\mathbf{G}}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{r}_k \otimes \mathbf{r}_\ell) \\
&= -\boldsymbol{\ell}_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{E}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \mathbf{r}_k \\
&\quad - 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{E}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \otimes \tilde{\mathcal{K}}(\sigma_\ell)^{-1} \tilde{\mathbf{B}} \right) (\mathbf{r}_k \otimes \mathbf{r}_\ell) \\
&= - \left(\boldsymbol{\ell}_k^\top \tilde{\mathbf{C}} \tilde{\mathcal{K}}(\sigma_k)^{-1} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \tilde{\mathbf{M}} \left(\tilde{\mathcal{K}}(\sigma_k)^{-1} \otimes \tilde{\mathcal{K}}(\sigma_\ell)^{-1} \tilde{\mathbf{B}} \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \right) \tilde{\mathbf{E}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \mathbf{r}_k \\
&= -\tilde{\mathbf{w}}_k^\top \tilde{\mathbf{E}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \mathbf{r}_k \quad (\text{by (6.41)}).
\end{aligned}$$

And so

$$-\tilde{\mathbf{w}}_k^\top \tilde{\mathbf{E}} \tilde{\mathcal{K}}(\sigma_k)^{-1} \tilde{\mathbf{B}} \mathbf{r}_k = -\tilde{\mathbf{w}}_k^\top \tilde{\mathbf{E}} \tilde{\mathbf{v}}_k = -\tilde{\mathbf{w}}_k^\top \mathbf{W}^\top \mathbf{E} \mathbf{V} \tilde{\mathbf{v}}_k \quad (\text{by (6.39)}).$$

By definition of $\tilde{\mathbf{w}}_k$ and $\tilde{\mathbf{v}}_k$ and the mixed-product property (2.9), at last we have that

$$\begin{aligned}
& -\tilde{\mathbf{w}}_k^\top \mathbf{W}^\top \mathbf{E} \mathbf{V} \tilde{\mathbf{v}}_k \\
&= - \left(\boldsymbol{\ell}_k^\top \mathbf{C} \mathcal{K}(\sigma_k)^{-1} + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \mathbf{M} \left(\mathcal{K}(\sigma_k)^{-1} \otimes \mathcal{K}(\sigma_\ell)^{-1} \mathbf{B} \right) (\mathbf{I}_m \otimes \mathbf{r}_\ell) \right) \mathcal{K}(\sigma_k)^{-1} \mathbf{B} \mathbf{r}_k \\
&= \boldsymbol{\ell}_k^\top \frac{d}{dS} \mathbf{G}_{\text{lo}}(\sigma_k) \mathbf{r}_k + 2 \sum_{\ell=1}^r \mathbf{q}_{k,\ell}^\top \frac{\partial}{\partial S_1} \mathbf{G}_{\text{qo}}(\sigma_k, \sigma_\ell) (\mathbf{r}_k \otimes \mathbf{r}_\ell).
\end{aligned}$$

Chaining all these equalities together proves (6.43), and thus the claim (6.38d). \square

In a vacuum, Theorem 6.11 offers a new strategy for the interpolatory model reduction of LQO systems by imposing the mixed-multipoint tangential interpolation conditions in (6.38c) and (6.38d). With regard to \mathcal{H}_2 -optimal model reduction, if we choose the interpolation data in Theorem 6.11 to be $\sigma_i = -\lambda_i$, $\mathbf{r}_i = \mathbf{b}_i$, $\boldsymbol{\ell}_i = \mathbf{c}_i$, and $\mathbf{q}_{j,k} = \mathbf{m}_{j,k}$, the poles and residue directions of a system that minimizes the \mathcal{H}_2 model error, the first-order optimality conditions from Theorem 6.9 will be satisfied by the reduced model. Note that (5.44) implies that the symmetry hypothesis imposed upon the left-tangential directions $\mathbf{q}_{j,k}$ is trivially satisfied for this choice.

This discussion reveals the same chicken-egg problem that we encountered in Section 6.3 with the Wilson framework; the optimal selection of interpolation points and tangent directions requires *a priori* knowledge of an \mathcal{H}_2 -optimal reduced model. Next, we introduce a second fixed-point algorithm that generalizes the iterative rational Krylov algorithm (IRKA) [97] for automatically determining the optimal interpolation data and enforcing the corresponding \mathcal{H}_2 -optimality conditions (6.32) in an iterative fashion.

6.4.2 An iterative rational Krylov algorithm for optimal- \mathcal{H}_2 approximation of linear quadratic-output systems

The interpolation-based optimality conditions of Theorem 6.9 along with the interpolatory projections offered by Theorem 6.11 suggest a fixed point procedure based on iteratively corrected interpolation that we present in Algorithm 6.4.1; this is precisely the idea of IRKA. At each step of Algorithm 6.4.1, the interpolation points and tangential directions are taken from the poles and residue directions of the previous reduced model iterate; the $3r + r^2$ tangential interpolation conditions in (6.38) are then enforced by Petrov-Galerkin projection using these data. The iteration repeats until the largest magnitude change in the reduced model poles between consecutive iterates falls below a user-specified tolerance. Thus, the interpolation-based \mathcal{H}_2 -optimality conditions in (6.32) will be satisfied up to this tolerance if Algorithm 6.4.1 converges. Because the construction of \mathbf{V} and \mathbf{W} in (6.36) and (6.37) requires only the solution of shifted linear systems and sparse matrix calculations involving the full-order matrix operators, the proposed method is suitable for large-scale problems.

We discuss here some implementation details specific to Algorithm 6.4.1.

Keeping it real.

A natural construction of the interpolatory model reduction bases \mathbf{V} and \mathbf{W} given by Theorem 6.11 involves first computing the requisite shifted linear system solves, populating the columns of \mathbf{V} and \mathbf{W} with the n -vectors in (6.36) and (6.37), and orthonormalizing. However, one will almost surely obtain complex-valued reduced models as an artifact of this primitive construction when complex-valued interpolation data is used, as is the case in Algorithm 6.4.1. This is significant because the optimality conditions derived in Theorem 6.9 assume that the approximating system (5.2) is real valued, and so it is imperative that Algorithm 6.4.1 produces real-valued approximations. Fortunately, one can guarantee the computation of real-valued intermediate models throughout the iteration of Algorithm 6.4.1 via the following alternative blueprint.

Lemma 6.12 (Real-valued reduced models from complex-valued interpolation data). Assume that we have the following interpolation data: distinct interpolation points $\sigma_1, \dots, \sigma_r \in \mathbb{C}$, right-tangential directions $\mathbf{r}_1, \dots, \mathbf{r}_r \in \mathbb{C}^m$, and left-tangential directions $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_r \in \mathbb{C}^p$ and $\mathbf{q}_{1,1}, \dots, \mathbf{q}_{r,r} \in \mathbb{C}^p$ that satisfy the hypotheses of Theorem 6.11. Suppose that the interpolation points are arranged into complex conjugate pairs so that $\bar{\sigma}_i = \sigma_{i+1}$ or σ_i is

Algorithm 6.4.1: Linear quadratic-output iterative rational Krylov algorithm (LQO-IRKA).

Input: $\mathbf{E}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{M}$ from (5.1), order $1 \leq r < n$, tolerance $\epsilon > 0$, maximum number of iteration steps $M \geq 1$, initial interpolation points $\lambda_1^{(0)}, \dots, \lambda_r^{(0)} \in \mathbb{C}$, and directions $\mathbf{b}_1^{(0)}, \dots, \mathbf{b}_r^{(0)} \in \mathbb{C}^m$, $\mathbf{c}_1^{(0)}, \dots, \mathbf{c}_r^{(0)} \in \mathbb{C}^p$, $\mathbf{m}_{1,1}^{(0)}, \dots, \mathbf{m}_{r,r}^{(0)} \in \mathbb{C}^p$ closed under complex conjugation such that $\lambda_k^{(0)} \mathbf{E} - \mathbf{A}$ is invertible and $\mathbf{m}_{j,k}^{(0)} = \mathbf{m}_{k,j}^{(0)}$ for all $j, k = 1, \dots, r$.

Output: $\tilde{\mathbf{E}}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{M}}$ —state-space matrices of the converged model (5.2).

1 Iteration count $i = 0$.

2 **while** $\max_j |\lambda_j^{(i+1)} - \lambda_j^{(i)}| > \epsilon$ **and** $i \leq M$ **do**

3 Compute interpolatory model reduction bases $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{n \times r}$ according to Lemma 6.12 such that

$$\mathbf{v}_k = \left(\lambda_k^{(i)} \mathbf{E} - \mathbf{A} \right)^{-1} \mathbf{B} \mathbf{b}_k^{(i)} \in \text{Range}(\mathbf{V}),$$

$$\left(\lambda_k^{(i)} \mathbf{E}^\top - \mathbf{A}^\top \right)^{-1} \left(2 \sum_{\ell=1}^r [\mathbf{M}_1 \mathbf{v}_\ell \cdots \mathbf{M}_p \mathbf{v}_\ell] \mathbf{m}_{k,\ell}^{(i)} + \mathbf{C}^\top \mathbf{c}_k^{(i)} \right) \in \text{Range}(\mathbf{W}).$$

4 Orthonormalize bases \mathbf{V} and \mathbf{W}

$$\mathbf{V} \leftarrow \text{orth}(\mathbf{V}), \quad \mathbf{W} \leftarrow \text{orth}(\mathbf{W}).$$

5 Compute reduced-order matrices by Petrov-Galerkin projection:

$$\begin{aligned} \tilde{\mathbf{E}}^{(i+1)} &= \mathbf{W}^\top \mathbf{E} \mathbf{V}, & \tilde{\mathbf{A}}^{(i+1)} &= \mathbf{W}^\top \mathbf{A} \mathbf{V}, & \tilde{\mathbf{B}}^{(i+1)} &= \mathbf{W}^\top \mathbf{B}, \\ \tilde{\mathbf{C}}^{(i+1)} &= \mathbf{C} \mathbf{V}, & \tilde{\mathbf{M}}^{(i+1)} &= \mathbf{M} (\mathbf{V} \otimes \mathbf{V}). \end{aligned}$$

6 Compute $\lambda_k^{(i+1)} \in \mathbb{C}$ and $\mathbf{b}_k^{(i+1)} \in \mathbb{C}^m$, $\mathbf{c}_k^{(i+1)} \in \mathbb{C}^p$, $\mathbf{m}_{i,j}^{(i+1)} \in \mathbb{C}^p$ according to (5.43) from the eigendecomposition of $s\tilde{\mathbf{E}} - \tilde{\mathbf{A}}$ and set $i \leftarrow i + 1$.

7 **end**

real-valued, and the corresponding tangential directions are arranged as follows:

$$\begin{aligned} \bar{\mathbf{r}}_k &= \begin{cases} \mathbf{r}_{k+1} & \text{if } \bar{\sigma}_k = \sigma_{k+1} \\ \mathbf{r}_k & \text{else,} \end{cases} & \bar{\boldsymbol{\ell}}_k &= \begin{cases} \boldsymbol{\ell}_{k+1} & \text{if } \bar{\sigma}_k = \sigma_{k+1} \\ \boldsymbol{\ell}_k & \text{else,} \end{cases} \\ \bar{\mathbf{q}}_{j,k} &= \begin{cases} \mathbf{q}_{j+1,k+1} & \text{if } \bar{\sigma}_j = \sigma_{j+1}, \quad \bar{\sigma}_k = \sigma_{k+1} \\ \mathbf{q}_{j+1,k} & \text{if } \bar{\sigma}_j = \sigma_{j+1}, \quad \text{Im}(\sigma_k) = 0 \\ \mathbf{q}_{j,k+1} & \text{if } \text{Im}(\sigma_j) = 0, \quad \bar{\sigma}_k = \sigma_{k+1} \\ \mathbf{q}_{j,k} & \text{else,} \end{cases} \end{aligned} \tag{6.44}$$

for every other j, k . Let $\mathbf{v}_k \in \mathbb{C}^n$ and $\mathbf{w}_k \in \mathbb{C}^n$ be defined as in (6.36) and (6.37). Suppose that the matrices $\mathbf{V} \in \mathbb{C}^{n \times r}$ and $\mathbf{W} \in \mathbb{C}^{n \times r}$ are constructed as

$$\begin{aligned} \mathbf{V}(:, k) &= \mathbf{v}_k, & \text{if } \text{Im}(\sigma_k) = 0, \\ \mathbf{V}(:, k: k+1) &= [\text{Re}(\mathbf{v}_k) \quad \text{Im}(\mathbf{v}_k)] & \text{else,} \end{aligned} \quad (6.45)$$

$$\begin{aligned} \mathbf{W}(:, k) &= \mathbf{w}_k & \text{if } \text{Im}(\sigma_k) = 0, \\ \mathbf{W}(:, k: k+1) &= [\text{Re}(\mathbf{w}_k) \quad \text{Im}(\mathbf{w}_k)] & \text{else,} \end{aligned} \quad (6.46)$$

for every other k . Then, \mathbf{V} and \mathbf{W} are real valued, and it holds that

$$\text{Range}(\mathbf{V}) = \text{Range}(\mathbf{V}_p) \quad \text{and} \quad \text{Range}(\mathbf{W}) = \text{Range}(\mathbf{W}_p),$$

where $\mathbf{V}_p = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_r] \in \mathbb{C}^{n \times r}$ and $\mathbf{W}_p = [\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_r] \in \mathbb{C}^{n \times r}$. \diamond

Proof of Lemma 6.12. Note that \mathbf{V} and \mathbf{W} are real-valued by construction. To prove that, e.g., $\text{Range}(\mathbf{W}) = \text{Range}(\mathbf{W}_p)$, it suffices to show that the columns of \mathbf{W}_p are closed under complex conjugation. If this holds true, then by the construction of (6.46), \mathbf{W} and \mathbf{W}_p are related according to $\mathbf{W} = \mathbf{W}_p \mathbf{Q}$, where $\mathbf{Q} \in \mathbb{C}^{r \times r}$ is the block-diagonal matrix with blocks equal to $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$ if $\bar{\sigma}_k = \sigma_{k+1}$ and 1 otherwise. Because \mathbf{Q} is an orthogonal matrix, \mathbf{W} and \mathbf{W}_p have the same range; this same logic applies to \mathbf{V} and \mathbf{V}_p . Consider a fixed k such that $\text{Im}(\sigma_k) = 0$, and hence $\mathbf{r}_k \in \mathbb{R}^m$. Because \mathbf{E} , \mathbf{A} , and \mathbf{B} appearing in \mathbf{v}_k in (6.36) are real valued, obviously $\bar{\mathbf{v}}_k = \mathbf{v}_k$ in this case. Moreover, the subset of tangential directions $\{\mathbf{q}_{k,1}, \dots, \mathbf{q}_{k,r}\}$ is closed under conjugation for such k . For indices k such that $\text{Im}(\sigma_k) \neq 0$ and so $\bar{\sigma}_k = \sigma_{k+1}$, it holds that $\bar{\mathbf{r}}_k = \mathbf{r}_{k+1}$, and so

$$\bar{\mathbf{v}}_k = (\bar{\sigma}_k \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \bar{\mathbf{r}}_k = (\sigma_{k+1} \mathbf{E} - \mathbf{A})^{-1} \mathbf{B} \mathbf{r}_{k+1} = \mathbf{v}_{k+1}.$$

Moreover, the organization scheme in (6.44) guarantees that the left tangential directions satisfy $\{\bar{\mathbf{q}}_{k,1}, \dots, \bar{\mathbf{q}}_{k,r}\} = \{\mathbf{q}_{k+1,1}, \dots, \mathbf{q}_{k+1,r}\}$ for such k . Then, it is a straightforward consequence of these facts that the sum appearing in the construction of the columns of \mathbf{W} (6.37) satisfies

$$\begin{aligned} \sum_{i=1}^r [\mathbf{M}_1 \bar{\mathbf{v}}_i \quad \cdots \quad \mathbf{M}_p \bar{\mathbf{v}}_i] \bar{\mathbf{q}}_{k,i} &= \sum_{i=1}^r [\mathbf{M}_1 \mathbf{v}_i \quad \cdots \quad \mathbf{M}_p \mathbf{v}_i] \mathbf{q}_{k,i} & \text{if } \text{Im}(\sigma_k) = 0, \\ \sum_{i=1}^r [\mathbf{M}_1 \bar{\mathbf{v}}_i \quad \cdots \quad \mathbf{M}_p \bar{\mathbf{v}}_i] \bar{\mathbf{q}}_{k,i} &= \sum_{i=1}^r [\mathbf{M}_1 \mathbf{v}_i \quad \cdots \quad \mathbf{M}_p \mathbf{v}_i] \mathbf{q}_{k+1,i} & \text{else.} \end{aligned}$$

Thus, for indices k such that $\text{Im}(\sigma_k) = 0$ it holds that

$$\bar{\mathbf{w}}_k = (\bar{\sigma}_k \mathbf{E}^\top - \mathbf{A}^\top)^{-1} \left(2 \sum_{i=1}^r [\mathbf{M}_1 \bar{\mathbf{v}}_i \quad \cdots \quad \mathbf{M}_p \bar{\mathbf{v}}_i] \bar{\mathbf{q}}_{k,i} + \mathbf{C}^\top \bar{\boldsymbol{\ell}}_k \right) = \mathbf{w}_k,$$

since $\bar{\ell}_k = \ell_k$ in this case by (6.44). For indices k such that $\text{Im}(\sigma_k) \neq 0$, it holds that

$$\begin{aligned}\bar{\mathbf{w}}_k &= (\bar{\sigma}_k \mathbf{E}^\top - \mathbf{A}^\top)^{-1} \left(2 \sum_{i=1}^r [\mathbf{M}_1 \bar{\mathbf{v}}_i \quad \cdots \quad \mathbf{M}_p \bar{\mathbf{v}}_i] \bar{\mathbf{q}}_{k,i} + \mathbf{C}^\top \bar{\ell}_k \right) \\ &= (\sigma_{k+1} \mathbf{E}^\top - \mathbf{A}^\top)^{-1} \left(2 \sum_{i=1}^r [\mathbf{M}_1 \mathbf{v}_i \quad \cdots \quad \mathbf{M}_p \mathbf{v}_i] \mathbf{q}_{k+1,i} + \mathbf{C}^\top \ell_{k+1} \right) = \mathbf{w}_{k+1}.\end{aligned}$$

We have shown that the columns of $\mathbf{V}_p = [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_r]$ and $\mathbf{W}_p = [\mathbf{w}_1 \quad \cdots \quad \mathbf{w}_r]$ are closed under complex conjugation. This implies that $\text{Range}(\mathbf{V}) = \text{Range}(\mathbf{V}_p)$ and $\text{Range}(\mathbf{W}) = \text{Range}(\mathbf{W}_p)$ under the construction (6.45) and (6.46), thus completing the proof. \square

Lemma 6.12 shows how to construct real-valued interpolatory model reduction bases that satisfy the hypotheses of Theorem 6.11. This facilitates the computation of a real-valued interpolatory reduced model from a real-valued full-order model (5.1) that satisfies the interpolation conditions in (6.38).

The organizational structure imposed upon the interpolation data in Lemma 6.12 is meant to mimic that of the interpolation data computed during Algorithm 6.4.1, as well as the optimal data from Theorem 6.9. Consider a reduced model (5.2), the eigenvalues $\lambda_k \in \mathbb{C}$ and eigenvectors $\mathbf{t}_k, \mathbf{s}_k \in \mathbb{C}^r$ for $k = 1, \dots, r$ computed from the generalized eigendecomposition of $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{A}}$ are closed under complex conjugation since these matrices are real valued. Thus, the eigenvalues and eigenvectors can be organized into conjugate eigenpairs according to $\bar{\lambda}_k = \lambda_{k+1}$, $\bar{\mathbf{t}}_k = \mathbf{t}_{k+1}$, and $\bar{\mathbf{s}}_k = \mathbf{s}_{k+1}$. One can verify directly that the residue directions (5.43) used for the interpolatory projections throughout Algorithm 6.4.1 obey the organizational scheme laid out in (6.44).

Convergence monitoring, unstable poles, and initialization strategies.

The iteration in Algorithm 6.4.1 repeats until either the iteration count exceeds a maximum number of allowed steps $M \geq 1$, or the largest magnitude change in the reduced model poles between consecutive iterates falls below a user-specified tolerance $\epsilon > 0$. As already discussed for Algorithm 6.3.1, there exist many possibilities for monitoring convergence. We choose to use the change in the poles because this guarantees that the first-order optimality conditions in (6.32) will be satisfied if the iteration converges. In fact, this quantity is typically used to monitor convergence in the traditional IRKA iteration [97]. One benefit of Algorithm 6.4.1 compared to Algorithm 6.3.1 is that this criterion is numerically efficient, since the poles and residues of the current model iterate are solved via an $r \times r$ generalized eigenvalue problem. Moreover, these quantities need to be computed for the subsequent iteration, anyway. Because LQO-IRKA aims to solve the \mathcal{H}_2 minimization problem (6.1), one natural alternative is to monitor the system \mathcal{H}_2 error throughout the iteration as is done for Algorithm 6.3.1. However, as discussed in Remark 6.7, this would require one to solve a

large-scale Lyapunov equation on the way to computing the \mathcal{H}_2 error using the formulae in Corollary 6.2. While the convergence of LQO-IRKA is *not* guaranteed, in practice, IRKA and TSIA for linear problems consistently converge to local minima. We have observed the same behavior for LQO-IRKA, as illustrated in Section 6.6.

As with the original IRKA iteration, asymptotic stability is not guaranteed by Algorithm 6.4.1 but is typically maintained in practice. If an unstable intermediate model does appear, one can simply reflect the unstable pole across the imaginary axis to avoid interpolation at this point, and ensure the interpolatory first-order necessary conditions are satisfied upon convergence. In our experiments, we have never observed that LQO-IRKA converges to an unstable reduced model given a stable initialization.

The initialization of Algorithm 6.4.1 corresponds to an appropriate selection of complex interpolation points and tangential directions, and this selection will affect the quality of the final reduced model. However, as we illustrate in Section 6.6, LQO-IRKA is robust to different initialization strategies in practice. Because the optimal interpolation points are the mirror images of the reduced model poles, and one would expect these to lie in the numerical range of $\mathbf{E}^{-1}\mathbf{A}$, choosing r interpolation points in this region is usually an effective strategy. Alternatively, one could use a subset of the eigenvalues of the full-order system since this will eliminate the transfer function mismatch at the full-order poles in the \mathcal{H}_2 error expression (6.31). The boundaries of the numerical range can be computed via, e.g., iterative methods such as the Arnoldi iteration, which aim to find the extremal eigenpairs of a matrix. Other strategies for the initial IRKA iteration that transfer to our setting are discussed in [97, Sec. 4.2].

6.5 Comparing the two optimality frameworks

We have developed two independent optimality frameworks and a pair of iterative methods for the optimal- \mathcal{H}_2 approximation of LQO systems. The proofs of Theorems 6.5 and 6.9 are independent of one another. Thus, it is natural to wonder whether there is some connection between the Sylvester- and interpolation-based frameworks. Here, we show that the Wilson conditions in (6.23) directly imply the interpolatory conditions in (6.32) when the optimal- \mathcal{H}_2 approximation in question has simple poles. In other words: *the Wilson conditions are in fact interpolation conditions*. To prove this result, we introduce the following lemma that characterizes the interpolatory model reduction bases of Theorem 6.11 as solutions to a pair of generalized Sylvester equations.

Lemma 6.13. Suppose that \mathcal{G}_{lqo} is an asymptotically stable LQO system as in (5.1). Consider interpolation points $\sigma_1, \dots, \sigma_r \in \mathbb{C}$ such that $\sigma_i \mathbf{E} - \mathbf{A}$ are invertible for all $i = 1, \dots, r$, right-tangential directions $\mathbf{r}_1, \dots, \mathbf{r}_r \in \mathbb{C}^m$, and left-tangential directions $\mathbf{l}_1, \dots, \mathbf{l}_r \in \mathbb{C}^p$ and

$\mathbf{q}_{1,1}, \dots, \mathbf{q}_{r,r} \in \mathbb{C}^p$. Define the matrices of interpolation data

$$\begin{aligned}\mathbb{S} &\stackrel{\text{def}}{=} \text{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{C}^{r \times r}, \\ \mathbb{R} &\stackrel{\text{def}}{=} [\mathbf{r}_1 \quad \dots \quad \mathbf{r}_r] \in \mathbb{C}^{m \times r} \\ \mathbb{L} &\stackrel{\text{def}}{=} [\boldsymbol{\ell}_1 \quad \dots \quad \boldsymbol{\ell}_r] \in \mathbb{C}^{p \times r} \\ \mathbb{Q}_k &\stackrel{\text{def}}{=} [\mathbf{q}_{1,k} \quad \dots \quad \mathbf{q}_{r,k}] \in \mathbb{C}^{p \times r}, \quad k = 1, \dots, r.\end{aligned}\tag{6.47}$$

Then, the primitive interpolatory model reduction bases $\mathbf{V}_p \in \mathbb{C}^{n \times r}$ and $\mathbf{W}_p \in \mathbb{C}^{n \times r}$ defined as

$$\mathbf{V}_p = [\mathbf{v}_1 \quad \dots \quad \mathbf{v}_r] \quad \text{and} \quad \mathbf{W}_p = [\mathbf{w}_1 \quad \dots \quad \mathbf{w}_r],\tag{6.48}$$

where $\mathbf{v}_i, \mathbf{w}_i \in \mathbb{C}^{n \times r}$ are defined according to (6.36) and (6.37), uniquely satisfy the generalized Sylvester equations

$$\mathbf{A}\mathbf{V}_p - \mathbf{E}\mathbf{V}_p\mathbb{S} + \mathbf{B}\mathbb{R} = \mathbf{0}_{n \times r}\tag{6.49}$$

$$\text{and} \quad \mathbf{A}^\top \mathbf{W}_p - \mathbf{E}^\top \mathbf{W}_p \mathbb{S} - 2 \sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \dots \quad \mathbf{M}_p \mathbf{v}_k] \mathbb{Q}_k - \mathbf{C}^\top \mathbb{L} = \mathbf{0}_{n \times r}.\tag{6.50}$$

◇

Proof of Lemma 6.13. By the assumption that $\sigma_i \mathbf{E} - \mathbf{A}$ is nonsingular for all i , the spectra of the coefficient matrices in (6.49) and (6.50) are disjoint, and so the solutions to these equations are unique. We solve explicitly for the columns of the solutions to (6.49) and (6.50), and show that these are the columns of \mathbf{V}_p and \mathbf{W}_p . Observe that the i -th column of (6.49) can be written explicitly as

$$\mathbf{0}_n = \mathbf{A}\mathbf{V}_p \mathbf{e}_i - \mathbf{E}\mathbf{V}_p \mathbb{S} \mathbf{e}_i + \mathbf{B}\mathbb{R} \mathbf{e}_i = \mathbf{A}\mathbf{v}_i - \sigma_i \mathbf{E}\mathbf{v}_i + \mathbf{B}\mathbf{r}_i \quad \text{implies} \quad \mathbf{v}_i = (\sigma_i \mathbf{E} - \mathbf{A})^{-1} \mathbf{B}\mathbf{r}_i,$$

which proves that \mathbf{V}_p as in (6.48) is the solution to (6.49). The i -th column of (6.50) can be written explicitly as

$$\begin{aligned}\mathbf{0}_{n \times r} &= \mathbf{A}^\top \mathbf{W}_p \mathbf{e}_i - \mathbf{E}^\top \mathbf{W}_p \mathbb{S} \mathbf{e}_i - 2 \sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \dots \quad \mathbf{M}_p \mathbf{v}_k] \mathbb{Q}_k \mathbf{e}_i - \mathbf{C}^\top \mathbb{L} \mathbf{e}_i \\ &= \mathbf{A}^\top \mathbf{w}_i - \mathbf{E}^\top \mathbf{w}_i \sigma_i + 2 \sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \dots \quad \mathbf{M}_p \mathbf{v}_k] \mathbf{q}_{i,k} + \mathbf{C}^\top \boldsymbol{\ell}_i, \\ \text{implies} \quad \mathbf{w}_i &= (\sigma_i \mathbf{E}^\top - \mathbf{A}^\top)^{-1} \left(2 \sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \dots \quad \mathbf{M}_p \mathbf{v}_k] \mathbf{q}_{i,k} + \mathbf{C}^\top \boldsymbol{\ell}_i \right),\end{aligned}$$

and so \mathbf{W}_p as in (6.48) is the solution to (6.50). □

Consider a system $\tilde{\mathcal{G}}_{\text{lqo}}$ with simple poles $\lambda_1, \dots, \lambda_r$ and the realization (5.2). Let $\mathbf{S}, \mathbf{T} \in \mathbb{C}^{r \times r}$ contain the left and right generalized eigenvectors of $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{A}}$. If we consider the realization of (5.2) in the bases of \mathbf{S} and \mathbf{T} , the state-space matrices are transformed such that

$$\mathbf{T}^\top \tilde{\mathbf{E}} \mathbf{S} = \mathbf{I}_r, \quad \mathbf{T}^\top \tilde{\mathbf{A}} \mathbf{S} = \mathbf{\Lambda}, \quad \mathbf{T}^\top \tilde{\mathbf{B}} = \begin{bmatrix} \mathbf{b}_1^\top \\ \vdots \\ \mathbf{b}_r^\top \end{bmatrix}, \quad \tilde{\mathbf{C}} \mathbf{S} = [\mathbf{c}_1 \quad \cdots \quad \mathbf{c}_r], \quad (6.51)$$

$$\tilde{\mathbf{M}}(\mathbf{S} \otimes \mathbf{S}) = [\mathbf{m}_{1,1} \quad \cdots \quad \mathbf{m}_{1,r} \quad \cdots \quad \mathbf{m}_{r,1} \quad \cdots \quad \mathbf{m}_{r,r}],$$

with $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_r)$. Consider the equation for \mathbf{X} in (6.26). For the realization of (5.2) in the bases of \mathbf{S} and \mathbf{T} , this equation becomes

$$\begin{aligned} \mathbf{0} &= \mathbf{A} \mathbf{X} \left(\mathbf{S}^\top \tilde{\mathbf{E}}^\top \mathbf{T} \right) + \mathbf{E} \mathbf{X} \left(\mathbf{S}^\top \tilde{\mathbf{A}}^\top \mathbf{T} \right) + \mathbf{B} \left(\tilde{\mathbf{B}}^\top \mathbf{T} \right) \\ &= \mathbf{A} \mathbf{X} - \mathbf{E} \mathbf{X} \mathbf{\Lambda} + \mathbf{B} \left(\tilde{\mathbf{B}}^\top \mathbf{T} \right), \end{aligned}$$

which is just (6.49) for the interpolation data $\mathbb{S} = -\mathbf{\Lambda}$ and $\mathbb{R} = \mathbf{T}^\top \tilde{\mathbf{B}}$. Thus, $\mathbf{X} = \mathbf{V}_p$ in these bases by uniqueness of the solution to the generalized Sylvester equation. Similarly, some additional massaging of the equation for $2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}$ in (6.26) reveals that $2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}} = \mathbf{W}_p$ in an appropriate basis. To see this, observe for each $i = 1, \dots, r$ that

$$\begin{aligned} \sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \cdots \quad \mathbf{M}_p \mathbf{v}_k] \mathbf{m}_{i,k} &= \sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \cdots \quad \mathbf{M}_p \mathbf{v}_k] \begin{bmatrix} \mathbf{s}_k^\top \tilde{\mathbf{M}}_1 \mathbf{s}_i \\ \vdots \\ \mathbf{s}_k^\top \tilde{\mathbf{M}}_p \mathbf{s}_i \end{bmatrix} \\ &= \sum_{k=1}^r \left(\sum_{\ell=1}^p \left(\mathbf{s}_k^\top \tilde{\mathbf{M}}_\ell \mathbf{s}_i \right) \mathbf{M}_\ell \mathbf{v}_k \right) = \sum_{\ell=1}^p \mathbf{M}_\ell \sum_{k=1}^r \left(\mathbf{s}_k^\top \tilde{\mathbf{M}}_\ell \mathbf{s}_i \right) \mathbf{v}_k \end{aligned}$$

by reordering terms in the double summation. For each $\ell = 1, \dots, p$, we observe that

$$\mathbf{M}_\ell \sum_{k=1}^r \left(\mathbf{s}_k^\top \tilde{\mathbf{M}}_\ell \mathbf{s}_i \right) \mathbf{v}_k = \mathbf{M}_\ell [\mathbf{v}_1 \quad \cdots \quad \mathbf{v}_r] \begin{bmatrix} \mathbf{s}_1^\top \\ \vdots \\ \mathbf{s}_r^\top \end{bmatrix} \tilde{\mathbf{M}}_\ell \mathbf{s}_i = \mathbf{M}_\ell \mathbf{V} \left(\mathbf{S}^\top \tilde{\mathbf{M}}_\ell \mathbf{S} \mathbf{e}_i \right).$$

In aggregate, this implies

$$\sum_{k=1}^p [\mathbf{M}_1 \mathbf{v}_k \quad \cdots \quad \mathbf{M}_p \mathbf{v}_k] \mathbf{m}_{i,k} = \sum_{k=1}^p \mathbf{M}_k \mathbf{V}_p \left(\mathbf{S}^\top \tilde{\mathbf{M}}_k \mathbf{S} \mathbf{e}_i \right).$$

For the realization of (5.2) in the bases of \mathbf{S} and \mathbf{T} , recall that $\mathbf{X} = \mathbf{V}_p$. Using this, we observe that, in the bases of \mathbf{S} and \mathbf{T} , the i -th column in the equation for $\mathbf{Z}_{2\text{lqo}-\text{qo}} =$

$2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}$ in (6.26) becomes

$$\begin{aligned} \mathbf{0} &= \mathbf{A}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}} \left(\mathbf{T}^\top \tilde{\mathbf{E}} \mathbf{S} \right) \mathbf{e}_i + \mathbf{E}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}} \left(\mathbf{T}^\top \tilde{\mathbf{A}} \mathbf{S} \right) \mathbf{e}_i \\ &\quad - 2 \sum_{k=1}^p \mathbf{M}_k \mathbf{V}_p \left(\mathbf{S}^\top \tilde{\mathbf{M}}_k \mathbf{S} \right) \mathbf{e}_i - \mathbf{C}^\top \left(\tilde{\mathbf{C}} \mathbf{S} \mathbf{e}_i \right) \\ &= \mathbf{A}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}} \mathbf{e}_i + \mathbf{E}^\top \mathbf{Z}_{2\text{lqo}-\text{qo}} \lambda_i - 2 \sum_{k=1}^p \left[\mathbf{M}_1 \mathbf{v}_k \quad \cdots \quad \mathbf{M}_p \mathbf{v}_k \right] \mathbf{m}_{i,k} - \mathbf{C}^\top \mathbf{c}_i, \end{aligned}$$

which is the i -th column of (6.50) for the interpolation data $\mathbb{S} = -\mathbf{A}$, $\mathbb{L} = \tilde{\mathbf{C}} \mathbf{S}$, and $\mathbb{Q}_k = \tilde{\mathbf{M}} (\mathbf{S} \otimes \mathbf{S} \mathbf{e}_k)$. Thus, $\mathbf{Z}_{2\text{lqo}-\text{qo}} = 2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}} = \mathbf{W}_p$ in these bases. These observations yield the following result.

