Lehrstuhl für Regelungstechnik

Technische Universität München

$\mathcal{H}_2$ PSEUDO-OPTIMAL MODEL ORDER REDUCTION

Thomas Wolf

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. dr. ir. Daniel J. Rixen
Prüfer der Dissertation:	Univ.-Prof. Dr.-Ing. habil. Boris Lohmann
	Univ.-Prof. Dr. Tatjana Stykel, Universität Augsburg

Die Dissertation wurde am 28.08.2014 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 13.11.2014 angenommen.

Abstract

The cornerstone of the thesis is to motivate and comprehensively describe $\mathcal{H}_2$ pseudo-optimality in the linear model order reduction of large-scale dynamical systems based on projections onto rational Krylov subspaces. The prefix “pseudo” relates to global optimality within a certain subset of all possible reduced models. It is demonstrated how $\mathcal{H}_2$ pseudo-optimality may be enforced with marginal computational effort, and its consequences, benefits, and possible applications are thoroughly discussed. Moreover, necessary and sufficient conditions for $\mathcal{H}_2$ pseudo-optimality are formulated together with small-scale and easy-to-evaluate matrix equations, which in turn constitute the main tool for the analysis and construction of $\mathcal{H}_2$ pseudo-optimal reduced models.

$\mathcal{H}_2$ pseudo-optimality is shown to be a natural extension of a cumulative framework for model order reduction, denoted as “CURE”, which permits the accumulation of independently reduced models and at the same time the preservation of the flexibility that projections onto rational Krylov subspaces offer. The additional numerical effort of CURE is marginal compared to the computation of Krylov subspaces and the main benefit of $\mathcal{H}_2$ pseudo-optimal reductions within CURE is that this ensures a monotonic decrease of the approximation error.

Although the results of this research are intended to improve model order reduction using projections onto rational Krylov subspaces, they may also be exploited to approximately solve large-scale Lyapunov equations. It is shown that applying the ideas of $\mathcal{H}_2$ pseudo-optimality to Lyapunov equations actually results in the same approximation as one would obtain from the widely-used ADI iteration. This thesis therefore not only provides a novel view on the ADI iteration, but it also offers tools for the analysis and improvement of the ADI iteration, which include a low-rank formulation of the residual and also the generalization of the ADI iteration to so-called tangential interpolation.

The main tool in this work certainly are particular Sylvester equations, which constitute some kind of duality to bases of rational Krylov subspaces, because basically all proofs in this thesis emanate from the detailed understanding of these equations.

Acknowledgements

I would like to thank all those who made this doctoral thesis possible. First of all, many thanks go to my adviser Prof. Boris Lohmann for his helpful advices. The freedom in the choice of the research direction and his encouragement of individual strength made this thesis possible. I also would like to thank Prof. Tatjana Stykel for her interest in this work and for being my second examiner.

My deepest gratitude goes out to my second to none co-worker and colleague Heiko Panzer. He is the one who contributed the most to this thesis, due to the endless conversations we had over the past few years, both on and off topic, while sharing an office—all of which is gratefully acknowledged. It was a pleasure to share the office also with Dr. Rudy Eid, who introduced Heiko and me to the field of model order reduction. Both of them contributed significantly to my education, for which I am very grateful. I would also like to thank Alessandro Castagnotto for carefully reading the manuscript, and furthermore, I thank all other colleagues for turning the Institute of Automatic Control into the extraordinary place it has been throughout these years.

I would also like to thank Dr. Jens Saak, Patrick Kürschner, and Matthias Voigt from the Max Planck Institute in Magdeburg for pointing me to the ADI iteration, and for all the fruitful discussions we had during our meetings at various conferences and workshops.

Last but definitely not least, I would like to thank Fanny & Alvin, for all their love and support they gave to me.

Contents

Preface

A tremendous technical progress has been made throughout the past decades in various scientific domains, such as mechanical and electrical engineering, physics, chemistry and economics. Dynamical models that are described in a particular mathematical structure made an important contribution to this progress: they allow the analysis and numerical simulation of physical phenomena, whereby time-consuming and costly experiments or prototypes may be avoided.

The enduring process of this technical progress generates the need for ever detailed models of increasing complexity. This can render their numerical simulation impossible because of limited accuracy and storage capabilities. One remedy is to employ techniques of model order reduction that try to approximate accurate large-scale models by ones of reduced order, and thereby gather their most dominant characteristics.

This work is concerned with the efficient computation of reduced models for given linear and large-scale dynamical systems. The cornerstone of this work is to motivate and comprehensively describe a new concept for linear model order reduction: $\mathcal{H}_2$ *pseudo-optimality*. The notation "pseudo" stems from the fact, that optimality in a particular subset can always be achieved for reduced order models—with negligible computational effort. This does not mean that a pseudo-optimal reduced model is a good approximation, it just implies that it is *the* optimal model in its respective subset. The benefit is that instead of searching for a "good" (whatever that means) reduced order model, one can search for an optimal subset—and then just pick the pseudo-optimal reduced model out of this subset. This thesis proposes a numerically efficient way to "pick the pseudo-optimal reduced model" out of a given subset, whereas the search for a "good" subset, which actually boils down to some sort of optimization technique, is not focussed on and reference is made instead to the literature, where appropriate.

Certainly, I do not claim the invention of *pseudo-optimality*. The basic idea goes back to at least the 1920s, and various labels, including "least-squares solution" and "sub-optimal reduction", have appeared in the rich literature on this topic. Even more, this research uncovers that prevailing methods for model order reduction have been implicitly using this concept—without even noticing, but nevertheless, in quite a sophisticated way. Most of the available literature, however, is not applicable to large-scale systems, because of numerically inappropriate algorithms. Furthermore, it seems like neither a detailed analysis of pseudo-optimality is available, nor has it been thoroughly embedded in a large-scale setting.

This work aims to fill this gap by integrating the concept of $\mathcal{H}_2$ pseudo-optimality in the framework of projective model order reduction using rational Krylov subspaces. To achieve this objective and ensure good readability, a simple label is essential. Therefore, the phrase "pseudo-optimality" should not be understood as a fact, it is instead used in this work to name a particular concept. The inaccuracy of the word "pseudo" is acknowledged, nevertheless, it is employed here for lack of a better alternative.

This dissertation would not be the same without the contributions of my colleague Heiko Panzer. All this started with the joint development of the error factorization presented in Section 3.1. In retrospect, it is hard to tell who added which jigsaw piece to the final formulation, but the idea of the resultant cumulative framework with its convenient formulation of the reduced matrices (as presented in Section 3.2) was instead the effort of Heiko Panzer. It is honestly hard to pinpoint Heiko Panzer's contributions to other aspects of this thesis due to the continuous exchange of ideas over the past few years. On this account, this thesis and the one of Heiko Panzer, [148], are sort of kindred because both of them build upon the very same ideas. They, however, have a different focus: this thesis is concerned with the analysis of Sylvester equations and of pseudo-optimality, and the solution of large-scale Lyapunov equations, whereas the thesis of Heiko Panzer deals with the application of these results with aim of a fully automatic and error-controlled model order reduction scheme for large-scale systems.

This document is divided into three main parts. In Part I, model order reduction and relevant preliminaries are reviewed. In particular, the balanced truncation and projections by rational Krylov subspaces are discussed, and in the end, the problem which this thesis addresses is specified. The theoretical contributions of this dissertation are contained in Part II: the duality of rational Krylov subspaces and certain Sylvester equations is discussed, an error factorization is proposed that yields a cumulative framework for model order reduction, and the concept of $\mathcal{H}_2$ pseudo-optimality is thoroughly described. Finally, these contributions are then applied to the approximate solution of large-scale Lyapunov equations in Part III.