Corollary 6.14. Suppose that \mathcal{G}_{lqo} and $\tilde{\mathcal{G}}_{\text{lqo}}$ are asymptotically stable LQO systems as in (5.1) and (5.2), and that $\tilde{\mathcal{G}}_{\text{lqo}}$ has simple poles $\lambda_1, \dots, \lambda_r \in \mathbb{C}$. If $\tilde{\mathcal{G}}_{\text{lqo}}$ is computed by Petrov-Galerkin projection (5.24) using $\mathbf{V} = \mathbf{X}$ and $\mathbf{W} = 2\mathbf{Z}_{\text{lqo}} - \mathbf{Z}_{\text{lo}}$ in (6.26), then $\tilde{\mathcal{G}}_{\text{lqo}}$ satisfies the $3r + r^2$ interpolation conditions (6.32). \diamond

6.5.1 Comparing the two iterative algorithms

Corollary 6.14 yields further implications regarding LQO-TSIA (Algorithm 6.3.1) and LQO-IRKA (Algorithm 6.4.1). First, Corollary 6.14 shows that LQO-TSIA *performs interpolation at every step* (as long as the i -th reduced model iterate has simple poles). The interpolation points and tangential directions are the (mirrored) poles and residue directions of the previous model iterate. Moreover, if LQO-TSIA converges, the interpolation-based optimality conditions (6.32) will be satisfied. This implies that monitoring the change in the reduced-order model's poles can be used to monitor the convergence of LQO-TSIA, as well as LQO-IRKA.

Still, Algorithms 6.3.1 and 6.4.1 differ in one important way: the diagonalizability assumption. Theorem 6.9 is limited to \mathcal{H}_2 -optimal approximants with simple poles, while Theorem 6.5 does not have this limitation. In the linear \mathcal{H}_2 -optimal model reduction problem, it has been shown that the presence of higher-order poles can lead to numerical issues for IRKA, but not for TSIA; see [218]. Because the set of matrices that are not diagonalizable constitutes a set of measure zero, this problem is rarely encountered in practice. As we illustrate in Section 6.6, both LQO-IRKA and LQO-TSIA perform quite similarly.

6.6 Numerical examples

We consider here the performance of the proposed LQO-TSIA and LQO-IRKA on two benchmark problems. For the first example, we apply both LQO-TSIA and LQO-IRKA. Based on

Corollary 6.14, we expect these methods to produce nearly identical approximations. For the second example, we only apply LQO-IRKA in light of Corollary 6.14.

6.6.1 1D advection-diffusion equation

The first benchmark we consider is the 1D advection-diffusion PDE discussed in Section 5.2.3. Recall that governing equations for this PDE are

$$\begin{aligned} \frac{\partial}{\partial t}v(t, x) - \alpha \frac{\partial^2}{\partial x^2}v(t, x) + \beta \frac{\partial}{\partial x}v(t, x) &= 0, \\ v(t, 0) = u_0(t), \quad \alpha \frac{\partial}{\partial x}v(t, 1) = u_1(t), \quad v(0, x) &= 0, \end{aligned}$$

for $x \in (0, 1)$ and $t \in (0, T)$ and inputs $u_0, u_1 \in \mathcal{L}_2(0, T)$; the diffusion and advection coefficients are $\alpha > 0$ and $\beta \geq 0$, respectively. The output that we consider is

$$C(x, t) = \frac{1}{2} \int_0^1 |v(t, x) - 1|^2 dx.$$

Discretizing the equations in (5.6) using $n + 1$ equidistant spatial points yields an order- n LQO system (5.1) with $m = 2$ inputs (u_0 and u_1) and $p = 1$ output. Let $\mathbf{x}(t) \in \mathbb{R}^n$ denote the spatial discretization of $v(t, x)$, $h = 1/n$, and $\mathbf{1}_s \in \mathbb{R}^s$ the vector consisting of all ones. Then, the discretization provides an approximation to the quadratic cost function:

$$C(x, t) \approx \frac{h}{2} \|\mathbf{x}(t) - \mathbf{1}\|_2^2 = -h \mathbf{1}_n^\top \mathbf{x}(t) + \frac{h}{2} \text{vec}(\mathbf{I}_n)^\top (\mathbf{x}(t) \otimes \mathbf{x}(t)) + \frac{h}{2} \|\mathbf{1}_n\|_2^2 = y(t) + \frac{h}{2} \|\mathbf{1}_n\|_2^2.$$

Then, $\mathbf{C} = -h \mathbf{1}_n^\top \in \mathbb{R}^{1 \times n}$ and $\mathbf{M} = \frac{h}{2} \text{vec}(\mathbf{I}_n)^\top \in \mathbb{R}^{1 \times n^2}$ in (5.1). We study this example to investigate how well LQO-TSIA and LQO-IRKA reduced models recover *time-domain* output trajectories. Given the bound (6.2), we expect LQO-TSIA and LQO-IRKA to perform well in this regard. To obtain an LQO system in state-space form (5.1) from (5.6), an upwind finite-difference discretization of (5.6) is performed using $n + 1 = 3001$ spatial grid points; the diffusion and advection parameters are selected as $\alpha = 1$ and $\beta = 1$, respectively. For this example, $\mathbf{E} = \mathbf{I}_n$ by construction.

Experimental setup.

For LQO-TSIA and LQO-IRKA, two different initialization strategies are tested to assess the iterations' robustness to different starting points. In the case of LQO-TSIA, the iteration is initialized using a reduced model computed via the interpolatory strategies below, whereas LQO-IRKA is initialized with the interpolation data itself.

`eigs` uses the (mirrored) poles and residue directions of an initial reduced model computed by Galerkin projection $\mathbf{V} = \mathbf{W}$, where \mathbf{V} is the orthonormalized basis of the r -dimensional invariant subspace of \mathbf{A} corresponding to the eigenvalues with smallest magnitude, which are obtained using MATLAB's `eigs` command with a tolerance of 10^{-10} and the 'smallestabs' input option.

`imag` takes the initial interpolation points to be r points of the form $\sigma_k = iz_k$, where z_k are $r/2$ logarithmically spaced points from 10^0 to 10^3 ; these points are closed under complex conjugation. The tangential directions are chosen to be the leading canonical basis vectors of dimension r .

We refer to the LQO-TSIA and LQO-IRKA iterations that use the initializations `eigs` and `imag` by $\text{LQO-TSIA}_{\text{eigs}}$, $\text{LQO-TSIA}_{\text{imag}}$, and $\text{LQO-IRKA}_{\text{eigs}}$, $\text{LQO-IRKA}_{\text{imag}}$. For comparison, we use two different benchmark approaches.

LQO-BT is the balanced truncation model reduction algorithm for LQO systems proposed in [28] and reviewed in Section 5.3.4.

$\text{interp}_{\text{oneStep}}$ computes a (one-step) interpolatory reduced model using $\mathbf{V} \in \mathbb{R}^{n \times r}$ and $\mathbf{W} \in \mathbb{R}^{n \times r}$ as in Lemma 6.12 with non-optimal interpolation data. Specifically, the data are chosen according to `eigs` and `imag`. We refer to $\text{interp}_{\text{oneStep}}$ with these selection strategies as $\text{interp}_{\text{oneStep,eigs}}$ and $\text{interp}_{\text{oneStep,imag}}$. In either case, $\text{interp}_{\text{oneStep}}$ produces a reduced model that satisfies all the interpolation conditions of Theorem 6.11.

We test the performance of the computed reduced-order models in recovering the full-order (time-domain) output \mathbf{y} for particular choices of inputs. Because the system is single-output, we write $\mathbf{y} = y$. The time-domain simulations are implemented using MATLAB's `ode15i` using a fixed step size. To visibly compare the performance of the reduced models, we plot the full- and reduced-order outputs, as well as their pointwise relative error given by

$$\text{relerr}(t_i) \stackrel{\text{def}}{=} \frac{|y(t_i) - \tilde{y}(t_i)|}{|y(t_i)|}, \quad t_i \in [t_{\min}, t_{\max}], \quad (6.52)$$

where $t_i \in [t_{\min}, t_{\max}]$ are the N linearly equidistant time steps in the simulation. To assess the worst-case performance of the reduced models over the simulation window, we use an approximation of the relative \mathcal{L}_∞ error:

$$\text{relerr}_{\mathcal{L}_\infty} \stackrel{\text{def}}{=} \max_{t_i \in [t_{\min}, t_{\max}]} \frac{|y(t_i) - \tilde{y}(t_i)|}{|y(t_i)|}. \quad (6.53)$$

To assess the average performance of the reduced models over the simulation window, we use an approximation of the relative \mathcal{L}_2 error:

$$\text{relerr}_{\mathcal{L}_2} \stackrel{\text{def}}{=} \left(\frac{\sum_{i=1}^N |y(t_i) - \tilde{y}(t_i)|^2}{\sum_{i=1}^N |y(t_i)|^2} \right)^{1/2}, \quad (6.54)$$

Table 6.1: Relative errors (6.53)–(6.55) for the order $r = 30$ reduced models. The smallest error for each metric is highlighted in **boldface**.

	LQO-IRKA _{eigs}	LQO-IRKA _{imag}	LQO-TSIA _{eigs}	LQO-TSIA _{imag}	LQO-BT	interp _{oneStep,eigs}	interp _{oneStep,imag}
relerr $_{\mathcal{L}_\infty}$ (u_{sinc})	6.4082e-5	6.4082e-5	6.4559e-5	5.9411e-4	2.4916e-4	5.5440e-2	2.5442e0
relerr $_{\mathcal{L}_\infty}$ (u_{exp})	5.8897e-6	5.8897e-6	4.6381e-6	8.7611e-6	1.7226e-4	1.6854e-2	1.8087e0
relerr $_{\mathcal{L}_2}$ (u_{sinc})	3.5553e-6	3.5553e-6	3.6045e-6	4.8389e-6	4.2695e-5	9.4232e-3	4.7825e-1
relerr $_{\mathcal{L}_2}$ (u_{exp})	5.4745e-7	5.4745e-7	5.7794e-7	1.0624e-6	6.4736e-5	4.5368e-3	2.0120e-1
relerr $_{\mathcal{H}_2}$	4.1927e-7	4.2585e-7	4.6977e-7	1.4430e-5	7.5169e-7	9.9902e-1	9.5336e-1

where N is the number of time steps in the simulation. We also score the reduced model performance using the relative \mathcal{H}_2 system error:

$$\text{relerr}_{\mathcal{H}_2} \stackrel{\text{def}}{=} \frac{\|\mathcal{G}_{\text{lqo}} - \tilde{\mathcal{G}}_{\text{lqo}}\|_{\mathcal{H}_2}}{\|\mathcal{G}\|_{\mathcal{H}_2}}. \quad (6.55)$$

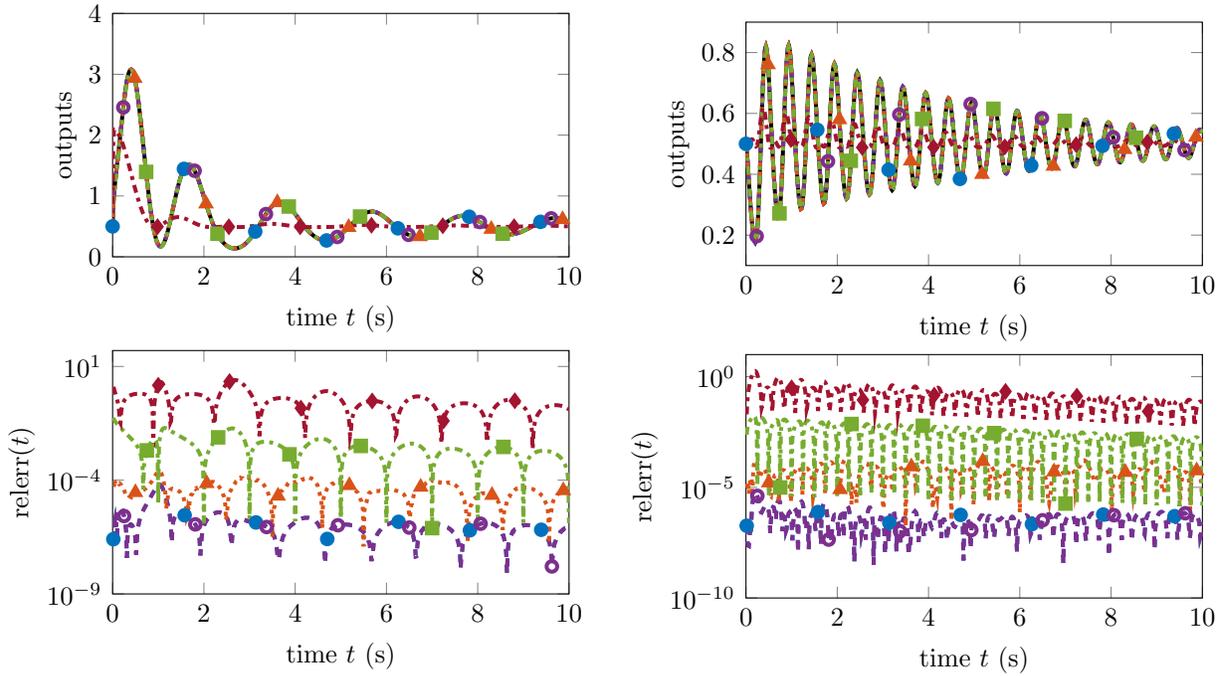
Discussion of results.

Seven reduced models of order $r = 30$ for the $n = 3000$ dimensional FOM are computed using LQO-IRKA and LQO-TSIA (each with the two initialization strategies **eigs** and **imag**), and LQO-BT, interp_{oneStep,eigs}, and interp_{oneStep,imag}. For the LQO-TSIA and LQO-IRKA iterations, the convergence tolerance is set to $\epsilon = 10^{-10}$ and the maximum number of allowed iterations is $M = 200$. The convergence tolerance is smaller in magnitude than one would typically use in practice; we choose this to investigate the long term convergence behavior of the iteration. Each iteration converged within the maximally allowed number of steps prescribed by M . The change in the reduced model poles is used to monitor the convergence of LQO-IRKA, while the change in the relative \mathcal{H}_2 errors (6.27) is used to monitor the convergence of LQO-TSIA. Although for LQO-IRKA we still compute the relative \mathcal{H}_2 error (6.55) throughout to investigate how (6.55) evolves throughout the iteration.

Time-domain simulations are performed using two different pairs of input signals; in either case, we enforce the Dirichlet boundary condition of $u_0(t) = v(t, 0) = 0$. The two different input signals used for u_1 are:

$$u_{\text{sinc}}(t) = 5 \frac{\sin(\pi t)}{\pi t} \quad \text{and} \quad u_{\text{exp}}(t) = e^{-t/5} \sin(4\pi t),$$

for $t \in [0, 10]$. Figure 6.1 plots the results of the LQO-IRKA reduced models against the benchmark approaches (LQO-BT and interp_{oneStep}) while Figure 6.2 plots results of the LQO-TSIA reduced models against the benchmark approaches. We plot the results of the LQO-IRKA and LQO-TSIA reduced models separately to avoid crowding the presentation, since these methods perform very similarly. In each figure, we plot full- and reduced-order outputs in response to u_{sinc} and u_{exp} , along with the associated relative pointwise approximation errors (6.52). The relative \mathcal{L}_∞ , \mathcal{L}_2 , and \mathcal{H}_2 errors according to (6.53), (6.54) and (6.55)



(a) Output magnitudes and pointwise relative errors (6.52) of the computed reduced-order models for input signals $u_0(t) = 0$ and $u_{\text{sinc}}(t) = \frac{5 \sin(\pi x)}{\pi x}$.

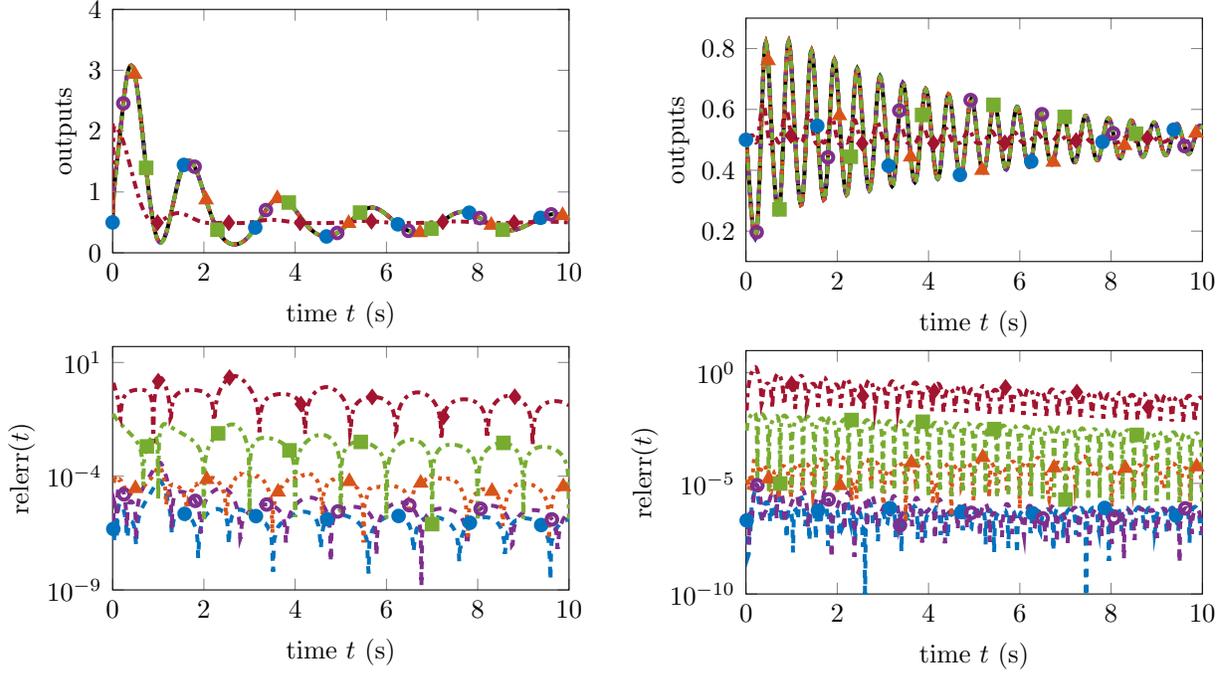
(b) Output magnitudes and pointwise relative errors (6.52) of the computed reduced-order models for input signals $u_0(t) = 0$ and $u_{\text{exp}}(t) = e^{-t/5} \sin(4\pi t)$.



Figure 6.1: Output magnitudes and pointwise relative errors (6.52) of the order $r = 30$ LQO-IRKA and benchmark reduced models in response to the input signals $u_0(t) = 0$, and $u_{\text{sinc}}, u_{\text{exp}}$.

induced by the order $r = 30$ reduced models are reported in Table 6.1. We observe that the LQO-IRKA, LQO-TSIA, and LQO-BT reduced models all produce very satisfactory reconstructions of the full-order output. While $\text{interp}_{\text{OneStep,eigs}}$ offers a reasonable approximation, $\text{interp}_{\text{OneStep,imag}}$ misses the output entirely. The \mathcal{H}_2 -optimal methods produce approximations that are a few orders of magnitude better than the benchmark approaches; this is also further supported by the relative output errors reported in Table 6.1. The reduced models computed by LQO-IRKA_{eigs}, LQO-IRKA_{imag}, and LQO-TSIA_{eigs} produce nearly indistinguishable approximations, while the LQO-TSIA_{imag} reduced model performs marginally worse in all metrics. This is likely due to the iteration converging prematurely for this initialization; we discuss this convergence behavior next.

Figure 6.3 plots the change in the relative \mathcal{H}_2 errors throughout the LQO-IRKA and LQO-



(a) Output magnitudes and pointwise relative errors (6.52) of the computed reduced-order models for input signals $u_0(t) = 0$ and $u_1(t) = u_{\text{sinc}}(t) = 5 \frac{\sin(\pi x)}{\pi x}$.

(b) Output magnitudes and pointwise relative errors (6.52) of the computed reduced-order models for input signals $u_0(t) = 0$ and $u_1(t) = u_{\text{exp}}(t) = e^{-t/5} \sin(4\pi t)$.



Figure 6.2: Output magnitudes and pointwise relative errors (6.52) of the order $r = 30$ LQO-TSIA and benchmark reduced models in response to the input signals $u_0(t) = 0$, and $u_{\text{sinc}}, u_{\text{exp}}$.

TSIA iterations. Although we emphasize that for LQO-IRKA, the maximal change in the reduced model poles is used to determine convergence. We observe that for both initialization strategies, LQO-IRKA finds a local minimum after approximately 15 iterations. The algorithm continues to iterate until the change in the poles of the reduced-order model iterates falls below the convergence tolerance $\epsilon = 10^{-10}$. On the other hand, LQO-TSIA_{eigs} converges to the *same* local minimum after 28 iterations, whereas LQO-TSIA_{imag} converges prematurely after 16 iterations, resulting in a reduced model with a larger \mathcal{H}_2 error. Based on the relative \mathcal{H}_2 errors, we conclude that the LQO-IRKA_{eigs}, LQO-IRKA_{imag}, and LQO-TSIA_{eigs} iterations all converge to the same local minimum/reduced-order model. Following the trend of convergence, we would expect LQO-TSIA_{imag} to ultimately converge to this local minimum as well if the iteration continued. These results suggest that monitoring convergence via the poles is more reliable, but may lead to unnecessary iterations even after a minimizer of the

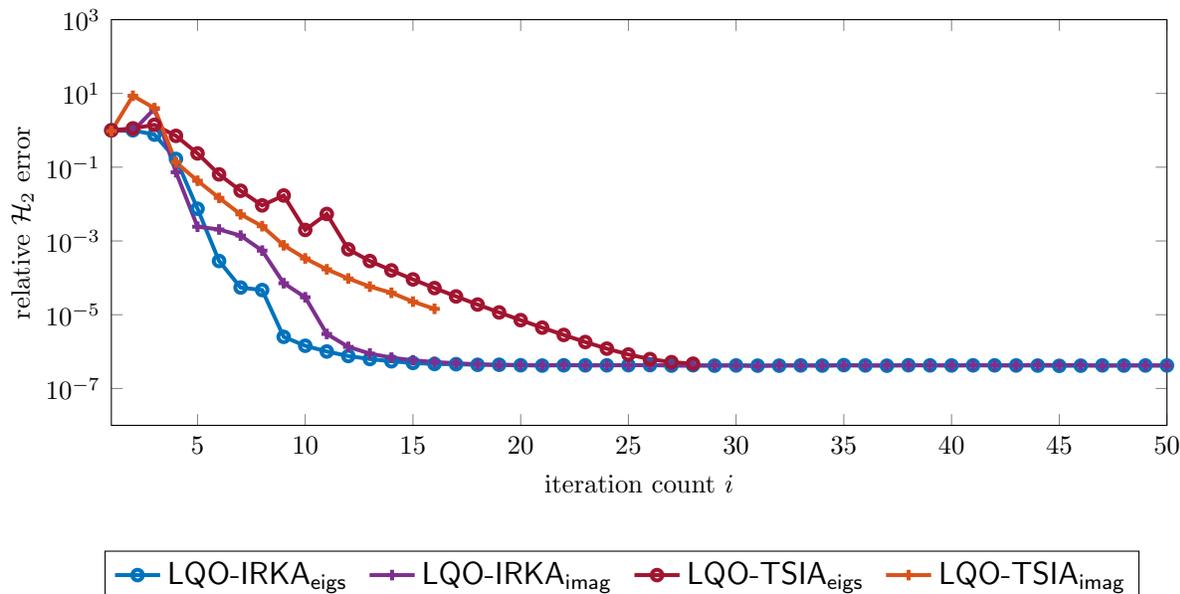


Figure 6.3: Relative \mathcal{H}_2 errors of the intermediate reduced models produced during the first 50 iterations of LQO-IRKA and LQO-TSIA.

\mathcal{H}_2 error is found. Whereas monitoring convergence using (6.27) may not be as robust and may cause the iteration to terminate early, it will not result in unnecessary computations. Interestingly, both the LQO-IRKA_{eigs} and LQO-IRKA_{imag} iterations seem to converge (with respect to (6.55)) in roughly half the number of steps compared to the LQO-TSIA_{eigs} iteration. In any case, LQO-IRKA and LQO-TSIA reduce the relative \mathcal{H}_2 model error by up to *six orders of magnitude* from the initial reduced model.

As a final experiment, we compute hierarchies of reduced models for orders $r = 2, 4, \dots, 30$ using LQO-IRKA, LQO-TSIA, and LQO-BT. We compute the relative \mathcal{H}_2 errors due to these approximations and plot them with respect to the increasing order r in Figure 6.4. The same experiment was performed for `interpOneStep`, and the computed reduced models all produced large relative \mathcal{H}_2 errors. We do not report these results here. Outside of LQO-TSIA_{imag}, for each method, the \mathcal{H}_2 error steadily decreases as the approximation order increases. For each order of reduction, the LQO-IRKA and LQO-TSIA_{eigs} reduced models produce the smallest errors, and are indistinguishable in the relative \mathcal{H}_2 error. The lines corresponding to these methods in Figure 6.4 are directly on top of each other. The LQO-TSIA_{imag} approach performs worse for orders $r = 16$ and higher, although this is likely due to the premature convergence observed in Figure 6.3. The LQO-BT reduced models also produce very small relative \mathcal{H}_2 errors, and are competitive with the \mathcal{H}_2 -optimal methods.

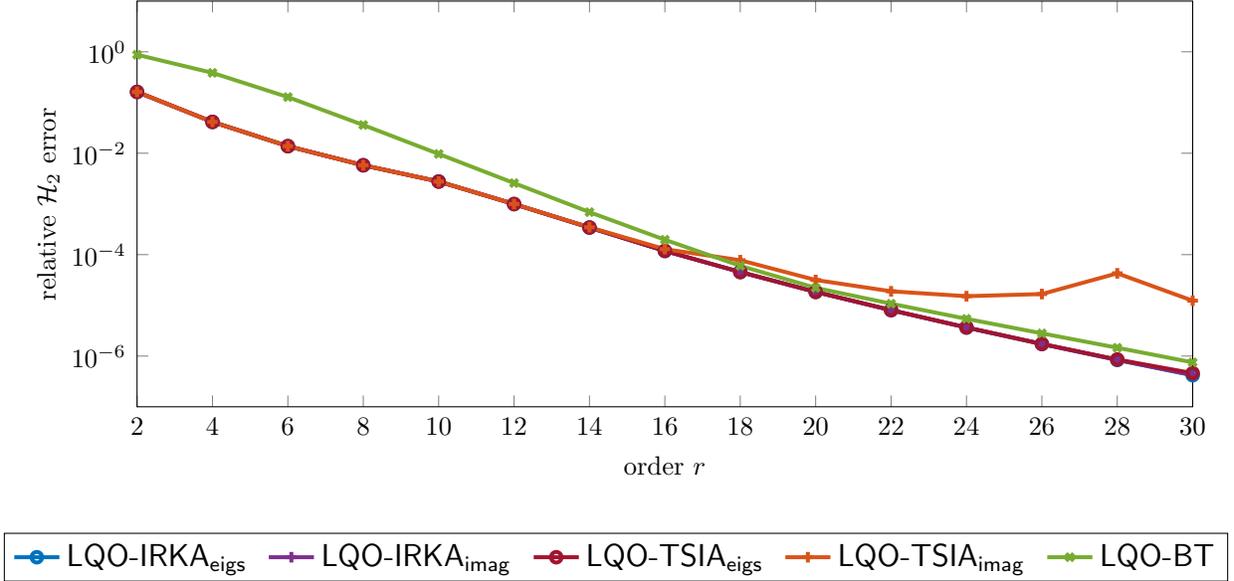


Figure 6.4: Relative \mathcal{H}_2 errors (6.55) due to the hierarchy of reduced models for orders $r = 2, 4, \dots, 30$.

6.6.2 Vibration of a plate with tuned vibration absorbers

The second benchmark we consider is the plate with TVAs discussed in Section 5.2.1. The dimension of this model in the first-order companion form (4.75) is $n = 403, 800$. The quantity of interest is the root mean squared displacement (5.3) with the reference state set to $\tilde{\mathbf{x}} = \mathbf{0}_n$ so that $\mathbf{G}_{\text{lo}} = \mathbf{0}_{p \times m}$. Thus, for $s \in i\mathbb{R}$ the root mean squared displacement in the frequency domain is specified by the system's quadratic-output transfer function (5.12b) with $s_1 = s$ and $s_2 = -s$. For this example, only LQO-IRKA is applied. We study this example to investigate how LQO-IRKA performs on a large-scale benchmark, as well as LQO-IRKA's performance in the frequency domain.

Experimental setup.

We assume a similar setup to Section 5.2.3. Due to the size of the benchmark, we only consider LQO-IRKA using the `imag` initialization. For this setting, `imag` takes r points of the form $\sigma_k = i z_k$, where z_k are $r/2$ linearly spaced points from 0 to 250 Hz; these points are closed under complex conjugation. The tangential directions are again chosen to be the leading canonical basis vectors of dimension r . For comparison, we only use the `interpOneStep,imag` method. The LQO-BT method is not feasible for this benchmark, given that it requires solving two n -dimensional Lyapunov equations (2.43) and (5.20). Additional results comparing LQO-IRKA to more sophisticated interpolatory model reduction strategies are published

Table 6.2: Relative \mathcal{H}_∞ errors (6.57) and \mathcal{H}_2 errors (6.58) for the order $r = 20$ and $r = 50$ reduced models of the plate with TVAs. The smallest error is highlighted in **boldface**.

	LQO-IRKA _{imag}	interp _{oneStep,imag}
relerr $_{\mathcal{H}_\infty}$ ($r = 20$)	2.9844e-1	3.4364e-1
relerr $_{\mathcal{H}_2}$ ($r = 20$)	1.2194e-1	9.2182e-2
relerr $_{\mathcal{H}_\infty}$ ($r = 50$)	1.6725e-1	3.1707e-1
relerr $_{\mathcal{H}_2}$ ($r = 50$)	3.5396e-2	7.8588e-2

in [188]. The LQO-IRKA reduced models presented therein were computed *without* using the realification strategy of Lemma 6.12. Because of this, LQO-IRKA did not converge within the prescribed number of iterations, resulting in lower-quality approximations.

We test the performance of the computed reduced-order models in recovering the quadratic-output frequency response \mathbf{G}_{qo} that models (5.3). To visibly compare the performance of the reduced models, we plot the full- and reduced-order transfer function response, as well as their pointwise relative error given by

$$\text{relerr}(\dot{i}\omega_i) \stackrel{\text{def}}{=} \frac{|\mathbf{G}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i) - \tilde{\mathbf{G}}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i)|}{|\mathbf{G}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i)|}, \quad \omega_i \in \Omega, \quad (6.56)$$

where Ω is a collection of 500 equispaced points in the range of [1, 251] Hz. The complex modulus is used because the system is single-input, single-output. To assess the worst-case performance of the reduced models over the frequency range of interest, we use an approximation of the relative \mathcal{H}_∞ error:

$$\text{relerr}_{\mathcal{H}_\infty} \stackrel{\text{def}}{=} \frac{\max_{\omega_i \in \Omega} |\mathbf{G}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i) - \tilde{\mathbf{G}}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i)|}{\max_{\omega_i \in \Omega} |\mathbf{G}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i)|}. \quad (6.57)$$

To assess the average performance of the reduced models over the frequency range of interest, we use an approximation of the relative \mathcal{H}_2 error:

$$\text{relerr}_{\mathcal{H}_2} \stackrel{\text{def}}{=} \left(\frac{\sum_{i=1}^N |\mathbf{G}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i) - \tilde{\mathbf{G}}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i)|^2}{\sum_{i=1}^N |\mathbf{G}_{\text{qo}}(\dot{i}\omega_i, -\dot{i}\omega_i)|^2} \right)^{1/2}. \quad (6.58)$$

for $\omega_i \in \Omega$ as defined above.

Discussion of results.

Reduced-order models of orders $r = 20$ and $r = 50$ are computed using LQO-IRKA_{imag} and interp_{oneStep,imag}. For the LQO-IRKA iterations, we use the convergence parameters $\epsilon = 10^{-6}$ and $M = 200$. The iteration converged within the prescribed number of M steps for each

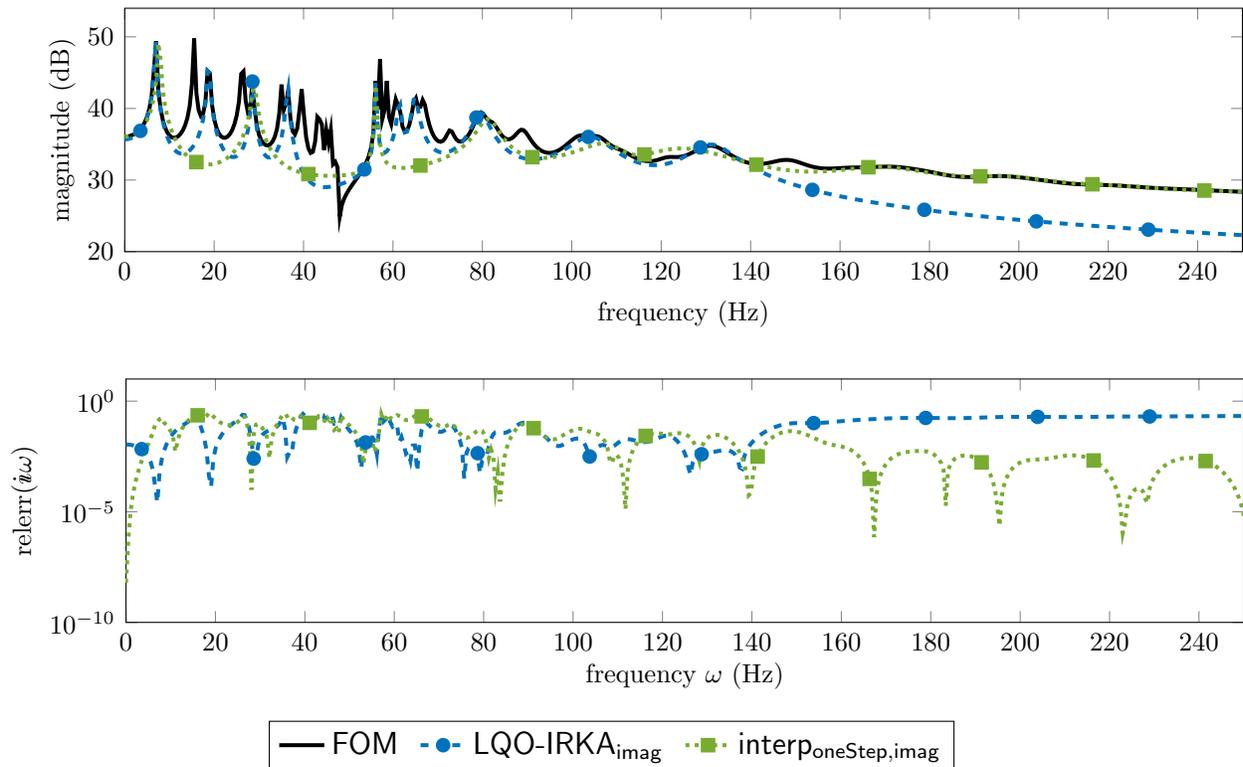


Figure 6.5: Frequency response magnitude and pointwise relative errors (6.56) for the order $r = 20$ reduced models of the plate with TVAs.

order r . The frequency response of the full- and reduced-order quadratic-output transfer functions, as well as the pointwise relative errors (6.56) of the order $r = 20$ and $r = 50$ reduced models are plotted in Figures 6.5 and 6.6. The frequency response is evaluated by taking the modulus of the full- and reduced-order quadratic-output transfer functions (5.12b) evaluated at $s_1 = i\omega_i$ and $s_2 = -si\omega_i$, for 500 equispaced points $i\omega_i \in [1, 251]$ Hz. The (approximate) relative \mathcal{H}_∞ and \mathcal{H}_2 errors (6.57) and (6.58) are reported in Table 6.2. Quantitatively, Table 6.2 suggests that both methods perform similarly. However, Figures 6.5 and 6.6 illustrate that the LQO-IRKA_{imag} reduced models do a much better job of capturing the interesting response behavior of the full-order transfer function in the low frequency range. Both order $r = 20$ approximations computed by LQO-IRKA_{imag} and interp_{oneStep,imag} are insufficient to completely realize the response behavior of the full-order transfer function. The LQO-IRKA_{imag} reduced model captures more peaks in the low frequency range, whereas the interp_{oneStep,imag} provides a better approximation at higher frequencies. The poor approximation of LQO-IRKA_{imag} from 140 Hz onward is what causes the lower relative \mathcal{H}_2 error exhibited by interp_{oneStep,imag} for this order of reduction. For the order $r = 50$ approximations, similar behavior is observed. In this case, the LQO-IRKA_{imag} reduced model captures almost all of the peaks of the full-order frequency response, except for the dip around 48 Hz (which

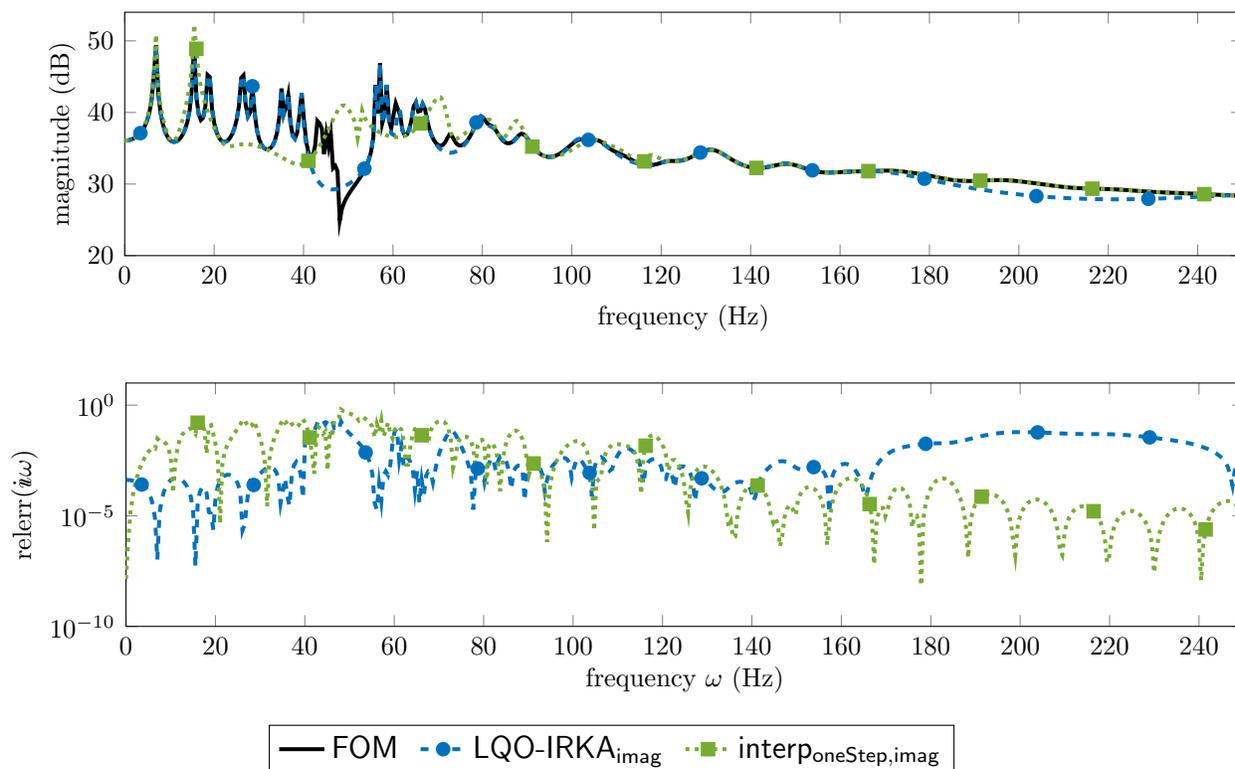


Figure 6.6: Frequency response magnitude and pointwise relative errors (6.56) for the order $r = 50$ reduced models of the plate with TVAs.

is difficult to match due to the use of TVAs to damp this frequency). The `interponeStep,imag` reduced model again matches the response behavior at high frequencies, but misses several response peaks in the range of 20 to 60 Hz.

6.7 Conclusions

In this chapter, the \mathcal{H}_2 -optimal approximation problem (6.1) for linear quadratic-output systems (5.1) is studied. Two novel \mathcal{H}_2 -optimality frameworks based on distinct sets of first-order necessary conditions for optimality were developed. The first is based on the solutions to generalized Sylvester equations (6.26) and the linear quadratic-output system Gramians, whereas the second is based on the multivariate rational interpolation of the linear- and quadratic-output transfer functions (5.12). In each case, it is shown how to enforce the necessary optimality conditions using a Petrov-Galerkin projection. It is also proven that the Sylvester equation-based framework enforces the interpolatory optimality conditions when the reduced model has simple poles. These results establish the Wilson and Meier-Leunberger \mathcal{H}_2 -optimality frameworks for the linear quadratic-output setting. Two

iterative algorithms based on repeated projection are proposed for computing \mathcal{H}_2 -optimal reduced models. The effectiveness of the proposed methods was investigated using two benchmark examples from the literature.

Chapter 7

Interpolatory Matrix Factorizations for Phasor Measurement Unit Data

7.1 Introduction

This chapter considers a separate problem in the data-driven modeling of dynamical systems and adopts a fresh set of notation. Unlike the content of Chapter 4, where the data are transfer function evaluations of linear dynamical systems, the data we consider in this chapter come from full or partial state observations. Specifically, we explore the use of low-rank *interpolatory* matrix decompositions (IMDs) [63, 126, 138, 206] and variants of the *discrete empirical interpolation method* (DEIM) [17, 52, 64, 165] for the sparse reconstruction of high-dimensional data sets. Our application of interest is the use of these approximations for reducing the scale of *Phasor Measurement Unit* (PMU) data used to monitor electrical power networks, as well as detect and localize disturbances, e.g., power line trips or outages.

7.1.1 Background and motivation

Electrical power networks are physical infrastructure that are particularly susceptible to low-probability, high-impact events [70]. Gone unchecked, local disturbances can cascade into wide-area, regional outages. For instance, the 2003 Northeastern United States blackout began as a series of local outages when high voltage power lines came into contact with overgrown tree branches [155], and system operators were unable to initiate corrective measures due to the lack of adequate notification systems. Simultaneously, the past two decades have seen the widespread incorporation of PMUs into wide-area monitoring systems (WAMS). PMUs are *in situ* sensor devices that provide global positioning system (GPS)-synchronized phasor readings of grid quantities such as nodal voltages, nodal currents, line currents, and their time derivatives, at a rate of 60–120 samples per second. These (streaming) real-time measurements offer an accurate reflection of the network’s current operating condition, although data accumulation presents a significant roadblock to real-time operational benefits. As a simple example, a network consisting of 100 PMUs each with a sampling rate of 120 Hz generates 200 gigabytes of data *per day* [79, 118]. Moreover, a significant amount of communication bandwidth is required to transmit these data from local PMU substations to regional control centers [57, 58, 59]. Thus, there is a need for the development of fast and

reliable *data-driven* methods for wide-area monitoring, so that system operators can monitor network performance, detect disturbances, and initiate corrective measures in real time.

This discussion motivates our investigation of two related research questions:

1. How can one effectively manage, analyze, and reduce the scale of large amounts of high-dimensional, streaming PMU data?
2. Can this reduction be reliably implemented in real time to enable the development of *data-driven* methods for wide-area monitoring, event detection, and localization?

It is well-documented in theory and industry practice that matrices of PMU data exhibit an underlying (approximate) low-rank structure under normal operating conditions and even when significant deviations from equilibrium occur; see, e.g., [56, 125, 222, 224, 235] and the references therein. Data-driven methods exploiting such linear dependencies in PMU data matrices have been successfully applied to a range of grid monitoring tasks, such as detection and localization of disturbance events [39, 119, 125, 130, 132, 180, 190, 225, 235, 246], recovery of missing and corrupted data [82, 102, 103, 104], and coherency identification [3, 129]. We refer to Wang et al. [224] for a recent survey of low-rank methods in power systems analysis.

Low-rank representations of PMU data are typically computed using methods that decompose the data into orthogonal components, e.g., the Singular Value Decomposition (SVD) [86, Sect. 2.4] or the closely related Principal Component Analysis (PCA) [113]. These methods provide *optimal* low-rank approximations to the full dataset by blending information from *all* of its rows and columns. However, this requires the transmission of large amounts of PMU data across communication networks before dimensionality reduction can be applied at a central location (such as a regional control center). Thus, these methods are often ill-suited for time-sensitive and bandwidth-limited applications. Moreover, certain applications in WAM do not explicitly seek an optimal reconstruction of the data, but rather aim to reveal a small subset of rows and columns that reveal the low-dimensional structure of a PMU data matrix, and correspond to points of interest in the network's operating history.

7.1.2 Chapter contents

As an alternative to PCA or the SVD, in this chapter we propose using the framework of *interpolatory matrix decompositions* (IMDs) [63, 126, 138, 206] and the *discrete empirical interpolation method* (DEIM) [17, 52, 64, 165] for the real-time dimensionality reduction (compression) of PMU data to enable fast and reliable methods for wide-area monitoring. In contrast to PCA or the SVD, IMDs reconstruct the entire matrix using only the information contained in a few columns and rows; the remaining features are approximated as linear combinations of the interpolative components. While necessarily suboptimal with respect to approximation quality, the more flexible IMDs possess several useful qualities that are

beneficial for the analysis of large PMU data matrices, and streaming data in general; see our discussion in Section 7.2.2. Moreover, this dimensionality reduction is achievable *in real time*. After reviewing the basics of IMDs in the context of PMU data reduction, in Section 7.2 we elaborate on an idea proposed by Xie et al. [235], and describe how these low-rank factorizations can be used for the sparse reconstruction of streaming PMU data. Specifically, we show how one can recover measurements by interacting only with a significantly reduced number of *pilot PMUs* [235], or collecting fewer down-sampled time snapshots [130]. Drawing upon the numerical linear algebra literature, framing these sparse reconstructions in the mathematical framework of IMDs enables us to state a rigorous, computable error bound on the interpolatory reconstruction error. This bound can be used to certify whether the chosen pilots or time instances truly capture the low rank character of the data, and leveraged towards various operational benefits that we describe in Sections 7.2.2 and 7.4. The success of the interpolatory approximation hinges on selecting the correct few rows or columns to use as the basis for the reconstruction, and there exists a variety of methods for this selection; see, e.g., [63, 126, 138, 206]. In contrast to the PCA- or energy-based selection approaches of [125, 225, 235], in Section 7.3 we propose using the *discrete empirical interpolation method* (DEIM) [17, 52] and its Q-DEIM variant [64] for selecting the rows or columns of a PMU data matrix, which form the basis of an interpolatory compression. These are *greedy* algorithms designed to minimize the computable interpolatory error bound. Based on the interpolatory approximations of Section 7.2 and the discrete empirical interpolation method of Section 7.3, in Section 7.4 we describe a joint IMD-DEIM-based framework for the data-driven and real-time monitoring of electrical power networks using a reduced number of pilot PMUs. The proposed framework builds upon the work of Xie et al. [235]; the key differences are that using the technology of IMDs provide us with an effective error estimator during online operations, and the DEIM algorithm provides a robust, adaptive method for pilot PMU selection and event localization. Numerical experiments are included throughout to validate the proposed methods using synthetically generated PMU data.

7.2 Low-rank matrix factorizations of Phasor Measurement Unit data

Briefly, we describe the problem setting and our assumptions on the collected data. We also introduce the interpolatory matrix decompositions and associated theory that are the focus of this chapter. Throughout, we use the following notation to index a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$: the (i, j) -th entry of \mathbf{X} is denoted $\mathbf{X}_{i,j} \in \mathbb{R}$; the i -th row of \mathbf{X} is denoted $\mathbf{X}_{i,:} \in \mathbb{R}^{1 \times n_2}$; the j -th column of \mathbf{X} is denoted $\mathbf{X}_{:,j} \in \mathbb{R}^{n_1}$ and occasionally $\mathbf{x}_j \in \mathbb{R}^{n_1}$. For a collection of

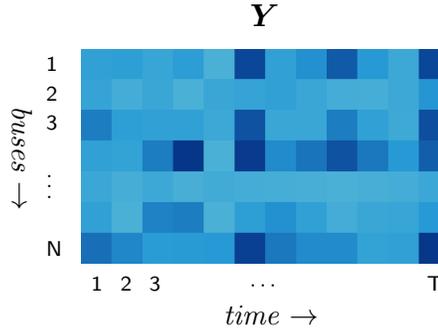


Figure 7.1: PMU data organized in an $N \times T$ matrix; each of the N rows corresponds to a bus; each of the T columns is a snapshot of the system in time.

indices $\mathbf{k} = \{k_1, \dots, k_m\}$, we define the notation

$$\mathbf{X}_{:, \mathbf{k}} \stackrel{\text{def}}{=} [\mathbf{X}_{:, k_1} \quad \dots \quad \mathbf{X}_{:, k_m}] \in \mathbb{R}^{n_1 \times m} \quad \text{and} \quad \mathbf{X}_{\mathbf{k}, :} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{X}_{k_1 :} \\ \vdots \\ \mathbf{X}_{k_m :} \end{bmatrix} \in \mathbb{R}^{m \times n_2},$$

to refer to the columns and rows of \mathbf{X} indexed by \mathbf{k} , respectively.

7.2.1 Basic setup and the Singular Value Decomposition

Suppose that a system operator collects data from $N \in \mathbb{Z}_{>0}$ PMUs. To simplify the exposition, we assume that each measured bus in the network is instrumented with a single PMU, and each PMU records a single grid quantity, such as a nodal voltage magnitude. Thus, we henceforth use the terms PMU and bus interchangeably when referring to a location in the network. The time-series data are collected into a (rectangular) matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$ as illustrated by Figure 7.1; each of the N rows contains a time series collected from a *single* PMU datastream, while each of the T columns contains a single snapshot in time of measurements collected from *every* PMU datastream. The assumption that \mathbf{Y} is rectangular is not necessary, but is used because typically $T \geq N$ for the problems we consider.

The underlying dimensionality of PMU data (and compression of that data) has been considered from a variety of perspectives; see, e.g. [56, 57, 58, 59, 83, 119, 222, 223, 224, 235]. It has been well-documented in theory and industry practice that matrices of PMU data exhibit an underlying low-rank structure; this phenomenon holds regardless of whether the data are collected during ambient or irregular operating conditions. As an immediate consequence, the dimension of \mathbf{Y} can be reduced by retaining only its dominant components computed via PCA [113] or the SVD [86, Sect. 2.4]. The fewer low-rank factors can be stored more efficiently, and expedite any subsequent computations and analysis involving the reconstructed data. Briefly, we recall how a matrix of PMU data can be approximated via its truncated SVD according to Theorems 2.2 and 2.3.

Consider a matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$ with $\text{rank}(\mathbf{Y}) = R \leq \min\{N, T\}$. The SVD of \mathbf{Y} has the dyadic form

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{k=1}^R \sigma_k \mathbf{u}_k \mathbf{v}_k^\top, \quad (7.1)$$

where $\mathbf{U} \in \mathbb{R}^{N \times R}$ and $\mathbf{V} \in \mathbb{R}^{T \times R}$ have orthonormal columns, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_R) \in \mathbb{R}^{R \times R}$ carries the singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R > 0$ of \mathbf{Y} . From Theorem 2.3, for any $1 \leq K < R$ the best rank- K approximation to \mathbf{Y} is given by

$$\mathbf{Y}_K \stackrel{\text{def}}{=} \sum_{k=1}^K \sigma_k \mathbf{u}_k \mathbf{v}_k^\top, \quad (7.2)$$

which is obtained by truncating the trailing $R - K$ components of the full SVD in (7.1). Moreover, approximation error due to (7.2) is given by

$$\sigma_{K+1} = \|\mathbf{Y} - \mathbf{Y}_K\|_2 \quad \text{and} \quad \sum_{i=K+1}^R \sigma_i^2 = \|\mathbf{Y} - \mathbf{Y}_K\|_F^2.$$

Evidently, \mathbf{Y}_K approximates \mathbf{Y} well if the $R - K$ trailing singular values are sufficiently small. In practice, the rank K is selected to deliver a relative approximation error below a certain threshold $0 < \alpha < 1$; for example

$$\frac{\|\mathbf{Y}_K\|_F^2}{\|\mathbf{Y}\|_F^2} = \frac{\sum_{k=1}^K \sigma_k^2}{\sum_{k=1}^R \sigma_k^2} \geq \alpha. \quad (7.3)$$

This compression via the SVD is akin to keeping the K principal components of a matrix, as practiced in [235].

Remark 7.1. Strictly speaking, the data would first be prepared for PCA by subtracting the mean of each row from every entry in that row, replacing \mathbf{Y} with $\mathbf{Y} - \boldsymbol{\mu}\mathbf{1}^\top$, where $\boldsymbol{\mu} \in \mathbb{R}^N$ has as its entries $\mu_j = (y_{j,1} + \dots + y_{j,T})/T$, the mean of the j -th row of \mathbf{Y} , and $\mathbf{1} \in \mathbb{R}^T$ is the vector of all ones. As in [235], we do not do any such preprocessing of \mathbf{Y} , and thus take PCA to be synonymous with the SVD. \diamond

As we highlighted in the introduction, computing the SVD requires blending information from *all* PMUs (buses) at *all* times, which carries with it a significant communication cost. Thus, the SVD is not typically feasible for time-sensitive applications, and is better suited for offline tasks, such as post-event analysis.

7.2.2 Interpolatory approximations of PMU data

As an alternative to PCA and the SVD, we propose the framework of *interpolatory matrix approximations* (IMDs) [63, 138, 206, 207] for reducing the dimensionality of PMU data.

These are low-rank factorizations expressed explicitly in terms of fewer actual rows and columns of the original matrix.

Definition 7.2 (Interpolatory matrix decompositions). Consider a matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$. For $1 \leq K < \max\{N, T\}$, an *interpolatory matrix approximation* is a low-rank factorization $\mathbf{Y}_s^t \in \mathbb{R}^{N \times T}$ of the form

$$\mathbf{Y} \approx \mathbf{Y}_s^t \stackrel{\text{def}}{=} \mathbf{C}^t \mathbf{X}_s^t \mathbf{R}_s, \quad (7.4)$$

where $\mathbf{X}_s^t \in \mathbb{R}^{K \times K}$. The matrices $\mathbf{C}^t \in \mathbb{R}^{N \times K}$ and $\mathbf{R}_s \in \mathbb{R}^{K \times T}$, defined by

$$\mathbf{C}^t \stackrel{\text{def}}{=} \mathbf{Y}_{:,t} = [\mathbf{Y}_{:,t_1} \quad \mathbf{Y}_{:,t_2} \quad \cdots \quad \mathbf{Y}_{:,t_k}] \quad \text{and} \quad \mathbf{R}_s \stackrel{\text{def}}{=} \mathbf{Y}_{s,:} = \begin{bmatrix} \mathbf{Y}_{s_1,:} \\ \mathbf{Y}_{s_2,:} \\ \vdots \\ \mathbf{Y}_{s_K,:} \end{bmatrix}, \quad (7.5)$$

contain a subset of the *columns* and *rows* of \mathbf{Y} indexed by $\mathbf{t} = \{t_1, t_2, \dots, t_K\} \subset \{1, 2, \dots, T\}$ and $\mathbf{s} = \{s_1, s_2, \dots, s_K\} \subset \{1, 2, \dots, N\}$. \diamond

In this setting where \mathbf{Y} is a matrix of PMU data as in Figure 7.1, \mathbf{C}^t contains a few select *snapshots* of the network in time across *every bus*, whereas \mathbf{R}_s contains datastreams from a few select *buses* in the network at *all sampled times*. The interpolatory approximation in (7.4) aims to use only the information contained in the select columns \mathbf{C}^t and rows \mathbf{R}_s of \mathbf{Y} to recover the full data; the small matrix $\mathbf{X}_s^t \in \mathbb{R}^{K \times K}$ is chosen to ensure that (7.4) produces a satisfactory reconstruction of \mathbf{Y} . The success of the approximation thus hinges on selecting the fewer rows and columns in \mathbf{R}_s and \mathbf{C}^t to use as the bases for the reconstruction. There exist various strategies for column and row selection in the numerical linear algebra literature; see, e.g., [63, 126, 138, 206]. In Section 7.3, we introduce a greedy selection strategy from [17, 52, 206] for iteratively choosing the rows and columns used in (7.4).

One-sided interpolatory approximations.

The formulation in (7.4) is the most general; one may also consider interpolatory approximations that expose only the rows *or* columns of \mathbf{Y} . Utilizing information contained only in certain *rows* of \mathbf{Y} amounts to using data (collected from $K < N$ fewer PMUs or buses) contained in \mathbf{R}_s to approximate the data from all other PMUs. At the matrix level, this amounts to finding a matrix $\mathbf{Z}_s \in \mathbb{R}^{N \times K}$ such that

$$\mathbf{Y} \approx \mathbf{Y}_s \stackrel{\text{def}}{=} \mathbf{Z}_s \mathbf{R}_s \in \mathbb{R}^{N \times T}. \quad (7.6)$$

We refer to the buses (PMUs) indicated by \mathbf{s} as *pilot buses* or *pilot PMUs* interchangeably, adopting terminology from [235]. The i -th row of \mathbf{Z}_s contains the weights that specify how

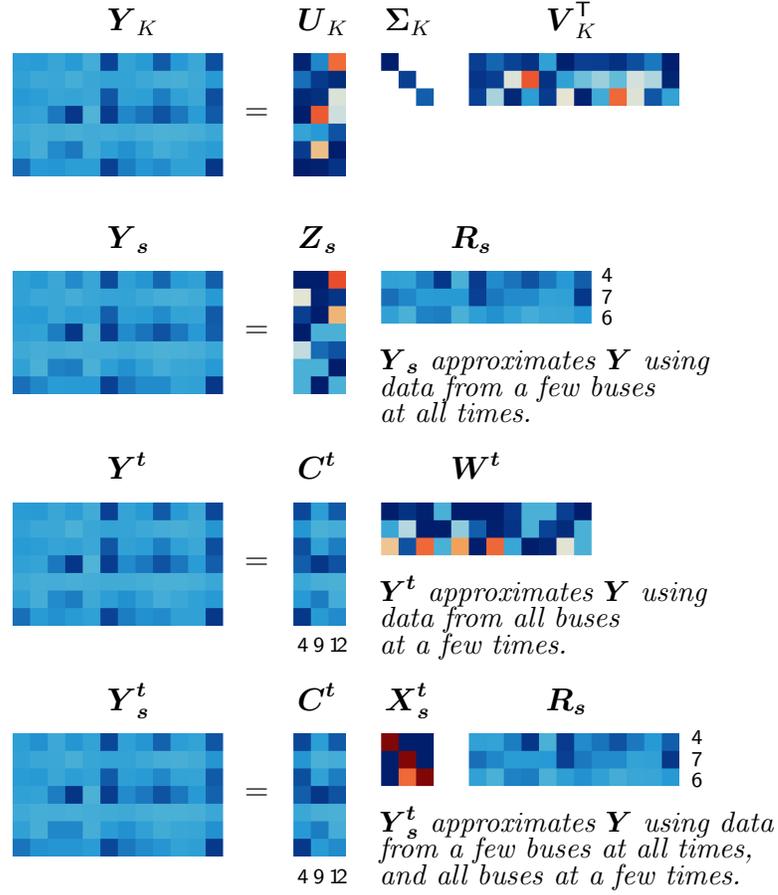


Figure 7.2: Visual illustration of low-rank approximations \mathbf{Y}_K , \mathbf{Y}_s , \mathbf{Y}^t , and \mathbf{Y}_s^t to the PMU data matrix \mathbf{Y} .

the data collected from the K pilots in \mathbf{R}_s should be combined to approximately recover the data $\mathbf{Y}_{i,:}$ at the i -th *non-pilot* bus, i.e.,

$$\mathbf{Y}_{i,:} \approx (\mathbf{Z}_s \mathbf{R}_s)_{i,:} = \sum_{k=1}^K (\mathbf{Z}_s)_{i,k} (\mathbf{R}_s)_{k,:}, \quad i \notin \mathbf{s}. \quad (7.7)$$

Using information contained only in certain *columns* of \mathbf{Y} amounts to using the $K \leq T$ *time snapshots* contained in \mathbf{C}^t to recover the full time series. This corresponds to finding a matrix $\mathbf{W}^t \in \mathbb{R}^{K \times T}$ such that

$$\mathbf{Y} \approx \mathbf{Y}^t \stackrel{\text{def}}{=} \mathbf{C}^t \mathbf{W}^t \in \mathbb{R}^{N \times T}. \quad (7.8)$$

As with the row-based approximations, the j -th column of \mathbf{W}^t contains the weights that describe how the selected time snapshots in \mathbf{C}^t should be combined to produce an approxi-

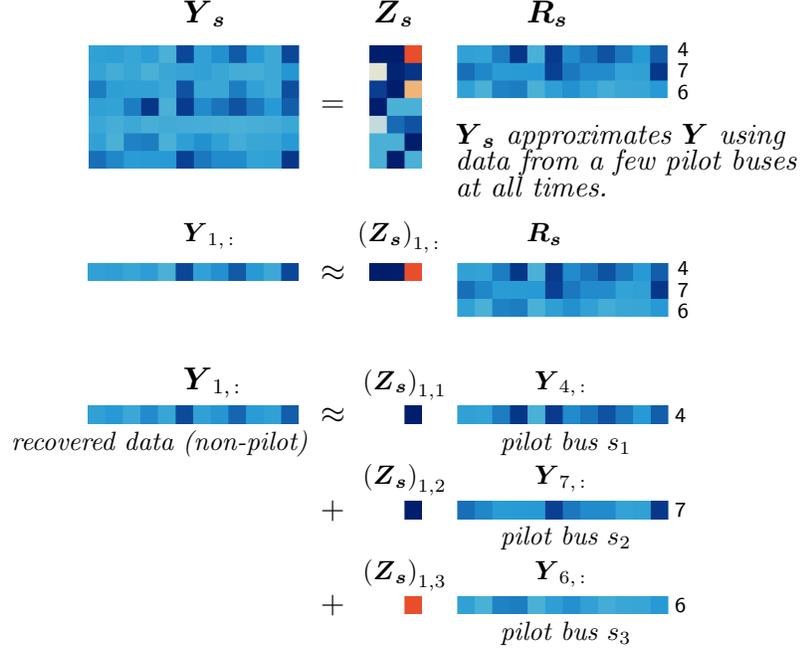


Figure 7.3: Visual illustration of the online pilot-based reconstruction of non-pilots described in Algorithm 7.2.1.

mation of the data $\mathbf{Y}_{:,j}$ at time t_j , i.e.,

$$\mathbf{Y}_{:,j} \approx (\mathbf{C}^t \mathbf{W}^t)_{:,j} = \sum_{k=1}^K (\mathbf{W}^t)_{j,k} (\mathbf{C}^t)_{:,k}, \quad j \notin \mathbf{t}. \quad (7.9)$$

Figure 7.2 provides a schematic illustration of the different interpolatory matrix approximation regimes compared to the SVD. We also refer to the approximations in (7.6) and (7.8) as *sparse reconstructions* of the data; this is because they use a “sparse” set of $K \ll N$ ($T \ll N$) measurements collected from selected buses (at selected times) to recover measurements at all other buses (times). While we consider both of the one-sided approximations (7.6) and (7.8), we will focus primarily on (7.6) in the later sections.