August 2014
Thomas Wolf

Glossary

Abbreviations

ADI	Alternating direction implicit
DAE	Differential algebraic equations
EKSM	Extended Krylov subspace method
LSE	Linear system of equations
LTI	Linear time-invariant
MIMO	Multiple-input multiple-output
MOR	Model order reduction
RK	Rational Krylov
RKSM	Rational Krylov subspace method
SISO	Single-input single-output
TBR	Truncated balanced realization

List of Symbols

$\mathbb{N}$	Set of natural numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{C}$	Set of complex numbers
$\mathrm{Re}(\alpha)$	Real part of a complex number α
$\mathrm{Im}(\alpha)$	Imaginary part of a complex number α
$\imath$	Complex unit, $\imath = \sqrt{-1}$
s	Complex number in $\mathbb{C}$

$\mathbf{A}$	Matrix of the dynamical system
$\mathbf{A}_r$	Reduced approximation of $\mathbf{A}$
$\mathbf{A}_s$	Shifted matrix $\mathbf{A} - s\mathbf{E}$ with $s \in \mathbb{C}$
$\mathbf{A}_{ij}$	Entry of the matrix $\mathbf{A}$ in row i and column j
$\overline{\mathbf{A}}$	Complex conjugate of $\mathbf{A}$
$\mathbf{A}^T$	Transpose of $\mathbf{A}$
$\mathbf{A}^*$	Complex conjugate transpose of $\mathbf{A}$
$\lambda_i\left(\mathbf{A}\right)$	i^{th} eigenvalue of the square matrix $\mathbf{A}$
$\Lambda\left(\mathbf{A}\right)$	Set of eigenvalues of the square matrix $\mathbf{A}$
$\operatorname{rank}\left(\mathbf{A}\right)$	Rank of $\mathbf{A}$
$\operatorname{span}\left(\mathbf{A}\right)$	Subspace generated by the columns of $\mathbf{A}$
$\operatorname{trace}\left(\mathbf{A}\right)$	Trace of the square matrix $\mathbf{A}$
$\operatorname{diag}\left(a_1, a_2, \ldots, a_N\right)$	Square matrix of dimension $N \in \mathbb{N}$, with the a_i's on the diagonal, and zero elsewhere
$\boldsymbol{G}$	Original model of an LTI dynamical system, order N
$\boldsymbol{G}_r$	Reduced order model, order $n \ll N$
$\boldsymbol{G}_e$	Error model $\boldsymbol{G}_e(s) = \boldsymbol{G}(s) - \boldsymbol{G}_r(s)$, order $N + n$
$\boldsymbol{G}_\perp$	Model in the error factorization $\boldsymbol{G}_e(s) = \boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$, with perpendicular input, order N
$\boldsymbol{G}_f$	Feed-through model in error factorization $\boldsymbol{G}_e(s) = \boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$, order n
$\dim(\boldsymbol{G})$	Order of any minimal realization of $\boldsymbol{G}(s)$ in state-space, or equivalently, the McMillan degree of $\boldsymbol{G}(s)$
$\lVert\cdot\rVert_2$	Euclidean norm of a vector, or the induced norm of a matrix
$\lVert\cdot\rVert_{\mathrm{F}}$	Frobenius norm of a matrix
$\lVert\cdot\rVert_{\mathcal{H}_2}$	$\mathcal{H}_2$ norm of a dynamical system
$\lVert\cdot\rVert_{\mathcal{H}_\infty}$	$\mathcal{H}_\infty$ norm of a dynamical system

Part I

Preliminary Results

1 Introduction

For the numerical simulation of dynamical systems in an engineering field, it is often sufficiently accurate to employ *linear time-invariant* (LTI) models, which are usually derived from linearisation at the operating point of interest. In this work, only LTI models are considered. To this end, the notation for LTI models is introduced in this chapter, and two widely-used methods for their reduction are reviewed: *Truncated Balanced Realization* (TBR) and *Moment Matching* via *Krylov subspaces.*

1.1 Linear Time-Invariant (LTI) Systems

The time-domain realization of those LTI dynamical systems that are considered in this work generally reads as