While the decompositions thus far have been written in matrix format, it is straightforward to see how the one-sided approximation in (7.6) can be adapted for a streaming setting where one measurement becomes available at a time t_i , $i = 1, \dots, T$. Notationally, take $y_k(t_i)$ to be the measurement collected from bus $k \in \{1, \dots, N\}$ at the i -th timestamp $t_i \in \{t_1, \dots, t_T\}$ in a finite monitoring window. Algorithm 7.2.1 describes a pilot-based (or row-based) reconstruction for recovering measurements at *all* locations in the network from a pre-selected set of $K < N$ pilot PMUs indexed by \mathbf{s} , and pre-computed recovery weights \mathbf{Z}_s . We emphasize that the dimensionality reduction of the underlying data is baked right into Algorithm 7.2.1. Indeed, the (online) reconstruction at bus $j \notin \mathbf{s}$ is computed *without ever interacting with that bus*, or any other of the non-pilot PMUs. This is one of the virtues of interpolatory approximations (7.6); once \mathbf{s} and \mathbf{Z}_s are computed, only the K rows in \mathbf{R}_s enter

Algorithm 7.2.1: Row-based interpolatory reconstruction of streaming PMU data.

Input: Recovery weights $\mathbf{Z}_s \in \mathbb{R}^{N \times K}$, pilot PMUs indexed by

$$\mathbf{s} = \{1, \dots, s_K\} \subset \{1, \dots, N\}, \text{ finite time-window } t_1, t_2, \dots, t_T \in \mathbb{R}_{\geq 0}.$$

Output: Pilot-based (interpolatory) reconstruction of $\tilde{y}_j(t_i)$ the measurement at non-pilot PMUs $j \notin \mathbf{s}$ at discrete timestamps t_1, t_2, \dots, t_T .

```

1 for  $t_i \in \{t_1, t_2, \dots, t_T\}$  do
2   Collect data  $y_{s_1}(t_i), \dots, y_{s_K}(t_i)$  from  $K$  pilot PMUs.
3   for Non-pilots  $j \notin \mathbf{s}$  do
4     
$$\tilde{y}_j(t_i) \approx \sum_{k=1}^K (\mathbf{Z}_s)_{j,k} y(t_i)_{s_k}.$$

5   end
6 end

```

into the approximation. This is illustrated in Figure 7.3 for multiple time instances: The data for the *non-pilot* bus 1 (first row of \mathbf{Y}) over a finite window t_1, \dots, t_T is approximated as a linear combination of the data collected from three pilot buses ($s_1 = 4$, $s_2 = 7$, and $s_3 = 6$). This pilot bus data, i.e., the true rows of \mathbf{Y} , are stored in \mathbf{R}_s . The coefficients in the first row of \mathbf{Z}_s provide the weights that reveal how much each pilot s_1, s_2 , and s_3 should contribute to the approximation. Lastly, we mention that the sparse approximations of PMU data proposed in [130, 235] can be interpreted as the streaming-based interpolatory approximations (7.6) and (7.8).

Comparisons with the Singular Value Decomposition for wide-area monitoring.

What are the benefits of using IMDs over the SVD for computing low-rank decompositions of PMU data? While necessarily suboptimal with respect to the approximation error, the error due to an IMD can still be analyzed in a rigorous way; see the forthcoming discussion in Section 7.2.3. Once both decompositions (an IMD and the SVD) are computed, the storage requirements for an IMD, i.e., the memory required to store \mathbf{C}^t , \mathbf{X}_s^t , and \mathbf{R}_s , are similar to that of the SVD. Moreover, like the optimal approximation \mathbf{Y}_K computed via the SVD, the IMDs \mathbf{Y}_s^t , \mathbf{Y}_s , and \mathbf{Y}^t are rank K or less, since $\mathbf{R}_s \in \mathbb{R}^{K \times T}$ and $\mathbf{C}^t \in \mathbb{R}^{N \times K}$. For the particular application of wide-area monitoring, there are two areas where IMDs outshine the SVD.

1. Structure preservation. As depicted in Figure 7.2, the orthonormal components determined by PCA or the SVD are not obviously related to the original data, and lack any physical interpretation with respect to grid quantities such as voltages. In the power systems literature, low-rank matrix decompositions are often used for event monitoring [119,

125, 130, 225, 235], and previous work has shown that the unique sparsity pattern of bus voltages following a disturbance can be advantageous for detecting events [119]. On the other hand, the low-rank factors in an interpolatory decomposition are drawn from true columns and rows of the data matrix, i.e., they are actual PMU voltages. Because of this, IMDs preserve important structural features of the data, such sparsity patterns characteristic of PMUs voltages during system disturbances [119].

2. Economy. The optimal rank- K approximation (7.2) computed by the SVD is a blend of *all* $N \gg K$ rows and $T \gg K$ columns of \mathbf{Y} ; in other words, computing the SVD requires information from *all* PMU time series *simultaneously*. Thus, to compute an optimal rank- K decomposition via the SVD, large amounts of PMU data from local substations must be transmitted using dedicated synchrophasor communication links to a central location, e.g., a regional control center. This requires a significant amount of communication bandwidth; in fact, the bandwidth requirement scales linearly with the number N of installed PMUs, and the sampling rates of PMUs are primarily limited by bandwidth constraints [57, 58, 59]. In other words, dimensionality reduction via the SVD is ill-suited for time-sensitive and bandwidth-limited applications, such as event monitoring and detection. By contrast, interpolatory approximations \mathbf{Y}_s and \mathbf{Y}^t in (7.6) and (7.8) perform dimensionality reduction by blending only $K = |\mathbf{s}| = |\mathbf{t}|$ rows and columns of \mathbf{Y} , i.e., information from $K \ll N$ PMU datastreams and/or $K \ll T$ system snapshots. Thus, once the matrices \mathbf{Z}_s and \mathbf{W}^t containing the recovery weights have been specified, a low-rank approximation to the full data \mathbf{Y} can be formulated *in real time while only interacting with $K \leq N$ pilot buses or aggregating $K \leq T$ time samples*. As observed in, e.g., [57, 58, 59, 235], this lowers the bandwidth requirement of synchrophasor communication networks significantly.

7.2.3 Analyzing the Interpolatory Approximation Error

Because \mathbf{Y}_K is the optimal rank- K approximation to \mathbf{Y} , the interpolatory approximation \mathbf{Y}_s^t cannot be any better:

$$\sigma_{K+1} = \|\mathbf{Y} - \mathbf{Y}_K\|_2 \leq \|\mathbf{Y} - \mathbf{Y}_s^t\|_2. \quad (7.10)$$

The same holds for approximations \mathbf{Y}_s and \mathbf{Y}^t . How close is \mathbf{Y}_s^t to the best approximation from PCA or the SVD? For the moment, we assume that the K row and column indices in \mathbf{s} and \mathbf{t} are given. To get an upper bound on the error $\|\mathbf{Y} - \mathbf{Y}_s^t\|_2$, we must address how to compute the matrix \mathbf{X}_s^t . One natural choice [138, 207] is

$$\mathbf{X}_s^t = (\mathbf{C}^t)^\dagger \mathbf{Y} (\mathbf{R}_s)^\dagger,$$

where \mathbf{M}^\dagger denotes the *Moore–Penrose pseudoinverse* [86, Section 5.5.2] of a matrix \mathbf{M} . Assuming the rows of \mathbf{R}_s and the columns of \mathbf{C}^t are linearly independent, we have that

$$(\mathbf{R}_s)^\dagger = \mathbf{R}_s^\top (\mathbf{R}_s \mathbf{R}_s^\top)^{-1} \quad \text{and} \quad (\mathbf{C}^t)^\dagger = \left((\mathbf{C}^t)^\top \mathbf{C}^t \right)^{-1} \mathbf{C}^{t\top}. \quad (7.11)$$

Then

$$\mathbf{Y}_s^t = \mathbf{C}^t \mathbf{X}_s^t \mathbf{R}_s = \left(\mathbf{C}^t (\mathbf{C}^t)^\dagger \right) \mathbf{Y} \left((\mathbf{R}_s)^\dagger \mathbf{R}_s \right),$$

is obtained by first projecting all the columns of \mathbf{Y} onto the column space of \mathbf{C}^t , and then projecting the result onto the row space of \mathbf{R}_s . Because $\mathbf{C}^t (\mathbf{C}^t)^\dagger$ and $(\mathbf{R}_s)^\dagger \mathbf{R}_s$ are the *orthogonal projectors* onto these subspaces, each step of this twofold projection is optimal with respect to the spectral norm. This interpretation suggests why \mathbf{Y}_s^t can be an effective way to compress the entire set of PMU data contained in \mathbf{Y} . With regards to the one-sided interpolatory approximations \mathbf{Y}_s and \mathbf{Y}^t , the same idea can be applied to obtain \mathbf{Z}_s and \mathbf{W}^t . In other words, we choose

$$\mathbf{W}^t = (\mathbf{C}^t)^\dagger \mathbf{Y} \quad \text{and} \quad \mathbf{Z}_s = \mathbf{Y} (\mathbf{R}_s)^\dagger.$$

In fact, this formulation shows that the i -th row (column) of \mathbf{Y}_s (\mathbf{Y}^t) is the least-squares approximation to the i -th row (column) of \mathbf{Y} from the span of the s rows (t columns) of \mathbf{Y} . Specifically, \mathbf{Z}_s and \mathbf{W}^t are the solution to:

$$\mathbf{Z}_s = \arg \min_{\mathbf{Z} \in \mathbb{R}^{N \times K}} \|\mathbf{Y} - \mathbf{Z} \mathbf{R}_s\|_F \quad \text{and} \quad \mathbf{W}^t = \arg \min_{\mathbf{W} \in \mathbb{R}^{K \times T}} \|\mathbf{Y} - \mathbf{C}^t \mathbf{W}\|_F. \quad (7.12)$$

Assuming the rows (columns) of \mathbf{Y} are linearly independent, the solutions \mathbf{Z}_s (\mathbf{W}^t) are unique. The pilot-based reconstruction proposed by Xie et al. [235] can be viewed as an instance of the row-based interpolatory approximation (7.6) with this choice of \mathbf{Z}_s . This is just one possibility for choosing \mathbf{Z}_s , \mathbf{W}^t , and \mathbf{X}_s^t ; an alternative strategy that is sub-optimal, but recovers the rows and columns specified by s and t of the matrix \mathbf{Y} exactly, is discussed in [206, Section 2].

The quality of this approximation can be assessed in a more quantitative way. In what follows, define $\mathbf{S} \stackrel{\text{def}}{=} \mathbf{I}_{:,s} \in \mathbb{R}^{N \times K}$ and $\mathbf{T} \stackrel{\text{def}}{=} \mathbf{I}_{:,t} \in \mathbb{R}^{T \times K}$ to be the matrices containing the K columns of the $N \times N$ and $T \times T$ identity matrices indexed by s and t .

Theorem 7.3 (Interpolatory error bound [206, Theorem 4.1]). Consider a matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$ and $1 \leq K < \min\{N, T\}$. Let $\mathbf{R}_s \in \mathbb{R}^{K \times T}$ and $\mathbf{C}^t \in \mathbb{R}^{N \times K}$ be given as in (7.5). Then, if $\mathbf{X}_s^t = (\mathbf{Z}_s)^\dagger \mathbf{A} (\mathbf{R}_s)^\dagger$, and $\mathbf{S}^\top \mathbf{U}_K$ and $\mathbf{T}^\top \mathbf{V}_K$ are nonsingular, we have that

$$\sigma_{K+1} \leq \|\mathbf{Y} - \mathbf{Y}_s^t\|_2 \leq (\eta_s + \eta_t) \sigma_{K+1}, \quad (7.13)$$

where *Lebesgue constants* $\eta_s, \eta_t \geq 1$ are given by

$$\eta_s \stackrel{\text{def}}{=} \left\| (\mathbf{S}^\top \mathbf{U}_K)^{-1} \right\|_2 \quad \text{and} \quad \eta_t \stackrel{\text{def}}{=} \left\| (\mathbf{T}^\top \mathbf{V}_K)^{-1} \right\|_2, \quad (7.14)$$

and $\mathbf{U}_K \in \mathbb{R}^{N \times K}$ and $\mathbf{V}_K \in \mathbb{R}^{T \times K}$ are the leading K left and right singular vectors of \mathbf{Y} . \diamond

The submatrices $\mathbf{S}^\top \mathbf{U}_K$ and $\mathbf{T}^\top \mathbf{V}_K$ are guaranteed to be nonsingular for certain row and column selection schemes; see [206, Lemma 3.2]. For the one-sided interpolatory approximations \mathbf{Y}_s and \mathbf{Y}^t , we have the simplified bounds:

$$\sigma_{K+1} \leq \|\mathbf{Y} - \mathbf{Y}_s\|_2 \leq \eta_s \sigma_{K+1} \quad \text{and} \quad \sigma_{K+1} \leq \|\mathbf{Y} - \mathbf{Y}^t\|_2 \leq \eta_t \sigma_{K+1}; \quad (7.15)$$

see [206, Lemma 4.2] and the earlier equivalent formulation [110, Theorem 1.5]. Let us unpack the constants $\eta_{\mathbf{s}}$ and $\eta_{\mathbf{t}}$: The matrix $\mathbf{S}^T \mathbf{U}_K = (\mathbf{U}_K)_{\mathbf{s},:}$ is a $K \times K$ submatrix of \mathbf{U}_K . The *columns* of \mathbf{U}_K , which are the leading left singular vectors of \mathbf{Y} , are orthonormal by design; $\eta_{\mathbf{s}}$ measures how far from orthonormal the \mathbf{s} rows of \mathbf{U}_K are. Thus, the Lebesgue constant $\eta_{\mathbf{s}}$ will get smaller in magnitude as the rows of \mathbf{U}_K specified by \mathbf{s} are “more linearly independent”, i.e., as the submatrix $(\mathbf{U}_K)_{\mathbf{s},:}$ is better conditioned. Likewise, $\eta_{\mathbf{t}}$ measures how far from orthonormal the \mathbf{t} rows of \mathbf{V}_K are. Notice that the *order* in which the indices are presented in \mathbf{s} and \mathbf{t} does not affect the values of $\eta_{\mathbf{s}}$ and $\eta_{\mathbf{t}}$.

The interpolatory error bounds in (7.13) and (7.15) hold for *any* collection of indices \mathbf{s} or \mathbf{t} . In an operational setting where PMU data are recovered using an interpolatory approximation, one can leverage these bounds to realize real-time computational and theoretical benefits. We describe these for the one-sided approximations (7.6) and (7.8), although the ideas apply to the more general two-sided formulation (7.4) as well.

1. **Fast error monitoring.** Because $\mathbf{S}^T \mathbf{U}_K$ and $\mathbf{T}^T \mathbf{V}_K$ are small $K \times K$ matrices, the error indicators $\eta_{\mathbf{s}}$ and $\eta_{\mathbf{t}}$ will be much quicker to compute than the full approximation errors $\|\mathbf{Y} - \mathbf{Y}_{\mathbf{s}}\|$ or $\|\mathbf{Y} - \mathbf{Y}^{\mathbf{t}}\|$ when N and T are large. Indeed, one does not even need to compute explicitly the low-rank approximations $\mathbf{Y}_{\mathbf{s}}$ or $\mathbf{Y}^{\mathbf{t}}$ in (7.6) and (7.8), and hence $\mathbf{Z}_{\mathbf{s}}$ or $\mathbf{W}^{\mathbf{t}}$, to evaluate $\eta_{\mathbf{s}}$ and $\eta_{\mathbf{t}}$. This allows for fast *a priori* estimation of the interpolatory approximation error, or enables the error constants to be monitored *during* the process of selecting the pilot PMUs \mathbf{s} or time snapshots \mathbf{t} .
2. **Pilot bus or time snapshot certification.** In an (online) operational setting, one can apply any desired strategy for selecting pilot PMUs or time snapshots used to construct (7.6) and (7.8). Then, the error constants $\eta_{\mathbf{s}}$ or $\eta_{\mathbf{t}}$ can be (quickly) computed to certify if the chosen indices capture the true rank- K nature of the PMU data matrix. If the error factor is below some threshold, e.g., $\eta_{\mathbf{s}}, \eta_{\mathbf{t}} \leq 100$, the selection \mathbf{s} or \mathbf{t} is accepted; otherwise, either replace some indices, or increase K and add some additional ones.

One could use minimization of $\eta_{\mathbf{s}}$ and $\eta_{\mathbf{t}}$ as a *design objective* to guide the selection of pilot PMUs \mathbf{s} or snapshot indices \mathbf{t} . However, explicit minimization of, e.g., $\eta_{\mathbf{s}}$, over all possible choices of K PMUs from the set of N possible PMUs is infeasible for practical problems, as one would need to test $N!/(K!(N-K)!)$ potential configurations. For example, choosing $K = 10$ pilots from a pool of $N = 50$ PMUs would require testing *over 10 billion pilot bus configurations*. This being unrealistic, in the next Section 7.3 we advocate for use of a *greedy* algorithm that attempts to control the growth of the error constants $\eta_{\mathbf{s}}$ and $\eta_{\mathbf{t}}$ as each new index is selected, one at a time. Henceforth, we refer to the problem of selecting the indices \mathbf{s} to use in a (row-based) interpolatory reconstruction of PMU data (7.6) as the *pilot bus selection problem*. Likewise, we refer to the problem of selecting \mathbf{t} as the *time snapshot selection problem*.

7.3 Strategies for the pilot bus and time snapshot selection problems

For computing solutions to the pilot bus and time snapshot selection problems that give favorable values of the error constant η_s and η_t in (7.13), we propose using the *discrete empirical interpolation method* (DEIM) index selection algorithm and its variants [17, 52, 64, 106, 206]. DEIM was originally developed for the model-order reduction of nonlinear dynamical systems [52], and is a method for constructing interpolatory (or sparse) approximations to vector-valued nonlinear functions. In [206], Sorensen and Embree propose a DEIM-based algorithm for computing interpolatory matrix decompositions (7.4).

The DEIM procedure iteratively parses the leading left and right singular vectors \mathbf{U}_K and \mathbf{V}_K of a matrix \mathbf{Y} to (independently) select the row and column indices \mathbf{s} and \mathbf{t} . Under the hood, DEIM attempts to minimize the growth of the error constants (7.14) as each new index is added to \mathbf{s} or \mathbf{t} . In practice, DEIM typically selects indices that yield small Lebesgue constants (7.14). For the numerical experiments in Section 7.3.3, we observe values for the error constants (7.14) resulting from DEIM that are on the order of $\eta_s, \eta_t \sim O(10^1)$ or less, whereas other, seemingly reliable, selection approaches produce Lebesgue constants on the order of $\eta_s, \eta_t \sim O(10^4)$. Thus, in conjunction with the approximation error (7.13), we expect DEIM to provide an effective strategy for determining pilot bus (time snapshot) configurations to be used in real-time recovery of PMU data.

7.3.1 The discrete empirical interpolation method

We sketch here how DEIM operates on \mathbf{U}_K to select $K \leq N$ pilot PMUs \mathbf{s} ; applying the same process to \mathbf{V}_K yields the time snapshots \mathbf{t} . Before doing so, we introduce some matrix machinery in the form of *interpolatory projectors*. These provide the computational backbone of the DEIM index selection algorithm.

Definition 7.4 (Interpolatory projectors [206, Definition 3.1]). Given a full rank matrix $\mathbf{U} \in \mathbb{R}^{N \times K}$, $N > K$, and set of distinct *interpolation indices* $\mathbf{s} = \{s_1, \dots, s_K\} \subset \{1, \dots, N\}$, the *interpolatory projector* for \mathbf{s} onto $\text{Range}(\mathbf{U})$ is defined to be

$$\mathbf{P} \stackrel{\text{def}}{=} \mathbf{U} (\mathbf{S}^\top \mathbf{U})^{-1} \mathbf{S}^\top \in \mathbb{R}^{N \times N}, \quad (7.16)$$

where $\mathbf{S} = [\mathbf{e}_{s_1} \ \cdots \ \mathbf{e}_{s_K}] \in \mathbb{R}^{N \times k}$ contains the K columns of the $N \times N$ identity matrix indexed by \mathbf{s} . \diamond

The invertibility of $\mathbf{S}^\top \mathbf{U} \in \mathbb{R}^{K \times K}$ is guaranteed by the assumption that \mathbf{U} is full rank [206, Lemma 3.2]. In general, \mathbf{P} in (7.16) is an oblique projector on $\text{Range}(\mathbf{U})$, and so $\mathbf{P}^2 = \mathbf{P}$. Significantly, the interpolatory projector (7.16) enjoys a very useful property involving the

interpolation indices \mathbf{s} . For any vector $\mathbf{x} \in \mathbb{R}^N$, the following interpolation property holds:

$$(\mathbf{P}\mathbf{x})_{\mathbf{s}} = \mathbf{S}^T \mathbf{P}\mathbf{x} = (\mathbf{S}^T \mathbf{U}) (\mathbf{S}^T \mathbf{U})^{-1} \mathbf{S}^T \mathbf{x} = \mathbf{x}_{\mathbf{s}}. \quad (7.17)$$

In other words, the indices \mathbf{s} of the projected $\mathbf{P}\mathbf{x}$ match those of \mathbf{x} .

Consider a matrix $\mathbf{Y} \in \mathbb{R}^{N \times T}$ and its best rank- K approximation via the SVD (7.2). In what follows, let $\mathbf{S}_k = [\mathbf{e}_{s_1} \ \cdots \ \mathbf{e}_{s_k}] \in \mathbb{R}^{N \times k}$ denote the k columns of the $N \times N$ identity corresponding to the indices in $\mathbf{s}_k = \{s_1, \dots, s_k\}$, and let $\mathbf{U}_k = \mathbf{U}_{:,1:k} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_k] \in \mathbb{R}^{N \times k}$ denote the leading k columns of the matrix $\mathbf{U} \in \mathbb{R}^{N \times R}$ of \mathbf{Y} 's left singular vectors.

DEIM operates on the K columns of \mathbf{U}_K one at a time, $k = 1, 2, \dots, K$, as follows. Start with the selection of the first pilot, s_1 , corresponding to $k = 1$. In this simple case the error constant $\eta_{\mathbf{s}}$ in (7.14) reduces to

$$\eta_{s_1} = \|(\mathbf{S}_1^T \mathbf{U}_1)^{-1}\|_2 = \frac{1}{|(\mathbf{u}_1)_{s_1}|},$$

i.e., the magnitude of the reciprocal of the s_1 entry of the leading singular vector \mathbf{u}_1 . Thus, to minimize η_{s_1} and make the reciprocal above as small as possible, choose the PMU $s_1 \in \{1, \dots, N\}$ to be the one corresponding to the *largest* magnitude entry of \mathbf{u}_1 .

- **Step 1.** Choose s_1 as the index of the largest magnitude entry of \mathbf{u}_1 :

$$s_1 = \arg \max_{1 \leq i \leq K} |(\mathbf{u}_1)_i|, \quad \mathbf{u}_1 = \begin{bmatrix} \times \\ \color{red}{\times} \\ \times \\ \times \\ \times \end{bmatrix} \leftarrow s_1$$

i.e., $|(\mathbf{u}_1)_{s_1}| \geq |(\mathbf{u}_1)_i|$ for $i = 1, \dots, N$.

The choice of the second pilot PMU s_2 is more subtle. Obviously, we do not want to accidentally choose the same pilot PMU twice ($s_2 = s_1$) as this would result in an infinite Lebesgue constant $\eta_{\mathbf{s}} = \|(\mathbf{S}_2^T \mathbf{U}_2)^{-1}\|_2$. Using the intuition that $\eta_{\mathbf{s}} = \|(\mathbf{S}_k^T \mathbf{U}_k)^{-1}\|$ is small if the rows selected by \mathbf{S}_k are quite distinct, we choose s_2 so that the two rows $(\mathbf{U}_2)_{s_2,:}$ are as *independent as possible* for $\mathbf{s}_2 = \{s_1, s_2\}$. To guarantee that we choose s_2 such that $s_2 \neq s_1$, i.e., that we select a new PMU, we remove a multiple of \mathbf{u}_1 from \mathbf{u}_2 , so as to zero out the s_1 entry,

$$\mathbf{r}_2 \stackrel{\text{def}}{=} \mathbf{u}_2 - \frac{(\mathbf{u}_2)_{s_1}}{(\mathbf{u}_1)_{s_1}} \mathbf{u}_1,$$

giving $(\mathbf{r}_2)_{s_1} = 0$ by construction. Then, we select s_2 to be the index of the largest-magnitude entry of \mathbf{r}_2 . To assist the formulation of later steps, it is useful to formulate the residual \mathbf{r}_2

in terms of an interpolatory projector (7.16). Specifically, define the interpolatory projector (7.16) for $\mathbf{s}_1 = \{s_1\}$ onto $\text{Range}(\mathbf{u}_1)$ by:

$$\mathbf{P}_1 \stackrel{\text{def}}{=} \mathbf{u}_1(\mathbf{S}_1^\top \mathbf{u}_1)^{-1} \mathbf{S}_1^\top. \quad (7.18)$$

Note that $\mathbf{r}_2 = \mathbf{u}_2 - \mathbf{P}_1 \mathbf{u}_2$, and that $(\mathbf{r}_2)_{s_1} = 0$ by the interpolation property (7.17). Using (7.18), we summarize the next step of DEIM as follows.

- **Step 2.** Construct the interpolatory projector (7.18) for s_1 onto the span of \mathbf{u}_1 . Compute the residual of the interpolatory projection of \mathbf{u}_2 onto $\text{Range}(\mathbf{u}_1)$:

$$\mathbf{r}_2 = \mathbf{u}_2 - \mathbf{P}_1 \mathbf{u}_2.$$

Choose s_2 as the largest-magnitude entry of \mathbf{r}_2 :

$$s_2 = \arg \max_{1 \leq i \leq N} |(\mathbf{r}_2)_i|, \quad \mathbf{r}_2 = \mathbf{u}_2 - \mathbf{P}_1 \mathbf{u}_2 = \begin{bmatrix} \star \\ 0 \\ \star \\ \star \\ \star \end{bmatrix} \leftarrow s_2$$

(The \star indicates a modified entry from the previous, $k = 1$, step.)

Subsequent steps, $k = 3, \dots, K$, follow this same template.

- **Step k .** Construct the interpolatory projector for buses s_1, \dots, s_{k-1} onto the span of $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$ according to Definition 7.4:

$$\mathbf{P}_{k-1} = \mathbf{U}_{k-1}(\mathbf{S}_{k-1}^\top \mathbf{U}_{k-1})^{-1} \mathbf{S}_{k-1}^\top.$$

Compute the residual

$$\mathbf{r}_k = \mathbf{u}_k - \mathbf{P}_{k-1} \mathbf{u}_k,$$

such that $(\mathbf{r}_k)_{s_1} = \dots = (\mathbf{r}_k)_{s_{k-1}} = 0$. Choose s_k to be the index of the largest-magnitude entry of \mathbf{r}_k :

$$s_k = \arg \max_{1 \leq i \leq N} |(\mathbf{r}_k)_i|.$$

At this last step, we are always assured that s_k is a new bus that differs from s_1, \dots, s_{k-1} , since $(\mathbf{r}_k)_{s_1} = \dots = (\mathbf{r}_k)_{s_{k-1}} = 0$ but $\mathbf{r}_k \neq \mathbf{0}$. (Note that $\mathbf{r}_k = \mathbf{0}$ would imply that \mathbf{u}_k is a linear combination of $\mathbf{u}_1, \dots, \mathbf{u}_{k-1}$, which is impossible since the singular vectors are orthogonal.)

Algorithm 7.3.1: The discrete empirical interpolation method (DEIM) [17, 52].

Input: Matrix with orthonormal columns $\mathbf{U}_K = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_K] \in \mathbb{R}^{N \times K}$, $1 \leq K < N$.

Output: Indices $\mathbf{s} = \{s_1, s_2, \dots, s_K\} \subset \{1, 2, \dots, N\}$.

1 Choose the first index s_1 :

$$s_1 = \arg \max_{1 \leq i \leq K} |(\mathbf{u}_1)_i|.$$

2 Take $\mathbf{s}_1 = \{s_1\}$.

3 **for** $k = 2, \dots, K$ **do**

4 Compute the residual by solving a k -dimensional linear system:

$$\mathbf{r}_k = \mathbf{u}_k - \mathbf{U}_{k-1} (\mathbf{S}_{k-1}^\top \mathbf{U}_{k-1})^{-1} \mathbf{S}_{k-1}^\top \mathbf{u}_k.$$

5 Choose the next index s_k :

$$s_k = \arg \max_{1 \leq i \leq K} |(\mathbf{r}_k)_i|.$$

6 Take $\mathbf{s}_k = \{\mathbf{s}_{k-1}, s_k\}$.

7 **end**

We mention that, while this sketch and Algorithm 7.3.1 are formulated to select K indices, over-sampling can be performed as well; see, e.g., [165, 228]. The previously outlined procedure is summarized in Algorithm 7.3.1. Note that, in an efficient implementation, the interpolatory projectors \mathbf{P}_k are never explicitly constructed since they are large, dense matrices. Instead, one computes the action of \mathbf{P}_k on the singular vectors \mathbf{u}_k in Algorithm 7.3.1 without constructing the projectors explicitly. At every iteration, the DEIM selection algorithm is trying to minimize the incremental growth of the objective error constant $\eta_{\mathbf{s}}$ in the bound (7.15); see [52, Lemma 3.2] for a proof. This explains why DEIM is an effective choice for computing the index sets \mathbf{s} and \mathbf{t} , as illustrated in [206, Section 6].

Error bounds and the Q-DEIM variant

There is a worst-case error analysis for the Lebesgue constants produced by DEIM; see, e.g., [52, Lemma 3.2] and [206, Lemma 4.4]. Although, as illustrated in Section 7.3.3 and [206, Section 6], these bounds are highly pessimistic, and Lebesgue constants produced by DEIM in practice are of a much smaller magnitude. Nonetheless, we state a version of this error analysis in the interest of being complete.

Lemma 7.5 (Bounds on the Lebesgue constants (7.14) [206, Lemma 4.4]). For the DEIM index selection procedure in Algorithm 7.3.1, the Lebesgue constants (7.14) satisfy:

$$\eta_{\mathbf{s}} \leq \sqrt{\frac{NK}{3}} 2^K \quad \text{and} \quad \eta_{\mathbf{t}} \leq \sqrt{\frac{TK}{3}} 2^K.$$

◇

The DEIM algorithm is directly linked to the selection strategy used in LU with partial pivoting; see [206, Section 3]. One alternative to the traditional DEIM index selection algorithm is its Q-DEIM variant [64]. Q-DEIM is closely a related variant of DEIM that identifies the pilots \mathbf{s} by applying a rank-revealing QR factorization to the rows of \mathbf{U}_K . The Q-DEIM approach yields a lower theoretical error bound on $\eta_{\mathbf{s}}$, and the ultimate set of pilots \mathbf{s} is invariant under permutations of the columns of \mathbf{U}_K . In practice, however, DEIM and Q-DEIM perform similarly. One drawback is that Q-DEIM is not iterative, and so the number of desired indices must be specified in advance. Thus, DEIM is better suited for an adaptive selection strategy that seeks to select \mathbf{s} so that $\eta_{\mathbf{s}} \leq \tau$, for some tolerance $\tau > 0$.

Finally, we comment on the fact that both DEIM and Q-DEIM require access to \mathbf{U}_K , the K leading left singular vectors of \mathbf{Y} . Indeed, computing \mathbf{U}_K is the dominant cost of these algorithms. The same quantity is required for other row selection strategies found in the literature, such as the PCA-based pilot bus selection of [235], or the leverage score sampling from [138]. If K is known in advance, one need not compute all the singular vectors; only those associated with the dominant singular values are needed. Modern algorithms can compute this SVD very quickly, especially when either the number of buses N , or time snapshots T , is not extremely large. For truly large PMU data matrices, it may be advantageous to estimate these singular vectors using randomized algorithms; see [63, 101] for details. Other strategies exist for row and column selection that do not require access to \mathbf{U}_K or \mathbf{V}_K , for instance, those based on row and column pivoted QR factorizations [53, 95, 207, 220].

7.3.2 Alternative strategies from the power systems literature

The DEIM is one possibility for the pilot PMU and time snapshot selection problems. The error analysis and insights of Section 7.2.3 apply to *any* selection strategy. Here, we review two other strategies for bus selection that appear in the power systems literature.

The strategy of Xie et al. [235]

To our knowledge, Xie et al. [235] were the first to advocate for the use of pilot-based monitoring and reconstructions. Indeed, their PCA-based approach for reducing the dimensionality of PMU data and event monitoring in [235, Section 2] can be seen as a row-based interpolatory matrix approximation (7.6). Although they do not use this language in the original work, this enables us to analyze their pilot PMU selection strategy using the framework of Section 7.2.2. Their objective was to minimize communication bandwidth during real-time event monitoring. Because the work of this chapter builds heavily upon [235], their pilot-selection strategy is a benchmark for the DEIM-based strategy that we propose, and so we describe it here.

Consider a matrix of PMU data $\mathbf{Y} \in \mathbb{R}^{N \times T}$ organized as in Figure 7.1. Xie et al. [235] performs a pre-processing step not included in our approach. This is accomplished by projecting \mathbf{Y} onto its leading $m \leq \max\{N, T\}$ principal components computed via PCA as follows: Form the *covariance matrix* $\mathbf{C} \in \mathbb{R}^{N \times N}$ defined as $\mathbf{C} \stackrel{\text{def}}{=} \mathbf{Y}\mathbf{Y}^\top$. By construction, \mathbf{C} is symmetric positive semi-definite. An eigendecomposition of $\mathbf{C} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^\top$ is computed, and then the leading m principal components $\mathbf{X}_{:,1:m} \in \mathbb{R}^{N \times m}$ are chosen so that the variance exceeds a specific tolerance $\tau > 0$. Then \mathbf{Y} is projected onto these leading components to form $\tilde{\mathbf{Y}} = \mathbf{X}_{:,1:m} \mathbf{X}_{:,1:m}^\top \mathbf{Y}$.

Then, among all possible $(N-1)N/2$ pairs of the rows of $\tilde{\mathbf{Y}}$, N rows of the projected matrix $\tilde{\mathbf{Y}}$, K rows $\mathbf{s} = \{s_1, \dots, s_K\}$, or pilot PMUs, are chosen so that they are as orthogonal as possible, i.e.,

$$\cos(\theta_{s_i, s_j}) = \frac{\tilde{\mathbf{Y}}_{s_i, :}^\top \tilde{\mathbf{Y}}_{s_j, :}}{\left| \tilde{\mathbf{Y}}_{s_i, :} \right| \left| \tilde{\mathbf{Y}}_{s_j, :} \right|} \approx 1, \quad i, j = 1, \dots, K, \quad (7.19)$$

where $|\mathbf{X}|$ takes the entrywise absolute values of a matrix. In other words, the pilot PMUs \mathbf{s} are chosen to be as orthogonal to each other as possible. In [235], it is not specified how to actually solve the optimization problem described by (7.19). Here, we use a mixed-integer linear program to solve (7.19).

The strategy of Li et al. [125]

Another strategy for pilot bus selection can be found in the work by Li et al. [125]. The authors rank buses according to the normalized *energies*:

$$\mathcal{E}_k \stackrel{\text{def}}{=} \|\mathbf{Y}_{k, :} - \boldsymbol{\mu} \mathbf{1} \mathbf{1}^\top\|_2^2, \quad (7.20)$$

where $\boldsymbol{\mu} \in \mathbb{R}^N$ contains the means of the rows of \mathbf{Y} and $\mathbf{1} \in \mathbb{R}^T$ is the vector of all ones. This is closely related to the idea of *leverage score sampling* for row and column selection; see [138]. We mention, however, the goal therein is to identify rows of a PMU matrix that correspond to locations of a fault. Nonetheless, there is a common matrix approximation problem at play: *Identify a small subset of rows and columns that reveal the low-dimensional structure of a PMU data matrix.*

7.3.3 Numerical experiments

At this point, we test the ability of DEIM-based IMDs to reduce the dimensionality of matrices of PMU data. Because actual PMU data are hard to obtain due to security concerns, we use synthetically generated PMU data. We compute several row- and column-based interpolatory approximations of the data according to (7.6) and (7.8) using different methods for selecting \mathbf{s} and \mathbf{t} . Specifically, the following methods are considered.

DEIM is the DEIM index selection procedure in Algorithm 7.3.1.

Q-DEIM is the Q-DEIM variant of DEIM from [64].

MILP is the strategy proposed by Xie et al. [235] that choose indices subject to (7.19). For implementing the selection, we formulate it as a mixed integer linear program, and pass it to MATLAB's 'intlinprog' function.

Rand is a random selection of indices using MATLAB's 'randi' command.

Once the indices \mathbf{s} and \mathbf{t} are computed, the matrices \mathbf{Z}_s and \mathbf{W}^t are computed to be the solutions of the least-squares problem in (7.12). For the presentation of the results, we compute the relative error in the 2-norm, i.e.,

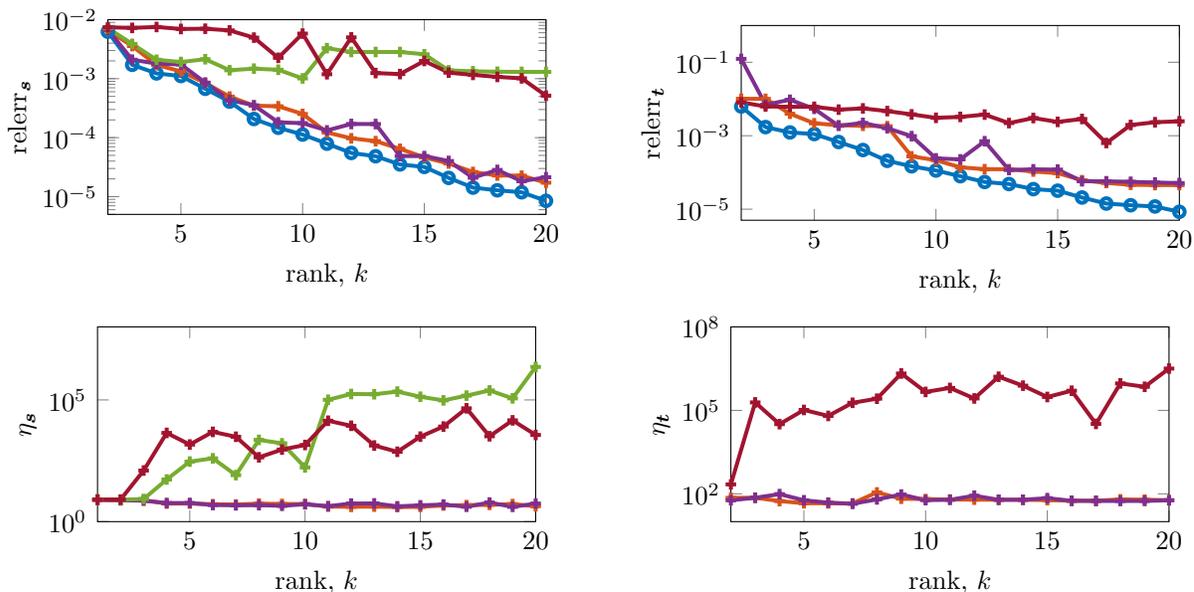
$$\text{relerr}_s \stackrel{\text{def}}{=} \frac{\|\mathbf{Y} - \mathbf{Y}_s\|_2}{\|\mathbf{Y}\|_2} \quad \text{and} \quad \text{relerr}_t \stackrel{\text{def}}{=} \frac{\|\mathbf{Y} - \mathbf{Y}_t\|_2}{\|\mathbf{Y}\|_2}. \quad (7.21)$$

These relative errors are compared against the (relative) best possible rank- k approximation error σ_{k+1}/σ_1 from the SVD. We also compute the associated Lebesgue constants η_s and η_t for each selection strategy.

68-bus 16-machine test system

The first set of synthetic data we consider are obtained from transient simulations performed using MATLAB's Power Systems Toolbox (PST) [54]. The data are generated from the NETS-NYPS 68-bus 16-machine test system [162, Ch. 4]. Voltage magnitudes are collected at every bus over a 100s window at a sampling rate of 100 Hz; after 30s, a three-phase line fault is applied between buses 28 and 29 and cleared after 0.2s. The system is driven by Gaussian white noise to mimic real-world conditions. These data are placed in a 68×6000 dimensional matrix \mathbf{Y} . For the column-based approximations, we do not use MILP, due to the fact that the matrices required for solving the program cannot fit in random access memory.

Rank $k = 1, 2, \dots, 20$ row- and column-based interpolatory approximations to \mathbf{Y} are computed using the selection strategies outlined above. The approximation errors (7.21) and Lebesgue constants (7.14) are plotted in Figure 7.4; specifically, Figures 7.4a and 7.4b contain the results for the row- and column-based approximations. We observe that, for all the selection strategies, the DEIM- and Q-DEIM-based approximations produce high-fidelity approximations. As the rank k increases, these IMDs are able to (for the most part) track with the optimal reconstruction error σ_{k+1} . The row-based approximations by DEIM and Q-DEIM are comparatively better than the column-based ones. This is likely due to the fact that there are more indices to choose from (6000 vs. 68) for the column-based approximations. For the row-based approximations, the MILP-based approximations perform poorly, even worse than the Rand-based approximations for certain ranks k . These results suggest that,



(a) Relative (row-based) interpolatory approximation error and associated Lebesgue constants η_s in (7.14) (b) Relative (col-based) interpolatory approximation error and associated Lebesgue constants η_t in (7.14).

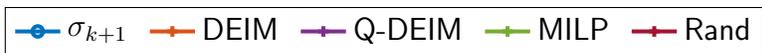


Figure 7.4: Relative errors (7.21) and associated Lebesgue constants for rank $k = 2, 3 \dots, 20$ interpolatory matrix approximations \mathbf{Y}_s and \mathbf{Y}^t of the data generated using the 68-bus 16-machine NETS-NYPS test system.

when paired with a reliable index selection strategy, IMDs are an effective tool for reducing the dimensionality of matrices of PMU data. Moreover, the DEIM-based approaches produce very small Lebesgue constants, providing a theoretical guarantee on the reconstruction error. On the other hand, the Lebesgue constants corresponding to the MILP- and Rand-based selections oscillated by orders of magnitude as the rank k grows, and are in general large in magnitude.

274-bus far west region of synthetic Texas grid

The second set of synthetic data is taken from the ACTIVsg-2000 test case [41, 236]. This is a 2000-bus test case built on the footprint of the Electric Reliability Council of Texas. The data we use are taken from the 274-bus far-west region of the ACTIVsg-2000 grid. For the first experiment, the data are bus voltage magnitudes and generator rotor angles collected over 500s window at a sampling rate of 30 Hz and placed into a matrix $\mathbf{Y} \in \mathbb{R}^{250 \times 250}$. This is to investigate how well IMDs can be used to reconstruct grid quantities other than voltages. We

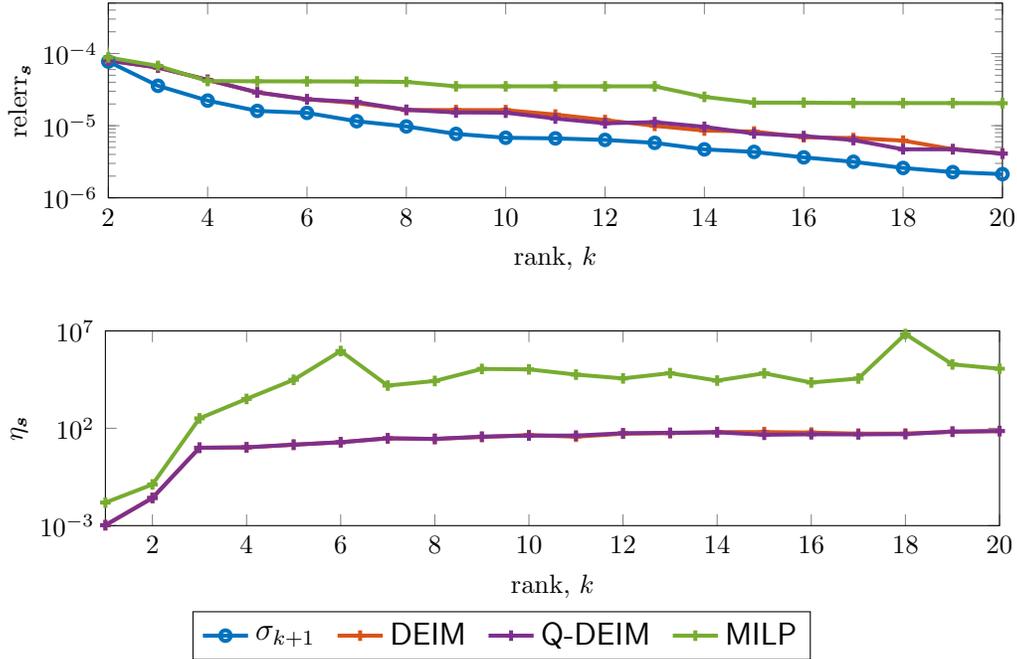


Figure 7.5: Relative errors (7.21) and associated Lebesgue constants for rank $k = 1, 2, \dots, 20$ interpolatory matrix approximations \mathbf{Y}_s and \mathbf{Y}^t of the data generated from the far-west region of the ACTIVsg-2000 test case.

only compute row-based approximations (7.6) using the methods DEIM, Q-DEIM, and MILP. Approximations using Rand were computed, but performed very poorly. The errors (7.21) and Lebesgue constants (7.14) for the computed IMDs of rank $k = 1, 2, \dots, 30$ are recorded in Figure 7.5. As the rank k of the approximation increases, the DEIM- and Q-DEIM-based approximations perform almost a full order of magnitude better than the MILP-based approximations. However, the magnitude of the Lebesgue constant η_s (7.14) is roughly 3 orders of magnitude smaller for the DEIM- and Q-DEIM-based approximations.

7.4 Data-driven monitoring with ID-DEIM

The interpolatory approximations of Section 7.2 and the discrete empirical interpolation method of Section 7.3 suggest a joint IMD-DEIM-based framework for the data-driven and real-time monitoring of electrical power networks using a reduced number of pilot PMUs. This framework builds upon the method proposed by Xie et al. [235]; the key difference is that the framework of IMDs and the bound in (7.15) provide us with an effective error estimator during online operations, and the DEIM algorithm provides a robust, adaptive method for pilot PMU selection and event localization. In this section, we elaborate on some

of the ideas already touched on in Section 7.2.2, and describe how IMDs and DEIM can be combined in an operational scenario to yield various theoretical and computational benefits. Numerical experiments are interspersed throughout.

7.4.1 Adaptive DEIM-based training of pilot bus configurations

At this point, we differentiate between *online* data $\mathbf{Y}_o \in \mathbb{R}^{N \times T_o}$ and offline, or *training* data $\mathbf{Y}_t \in \mathbb{R}^{N \times T_t}$. The positive integers T_o, T_t dictate the size of the online monitoring window and the training window. We can take T_o to be as big as we want; typically, T_t is chosen to contain a ~ 120 s worth of data collected during ambient operating conditions. The online data \mathbf{Y}_o are recovered via a row-based interpolatory approximation (7.6); this can be accomplished in a (matrix) batched format, or one sample at a time according to Algorithm 7.2.1.

For an online pilot bus-based (interpolatory) reconstruction of the data in \mathbf{Y}_o , two things need to be computed offline using \mathbf{Y}_t .

1. The pilot bus configuration $\mathbf{s} = \{s_1, s_2, \dots, s_K\}$ underlying the approximation (7.6).
2. The matrix of weights $\mathbf{Z}_s \in \mathbb{R}^{N \times K}$ that specifies how the data from the pilot buses should be combined to recover measurements at all non-pilot buses.

For choosing the pilots in \mathbf{s} , we use the DEIM index selection algorithm applied to the leading K left singular vectors of \mathbf{Y}_t . More specifically, given a user-specified tolerance $\tau > 0$, DEIM operates on the singular vectors of \mathbf{Y}_t until the interpolatory error bound in (7.15) at step k falls below τ , i.e., $\eta_{\mathbf{s}_k} \sigma_{k+1} \leq \tau$, where σ_{k+1} is the $(k+1)$ -st singular value of \mathbf{Y}_t . As discussed in Section 7.2.3, the error indicator $\eta_{\mathbf{s}_k}$ can be evaluated very cheaply at every step of the training procedure, given that it is computed from the 2-norm of a small $k \times k$ matrix. In our experiments, this can be accomplished for very small values of k , e.g., $k = 3$ or less. For redundancy, one can set a minimum or maximum number of iterations for DEIM to run. Other constraints could be incorporated into the DEIM-based training procedure; e.g., geographic constraints. Once \mathbf{s} is fixed, the computation of \mathbf{Z}_s proceeds according to (7.12) applied to the training data \mathbf{Y}_t .

This adaptive strategy for pilot bus selection offers a few significant advantages. For starters, so long as the online data spans the same low-dimensional subspace as the training data, then the pilot configuration \mathbf{s} is expected to provide a very satisfactory reconstruction during online operations. This is jointly due to the interpolatory error bound (7.15) and the small value of the Lebesgue constant η_s generated by DEIM. Violation of this bound can also be informative; see the discussion in the subsequent Section 7.4.2.

Secondly, we comment on re-training: Instead of periodically re-training pilot bus configurations, after a fixed amount of time, the leading singular vectors \mathbf{U}_K of the most recent batch

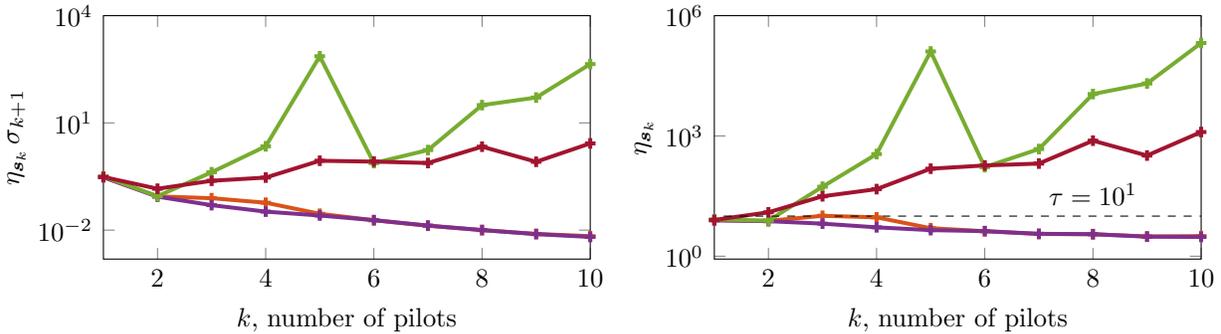
(a) Error bound $\eta_{s_k} \sigma_{k+1}$ during training.(b) Lebesgue constant η_{s_k} during training..

Figure 7.6: Evolution of the Lebesgue constant η_{s_k} and the interpolatory error bound (7.15) throughout the adaptive training as more buses are added.

of online data are computed. Then, the Lebesgue constant η_{s_k} is re-evaluated for the current set of pilots and these singular vectors. If η_{s_k} still lies below some acceptable threshold, then the current configuration \mathbf{s} of pilots is accepted and monitoring continues as normal. If not, a new pilot bus configuration is adaptively selected by DEIM.

Numerical experiments

We illustrate the adaptive DEIM-based training procedure just described (and its effect on the subsequent online reconstruction) using synthetically generated PMU data from the NETS-NYPS 68-bus 16-machine test system used in Section 7.3. For comparison to the DEIM-based approach, we also train the pilot bus configurations using the Q-DEIM, MILP, and Rand selection strategies described in Section 7.3.3. To illustrate the proposed training procedure and evolution of the Lebesgue constants, we fix each method to run for $K = 10$ steps. The iterative DEIM adds each new bus sequentially according to Algorithm 7.3.1. Because Q-DEIM, MILP, and Rand are not iterative, they are each run K times to compute different pilot bus configurations of size $k = 1, \dots, K$. The training data \mathbf{Y}_t are 120 s worth of voltages collected during ambient operating conditions at a sampling rate of 100 Hz. As in Section 7.3.3, white Gaussian noise is added to mimic realistic conditions.

Figure 7.6 shows the evolution of the Lebesgue constant η_{s_k} in (7.14) and associated interpolatory error bound $\eta_{s_k} \sigma_{k+1}$ throughout the adaptive training procedure as more pilots are selected. Had it not been forced to run for 10 iterations, the adaptive selection by DEIM would have terminated after 3 iterations. Moreover, as more pilots are added, the Lebesgue

constant η_s in (7.14) and associated interpolatory error bound $\eta_s \sigma_{k+1}$ continue to steadily decrease. Surprisingly, again, the MILP-based pilot bus selection performs worse than a random selection, from the perspective of this upper bound on the error. As more buses are added, both the Lebesgue constant and thus the error bound tend to *increase*.

7.4.2 Event detection using the error bound (7.15)

Because the leading singular vectors of the online data \mathbf{Y}_o are different from those of \mathbf{Y}_t , the error bound (7.15) is technically no longer valid. However, if we assume that the low-dimensional subspace spanned by \mathbf{Y}_o is similar to that of \mathbf{Y}_t , then the bound (7.15) provides an effective and reliable worst-case estimate of the online reconstruction error. This is because the actual data are reflective of the data used to train the pilots in this case. What happens if this error estimator $\eta_s \sigma_{k+1}$ is no longer accurate during real-time operations? In other words, what if the actual reconstruction error satisfies

$$\|\mathbf{Y}_s - \mathbf{Y}_o\|_2 > \eta_s \sigma_{k+1}, \quad (7.22)$$

where \mathbf{Y}_s is the interpolatory reconstruction of the online data, and σ_{k+1} is the $(k+1)$ -st singular value of \mathbf{Y}_t ? Necessarily, (7.22) would violate the assumption that the low-dimensional subspace spanned by \mathbf{Y}_o is similar to that of \mathbf{Y}_t , and suggests that there has been a fundamental change in the network's operating condition. Changes in the low-dimensional subspace of PMU data have been used to detect, identify, and localize disturbances in [125]. Based on these observations, we propose that upper bound (7.15) can be used as an error estimator during online operations, and deterioration of this estimate as in (7.22) can be used as a simple mechanism for detecting changes to the network's operating conditions. Necessarily, this requires monitoring some non-pilot PMUs to evaluate the error and check for the condition (7.22). To choose these, the DEIM-based training procedure can simply be set to run for m more iterations, where m is the number of non-pilots to be monitored.

Numerical experiments

Here, we investigate the ability of the error bound-based estimator (7.22) to detect disturbance events. For the offline training, DEIM is applied adaptively to 120s of ambient noisy voltages collected during the start of a 500s simulation window. Again, the simulated data are generated using the NETS-NYPS 68-bus 16-machine test system. The training tolerance is set to $\tau = 10^1$; DEIM selects three pilots $s_1 = 48$, $s_2 = 61$, and $s_3 = 50$ as the basis for the online reconstruction until $\eta_s \sigma_{k+1}$. Based on this selection, the (non-adaptive) Q-DEIM and MILP are run to identify three pilot buses each to use for reconstruction. These are recorded in Table 7.1 along with the associated value of the error bound (7.15). After selection, the pilots chosen by DEIM, Q-DEIM, and MILP are continuously monitored until the 450s mark, at which point a three-phase fault of the line between buses 28 and 29 is applied, and cleared .02s later.

Table 7.1: Pilot buses as chosen by DEIM, Q-DEIM and MILP and corresponding error bounds (7.15).

	s_1	s_2	s_3	$\eta_s \sigma_4$
DEIM	48	61	50	8.8667e-2
Q-DEIM	61	50	59	4.4814e-2
MILP	5	8	38	3.4832e0

The pilot-based reconstructions using the configurations identified by DEIM, Q-DEIM, and MILP at the non-pilot buses 28, 29, 26, and 47 are plotted in Figures 7.7, 7.8, and 7.9 respectively. Buses 28 and 29 are included because they are the source of the fault. We monitor buses 26 and 47 because they are, respectively, geographically close to the faulted line. In each case, the approximation quality of each pilot-based reconstruction is satisfactory up until the occurrence of the fault. At this point, all three degrade in quality. This holds for all four of the monitored buses. Although, in all three cases, the degradation in quality is not as poor for bus 47. This phenomenon is likely due to bus 48 being geographically far away from the fault. For the DEIM- and Q-DEIM-based reconstructions, at the time instance immediately after the fault, the bound is violated (7.22) at all four buses. In fact, for the Q-DEIM-based reconstruction, the error at bus 26 violates the error bound before the fault. This could lead to a false warning or indicate the need for retraining. On the other hand, the MILP-based reconstructions do not violate the corresponding error bound; the degradation in the approximation quality is still within acceptable operating conditions according to $\eta_s \sigma_4$.

7.4.3 Event localization using DEIM

Following a disturbance, it is imperative that its source—e.g., the buses adjacent to a faulted line—be located quickly so system operators can take corrective action to prevent cascading failures. A variety of works have tackled the event location problem specifically; see, e.g. [39, 125, 132, 225]. One common approach is to score the affected buses using some energy-based criterion. For instance, Li et al. [125] and Wang et al. [225] use the energy-based scoring (7.20) discussed in Section 7.3. Following a disturbance, buses are ranked in descending order with respect to the energies (7.20); those that produce the highest energies are considered to be the most affected, and are used as a proxy for localizing the source of the disturbance.

In contrast to the energy-based approaches considered in the aforementioned works, we propose using DEIM to localize the source of system disturbance events. Specifically, once a disturbance event is detected using (7.22) or any other mechanism for detection, e.g., those in [125, 235], DEIM is applied to a few seconds of data collected directly after. (In our experiments, as little as 1 s is needed to localize the source of the event.) These data contain the transient system response due to the underlying disturbance. Because DEIM

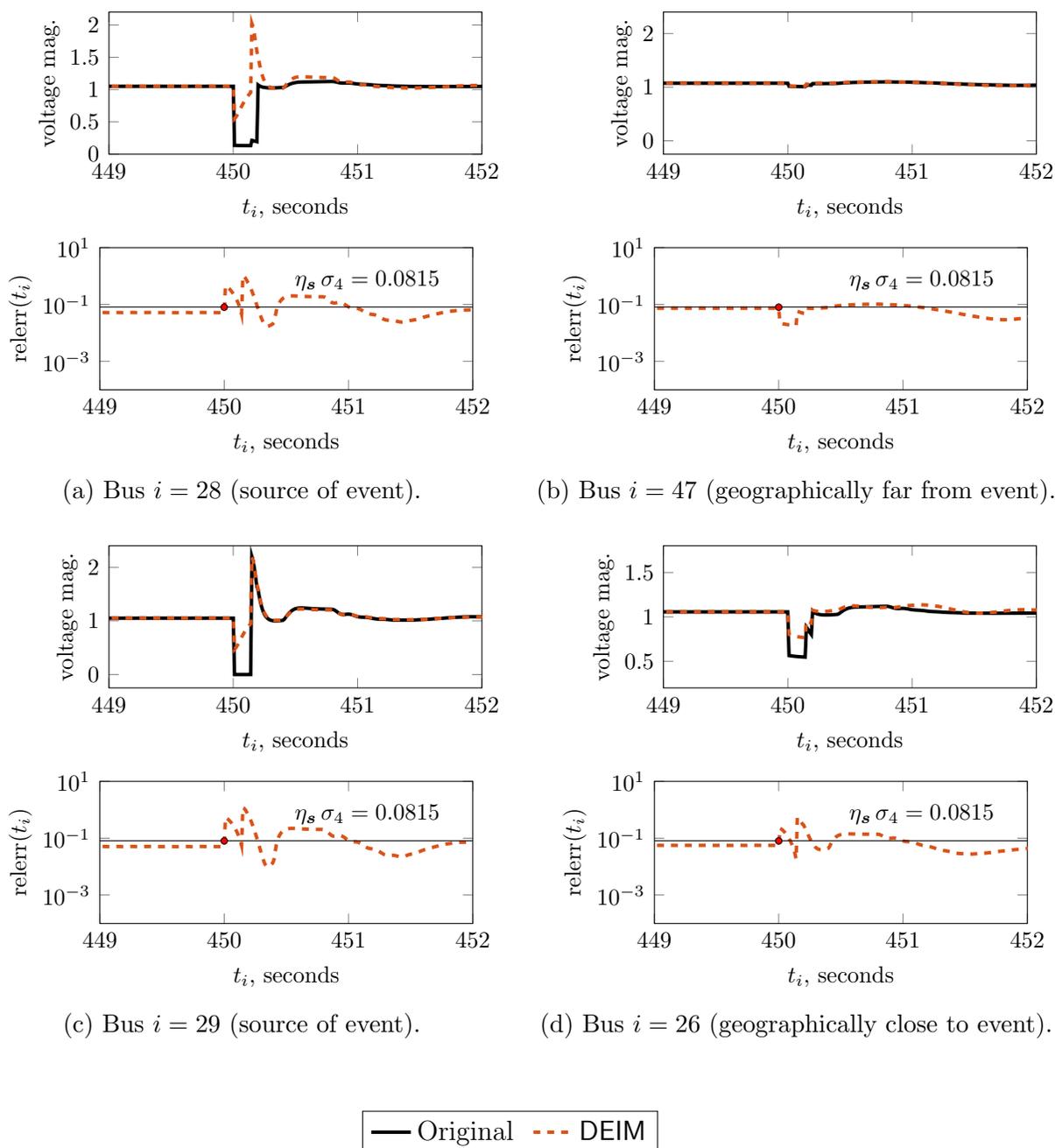


Figure 7.7: Interpolatory reconstructions of various (pilot and non-pilot) PMU datastreams using pilots chosen by DEIM during a three-phase fault of the line between buses 28 and 29.

attempts to extract rows from the singular vectors \mathbf{U}_K that are as linearly independent as possible, we expect DEIM to be able to identify rows (buses) that deviate significantly from normal operating conditions, or unaffected parts of the network. These new DEIM-identified

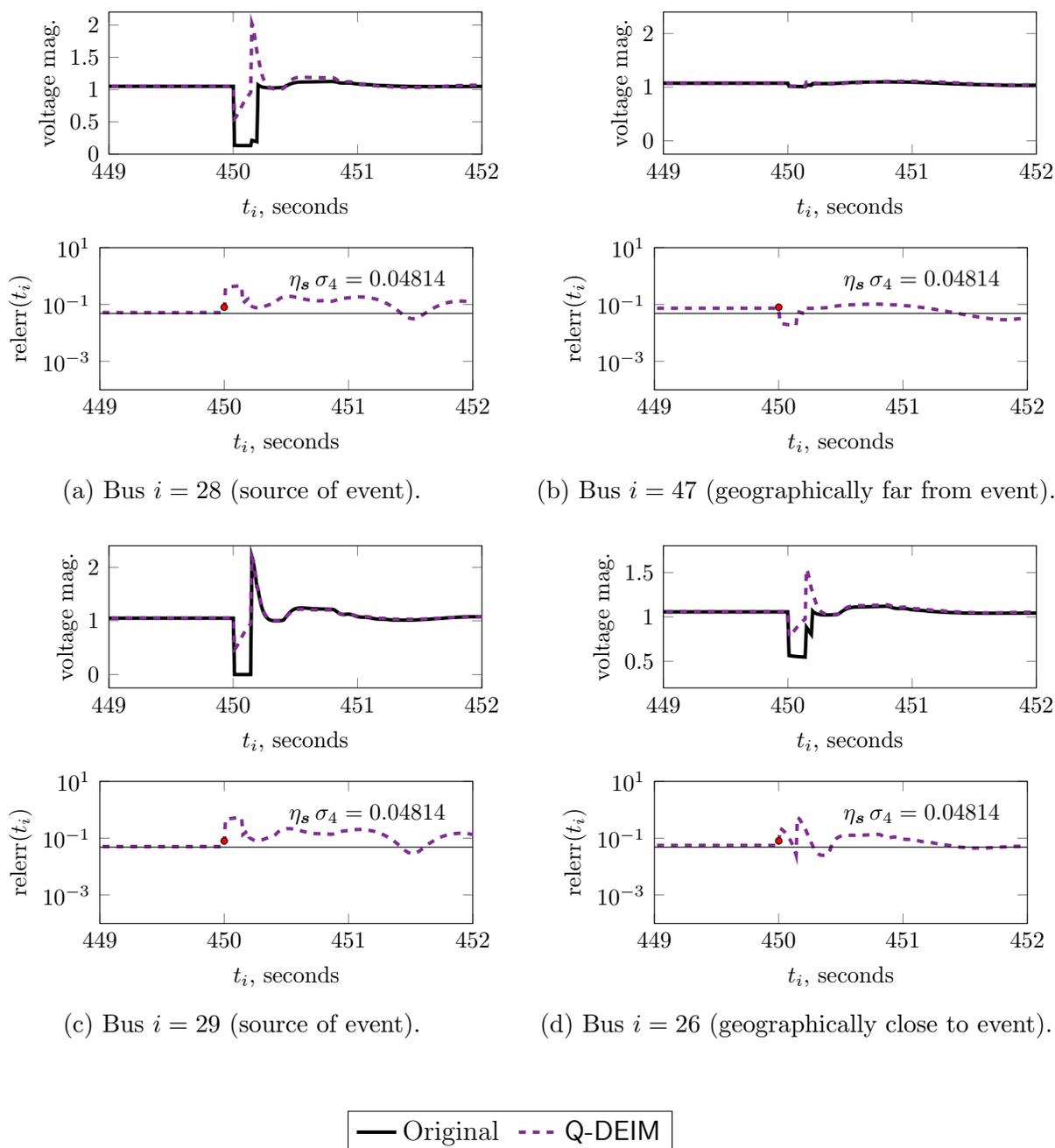


Figure 7.8: Interpolatory reconstructions of various (pilot and non-pilot) PMU datastreams using pilots chosen by Q-DEIM during a three-phase fault of the line between buses 28 and 29.

locations show particular diagnostic power: we propose their use to localize the source of a disturbance more robustly compared to other purely data-driven methods.