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t), \end{aligned} \tag{1.1}$$

where $\mathbf{x}(t) \in \mathbb{R}^N$, $\mathbf{u}(t) \in \mathbb{R}^m$ and $\mathbf{y}(t) \in \mathbb{R}^p$ denote the state, input and output, respectively. In a large-scale setting, the order N reaches values of $N = 10^2, \ldots, 10^6$, whereas the number of inputs and outputs, m and p, respectively, is assumed to be small, i. e. up to a few tens. The model (1.1) then realizes a *multi-input multi-output* (MIMO) model, whereas the case $m = q = 1$ is denoted as *single-input single-output* (SISO). Consequently, the dynamics of a MIMO model are described by the matrices $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{N \times N}$, $\mathbf{B} \in \mathbb{R}^{N \times m}$ and $\mathbf{C} \in \mathbb{R}^{p \times N}$, whereas in the SISO case $\mathbf{B} \to \mathbf{b} \in \mathbb{R}^N$ and $\mathbf{C} \to \mathbf{c} \in \mathbb{R}^{1 \times N}$, and $\mathbf{u}(t) \to u(t) \in \mathbb{R}$ and $\mathbf{y}(t) \to y(t) \in \mathbb{R}$.

The matrix $\mathbf{E}$ is assumed non-singular, which means that the model does not contain algebraic constraints. Even though this would always allow to get rid of the $\mathbf{E}$ matrix by multiplying the state equation of (1.1) with its inverse from the left, this should usually be avoided in a large scale-setting due to numerical reasons, because all methods that are investigated in this work can also be generalized to an $\mathbf{E}$ matrix. As a consequence, only multiplications with $\mathbf{E}$ have to be carried out in the algorithms—instead of solving linear systems with $\mathbf{E}$. Working with the generalized form (1.1) thus may reduce the numerical effort and improve the conditioning. It should be noted that the inverse $\mathbf{E}^{-1}$

is still used in this work, but only in theoretical contexts, whereas applications are all implemented without it.

All large-scale models are assumed to be completely controllable and observable, or equivalently, the pair $(\mathbf{E}^{-1}\mathbf{A}, \mathbf{E}^{-1}\mathbf{B})$ is assumed controllable, whereas the pair $(\mathbf{C}, \mathbf{E}^{-1}\mathbf{A})$ is assumed observable. This means that (1.1) is a minimal realization. It is additionally assumed that the set of eigenvalues, denoted as $\Lambda\left(\mathbf{E}^{-1}\mathbf{A}\right)$, is contained in the open left half of the complex plane, or in other words: the system is asymptotically stable.

The transfer function $\boldsymbol{G}(s)$ of (1.1) in the Laplace domain is given by

$$\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B}, \tag{1.2}$$

and analogously, the impulse response of (1.1) in the time domain is

$$\boldsymbol{G}(t) = \mathbf{C}\mathrm{e}^{\mathbf{E}^{-1}\mathbf{A}t}\mathbf{E}^{-1}\mathbf{B}. \tag{1.3}$$

For the ease of presentation, notation is slightly abused by letting $\boldsymbol{G}$ denote the system itself, as well as its state-space realization (1.1), its transfer function (1.2) and its impulse response (1.3). The particular meaning should become clear from the context.

Excellent textbooks on LTI systems are available, such as e. g. [52, 114, 222], whose elaborateness cannot be achieved here. On this account, the reader is referred to the textbook of his choice for details on the important concepts of stability, controllability, observability, and minimality, and for additional aspects of linear systems theory.