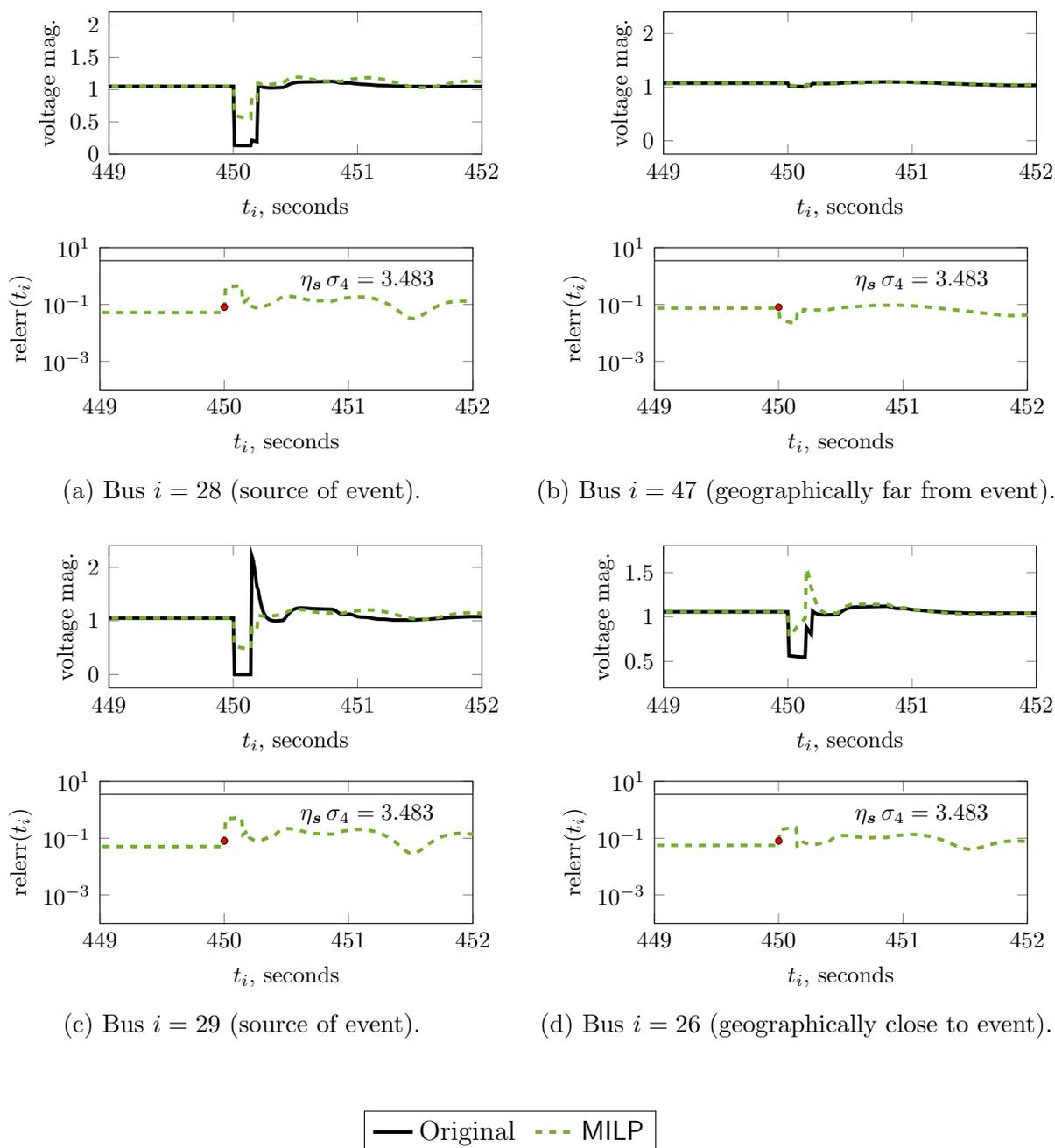


Figure 7.9: Interpolatory reconstructions of various (pilot and non-pilot) PMU datastreams using pilots chosen by MILP during a three-phase fault of the line between buses 28 and 29.

We test this hypothesis with a final experiment. For comparison, we use the energy-based and data-driven event localization method from [125]. As discussed in Section 7.3, in [125], buses are ranked according to (7.20) to localize the source of a disturbance. We call this

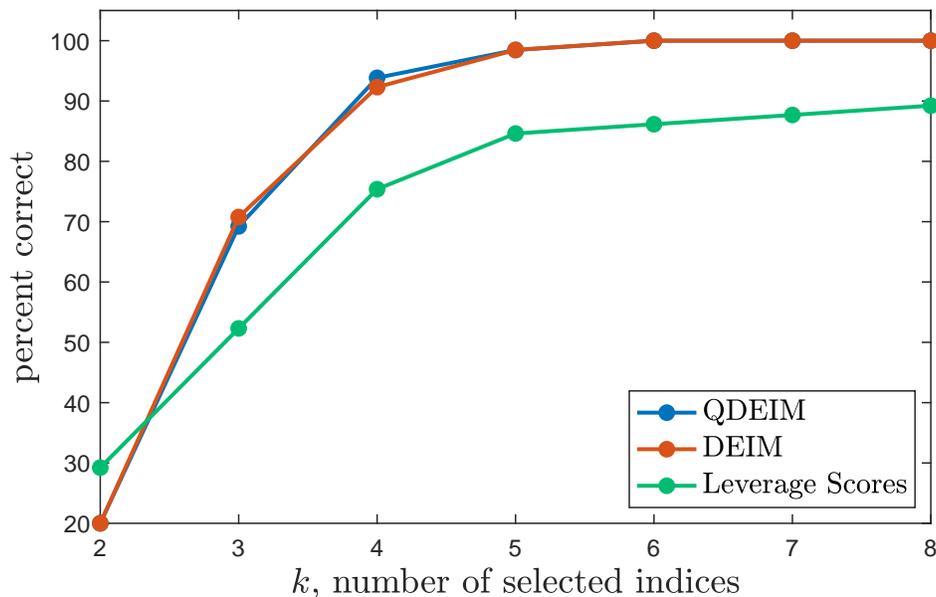


Figure 7.10: Percentage of event simulations in which the indicated method correctly identified both buses associated with the faulted line in the first k indices for $k = 2, \dots, 8$.

approach LS, due to its similarity with leverage-score sampling [138]. Event simulation data for 67 distinct scenarios are generated using the NETS-NYPS 68-bus 16-machine test system. The sampling rate is 100 Hz in all cases. For each scenario, data are generated for 5 s before, during, and after a three-phase line fault at a particular location in the network. In each case, the fault is cleared 0.2 s later. After each event, 0.5 s worth of data collected directly before the event and 1 s worth of data collected directly after the event are placed into a matrix $\mathbf{Y} \in \mathbb{R}^{N \times 150}$. These data are then pre-processed by removing the mean of the pre-event data as in (7.20). Then, for $k = 2, 3, \dots, 8$, DEIM, Q-DEIM, and LS are applied to the leading k singular vectors \mathbf{U}_k of the pre-processed data to select k row indices. If *both* buses connected to the faulted line are found within these k indices, we classify the method (DEIM, Q-DEIM, or LS) as having localized the source of the event. This is repeated for each of the 67 distinct scenarios. The percentages of these event scenarios for which DEIM, Q-DEIM, or LS correctly identified its source within k indices are plotted in Figure 7.10. For $k \geq 4$, DEIM and Q-DEIM are able to localize the source of the faulted line with greater than 90 percent accuracy, and 100 percent accuracy for $k \geq 5$. On the other hand, the prediction rate of LS approaches 90 percent accuracy for $k = 8$. Thus, given the freedom to select enough indices, our DEIM- and Q-DEIM-based localization strategies correctly identify both affiliate buses in all of the tested scenarios, and perform better on average than the reference approach in [125]. Very similar results are observed for other types of voltage disturbance events, e.g., line-to-line and line-to-ground faults.

7.5 Conclusions

In this section, we have investigated the use of interpolatory matrix approximations the the discrete empirical interpolation method for low-rank approximations of PMU data and disturbance event monitoring. We illustrate through our discussion and numerical experiments that IMDs are a very effective strategy for computing low-rank reconstructions of PMU data, particularly during time-sensitive and bandwidth-limited applications such as wide-area monitoring. Specifically, we illustrate how interpolatory approximations can be computed in real-time while interacting only with $K < N$ pilot buses. For identifying pilot bus configurations to be used during online operations, we employed DEIM [17, 52, 206], a greedy method designed to minimize the computable error bound. Finally, we proposed a joint IMD-DEIM-based framework for event monitoring; this can be viewed as a generalization of the method proposed in [235]. Like in [235], because monitoring is performed using $K < N$ pilot buses, communication bandwidth is minimized. Framing the reconstruction in the mathematical framework of IMDs allows us to state a rigorous upper bound on the online reconstruction error. DEIM is applied during an offline stage to adaptively select pilots until the computable error bound (7.15) falls below a user-specified tolerance, thus providing a rigorous estimate on the online reconstruction error. If this error estimate ever deteriorates, there necessarily has been a significant change to the network's operating condition compared to training conditions, making the error estimator provided by (7.15) a suitable mechanism for detecting disturbances. Finally, we illustrate that DEIM can be used to localize the source of disturbance events more robustly compared to other purely data-driven approaches.

Chapter 8

Conclusions and outlook

8.1 Summary of contributions

This dissertation investigates various problems in the system-theoretic model-order reduction, data-driven reduced-order monitoring, and real-time monitoring of large-scale and structured dynamical systems.

In Chapter 3, it is shown that the balanced truncation error bound (2.77) holds with equality for SISO systems satisfying a sign consistency (3.8) condition in the truncated part of the model. This analysis generalizes an earlier result for state-space symmetric systems from [131]. It is additionally shown that the sign parameters corresponding to a system's Hankel singular values can be determined by a generalized state-space symmetry property of the system. This result is strengthened for a special class of arrowhead systems, illustrated by a model of coherent generators in power systems dynamics that motivated our study.

In Chapter 4, data-driven formulations for various types of balanced truncation model reduction of linear first-order and second-order dynamical systems are developed. The results of Chapter 4 generalize the QuadBT framework of [89] to other types of BT model reduction, thus enabling the construction of various balancing-related reduced models directly from input-output invariant transfer function data. Specifically, the considered variants are balanced stochastic truncation [61, 92, 93], positive-real or passivity-preserving balanced truncation [61], bounded-real balanced truncation [159], frequency-weighted balanced truncation [69, 116, 244], and position-velocity balanced truncation [181]. For the linear first-order variants, it is shown that sampling certain spectral factors or frequency weights is required to compute the (approximate) quadrature-based reduced models. For position-velocity BT, it is shown that the data-based construction requires evaluations of the system's position- and velocity-output transfer functions. Numerical experiments are included to validate the data-driven reduced models against their intrusive counterparts.

Chapters 5 and 6 study the \mathcal{H}_2 -optimal approximation problem (6.1) for linear quadratic-output systems (5.1). Two novel \mathcal{H}_2 -optimality frameworks based on distinct sets of first-order necessary conditions for optimality are presented. The first is based upon the solutions to generalized Sylvester equations (6.23) and the linear quadratic-output system Gramians, whereas the second is based on the multivariate rational interpolation of the system's linear- and quadratic-output transfer functions (6.32). In either case, it is shown how to enforce the

established first-order optimality conditions using a Petrov-Galerkin projection. It is also proven that the Sylvester equation-based framework enforces the interpolatory optimality conditions when the reduced model has simple poles. These results establish the Wilson [217, 233] and Meier-Luenberger [97, 142] \mathcal{H}_2 -optimality frameworks for the linear quadratic-output setting. Based on the theoretical optimality frameworks, two iterative algorithms based on repeated projection are proposed for computing \mathcal{H}_2 -optimal reduced models. These generalize the two-sided iterative algorithm [30, 217, 237] and the iterative rational Krylov algorithm [97] from linear model-order reduction. The effectiveness of the proposed methods was illustrated using two benchmark examples from the literature.

Chapter 7 considers the dimensionality reduction of streaming Phasor Measurement Unit (PMU) data, collected from electrical power networks, using interpolatory matrix decompositions (IMDs) [138, 206] and the discrete empirical interpolation method (DEIM) [17, 52]. It is shown that IMDs are better suited for certain tasks in wide-area monitoring, and that the joint IMD-DEIM framework is an effective strategy for *pilot PMU* selection and sparse reconstruction [235]. Specifically, it is shown that IMDs constructed using DEIM are able to produce approximations of synthetic PMU data on par with the SVD. We propose an adaptive DEIM-based training procedure for identifying pilot buses to be monitored during online operations. An error bound-based indicator is also proposed for detecting changes in the network's operating conditions; this is shown to be able to detect disturbances in real time.

8.2 Opportunities for future research

There are several interesting opportunities for future work based on the results of this dissertation. In Chapter 3, the conditions proven in Theorem 3.4 for which the balanced truncation \mathcal{H}_∞ error bound (2.77) holds with equality are *sufficient*. It would be interesting if one could show that the class of systems satisfying the symmetry hypothesis (3.8) is the *only* class of systems for which the bound (2.77) holds with equality, thus proving that (3.8) is a necessary and sufficient condition.

In Chapter 4, the quadrature-based (data-driven) formulations of balanced stochastic truncation, positive-real balanced truncation, and bounded-real balanced truncation require sampling certain spectral factors associated with the transfer function of the underlying linear system. In a practical or real-world setup, it is not clear how to compute these data without an explicit computational model. A well-known iterative method for obtaining the minimal solution of an algebraic Riccati equation is the Newton-Kleinman iteration [117]. This solution procedure requires solving a Lyapunov equation at each iteration. This Lyapunov equation can be viewed as corresponding to a closed-loop system, where the state-feedback law is described by the previous solution iterate. Therefore, if it is possible to re-sample the input-to-output transfer function of the (linear) closed-loop system, it might be possible to obtain samples of the spectral factors (which correspond to the converged stabilizing

solution of the Riccati equations) by continuously re-sampling these closed-loop systems. Clearly, this will require *active* resampling. This is not surprising since, unlike the Lyapunov equations, the Riccati equations are nonlinear and need to be solved iteratively. Another interesting question for future research is whether the quadrature-based balanced truncation of [89] can be generalized to other linear systems with various internal (differential) structures. One possibility is the position-balancing for finite-delay systems from [112]. In this setting, the Gramians are solutions to *delay* Lyapunov equations. The ingredients required for a quadrature-based framework are present. Namely, there are contour integral formulations of the relevant Gramians in the frequency domain [171, Prop. 6.29], and one can resolve Loewner-like matrices from the resulting quadrature-based factors of these integrals [173, 201]. This does require a Rayleigh-like assumption on the delay matrices, e.g., $\mathbf{A} = \alpha \mathbf{E} + \beta \mathbf{A}_\tau$ where \mathbf{A}_τ models the delay term. There are also several open theoretical questions in the area of data-driven balancing. For instance, an affirmative answer to the following would be interesting: Is a data-driven BT reduced model computed by QuadBT the *exact* BT reduced model of a nearby linear system?

In Chapters 5 and 6, the optimal- \mathcal{H}_2 approximation of linear quadratic-output systems (5.1) is considered. With regard to the LQO-IRKA: It is known that the (linear) IRKA is a locally convergent fixed-point iteration when the full-order model is state-space symmetric [76]. Systems with this structure provoke a similar convergence behavior for LQO-IRKA; it would be interesting to investigate whether LQO-IRKA also has these nice convergence properties. As noted in Remark 6.10, the proof of Theorem 6.9 does not rely on the full-order model being linear time-invariant in the state equations. It should be possible to develop an extension of the TF-IRKA method [23], which only uses transfer function evaluations, to the setting of LQO systems. Moving away from optimality momentarily, the development of a structure-preserving interpolation theory akin to [22] for the model reduction of quadratic-output systems is another possible research problem. This is motivated by examples of quadratic-output systems with internal structure, such as the plate with tuned vibration absorbers from Section 5.2.1, that exhibit second-order mechanical structure. This should be possible given similar results for the structure-preserving interpolation of bilinear and quadratic-bilinear systems [226], which have even more complex transfer functions than the quadratic-output case. Another natural extension of the work of Chapter 6 is the \mathcal{H}_2 -optimal model reduction of linear systems with *polynomial* output functions. This is motivated by the following setting: Consider an output function of the form $\mathbf{y} = \mathbf{g}(t, \mathbf{x}(t), \mathbf{u}(t))$ for $\mathbf{g}: \mathbb{R}_{\geq 0} \times \mathbb{R}^n \times \mathbb{R}^m$ that is infinitely differentiable. Then, one can expand \mathbf{g} in a power series expansion, and approximate \mathbf{g} arbitrarily well with polynomial (output) terms. It may also be possible to deal with such an analytic output function directly. Then, the input-to-output response of the system will be described by *infinitely* many Volterra kernels in the time domain, and infinitely many transfer functions in the frequency domain. As already discussed in Chapter 6, \mathcal{H}_2 -optimal reduction theory has already been developed for bilinear and quadratic-bilinear systems [50, 77], which are described by infinitely many kernels. Finally, in Chapter 6 we have implicitly assumed that *direct* methods are used to solve the linear systems required to compute the interpolatory bases \mathbf{V} and \mathbf{W} in Theorem 6.11. For the linear case, Beattie

et al. [21] investigated the impact of (inexact) iterative solves on the resulting interpolatory reduced models. Specifically, [21] shows that employing a Petrov-Galerkin framework for the inexact solves yields a rational interpolant of a nearby full-order system, thus establishing a backward stability framework for interpolatory model reduction. It will be an interesting research direction to establish whether such a backward error result holds for the bases in Theorem 6.11 and the interpolatory model reduction of LQO systems.

Chapter 7 considers, in part, interpolatory matrix decompositions for reducing the dimensionality of PMU data. One obvious next step is to test the proposed algorithms on real-world PMU data collected from actual power networks, although such data are hard to obtain due to security concerns. The examples considered in Chapter 7 are small in size. For larger-scale networks, it would be interesting to investigate other, possibly randomized, algorithms for computing interpolatory decompositions [63, 101]. Actual PMU data are noisy, and it has been illustrated that the performance of DEIM degrades in the presence of noisy data [165]. It could be interesting to investigate the oversampling strategy of [165] in a large-scale setting. Lastly, the row- and column-based interpolatory approximations introduced in Chapter 7 correspond to sampling a reduced number of PMUs at *all* time, or sampling *all* PMUs at fewer time samples. If one were to consider data collected from fewer PMUs *and* fewer time samples, this would correspond to a matrix completion problem. Such a setting could be of interest for truly large-scale networks, where bandwidth is so limited that pilot-based reconstructions *and* down-sampling are required.

Bibliography

- [1] N. ALIYEV, P. BENNER, E. MENGI, P. SCHWERDTNER, AND M. VOIGT, *A greedy subspace method for computing the \mathcal{L}_∞ -norm*, Proceedings in Applied Mathematics and Mechanics, 17 (2017), pp. 751–752. 29
- [2] N. ALIYEV, P. BENNER, E. MENGI, AND M. VOIGT, *A subspace framework for \mathcal{H}_∞ -norm minimization*, SIAM Journal on Matrix Analysis and Applications, 41 (2020), pp. 928–956. 29
- [3] K. K. ANAPARTHI, B. CHAUDHURI, N. F. THORNHILL, AND B. C. PAL, *Coherency identification in power systems through principal component analysis*, IEEE Transactions on Power Systems, 20 (2005), pp. 1658–1660. 198
- [4] A. C. ANTOULAS, *Approximation of Large-Scale Dynamical Systems*, SIAM, Philadelphia, PA, 2005. 2, 17, 18, 19, 21, 22, 23, 24, 25, 27, 35, 36, 38, 39, 43, 50, 54, 68, 73, 76, 87, 126, 140
- [5] A. C. ANTOULAS, C. A. BEATTIE, AND S. GUGERCIN, *Interpolatory Methods for Model Reduction*, SIAM, Philadelphia, PA, 2020. 2, 17, 25, 27, 28, 29, 32, 136, 140, 164
- [6] A. C. ANTOULAS, P. BENNER, AND L. FENG, *Model reduction by iterative error system approximation*, Mathematical and Computer Modeling of Dynamical Systems, 24 (2018), pp. 103–118. 29
- [7] A. C. ANTOULAS, D. C. SORENSEN, AND Y. ZHOU, *On the decay rate of Hankel singular values and related issues*, Systems & Control Letters, 46 (2002), pp. 323–342. 38, 114
- [8] Q. AUMANN AND S. W. R. WERNER, *Code, data and results for numerical experiments in “Structured model order reduction for vibro-acoustic problems using interpolation and balancing methods” (version 1.1)*, Aug. 2022, <https://doi.org/10.5281/zenodo.6806016>. 109
- [9] Q. AUMANN AND S. W. R. WERNER, *Structured model order reduction for vibro-acoustic problems using interpolation and balancing methods*, Journal of Sound and Vibration, 543 (2023), p. 117363. xii, 94, 96, 109, 113, 114, 116, 117, 173
- [10] P. J. BADDOO, B. HERRMANN, B. J. MCKEON, N. J. KUTZ, AND S. L. BRUNTON, *Physics-informed dynamic mode decomposition*, Proceedings of the Royal Society A, 479 (2023), p. 20220576. 59

- [11] Z. BAI, K. MEERBERGEN, AND Y. SU, *Arnoldi methods for structure-preserving dimension reduction of second-order dynamical systems*, in Dimension Reduction of Large-Scale Systems: Proceedings of a Workshop held in Oberwolfach, Germany, October 19–25, 2003, Springer, 2005, pp. 173–189. [95](#)
- [12] Z. BAI AND Y. SU, *Dimension reduction of large-scale second-order dynamical systems via a second-order Arnoldi method*, SIAM Journal on Scientific Computing, 26 (2005), pp. 1692–1709. [95](#)
- [13] J. BAK AND D. J. NEWMAN, *Complex Analysis*, vol. 8, Springer, 2010. [14](#)
- [14] G. A. J. BAKER, *Essentials of Padé Approximants*, Elsevier, 1975. [28](#)
- [15] J. BAKER, M. EMBREE, AND J. SABINO, *Fast singular value decay for Lyapunov solutions with nonnormal coefficients*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 656–668. [38](#)
- [16] L. BALICKI AND S. GUGERCIN, *Energy-based approximation of linear systems with polynomial outputs*, e-prints 2409.19730, arXiv, 2024. [115](#), [123](#), [124](#), [127](#)
- [17] M. BARRAULT, Y. MADAY, N. C. NGUYEN, AND A. T. PATERA, *An empirical interpolation method: application to efficient reduced-basis discretization of partial differential equations*, Comptes Rendus Mathématique, 339 (2004), pp. 667–672. [6](#), [197](#), [198](#), [199](#), [202](#), [209](#), [212](#), [226](#), [228](#)
- [18] R. H. BARTELS AND G. W. STEWART, *Solution of the matrix equation $AX + XB = C$* , Communications of the ACM, 15 (1972), pp. 820–826. [38](#)
- [19] U. BAUR, P. BENNER, AND L. FENG, *Model order reduction for linear and non-linear systems: a system-theoretic perspective*, Archives of Computational Methods in Engineering, 21 (2014), pp. 331–358. [2](#)
- [20] C. BEATTIE AND P. BENNER, *\mathcal{H}_2 -optimality conditions for structured dynamical systems*, Preprint MPIMD/14-18, Max Planck Institute Magdeburg, (2014). [95](#)
- [21] C. BEATTIE, S. GUGERCIN, AND S. WYATT, *Inexact solves in interpolatory model reduction*, Linear Algebra and its Applications, 436 (2012), pp. 2916–2943. [230](#)
- [22] C. A. BEATTIE AND S. GUGERCIN, *Interpolatory projection methods for structure-preserving model reduction*, Systems & Control Letters, 58 (2009), pp. 225–232. [29](#), [95](#), [229](#)
- [23] C. A. BEATTIE AND S. GUGERCIN, *Realization-independent \mathcal{H}_2 -approximation*, in 51st IEEE Conference on Decision and Control (CDC), 2012, pp. 4953–4958. [32](#), [33](#), [229](#)

- [24] P. BENNER AND T. BREITEN, *Interpolation-based \mathcal{H}_2 -model reduction of bilinear control systems*, SIAM Journal on Matrix Analysis and Applications, 33 (2012), pp. 859–885. [129](#), [140](#), [156](#), [161](#)
- [25] P. BENNER AND T. BREITEN, *Model order reduction based on system balancing*, in Model Reduction and Approximation: Theory and Algorithms, SIAM, Philadelphia, PA, 2017, ch. 6, pp. 261–295. [68](#)
- [26] P. BENNER, P. GOYAL, AND S. GUGERCIN, *\mathcal{H}_2 -quasi-optimal model order reduction for quadratic-bilinear control systems*, SIAM Journal on Matrix Analysis and Applications, 39 (2018), pp. 983–1032. [156](#), [161](#)
- [27] P. BENNER, P. GOYAL, B. KRAMER, B. PEHERSTORFER, AND K. WILLCOX, *Operator inference for non-intrusive model reduction of systems with non-polynomial non-linear terms*, Computer Methods in Applied Mechanics and Engineering, 372 (2020), p. 113433. [59](#)
- [28] P. BENNER, P. GOYAL, AND I. PONTES DUFF, *Gramians, energy functionals, and balanced truncation for linear dynamical systems with quadratic outputs*, IEEE Transactions on Automatic Control, 67 (2021), pp. 886–893. [115](#), [120](#), [123](#), [124](#), [125](#), [126](#), [127](#), [128](#), [130](#), [131](#), [143](#), [187](#)
- [29] P. BENNER, P. GOYAL, AND P. VAN DOOREN, *Identification of port-Hamiltonian systems from frequency response data*, Systems & Control Letters, 143 (2020), p. 104741. [82](#)
- [30] P. BENNER, M. KÖHLER, AND J. SAAK, *Sparse-dense Sylvester equations in \mathcal{H}_2 -model order reduction*, Preprint MPIMD/11-11, Max Planck Institute Magdeburg, 2011. [31](#), [157](#), [159](#), [228](#)
- [31] P. BENNER, V. MEHRMANN, AND D. C. SORENSEN, *Dimension Reduction of Large-Scale Systems*, vol. 45 of Lecture Notes in Computational Science and Engineering, Springer, Berlin, Heidelberg, 2005. [2](#), [17](#), [27](#)
- [32] P. BENNER, M. OHLBERGER, A. COHEN, AND K. WILLCOX, *Model Reduction and Approximation: Theory and Algorithms*, SIAM, Philadelphia, PA, 2017. [2](#), [17](#), [27](#)
- [33] P. BENNER, E. S. QUINTANA-ORTI, AND G. QUINTANA-ORTI, *Efficient numerical algorithms for balanced stochastic truncation*, International Journal of Applied Mathematics and Computer Science, 11 (2001), pp. 1123–1150. [69](#), [70](#)
- [34] P. BENNER AND J. SAAK, *Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey*, GAMM-Mitteilungen, 36 (2013), pp. 32–52. [38](#), [128](#)

- [35] P. BENNER AND A. SCHNEIDER, *Balanced truncation model order reduction for LTI systems with many inputs or outputs*, in Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS, vol. 5, 2010. [59](#)
- [36] P. BENNER AND S. W. WERNER, *Frequency-and time-limited balanced truncation for large-scale second-order systems*, Linear Algebra and its Applications, 623 (2021), pp. 68–103. [82](#), [95](#), [99](#)
- [37] P. BENNER AND S. W. R. WERNER, *MORLAB—the model order reduction LABoratory*, in Model Reduction of Complex Dynamical Systems, Birkhäuser, 2021, pp. 393–415. [106](#)
- [38] D. S. BERNSTEIN, *Matrix Mathematics: Theory, Facts, and Formulas*, Princeton University Press, 2009. [102](#)
- [39] P. BHUI AND N. SENROY, *Online identification of tripped line for transient stability assessment*, IEEE Transactions on Power Systems, 31 (2015), pp. 2214–2224. [198](#), [221](#)
- [40] D. BILLGER, *The Butterfly Gyro*, in Dimension Reduction of Large-Scale Systems, P. Benner, V. Mehrmann, and D. C. Sorensen, eds., vol. 45 of Lecture Notes in Computational Science and Engineering, Springer, Berlin, 2005, pp. 349–352, https://doi.org/10.1007/3-540-27909-1_18. [xii](#), [94](#), [107](#)
- [41] A. B. BIRCHFIELD, T. XU, K. M. GEGNER, K. S. SHETYE, AND T. J. OVERBYE, *Grid structural characteristics as validation criteria for synthetic networks*, IEEE Transactions on Power Systems, 32 (2016), pp. 3258–3265. [216](#)
- [42] F. BLAABJERG, *Control of Power Electronic Converters and Systems: Volume 2*, Academic Press, London, 2018. [94](#)
- [43] S. BOCHNER AND K. CHANDRASEKHARAN, *Fourier Transforms*, Princeton University Press, 1949. [16](#), [23](#), [131](#)
- [44] T. BREITEN, *Interpolatory methods for model reduction of large-scale dynamical systems*, Dissertation, Otto-von-Guericke Universität Magdeburg, 2013. [156](#), [161](#)
- [45] T. BREITEN, *Structure-preserving model reduction for integro-differential equations*, SIAM Journal on Control and Optimization, 54 (2016), pp. 2992–3015. [59](#), [99](#)
- [46] T. BREITEN AND T. STYKEL, *Balancing-related model reduction methods*, in Model Order Reduction Volume 1: System-and Data-Driven Methods and Algorithms, Walter de Gruyter GmbH, Berlin, 2021, pp. 15–56. [27](#), [34](#), [68](#)
- [47] J. BREWER, *Kronecker products and matrix calculus in system theory*, IEEE Transactions on Circuits and Systems, 25 (1978), pp. 772–781. [10](#), [166](#)