Throughout this work, we use the following notation: $\mathbb{N}$ denotes the set of natural numbers, $\mathbb{R}$ the set of real numbers, and $\mathbb{C}$ the set of complex numbers. The real and imaginary part of a complex number are given by $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$, respectively, and $\imath = \sqrt{-1}$ denotes the complex unit. Matrices are denoted by upper case letters, and vectors by lower case letters, both in upright boldface type, like e. g. $\mathbf{A}$ and $\mathbf{a}$, respectively. Scalars are denoted by upper and lower case letters in italic type, like e. g. A and a. For a better distinction, transfer functions in Laplace domain are printed in italic boldface type, like e. g. $\boldsymbol{A}(s)$ and $\boldsymbol{a}(s)$, where $s \in \mathbb{C}$. The transpose of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}^T$, its complex conjugate by $\overline{\mathbf{A}}$, and $\mathbf{A}^*$ means transposition with complex conjugation. The entry in row i and column j of a matrix $\mathbf{A}$ is accessed by $\mathbf{A}_{ij}$. $\Lambda\left(\mathbf{A}\right)$ denotes the set of eigenvalues, whereas $\lambda_i\left(\mathbf{A}\right)$ refers only to the i^{th} eigenvalue. The rank of a matrix is denoted by $\mathrm{rank}(\mathbf{A})$, whereas $\mathrm{span}(\mathbf{A})$ denotes the subspace that is generated by the columns of $\mathbf{A}$. The trace of a square matrix is denoted by $\mathrm{trace}(\mathbf{A})$. The matrix $\mathbf{I}$ always denotes the identity matrix, and unless specified, its dimensions should become clear from the context. $\mathrm{diag}(\mathbf{A})$ denotes all entries on the diagonal of a square matrix, whereas diag $(a_1, a_2, \dots, a_n)$ equals a square matrix of dimension n, with a_i on the diagonal and zero elsewhere. Generally, data referring to a reduced model are denoted with an index "r", like e. g. $\mathbf{A}_r$.

1.2 Model Order Reduction

Model order reduction (MOR) seeks for a simpler model than the original one (1.1), while at the same time the dynamical behaviour should be well approximated. Simpler, in this context, means being of reduced order, i. e. having fewer state variables. Accordingly, the state-space realization of a reduced model generally takes the form

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

where the state $\mathbf{x}_r(t) \in \mathbb{R}^n$ is of lower dimension $n \ll N$, and the output $\mathbf{y}_r(t) \in \mathbb{R}^p$ is determined by the matrices $\mathbf{A}_r, \mathbf{E}_r \in \mathbb{R}^{n\times n}$, $\mathbf{B}_r \in \mathbb{R}^{n\times m}$ and $\mathbf{C}_r \in \mathbb{R}^{p\times n}$.

Remark 1.1. A dynamical system may additionally feature a feed-through term, such that the output equation in (1.1) would change to $\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)$, with $\mathbf{D} \in \mathbb{R}^{p\times m}$, whereas the transfer function would take the form $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B} + \mathbf{D}$. As MOR is concerned with the approximation of the transfer behaviour, a reduced model would usually feature an equal feed-through term $\mathbf{D}_r = \mathbf{D}$, that is, the output equation in (1.4) would change to $\mathbf{y}_r(t) = \mathbf{C}_r\mathbf{x}_r(t) + \mathbf{D}_r\mathbf{u}(t)$. Feed-through terms are therefore ignored, as they would not change the results in this work. It, however, should be noted, that $\mathbf{D}_r \neq \mathbf{D}$ may still be used as an additional degree of freedom in order to improve the approximation, cf. [66].

The benefit of the reduced model (1.4) is that it requires both less storage and evaluation time. In the context of control theory, the lower complexity of the reduced model (1.4) also facilitates designing a controller. Furthermore, a reduced control law can be faster evaluated, such that simpler and cheaper hardware may be used. The design of a controller based on a reduced model, however, requires additional attention, cf. [6, 20, 88, 186, 192], which is out of the scope of this work.

The development and parameter optimization of technical systems can be improved by methods of *parametric* MOR. There, not merely a single large-scale system has to be approximated, but rather a family of large-scale systems over a whole range of parameter values. Many researchers tackled this problem in the past few years, which, however, is also out of the scope of this work. The interested reader is instead referred to the recent survey [30] and the references therein.

On that account, the main goals of MOR in this work can be summarized as follows:

- First of all, $\mathbf{y}_r(t)$ should approximate $\mathbf{y}(t)$ well. Accordingly, the error $\boldsymbol{G}(s)$–$\boldsymbol{G}_r(s)$, where $\boldsymbol{G}_r(s) = \mathbf{C}_r \left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1} \mathbf{B}_r$ is the transfer function of (1.4), should be small. This is sometimes required only in a certain frequency range of interest, cf. [7, 59, 80, 82, 90, 202].
- Another goal of MOR is preserving structural properties of the original model in the reduced system, such as stability [86, 109], passivity [72, 145, 160], a second-order structure [45, 51, 159, 172] or a port-Hamiltonian representation [94, 206].
- Additionally, the method to be used should be numerically efficient and stable,
- and finally, a quantitative information on the error should be available; at least an upper bound is desirable unless the error itself may be computed.