- [48] N. BRUINSMA AND M. STEINBUCH, *A fast algorithm to compute the \mathcal{H}_∞ -norm of a transfer function matrix*, Systems & Control Letters, 14 (1990), pp. 287–293. [83](#)
- [49] Y.-P. BU, *Krylov subspace model order reduction of linear dynamical systems with quadratic output*, Transactions of the Institute of Measurement and Control, 47 (2024), pp. 827–838. [115](#), [128](#)
- [50] X. CAO, J. MAUBACH, W. SCHILDERS, AND S. WEILAND, *Interpolation-based model order reduction for quadratic-bilinear systems and \mathcal{H}_2 optimal approximation*, in Realization and Model Reduction of Dynamical Systems: A Festschrift in Honor of the 70th Birthday of Thanos Antoulas, Springer, Switzerland, 2022, pp. 117–135. [129](#), [140](#), [161](#), [172](#), [229](#)
- [51] Y. CHAHLAOUI, D. LEMONNIER, A. VANDENDORPE, AND P. VAN DOOREN, *Second-order balanced truncation*, Linear Algebra and its Applications, 415 (2006), pp. 373–384. [95](#), [98](#), [99](#)
- [52] S. CHATURANTABUT AND D. C. SORENSEN, *Nonlinear model reduction via discrete empirical interpolation*, SIAM Journal on Scientific Computing, 32 (2010), pp. 2737–2764. [6](#), [197](#), [198](#), [199](#), [202](#), [209](#), [212](#), [226](#), [228](#)
- [53] H. CHENG, Z. GIMBUTAS, P.-G. MARTINSSON, AND V. ROKHLIN, *On the compression of low rank matrices*, SIAM Journal on Scientific Computing, 26 (2005), pp. 1389–1404. [213](#)
- [54] J. H. CHOW AND K. W. CHEUNG, *A toolbox for power system dynamics and control engineering education and research*, IEEE Transactions on Power Systems, 7 (1992), pp. 1559–1564. [215](#)
- [55] R. COLEMAN, *Calculus on Normed Vector Spaces*, Springer, New York, NY, 2012. [12](#), [13](#), [14](#)
- [56] N. DAHAL, R. L. KING, AND V. MADANI, *Online dimension reduction of synchrophasor data*, in IEEE PES Transmission and Distribution Conference and Exposition 2012, 2012, pp. 1–7. [198](#), [200](#)
- [57] S. DAS, *Sub-Nyquist rate ADC sampling in digital relays and PMUs: Advantages and challenges*, in 2016 IEEE 6th International Conference on Power Systems (ICPS), IEEE, 2016, pp. 1–6. [197](#), [200](#), [206](#)
- [58] S. DAS AND T. SIDHU, *Application of compressive sampling in computer based monitoring of power systems*, Advances in Computer Engineering, 2014 (2014), p. 524740. [197](#), [200](#), [206](#)
- [59] S. DAS AND T. S. SIDHU, *Application of compressive sampling in synchrophasor data communication in WAMS*, IEEE Transactions on Industrial Informatics, 10 (2013), pp. 450–460. [197](#), [200](#), [206](#)

- [60] C. DE VILLEMPEGNE AND R. E. SKELTON, *Model reductions using a projection formulation*, International Journal of Control, 46 (1987), pp. 2141–2169. [28](#)
- [61] U. DESAI AND D. PAL, *A transformation approach to stochastic model reduction*, IEEE Transactions on Automatic Control, 29 (1984), pp. 1097–1100. [4](#), [27](#), [34](#), [58](#), [69](#), [70](#), [72](#), [74](#), [227](#)
- [62] A. N. DIAZ, M. HEINKENSCHLOSS, I. V. GOSEA, AND A. C. ANTOULAS, *Interpolatory model reduction of quadratic-bilinear dynamical systems with quadratic-bilinear outputs*, Advances in Computational Mathematics, 49 (2023), pp. 1–28. [115](#), [118](#), [121](#), [127](#), [128](#), [129](#)
- [63] Y. DONG AND P.-G. MARTINSSON, *Simpler is better: A comparative study of randomized algorithms for computing the CUR decomposition*, e-prints 2104.05877, arXiv, 2021. [197](#), [198](#), [199](#), [201](#), [202](#), [213](#), [230](#)
- [64] Z. DRMAC AND S. GUGERCIN, *A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions*, SIAM Journal on Scientific Computing, 38 (2016), pp. A631–A648. [197](#), [198](#), [199](#), [209](#), [213](#), [215](#)
- [65] Z. DRMAČ, S. GUGERCIN, AND C. BEATTIE, *Quadrature-based vector fitting for discretized \mathcal{H}_2 approximation*, SIAM Journal on Scientific Computing, 37 (2015), pp. A625–A652. [59](#), [95](#)
- [66] V. DRUSKIN AND V. SIMONCINI, *Adaptive rational Krylov subspaces for large-scale dynamical systems*, Systems & Control Letters, 60 (2011), pp. 546–560. [29](#)
- [67] G. E. DULLERUD AND F. PAGANINI, *A Course in Robust Control Theory: A Convex Approach*, Springer, New York, NY, 2000. [17](#), [19](#), [21](#), [22](#)
- [68] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part 1: General Theory*, vol. 10, John Wiley & Sons, Hoboken, NJ, 1988. [11](#)
- [69] D. F. ENNS, *Model reduction with balanced realizations: An error bound and a frequency weighted generalization*, in 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 127–132. [4](#), [27](#), [34](#), [37](#), [41](#), [44](#), [58](#), [60](#), [86](#), [88](#), [89](#), [227](#)
- [70] U. FAROOQ AND R. B. BASS, *Frequency event detection and mitigation in power systems: A systematic literature review*, IEEE Access, 10 (2022), pp. 61494–61519. [197](#)
- [71] K. FERNANDO AND H. NICHOLSON, *Minimality of SISO linear systems*, Proceedings of the IEEE, 70 (1982), pp. 1241–1242. [23](#)
- [72] K. FERNANDO AND H. NICHOLSON, *Singular perturbational model reduction in the frequency domain*, IEEE Transactions on Automatic Control, 27 (1982), pp. 969–970. [38](#)

- [73] K. FERNANDO AND H. NICHOLSON, *Singular perturbational model reduction of balanced systems*, IEEE Transactions on Automatic Control, 27 (1982), pp. 466–468. [38](#), [40](#)
- [74] K. FERNANDO AND H. NICHOLSON, *On a fundamental property of the cross-Gramian matrix*, IEEE Transactions on Circuits and Systems, 31 (1984), pp. 504–505. [23](#)
- [75] K. FERNANDO AND H. NICHOLSON, *On the cross-Gramian for symmetric MIMO systems*, IEEE Transactions on Circuits and Systems, 32 (1985), pp. 487–489. [24](#), [49](#)
- [76] G. FLAGG, C. BEATTIE, AND S. GUGERCIN, *Convergence of the iterative rational Krylov algorithm*, Systems & Control Letters, 61 (2012), pp. 688–691. [229](#)
- [77] G. FLAGG AND S. GUGERCIN, *Multipoint Volterra series interpolation and \mathcal{H}_2 optimal model reduction of bilinear systems*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 549–579. [129](#), [140](#), [161](#), [172](#), [229](#)
- [78] G. M. FLAGG, *Interpolation Methods for the Model Reduction of Bilinear Systems*, Dissertation, Virginia Tech, 2012. [129](#), [161](#), [172](#)
- [79] P. H. GADDE, M. BISWAL, S. BRAHMA, AND H. CAO, *Efficient compression of PMU data in WAMS*, IEEE Transactions on Smart Grid, 7 (2016), pp. 2406–2413. [6](#), [197](#)
- [80] K. GALLIVAN, A. VANDENDORPE, AND P. VAN DOOREN, *Model reduction of MIMO systems via tangential interpolation*, SIAM Journal on Matrix Analysis and Applications, 26 (2004), pp. 328–349. [28](#)
- [81] T. GAMELIN, *Complex Analysis*, Springer, New York, NY, 2003. [14](#), [15](#)
- [82] P. GAO, M. WANG, S. G. GHIOCEL, J. H. CHOW, B. FARDANESH, AND G. STEFOPOULOS, *Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements*, IEEE Transactions on Power Systems, 31 (2015), pp. 1006–1013. [198](#)
- [83] P. GAO, R. WANG, M. WANG, AND J. H. CHOW, *Low-rank matrix recovery from noisy, quantized, and erroneous measurements*, IEEE Transactions on Signal Processing, 66 (2018), pp. 2918–2932. [200](#)
- [84] W. GAWRONSKI AND J.-N. JUANG, *Model reduction in limited time and frequency intervals*, International Journal of Systems Science, 21 (1990), pp. 349–376. [86](#)
- [85] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an \mathcal{H}_∞ -norm bound and relations to risk sensitivity*, Systems & Control Letters, 11 (1988), pp. 167–172. [76](#)

- [86] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013. [8](#), [9](#), [10](#), [11](#), [52](#), [198](#), [200](#), [206](#)
- [87] I. V. GOSEA AND A. C. ANTOULAS, *A two-sided iterative framework for model reduction of linear systems with quadratic output*, in 2019 IEEE 58th Conference on Decision and Control (CDC), 2019, pp. 7812–7817. [115](#), [121](#), [128](#), [130](#)
- [88] I. V. GOSEA AND S. GUGERCIN, *Data-driven modeling of linear dynamical systems with quadratic output in the AAA framework*, Journal of Scientific Computing, 91 (2022), p. 16. [115](#), [121](#)
- [89] I. V. GOSEA, S. GUGERCIN, AND C. BEATTIE, *Data-driven balancing of linear dynamical systems*, SIAM Journal on Scientific Computing, 44 (2022), pp. A554–A582. [4](#), [58](#), [59](#), [60](#), [61](#), [63](#), [65](#), [66](#), [67](#), [86](#), [95](#), [101](#), [105](#), [112](#), [227](#), [229](#)
- [90] I. V. GOSEA, S. GUGERCIN, AND S. W. WERNER, *Structured barycentric forms for interpolation-based data-driven reduced modeling of second-order systems*, Advances in Computational Mathematics, 50 (2024), pp. 1–32. [59](#), [95](#)
- [91] P. K. GOYAL, *System-Theoretic Model Order Reduction for Bilinear and Quadratic-Bilinear Control Systems*, Dissertation, Otto-von-Guericke-Universität Magdeburg, 2018. [156](#), [161](#)
- [92] M. GREEN, *Balanced stochastic realizations*, Linear Algebra and its Applications, 98 (1988), pp. 211–247. [4](#), [27](#), [34](#), [58](#), [69](#), [70](#), [227](#)
- [93] M. GREEN, *A relative error bound for balanced stochastic truncation*, IEEE Transactions on Automatic Control, 33 (1988), pp. 961–965. [4](#), [58](#), [69](#), [70](#), [227](#)
- [94] E. J. GRIMME, *Krylov Projection Methods for Model Reduction*, Dissertation, University of Illinois, Urbana-Champaign, USA, 1997. [28](#)
- [95] M. GU AND S. C. EISENSTAT, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM Journal on Scientific Computing, 17 (1996), pp. 848–869. [213](#)
- [96] S. GUGERCIN AND A. C. ANTOULAS, *A survey of model reduction by balanced truncation and some new results*, International Journal of Control, 77 (2004), pp. 748–766. [27](#), [34](#), [68](#), [74](#), [82](#)
- [97] S. GUGERCIN, A. C. ANTOULAS, AND C. BEATTIE, *\mathcal{H}_2 model reduction for large-scale linear dynamical systems*, SIAM Journal on Matrix Analysis and Applications, 30 (2008), pp. 609–638. [5](#), [25](#), [26](#), [28](#), [29](#), [30](#), [31](#), [32](#), [33](#), [34](#), [136](#), [140](#), [142](#), [161](#), [172](#), [177](#), [181](#), [182](#), [228](#)

- [98] S. GUGERCIN, R. V. POLYUGA, C. BEATTIE, AND A. VAN DER SCHAFT, *Structure-preserving tangential interpolation for model reduction of port-Hamiltonian systems*, *Automatica*, 48 (2012), pp. 1963–1974. [110](#)
- [99] R. C. GUNNING AND H. ROSSI, *Analytic Functions of Several Complex Variables*, American Mathematical Society, Providence, 2022. [15](#)
- [100] B. GUSTAVSEN AND A. SEMLYEN, *Rational approximation of frequency domain responses by vector fitting*, *IEEE Transactions on Power Delivery*, 14 (1999), pp. 1052–1061. [59](#), [95](#)
- [101] N. HALKO, P.-G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, *SIAM Review*, 53 (2011), pp. 217–288. [213](#), [230](#)
- [102] J. HAO, R. J. PIECHOCKI, D. KALESHI, W. H. CHIN, AND Z. FAN, *Sparse malicious false data injection attacks and defense mechanisms in smart grids*, *IEEE Transactions on Industrial Informatics*, 11 (2015), pp. 1–12. [198](#)
- [103] Y. HAO, M. WANG, AND J. H. CHOW, *Modelless streaming synchrophasor data recovery in nonlinear systems*, *IEEE Transactions on Power Systems*, 35 (2019), pp. 1166–1177. [198](#)
- [104] Y. HAO, M. WANG, J. H. CHOW, E. FARANTATOS, AND M. PATEL, *Modelless data quality improvement of streaming synchrophasor measurements by exploiting the low-rank Hankel structure*, *IEEE Transactions on Power Systems*, 33 (2018), pp. 6966–6977. [198](#)
- [105] P. HARSHAVARDHANA, E. A. JONCKHEERE, AND L. M. SILVERMAN, *Stochastic balancing and approximation-stability and minimality*, in *Mathematical Theory of Networks and Systems: Proceedings of the MTNS-83 International Symposium Be'er Sheva, Israel, June 20–24, 1983, New York, NY, 1984, Springer*, pp. 406–420. [69](#), [70](#), [74](#)
- [106] E. P. HENDRYX LYONS, *The discrete empirical interpolation method in class identification and data summarization*, *WIREs Computational Statistics*, 16 (2024), p. e1653 (18pp). [209](#)
- [107] C. HIMPE, *emgr—The Empirical Gramian Framework*, *Algorithms*, 11 (2018), p. 91. [59](#)
- [108] C. HIMPE, *Comparing (empirical-Gramian-based) model order reduction algorithms*, in *Model Reduction of Complex Dynamical Systems*, Birkhäuser, Cham, 2021, pp. 141–164. [59](#)

- [109] T. HOLICKI, J. NICODEMUS, P. SCHWERDTNER, AND B. UNGER, *Energy matching in reduced passive and port-Hamiltonian systems*, e-prints 2309.05778, arXiv, 2023. [113](#), [118](#)
- [110] Y. P. HONG AND C.-T. PAN, *Rank-revealing QR factorizations and the singular value decomposition*, *Mathematics of Computation*, 58 (1992), pp. 213–232. [208](#)
- [111] D. HYLAND AND D. BERNSTEIN, *The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton, and Moore*, *IEEE Transactions on Automatic Control*, 30 (1985), pp. 1201–1211. [31](#), [161](#)
- [112] E. JARLEBRING, T. DAMM, AND W. MICHIELS, *Model reduction of time-delay systems using position balancing and delay Lyapunov equations*, *Mathematics of Control, Signals, and Systems*, 25 (2013), pp. 147–166. [229](#)
- [113] I. T. JOLLIFFE, *Principal Component Analysis*, Springer, New York, NY, second ed., 2002. [198](#), [200](#)
- [114] E. JONCKHEERE AND L. SILVERMAN, *A new set of invariants for linear systems—application to reduced order compensator design*, *IEEE Transactions on Automatic Control*, 28 (1983), pp. 953–964. [27](#), [34](#)
- [115] J.-N. JUANG AND R. S. PAPPAS, *An eigensystem realization algorithm for modal parameter identification and model reduction*, *Journal of Guidance, Control, and Dynamics*, 8 (1985), pp. 620–627. [59](#)
- [116] S. W. KIM, B. D. ANDERSON, AND A. G. MADIEVSKI, *Error bound for transfer function order reduction using frequency weighted balanced truncation*, *Systems & Control Letters*, 24 (1995), pp. 183–192. [4](#), [58](#), [60](#), [86](#), [88](#), [227](#)
- [117] D. KLEINMAN, *On an iterative technique for Riccati equation computations*, *IEEE Transactions on Automatic Control*, 13 (1968), pp. 114–115. [228](#)
- [118] R. KLUMP, P. AGARWAL, J. E. TATE, AND H. KHURANA, *Lossless compression of synchronized phasor measurements*, in *IEEE PES General Meeting*, IEEE, 2010, pp. 1–7. [197](#)
- [119] X. KONG, B. FOGGO, K. YAMASHITA, AND N. YU, *Online voltage event detection using synchrophasor data with structured sparsity-inducing norms*, *IEEE Transactions on Power Systems*, 37 (2021), pp. 3506–3515. [198](#), [200](#), [205](#), [206](#)
- [120] B. KRAMER, S. GUGERCIN, J. BORGGAARD, AND L. BALICKI, *Scalable computation of energy functions for nonlinear balanced truncation*, *Computer Methods in Applied Mechanics and Engineering*, 427 (2024), p. 117011. [127](#)
- [121] P. KÜRSCHNER, *Efficient Low-Rank Solution of Large-Scale Matrix Equations*, Dissertation, Otto-von-Guericke-Universität, Magdeburg, Germany, 2016. [38](#), [128](#)

- [122] S. LALL, J. E. MARSDEN, AND S. GLAVAŠKI, *Empirical model reduction of controlled nonlinear systems*, IFAC Proceedings Volumes, 32 (1999), pp. 2598–2603. [59](#)
- [123] A. LAUB, L. M. SILVERMAN, AND M. VERMA, *A note on cross-Grammians for symmetric realizations*, Proceedings of the IEEE, 71 (1983), pp. 904–905. [23](#)
- [124] A. J. LAUB, M. T. HEATH, C. C. PAIGE, AND R. C. WARD, *Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms*, IEEE Transactions on Automatic Control, 32 (1987), pp. 115–122. [38](#), [62](#), [99](#)
- [125] W. LI, M. WANG, AND J. H. CHOW, *Real-time event identification through low-dimensional subspace characterization of high-dimensional synchrophasor data*, IEEE Transactions on Power Systems, 33 (2018), pp. 4937–4947. [6](#), [198](#), [199](#), [206](#), [214](#), [220](#), [221](#), [224](#), [225](#)
- [126] E. LIBERTY, F. WOOLFE, P.-G. MARTINSSON, V. ROKHLIN, AND M. TYGERT, *Randomized algorithms for the low-rank approximation of matrices*, Proceedings of the National Academies of Science, 104 (2007), pp. 20167–20172. [197](#), [198](#), [199](#), [202](#)
- [127] B. LILJEGREN-SAILER AND I. V. GOSEA, *Data-driven and low-rank implementations of balanced singular perturbation approximation*, SIAM Journal on Scientific Computing, 46 (2024), pp. A483–A507. [40](#), [60](#)
- [128] C.-A. LIN AND T.-Y. CHIU, *Model reduction via frequency weighted balanced realization*, Control, Theory and Advanced Technology, 8 (1992), pp. 341–351. [86](#), [88](#)
- [129] Z. LIN, F. WEN, Y. DING, AND Y. XUE, *Data-driven coherency identification for generators based on spectral clustering*, IEEE Transactions on Industrial Informatics, 14 (2017), pp. 1275–1285. [198](#)
- [130] S. LIU, Y. ZHAO, Z. LIN, Y. LIU, Y. DING, L. YANG, AND S. YI, *Data-driven event detection of power systems based on unequal-interval reduction of PMU data and local outlier factor*, IEEE Transactions on Smart Grid, 11 (2019), pp. 1630–1643. [198](#), [199](#), [205](#), [206](#)
- [131] W. Q. LIU, V. SREERAM, AND K. L. TEO, *Model reduction for state-space symmetric systems*, Systems & Control Letters, 34 (1998), pp. 209–215. [41](#), [44](#), [56](#), [227](#)
- [132] X. LIU, D. LAVERTY, R. BEST, K. LI, D. MORROW, AND S. MCLOONE, *Principal component analysis of wide-area phasor measurements for islanding detection—a geometric view*, IEEE Transactions on Power Delivery, 30 (2015), pp. 976–985. [198](#), [221](#)
- [133] Y. LIU AND B. ANDERSON, *Singular perturbation approximation of balanced systems*, in Proceedings of the 28th IEEE Conference on Decision and Control, IEEE, 1989, pp. 1355–1360. [38](#)

- [134] Y. LIU AND B. D. O. ANDERSON, *Singular perturbation approximation of balanced systems*, International Journal of Control, 50 (1989), pp. 1379–1405. [38](#), [48](#)
- [135] T.-T. LU AND S.-H. SHIOU, *Inverses of 2×2 block matrices*, Computers & Mathematics with Applications, 43 (2002), pp. 119–129. [46](#)
- [136] J. M. MACIEJOWSKI AND R. J. OBER, *Balanced parametrizations and canonical forms for system identification*, IFAC Proceedings, 21 (1988), pp. 701–708. [43](#), [50](#)
- [137] J. R. MAGNUS AND H. NEUDECKER, *The commutation matrix: some properties and applications*, The Annals of Statistics, 7 (1979), pp. 381–394. [10](#), [11](#)
- [138] M. W. MAHONEY AND P. DRINEAS, *CUR matrix decompositions for improved data analysis*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 697–702. [6](#), [197](#), [198](#), [199](#), [201](#), [202](#), [206](#), [213](#), [214](#), [225](#), [228](#)
- [139] A. J. MAYO AND A. C. ANTOULAS, *A framework for the solution of the generalized realization problem*, Linear Algebra and its Applications, 425 (2007), pp. 634–662. [28](#), [59](#), [95](#)
- [140] G. P. MCCORMICK, *Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems*, Mathematical Programming, 10 (1976), pp. 147–175. [156](#)
- [141] V. MEHRMANN AND B. UNGER, *Control of port-Hamiltonian differential-algebraic systems and applications*, Acta Numerica, 32 (2023), pp. 395–515. [73](#), [117](#)
- [142] L. MEIER AND D. LUENBERGER, *Approximation of linear constant systems*, IEEE Transactions on Automatic Control, 12 (1967), pp. 585–588, <https://doi.org/10.1109/TAC.1967.1098680>. [5](#), [28](#), [29](#), [31](#), [33](#), [142](#), [161](#), [228](#)
- [143] D. G. MEYER AND S. SRINIVASAN, *Balancing and model reduction for second-order form linear systems*, IEEE Transactions on Automatic Control, 41 (1996), pp. 1632–1644. [95](#), [98](#), [99](#)
- [144] H. MIN, F. PAGANINI, AND E. MALLADA, *Accurate reduced order models for coherent synchronous generators*, in 57th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, 2019, pp. 316–317. [41](#), [42](#)
- [145] H. MIN, F. PAGANINI, AND E. MALLADA, *Accurate reduced-order models for heterogeneous coherent generators*, IEEE Control Systems Letters, 5 (2020), pp. 1741–1746. [41](#)
- [146] P. MLINARIĆ, *Structure-Preserving Model Order Reduction for Network Systems*, Dissertation, Otto-von-Guericke Universität Magdeburg, 2020. [30](#)

- [147] P. MLINARIĆ, C. A. BEATTIE, Z. DRMAČ, AND S. GUGERCIN, *IRKA is a Riemannian gradient descent method*, IEEE Transactions on Automatic Control, (2024). [29](#), [34](#)
- [148] P. MLINARIĆ, P. BENNER, AND S. GUGERCIN, *Interpolatory-optimality conditions for structured linear time-invariant systems*, SIAM Journal on Numerical Analysis, 63 (2025), pp. 949–975. [29](#), [34](#)
- [149] P. MLINARIĆ AND S. GUGERCIN, *\mathcal{L}_2 -optimal reduced-order modeling using parameter-separable forms*, SIAM Journal on Scientific Computing, 45 (2023), pp. A554–A578. [29](#), [34](#)
- [150] P. MLINARIĆ AND S. GUGERCIN, *A unifying framework for interpolatory \mathcal{L}_2 -optimal reduced-order modeling*, SIAM Journal on Numerical Analysis, 61 (2023), pp. 2133–2156. [29](#), [34](#)
- [151] B. MOORE, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Transactions on Automatic Control, 26 (1981), pp. 17–32. [27](#), [34](#), [35](#), [43](#), [59](#), [60](#), [127](#)
- [152] C. MULLIS AND R. A. ROBERTS, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Transactions on Circuits and Systems, 23 (1976), pp. 551–562. [27](#), [34](#), [35](#), [60](#), [127](#)
- [153] Y. NAKATSUKASA, O. SÈTE, AND L. N. TREFETHEN, *The AAA algorithm for rational approximation*, SIAM Journal on Scientific Computing, 40 (2018), pp. A1494–A1522. [59](#), [115](#)
- [154] H. NIJMEIJER AND A. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, vol. 175, Springer, New York, 1990. [124](#)
- [155] D. NOVOSEL, G. BARTOK, G. HENNEBERG, P. MYSORE, D. TZIOUVARAS, AND S. WARD, *IEEE PSRC report on performance of relaying during wide-area stressed conditions*, IEEE Transactions on Power Delivery, 25 (2009), pp. 3–16. [197](#)
- [156] R. OBER AND D. MCFARLANE, *Balanced canonical forms for minimal systems: A normalized coprime factor approach*, Linear Algebra and its Applications, 122 (1989), pp. 23–64. [43](#)
- [157] R. J. OBER, *Balanced realizations: canonical form, parametrization, model reduction*, International Journal of Control, 46 (1987), pp. 643–670. [42](#), [43](#)
- [158] OBERWOLFACH BENCHMARK COLLECTION, *Butterfly gyroscope*. Hosted at MOR-wiki – Model Order Reduction Wiki, 2005, http://modelreduction.org/index.php/Butterfly_Gyroscope. [xii](#), [107](#)

- [159] P. OPDENACKER AND E. JONCKHEERE, *A contraction mapping preserving balanced reduction scheme and its infinity norm error bounds*, IEEE Transactions on Circuits and Systems, 35 (1988), pp. 184–189. [4](#), [34](#), [58](#), [77](#), [227](#)
- [160] M. R. OPMEER AND T. REIS, *A lower bound for the balanced truncation error for MIMO systems*, IEEE Transactions on Automatic Control, 60 (2015), pp. 2207–2212. [47](#), [48](#)
- [161] F. PAGANINI AND E. MALLADA, *Global analysis of synchronization performance for power systems: Bridging the theory-practice gap*, IEEE Transactions on Automatic Control, 65 (2020), pp. 3007–3022. [42](#)
- [162] B. PAL AND B. CHAUDHURI, *Robust Control in Power Systems*, Springer, New York, NY, 2005. [215](#)
- [163] B. PASCUAL AND S. ADHIKARI, *Dynamic response of structures with frequency dependent damping models*, in 49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 2008. [96](#)
- [164] B. PEHERSTORFER, *Sampling low-dimensional Markovian dynamics for preasymptotically recovering reduced models from data with operator inference*, SIAM Journal on Scientific Computing, 42 (2020), pp. A3489–A3515. [59](#)
- [165] B. PEHERSTORFER, Z. DRMAC, AND S. GUGERCIN, *Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points*, SIAM Journal on Scientific Computing, 42 (2020), pp. A2837–A2864. [197](#), [198](#), [212](#), [230](#)
- [166] B. PEHERSTORFER, S. GUGERCIN, AND K. WILLCOX, *Data-driven reduced model construction with time-domain Loewner models*, SIAM Journal on Scientific Computing, 39 (2017), pp. A2152–A2178. [59](#)
- [167] B. PEHERSTORFER AND K. WILLCOX, *Data-driven operator inference for noninvasive projection-based model reduction*, Computer Methods in Applied Mechanics and Engineering, 306 (2016), pp. 196–215. [59](#), [95](#)
- [168] T. PENZL, *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case*, Systems & Control Letters, 40 (2000), pp. 139–144. [38](#)
- [169] L. PERNEBO AND L. SILVERMAN, *Model reduction via balanced state space representations*, IEEE Transactions on Automatic Control, 27 (1982), pp. 382–387. [37](#)
- [170] J. PHILLIPS, L. DANIEL, AND L. M. SILVEIRA, *Guaranteed passive balancing transformations for model order reduction*, in Proceedings of the 39th Annual Design Automation Conference, 2002, pp. 52–57. [72](#)

- [171] E. PLISCHKE, *Transient Effects of Linear Dynamical Systems*, Dissertation, Universität Bremen, 2005. [229](#)
- [172] I. PONTES DUFF, P. GOYAL, AND P. BENNER, *Data-driven identification of Rayleigh-damped second-order systems*, in *Realization and Model Reduction of Dynamical Systems*, C. Beattie, P. Benner, M. Embree, S. Gugercin, and S. Lefteriu, eds., Springer, Switzerland, 2022, pp. 255–272. [59](#), [95](#), [105](#)
- [173] I. PONTES DUFF, C. POUSSOT-VASSAL, AND C. SEREN, *Realization independent single time-delay dynamical model interpolation and \mathcal{H}_2 -optimal approximation*, in 2015 54th IEEE Conference on Decision and Control (CDC), IEEE, 2015, pp. 4662–4667. [173](#), [229](#)
- [174] J. L. PROCTOR, S. L. BRUNTON, AND J. N. KUTZ, *Dynamic mode decomposition with control*, *SIAM Journal on Applied Dynamical Systems*, 15 (2016), pp. 142–161. [59](#)
- [175] J. PRZYBILLA, I. PONTES DUFF, AND P. BENNER, *Model reduction for second-order systems with inhomogeneous initial conditions*, *Systems & Control Letters*, 183 (2024), p. 105671. [99](#)
- [176] J. PRZYBILLA, I. PONTES DUFF, P. GOYAL, AND P. BENNER, *Balanced truncation of descriptor systems with a quadratic output*, e-prints 2402.14716, arXiv, 2024. [115](#), [125](#), [128](#), [131](#)
- [177] R. PULCH, *Energy-based model order reduction for linear stochastic Galerkin systems of second order*, *Proceedings in Applied Mathematics and Mechanics*, 23 (2023). [113](#), [115](#)
- [178] R. PULCH AND A. NARAYAN, *Balanced truncation for model order reduction of linear dynamical systems with quadratic outputs*, *SIAM Journal on Scientific Computing*, 41 (2019), pp. A2270–A2295. [115](#)
- [179] E. QIAN, B. KRAMER, B. PEHERSTORFER, AND K. WILLCOX, *Lift & Learn: Physics-informed machine learning for large-scale nonlinear dynamical systems*, *Physica D: Nonlinear Phenom.*, 406 (2020), p. 132401. [59](#), [95](#), [156](#)
- [180] M. RAFFERTY, X. LIU, D. M. LAVERTY, AND S. MCLOONE, *Real-time multiple event detection and classification using moving window PCA*, *IEEE Transactions on Smart Grid*, 7 (2016), pp. 2537–2548. [198](#)
- [181] T. REIS AND T. STYKEL, *Balanced truncation model reduction of second-order systems*, *Mathematical and Computer Modeling of Dynamical Systems*, 14 (2008), pp. 391–406. [4](#), [34](#), [58](#), [61](#), [95](#), [96](#), [97](#), [98](#), [99](#), [100](#), [105](#), [227](#)

- [182] S. REITER, T. DAMM, M. EMBREE, AND S. GUGERCIN, *On the balanced truncation error bound and sign parameters from arrowhead realizations*, Advances in Computational Mathematics, 50 (2024), p. 10. [4](#), [34](#), [41](#), [45](#)
- [183] S. REITER, M. EMBREE, S. GUGERCIN, AND V. KEKATOS, *Interpolatory approximations for PMU data: Dimension reduction, pilot bus selection, and event monitoring*, in preparation, (2025). [6](#)
- [184] S. REITER, I. V. GOSEA, AND S. GUGERCIN, *Generalizations of data-driven balancing: what to sample for different balancing-based reduced models*, e-prints 2312.12561, arXiv, 2023. [4](#), [58](#), [63](#), [67](#)
- [185] S. REITER, I. V. GOSEA, I. PONTES DUFF, AND S. GUGERCIN, *\mathcal{H}_2 -optimal model reduction of linear quadratic-output systems by multivariate rational interpolation*, e-prints 2505.03057, arXiv, 2025. [5](#), [115](#), [142](#)
- [186] S. REITER, I. PONTES DUFF, I. V. GOSEA, AND S. GUGERCIN, *\mathcal{H}_2 -optimal model reduction of linear systems with multiple quadratic outputs*, e-prints 2405.05951, arXiv, 2024. [5](#), [115](#), [118](#), [130](#), [131](#), [142](#), [146](#), [147](#), [154](#), [160](#)
- [187] S. REITER AND S. W. R. WERNER, *Data-driven balanced truncation for second-order systems with generalized proportional damping*, tech. report, In preparation, 2025. [5](#), [59](#)
- [188] S. REITER AND S. W. R. WERNER, *Interpolatory model reduction of dynamical systems with root mean squared error*, IFAC-PapersOnLine, 59 (2025), pp. 385–390. [6](#), [113](#), [115](#), [128](#), [142](#), [143](#), [193](#)
- [189] S. J. REITER, *On the Tightness of the Balanced Truncation Error Bound with an Application to Arrowhead Systems*, master’s thesis, Virginia Tech, 2022. [4](#), [41](#)
- [190] G. ROVATSOS, X. JIANG, A. D. DOMINGUEZ-GARCIA, AND V. V. VEERAVALLI, *Statistical power system line outage detection under transient dynamics*, IEEE Transactions on Signal Processing, 65 (2017), pp. 2787–2797. [198](#)
- [191] C. W. ROWLEY, *Model reduction for fluids, using balanced proper orthogonal decomposition*, International Journal of Bifurcation and Chaos, 15 (2005), pp. 997–1013. [59](#)
- [192] W. RUDIN, *Function Theory in Polydiscs*, W. A. Benjamin Inc., New York, NY, 1969. [15](#)
- [193] W. RUDIN, *Function Theory in the Unit Ball of \mathbb{C}^n* , Springer, New York, NY, 2008. [15](#)
- [194] W. J. RUGH, *Nonlinear System Theory: The Volterra/Wiener Approach*, Johns Hopkins University Press, Baltimore, MD, 1981. [120](#), [121](#), [156](#)

- [195] A. RUHE, *Rational Krylov algorithms for nonsymmetric eigenvalue problems. II. Matrix pairs*, Linear Algebra and its Applications, 197 (1994), pp. 283–295. [28](#)
- [196] J. SAAK, D. SIEBELTS, AND S. W. R. WERNER, *A comparison of second-order model order reduction methods for an artificial fishtail*, at-Automatisierungstechnik, 67 (2019), pp. 648–667. [95](#), [99](#)
- [197] D. K. SALKUYEH AND F. P. A. BEIK, *An explicit formula for the inverse of arrow-head and doubly arrow matrices*, International Journal of Applied and Computational Mathematics, 4 (2018), pp. 1–8. [52](#)
- [198] J. M. SCHERPEN, *Balancing for nonlinear systems*, Systems & Control Letters, 21 (1993), pp. 143–153. [124](#), [127](#)
- [199] J. M. SCHERPEN AND W. S. GRAY, *Minimality and local state decompositions of a nonlinear state space realization using energy functions*, IEEE Transactions on Automatic Control, 45 (2000), pp. 2079–2086. [124](#), [127](#)
- [200] P. J. SCHMID, *Dynamic mode decomposition of numerical and experimental data*, Journal of Fluid Mechanics, 656 (2010), pp. 5–28. [59](#)
- [201] P. SCHULZE AND B. UNGER, *Data-driven interpolation of dynamical systems with delay*, Systems & Control Letters, 97 (2016), pp. 125–131. [229](#)
- [202] H. SHARMA AND B. KRAMER, *Preserving Lagrangian structure in data-driven reduced-order modeling of large-scale dynamical systems*, Physica D: Nonlinear Phenomena, 462 (2024), p. 134128. [59](#), [95](#)
- [203] V. SIMONCINI, *Computational methods for linear matrix equations*, SIAM Review, 58 (2016), pp. 377–441. [38](#), [128](#)
- [204] J. R. SINGLER AND B. A. BATTEN, *A proper orthogonal decomposition approach to approximate balanced truncation of infinite dimensional linear systems*, International Journal of Computer Mathematics, 86 (2009), pp. 355–371. [59](#)
- [205] Q.-Y. SONG, U. ZULFIQAR, Z.-H. XIAO, M. M. UDDIN, AND V. SREERAM, *Balanced truncation of linear systems with quadratic outputs in limited time and frequency intervals*, e-prints 2402.11445, arXiv, 2024, <https://arxiv.org/abs/2402.11445>. [115](#)
- [206] D. C. SORENSEN AND M. EMBREE, *A DEIM induced CUR factorization*, SIAM Journal on Scientific Computing, 38 (2016), pp. A1454–A1482. [6](#), [197](#), [198](#), [199](#), [201](#), [202](#), [207](#), [208](#), [209](#), [212](#), [213](#), [226](#), [228](#)
- [207] G. W. STEWART, *Four algorithms for the efficient computation of truncated pivoted QR approximations to a sparse matrix*, Numerische Mathematik, 83 (1999), pp. 313–323. [201](#), [206](#), [213](#)

- [208] M. S. TOMBS AND I. POSTLETHWAITE, *Truncated balanced realization of a stable non-minimal state-space system*, International Journal of Control, 46 (1987), pp. 1319–1330. [38](#), [62](#)
- [209] Z. TOMLJANOVIĆ, C. BEATTIE, AND S. GUGERCIN, *Damping optimization of parameter dependent mechanical systems by rational interpolation*, Advances in Computational Mathematics, 44 (2018), pp. 1797–1820. [99](#), [113](#)
- [210] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005. [19](#)
- [211] L. N. TREFETHEN AND J. WEIDEMAN, *The exponentially convergent trapezoidal rule*, SIAM Review, 56 (2014), pp. 385–458. [81](#)
- [212] N. TRUHAR AND K. VESELIĆ, *An efficient method for estimating the optimal dampers' viscosity for linear vibrating systems using Lyapunov equation*, SIAM Journal on Matrix Analysis and Applications, 31 (2009), pp. 18–39. [110](#)
- [213] J. H. TU, *Dynamic Mode Decomposition: Theory and Applications*, Dissertation, Princeton University, 2013. [59](#)
- [214] R. VAN BEEUMEN AND K. MEERBERGEN, *Model reduction by balanced truncation of linear systems with a quadratic output*, AIP Conf. Proc., 1281 (2010), pp. 2033–2036. [113](#), [114](#), [115](#)
- [215] R. VAN BEEUMEN, K. VAN NIMMEN, G. LOMBAERT, AND K. MEERBERGEN, *Model reduction for dynamical systems with quadratic output*, International Journal for Numerical Methods in Engineering, 91 (2012), pp. 229–248. [3](#), [114](#), [115](#), [116](#), [128](#)
- [216] A. VAN DER SCHAFT, *Port-Hamiltonian systems: an introductory survey*, in Proceedings of the International Congress of Mathematicians, vol. 3, European Mathematical Society, 2006, pp. 1339–1365. [73](#), [117](#)
- [217] P. VAN DOOREN, K. A. GALLIVAN, AND P.-A. ABSIL, *\mathcal{H}_2 -optimal model reduction of MIMO systems*, Applied Mathematics Letters, 21 (2008), pp. 1267–1273. [5](#), [28](#), [29](#), [30](#), [31](#), [33](#), [142](#), [161](#), [172](#), [228](#)
- [218] P. VAN DOOREN, K. A. GALLIVAN, AND P.-A. ABSIL, *\mathcal{H}_2 -optimal model reduction with higher-order poles*, SIAM Journal on Matrix Analysis and Applications, 31 (2010), pp. 2738–2753. [28](#), [29](#), [31](#), [135](#), [161](#), [185](#)
- [219] A. VARGA, *On stochastic balancing related model reduction*, in Proceedings of the 39th IEEE Conference on Decision and Control, vol. 3, 2000, pp. 2385–2390. [69](#), [70](#)

- [220] S. VORONIN AND P.-G. MARTINSSON, *Efficient algorithms for CUR and interpolative matrix decompositions*, Advances in Computational Mathematics, 43 (2017), pp. 495–516. [213](#)
- [221] G. WANG, V. SREERAM, AND W. LIU, *A new frequency-weighted balanced truncation method and an error bound*, IEEE Transactions on Automatic Control, 44 (1999), pp. 1734–1737. [86](#), [88](#)
- [222] M. WANG, J. H. CHOW, P. GAO, X. T. JIANG, Y. XIA, S. G. GHIOCEL, B. FARDANESH, G. STEFOPOLOUS, Y. KOKAI, N. SAITO, AND OTHERS, *A low-rank matrix approach for the analysis of large amounts of power system synchrophasor data*, in 2015 48th Hawaii International Conference on System Sciences, 2015, pp. 2637–2644. [198](#), [200](#)
- [223] M. WANG, J. H. CHOW, Y. HAO, S. ZHANG, W. LI, R. WANG, P. GAO, C. LACKNER, E. FARANTATOS, AND M. PATEL, *A low-rank framework of PMU data recovery and event identification*, in 2019 International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA), 2019, pp. 1–9. [200](#)
- [224] M. WANG, J. H. CHOW, D. OSIPOV, S. KONSTANTINOPOULOS, S. ZHANG, E. FARANTATOS, AND M. PATEL, *Review of low-rank data-driven methods applied to synchrophasor measurement*, IEEE Open Access Journal of Power and Energy, 8 (2021), pp. 532–542. [198](#), [200](#)
- [225] Z. WANG, Y. ZHANG, AND J. ZHANG, *Principal components fault location based on WAMS/PMU measure system*, in 2011 IEEE Power and Energy Society General Meeting, IEEE, 2011, pp. 1–5. [198](#), [199](#), [206](#), [221](#)
- [226] S. W. R. WERNER, *Structure-Preserving Model Reduction for Mechanical Systems*, Dissertation, Otto-von-Guericke-Universität, Magdeburg, Germany, 2021. [3](#), [94](#), [95](#), [98](#), [99](#), [173](#), [229](#)
- [227] S. W. R. WERNER, I. V. GOSEA, AND S. GUGERCIN, *Structured vector fitting framework for mechanical systems*, IFAC-PapersOnLine, 55 (2022), pp. 163–168. [59](#), [95](#)
- [228] K. WILLCOX, *Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition*, Computers & Fluids, 35 (2006), pp. 208–226. [212](#)
- [229] K. WILLCOX AND J. PERAIRE, *Balanced model reduction via the proper orthogonal decomposition*, AIAA Journal, 40 (2002), pp. 2323–2330. [59](#)
- [230] J. C. WILLEMS, *Dissipative dynamical systems part I: General theory*, Archive for Rational Mechanics and Analysis, 45 (1972), pp. 321–351. [68](#)

- [231] J. C. WILLEMS, *Dissipative dynamical systems part II: Linear systems with quadratic supply rates*, Archive for Rational Mechanics and Analysis, 45 (1972), pp. 352–393. [68](#)
- [232] D. WILSON AND A. KUMAR, *Symmetry properties of balanced systems*, IEEE Transactions on Automatic Control, 28 (1983), pp. 927–929. [42](#)
- [233] D. A. WILSON, *Optimum solution of model-reduction problem*, in Proceedings of the Institution of Electrical Engineers, vol. 117(6), 1970, pp. 1161–1165. [5](#), [28](#), [29](#), [30](#), [142](#), [161](#), [228](#)
- [234] S. A. WYATT, *Issues in Interpolatory Model Reduction: Inexact Solves, Second-Order Systems and DAEs*, Dissertation, Virginia Tech, 2012. [95](#)
- [235] L. XIE, Y. CHEN, AND P. R. KUMAR, *Dimensionality reduction of synchrophasor data for early event detection: Linearized analysis*, IEEE Transactions on Power Systems, 29 (2014), pp. 2784–2794. [6](#), [198](#), [199](#), [200](#), [201](#), [202](#), [205](#), [206](#), [207](#), [213](#), [214](#), [215](#), [217](#), [221](#), [226](#), [228](#)
- [236] T. XU, A. B. BIRCHFIELD, K. S. SHETYE, AND T. J. OVERBYE, *Creation of synthetic electric grid models for transient stability studies*, in The 10th Bulk Power Systems Dynamics and Control Symposium (IREP 2017), 2017, pp. 1–6. [216](#)
- [237] Y. XU AND T. ZENG, *Optimal \mathcal{H}_2 model reduction for large scale MIMO systems via tangential interpolation*, International Journal of Numerical Analysis and Modeling, 8 (2011), pp. 174–188. [29](#), [31](#), [32](#), [142](#), [157](#), [161](#), [228](#)
- [238] W.-Y. YAN AND J. LAM, *An approximate approach to H_2 optimal model reduction*, IEEE Transactions on Automatic Control, 44 (1999), pp. 1341–1358. [147](#)
- [239] A. YOUSUFF AND R. SKELTON, *Covariance equivalent realizations with application to model reduction of large-scale systems*, Control and Dynamic Systems, 22 (1985), pp. 273–348. [28](#)
- [240] A. YOUSUFF, D. WAGIE, AND R. SKELTON, *Linear system approximation via covariance equivalent realizations*, Journal of Mathematical Analysis and Applications, 106 (1985), pp. 91–115. [28](#)
- [241] Y. YUE AND K. MEERBERGEN, *Using Krylov-Padé model order reduction for accelerating design optimization of structures and vibrations in the frequency domain*, International Journal for Numerical Methods in Engineering, 90 (2012), pp. 1207–1232. [94](#), [113](#), [116](#), [118](#)
- [242] Y. YUE AND K. MEERBERGEN, *Accelerating optimization of parametric linear systems by model order reduction*, SIAM Journal on Optimization, 23 (2013), pp. 1344–1370. [113](#), [116](#)

- [243] L. ZHANG AND J. LAM, *On H_2 model reduction of bilinear systems*, *Automatica*, 38 (2002), pp. 205–216. [156](#), [161](#)
- [244] K. ZHOU, *Frequency-weighted \mathcal{L}_∞ norm and optimal Hankel norm model reduction*, *IEEE Transactions on Automatic Control*, 40 (1995), pp. 1687–1699. [4](#), [34](#), [58](#), [60](#), [86](#), [89](#), [91](#), [227](#)
- [245] K. ZHOU, J. C. DOYLE, AND K. GLOVER, *Robust and Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1996. [4](#), [15](#), [16](#), [17](#), [20](#), [21](#), [24](#), [25](#), [52](#), [68](#), [69](#), [73](#), [74](#), [75](#), [77](#), [78](#)
- [246] H. ZHU AND G. B. GIANNAKIS, *Sparse overcomplete representations for efficient identification of power line outages*, *IEEE Transactions on Power Systems*, 27 (2012), pp. 2215–2224. [198](#)
- [247] D. ŽIGIĆ, L. T. WATSON, AND C. BEATTIE, *Contragredient transformations applied to the optimal projection equations*, *Linear Algebra and its Applications*, 188 (1993), pp. 665–676. [31](#)

Appendices

Appendix A

Error calculations in the Proof of Theorem 6.9.

In this appendix, we seek to show that $\|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2 = O(\varepsilon^2)$ and $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 = O(\varepsilon^2)$ in the context of deriving the Hermite interpolation conditions (6.32d) in the proof of Theorem 6.9. Apply (5.46) and substitute directly into $\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}$ in (6.35b) to get

$$\begin{aligned}
\|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2 &= \mathbf{c}_k^\top \left(\tilde{\mathbf{G}}_{\text{lo}}(-\lambda_k) - \check{\mathbf{G}}_{\text{lo}}(-\lambda_k) - \left(\tilde{\mathbf{G}}_{\text{lo}}(-\eta_k) - \check{\mathbf{G}}_{\text{lo}}(-\eta_k) \right) \right) \mathbf{b}_k \\
&= \|\mathbf{c}_k\|_2^2 \|\mathbf{b}_k\|_2^2 \left(\frac{1}{-2\operatorname{Re}(\lambda_k)} - \frac{1}{-\bar{\lambda}_k - \eta_k} - \frac{1}{-\lambda_k - \bar{\eta}_k} + \frac{1}{-2\operatorname{Re}(\eta_k)} \right) \\
&= \|\mathbf{c}_k\|_2^2 \|\mathbf{b}_k\|_2^2 \left(\frac{\operatorname{Re}(\lambda_k + \eta_k) (4\operatorname{Re}(\lambda_k)\operatorname{Re}(\eta_k) - |\lambda_k + \bar{\eta}_k|^2)}{2\operatorname{Re}(\lambda_k)\operatorname{Re}(\eta_k)|\lambda_k + \bar{\eta}_k|^2} \right) \\
&= -\|\mathbf{c}_k\|_2^2 \|\mathbf{b}_k\|_2^2 \left(\frac{\operatorname{Re}(\lambda_k + \eta_k)|\lambda_k - \eta_k|^2}{2\operatorname{Re}(\lambda_k)\operatorname{Re}(\eta_k)|\lambda_k + \bar{\eta}_k|^2} \right) = O(\varepsilon^2),
\end{aligned}$$

since $|\lambda_k + \bar{\eta}_k|^2 - 4\operatorname{Re}(\lambda_k)\operatorname{Re}(\eta_k) = |\lambda_k - \eta_k|^2 = \varepsilon^2$ by our choice of η_k . Note that this term is in fact positive since $\operatorname{Re}(\lambda_k + \eta_k) < 0$ by asymptotic stability. We next show that $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 = O(\varepsilon^2)$. To make the calculations more compact, we introduce the notation $\mathbf{H}_{\text{qo}} \stackrel{\text{def}}{=} \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}$ and $\mathbf{b}_{i,j} \stackrel{\text{def}}{=} (\mathbf{b}_i \otimes \mathbf{b}_j) \in \mathbb{C}^{1 \times m^2}$. Observe that

$$\begin{aligned}
\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2 &= \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \left(\bar{\mathbf{H}}_{\text{qo}}(-\lambda_i, -\lambda_k) - \bar{\mathbf{H}}_{\text{qo}}(-\lambda_i, -\eta_k) \right) \mathbf{b}_{i,k} \\
&\quad + \sum_{j \neq k}^r \mathbf{m}_{j,k}^\top \left(\bar{\mathbf{H}}_{\text{qo}}(-\lambda_k, -\lambda_j) - \bar{\mathbf{H}}_{\text{qo}}(-\lambda_k, -\eta_j) \right) \mathbf{b}_{k,j} \\
&\quad + \mathbf{m}_{k,k}^\top \left(\bar{\mathbf{H}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \bar{\mathbf{H}}_{\text{qo}}(-\eta_k, -\eta_k) \right) \mathbf{b}_{k,k} \tag{1} \\
&= 2 \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \left(\bar{\mathbf{H}}_{\text{qo}}(-\lambda_i, -\lambda_k) - \bar{\mathbf{H}}_{\text{qo}}(-\lambda_i, -\eta_k) \right) \mathbf{b}_{i,k} \\
&\quad + \mathbf{m}_{k,k}^\top \left(\bar{\mathbf{H}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \bar{\mathbf{H}}_{\text{qo}}(-\eta_k, -\eta_k) \right) \mathbf{b}_{k,k}
\end{aligned}$$

by (5.44) and (5.14). The notation $\sum_{i \neq k}^r$ means that the summation runs over all $i = 1, \dots, k-1, k+1, \dots, r$. Substituting directly into the expression for $\mathbf{H}_{\text{qo}} = \tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}$

in (6.35b), the first term in the expansion above becomes

$$\begin{aligned}\gamma_\star &\stackrel{\text{def}}{=} 2 \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \left(\overline{\mathbf{H}}_{\text{qo}}(-\lambda_i, -\lambda_k) - \overline{\mathbf{H}}_{\text{qo}}(-\lambda_i, -\eta_k) \right) \mathbf{b}_{i,k} \\ &= 2 \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \left[\gamma_1 \sum_{i \neq k}^r \mathbf{m}_{i,k}^\top \sum_{\ell \neq k}^r \frac{\overline{\mathbf{m}}_{\ell,k} \overline{\mathbf{b}}_{\ell,k}^\top}{-\lambda_i - \overline{\lambda}_\ell} + \sum_{\ell \neq k}^r \mathbf{m}_{k,\ell} \mathbf{b}_{k,\ell}^\top \gamma_2^{(i,\ell)} + \gamma_3^{(i)} \overline{\mathbf{m}}_{k,k} \overline{\mathbf{b}}_{k,k}^\top \right] \mathbf{b}_{i,k},\end{aligned}\tag{2a}$$

where the terms $\gamma_1, \gamma_2^{(i,\ell)}$, and $\gamma_3^{(i)}$ are given by

$$\gamma_1 = \frac{1}{-2 \operatorname{Re}(\lambda_k)} - \frac{1}{-\overline{\lambda}_k - \eta_k} - \frac{1}{-\lambda_k - \overline{\eta}_k} + \frac{1}{-2 \operatorname{Re}(\eta_k)},\tag{2b}$$

$$\gamma_2^{(i,\ell)} = \left(\frac{1}{-\lambda_i - \overline{\lambda}_k} - \frac{1}{-\lambda_i - \overline{\eta}_k} \right) \left(\frac{1}{-\lambda_k - \overline{\lambda}_\ell} - \frac{1}{-\eta_k - \overline{\lambda}_\ell} \right),\tag{2c}$$

$$\gamma_3^{(i)} = \frac{1}{-\lambda_i - \overline{\lambda}_k} \left(\frac{1}{-2 \operatorname{Re}(\lambda_k)} - \frac{1}{-\overline{\lambda}_k - \eta_k} \right) + \frac{1}{-\lambda_i - \overline{\eta}_k} \left(\frac{1}{-2 \operatorname{Re}(\eta_k)} - \frac{1}{-\lambda_k - \overline{\eta}_k} \right),\tag{2d}$$

for all $i \neq k$ and $\ell \neq k$. Likewise, the second term in (1) can be expressed as

$$\begin{aligned}\xi_\star &\stackrel{\text{def}}{=} \mathbf{m}_{k,k}^\top \left(\overline{\mathbf{H}}_{\text{qo}}(-\lambda_k, -\lambda_k) - \overline{\mathbf{H}}_{\text{qo}}(-\eta_k, -\eta_k) \right) \mathbf{b}_{k,k} \\ &= \mathbf{m}_{k,k}^\top \left[\sum_{\ell \neq k}^r \left(\overline{\mathbf{m}}_{\ell,k} \overline{\mathbf{b}}_{\ell,k}^\top + \overline{\mathbf{m}}_{k,\ell} \overline{\mathbf{b}}_{k,\ell}^\top \right) \xi_1^{(\ell)} + \overline{\mathbf{m}}_{k,k} \overline{\mathbf{b}}_{k,k}^\top \xi_2 \right] \mathbf{b}_{k,k},\end{aligned}\tag{2e}$$

where the terms $\xi_1^{(\ell)}$ and ξ_2 are given by

$$\xi_1^{(\ell)} = \left(\frac{1}{-2 \operatorname{Re}(\lambda_k)} - \frac{1}{-\lambda_k - \overline{\eta}_k} \right) \frac{1}{-\lambda_k - \overline{\lambda}_\ell} + \left(\frac{1}{-2 \operatorname{Re}(\eta_k)} - \frac{1}{-\overline{\lambda}_k - \eta_k} \right) \frac{1}{-\eta_k - \overline{\lambda}_\ell},\tag{2f}$$

$$\xi_2 = \frac{1}{4 \operatorname{Re}(\lambda_k)^2} - \frac{1}{(-\lambda_k - \overline{\eta}_k)^2} - \frac{1}{(-\eta_k - \overline{\lambda}_k)^2} + \frac{1}{4 \operatorname{Re}(\eta_k)^2},\tag{2g}$$

for $\ell \neq k$. The calculations required to resolve $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2$ as $O(\varepsilon^2)$ are direct but tedious. We do so by proving that the factors $\gamma_1, \gamma_2^{(i,\ell)}, \gamma_3^{(i)}, \xi_1^{(i)}, \xi_2$ defined in (2b)–(2g) are all $O(\varepsilon^2)$ for each $i, \ell \neq k$. Because every term in the expansion of the error $\|\tilde{\mathbf{G}}_{\text{qo}} - \check{\mathbf{G}}_{\text{qo}}\|_{\mathcal{H}_2}^2$ is a multiple of one of these, it follows that the error is $O(\varepsilon^2)$. We begin by observing that γ_1 in (2b) is precisely the term appearing in $\|\tilde{\mathbf{G}}_{\text{lo}} - \check{\mathbf{G}}_{\text{lo}}\|_{\mathcal{H}_2}^2$, and so $\gamma_1 = O(\varepsilon^2)$ by this previous calculation. For $\gamma_2^{(i,\ell)}$ in (2c), observe first that

$$\begin{aligned}\frac{1}{-\lambda_i - \overline{\lambda}_k} - \frac{1}{-\lambda_i - \overline{\eta}_k} &= \frac{\overline{\lambda}_k - \overline{\eta}_k}{(-\lambda_i - \overline{\lambda}_k)(-\lambda_i - \overline{\eta}_k)} \\ \text{and } \frac{1}{-\lambda_k - \overline{\lambda}_\ell} - \frac{1}{-\eta_k - \overline{\lambda}_\ell} &= \frac{\lambda_k - \eta_k}{(-\lambda_k - \overline{\lambda}_\ell)(-\eta_k - \overline{\lambda}_\ell)}\end{aligned}$$

for all $i \neq k$ and $\ell \neq k$. Thus,

$$\begin{aligned}\gamma_2^{(i,\ell)} &= \left(\frac{1}{-\lambda_i - \bar{\lambda}_k} - \frac{1}{-\lambda_i - \bar{\eta}_k} \right) \left(\frac{1}{-\lambda_k - \bar{\lambda}_\ell} - \frac{1}{-\eta_k - \bar{\lambda}_\ell} \right) \\ &= \frac{|\lambda_k - \eta_k|^2}{(-\lambda_i - \bar{\lambda}_k)(-\lambda_i - \bar{\eta}_k)(-\lambda_k - \bar{\lambda}_\ell)(-\eta_k - \bar{\lambda}_\ell)} = O(\varepsilon^2).\end{aligned}$$

For $\gamma_3^{(i)}$ in (2d), first define

$$\begin{aligned}\gamma_4 &\stackrel{\text{def}}{=} \frac{1}{-2 \operatorname{Re}(\lambda_k)} - \frac{1}{-\bar{\lambda}_k - \eta_k} = \frac{|\lambda_k + \bar{\eta}_k|^2 - 2 \operatorname{Re}(\lambda_k) (\lambda_k + \bar{\eta}_k)}{2 \operatorname{Re}(\lambda_k) |\lambda_k + \bar{\eta}_k|^2} \\ \gamma_5 &\stackrel{\text{def}}{=} \frac{1}{-2 \operatorname{Re}(\eta_k)} - \frac{1}{-\bar{\eta}_k - \lambda_k} = \frac{|\lambda_k + \bar{\eta}_k|^2 - 2 \operatorname{Re}(\eta_k) (\bar{\lambda}_k + \eta_k)}{2 \operatorname{Re}(\eta_k) |\lambda_k + \bar{\eta}_k|^2}.\end{aligned}$$

Now $\gamma_3^{(i)}$ can be written as

$$\gamma_3^{(i)} = \frac{\gamma_4}{-\lambda_i - \bar{\lambda}_k} + \frac{\gamma_5}{-\lambda_i - \bar{\eta}_k} = \frac{(-\lambda_i (\gamma_4 + \gamma_5) - (\bar{\eta}_k \gamma_4 + \bar{\lambda}_k \gamma_5))}{(-\lambda_i - \bar{\lambda}_k) (-\lambda_i - \bar{\eta}_k)}.$$

Direct calculations reveal that

$$\begin{aligned}-\lambda_i (\gamma_4 + \gamma_5) &= \frac{\lambda_i \operatorname{Re}(\lambda_k + \eta_k) (|\lambda_k + \bar{\eta}_k|^2 - 4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k))}{2 |\lambda_k + \bar{\eta}_k|^2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k)} = \frac{\lambda_i \operatorname{Re}(\lambda_k + \eta_k) |\lambda_k - \eta_k|^2}{2 |\lambda_k + \bar{\eta}_k|^2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k)} \\ &= O(\varepsilon^2).\end{aligned}$$

More involved, but very similar calculations using the fact that $\eta_k = \lambda_k + \varepsilon e^{i\theta}$ reveal that

$$\begin{aligned}\bar{\eta}_k \gamma_4 + \bar{\lambda}_k \gamma_5 &= \frac{\bar{\lambda}_k \operatorname{Re}(\lambda_k + \eta_k) (4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) - |\lambda_k + \bar{\eta}_k|^2)}{2 |\lambda_k + \bar{\eta}_k|^2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k)} \\ &\quad + \varepsilon e^{-i\theta} \frac{\operatorname{Re}(\eta_k) (\lambda_k + \bar{\eta}_k) (2 \operatorname{Re}(\lambda_k) - (\bar{\lambda}_k + \eta_k))}{2 |\lambda_k + \bar{\eta}_k|^2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k)} \\ &= \frac{\bar{\lambda}_k \operatorname{Re}(\lambda_k + \eta_k) |\lambda_k - \eta_k|^2}{2 |\lambda_k + \bar{\eta}_k|^2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k)} \\ &\quad + \varepsilon e^{-i\theta} \frac{\operatorname{Re}(\eta_k) (\lambda_k + \bar{\eta}_k) (\lambda_k - \eta_k)}{2 |\lambda_k + \bar{\eta}_k|^2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k)} = O(\varepsilon^2),\end{aligned}$$

because $|\lambda_k - \eta_k|^2 = \varepsilon^2$ and $\lambda_k - \eta_k = \varepsilon e^{i\theta}$. This proves that $\gamma_3^{(i)}$ in (2d) is $O(\varepsilon^2)$ and thus γ_* in (2a) is $O(\varepsilon^2)$. We observe that $\xi_1^{(\ell)}$ in (2f) is the complex conjugate of $\gamma_3^{(i)}$ in (2d) with λ_ℓ taking the place of λ_i , and so $\xi_1^{(\ell)}$ is $O(\varepsilon^2)$ for all $\ell \neq k$. This just leaves ξ_2 in (2g). We

start by combining the individual terms in (2g) over a single denominator

$$\begin{aligned}\xi_2 &= \frac{1}{4 \operatorname{Re}(\lambda_k)^2} - \frac{1}{(-\lambda_k - \bar{\eta}_k)^2} - \frac{1}{(-\eta_k - \bar{\lambda}_k)^2} + \frac{1}{4 \operatorname{Re}(\eta_k)^2} \\ &= \frac{|\lambda_k + \bar{\eta}_k|^4 (\operatorname{Re}(\lambda_k)^2 + \operatorname{Re}(\eta_k)^2) - 32 \operatorname{Re}(\lambda_k)^3 \operatorname{Re}(\eta_k)^3 - 8 |\eta_k - \lambda_k|^2 \operatorname{Re}(\lambda_k)^2 \operatorname{Re}(\eta_k)^2}{4 \operatorname{Re}(\lambda_k)^2 \operatorname{Re}(\eta_k)^2 |\lambda_k + \bar{\eta}_k|^4}.\end{aligned}$$

One can expand $|\lambda_k + \bar{\eta}_k|^4 = (4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) + |\eta_k - \lambda_k|^2)^2$, and so the numerator in the above expression can be written as

$$\begin{aligned}& |\lambda_k + \bar{\eta}_k|^4 (\operatorname{Re}(\lambda_k)^2 + \operatorname{Re}(\eta_k)^2) - 32 \operatorname{Re}(\lambda_k)^3 \operatorname{Re}(\eta_k)^3 - 8 |\eta_k - \lambda_k|^2 \operatorname{Re}(\lambda_k)^2 \operatorname{Re}(\eta_k)^2 \\ &= (4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) + |\eta_k - \lambda_k|^2)^2 (\operatorname{Re}(\lambda_k)^2 + \operatorname{Re}(\eta_k)^2) \\ &\quad - 8 (4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) + |\eta_k - \lambda_k|^2) (\operatorname{Re}(\lambda_k)^2 \operatorname{Re}(\eta_k)^2) \\ &= (4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) + |\eta_k - \lambda_k|^2) ((4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) + |\eta_k - \lambda_k|^2) \\ &\quad \times (\operatorname{Re}(\lambda_k)^2 + \operatorname{Re}(\eta_k)^2) - 8 \operatorname{Re}(\lambda_k)^2 \operatorname{Re}(\eta_k)^2).\end{aligned}$$

Thus, the numerator in the expression for ξ_2 becomes

$$\begin{aligned}& (4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) + |\eta_k - \lambda_k|^2) (\operatorname{Re}(\lambda_k)^2 + \operatorname{Re}(\eta_k)^2) - 8 \operatorname{Re}(\lambda_k)^2 \operatorname{Re}(\eta_k)^2 \\ &= O(\varepsilon^2) + 4 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k) \underbrace{(\operatorname{Re}(\lambda_k)^2 + \operatorname{Re}(\eta_k)^2 - 2 \operatorname{Re}(\lambda_k) \operatorname{Re}(\eta_k))}_{= \operatorname{Re}(\lambda_k - \eta_k)^2 = O(\varepsilon^2)}.\end{aligned}$$

The $O(\varepsilon^2)$ term comes from those multiplied by $|\eta_k - \lambda_k|^2$. Thus, ξ_2 in (2g) is $O(\varepsilon^2)$, and we have that $\|\tilde{\mathbf{G}}_{\mathbf{q}_0} - \check{\mathbf{G}}_{\mathbf{q}_0}\|_{\mathcal{H}_2}^2 = O(\varepsilon^2)$ as claimed.