It is a hard task to entirely achieve all these goals. Nevertheless, the state-of-the-art methods for model order reduction can at least partly achieve the above mentioned objectives; and as they may be subsumed in a projective framework, the following section reviews model order reduction based on projections.

1.3 Projective Model Order Reduction

Assume there exists an n-dimensional subspace, with $n \ll N$, that contains the most dominant dynamics of (1.1). Let $\mathbf{V} \in \mathbb{R}^{N \times n}$ have full column rank and assume that the columns of $\mathbf{V}$ form a basis of this subspace, then the approximation $\mathbf{x}(t) \approx \mathbf{V}\mathbf{x}_r(t)$ is admissible. The resulting error $\mathbf{e}(t)$ is defined by $\mathbf{x}(t) = \mathbf{V}\mathbf{x}_r(t) + \mathbf{e}(t)$, which can be substituted in the state equation of the original model (1.1),

$$\mathbf{E}\mathbf{V}\dot{\mathbf{x}}_r(t) = \mathbf{A}\mathbf{V}\mathbf{x}_r(t) + \mathbf{B}\mathbf{u}(t) + \boldsymbol{\epsilon}(t), \tag{1.5}$$

where the residual $\boldsymbol{\epsilon}(t) = \mathbf{A}\mathbf{e}(t)$–$\mathbf{E}\dot{\mathbf{e}}(t)$ contains all errors. The differential equation (1.5) is overdetermined: N equations for the n unknowns in $\mathbf{x}_r(t)$. This may be resolved by projecting the whole equation onto the subspace span($\mathbf{E}\mathbf{V}$). To this end, let $\mathbf{W} \in \mathbb{R}^{N \times n}$ have full column rank and assume that $\mathbf{W}^T\mathbf{E}\mathbf{V}$ is non-singular. Then a projector $\mathbf{\Pi} = \mathbf{\Pi}^2 \in \mathbb{R}^{N \times N}$ can be defined by $\mathbf{\Pi} = \mathbf{E}\mathbf{V}\left(\mathbf{W}^T\mathbf{E}\mathbf{V}\right)^{-1}\mathbf{W}^T$. In this respect, the matrices $\mathbf{V}$ and $\mathbf{W}$ will be called "projection matrices" hereafter because they generate the projector $\mathbf{\Pi}$. Multiplying (1.5) with $\mathbf{\Pi}$ from the left, yields

$$\mathbf{\Pi}\mathbf{E}\mathbf{V}\dot{\mathbf{x}}_r(t) = \mathbf{\Pi}\mathbf{A}\mathbf{V}\mathbf{x}_r(t) + \mathbf{\Pi}\mathbf{B}\mathbf{u}(t) + \mathbf{\Pi}\boldsymbol{\epsilon}(t). \tag{1.6}$$

The so-called *Petrov-Galerkin condition* is defined as $\boldsymbol{\epsilon}(t) \perp \mathbf{W}$, which implies $\boldsymbol{\Pi}\boldsymbol{\epsilon}(t) = \mathbf{0}$. Consequently, by imposing the Petrov-Galerkin condition on (1.6), one may factor out $\mathbf{EV}\left(\mathbf{W}^T\mathbf{EV}\right)^{-1}$ on the left hand side of each product, which finally leads to a reduced model of the form (1.4),

$$\begin{aligned} \overbrace{\mathbf{W}^T\mathbf{EV}}^{\mathbf{E}_r}\,\dot{\mathbf{x}}_r(t) &= \overbrace{\mathbf{W}^T\mathbf{AV}}^{\mathbf{A}_r}\,\mathbf{x}_r(t) + \overbrace{\mathbf{W}^T\mathbf{B}}^{\mathbf{B}_r}\,\mathbf{u}(t), \\ \mathbf{y}_r(t) &= \underbrace{\mathbf{CV}}_{\mathbf{C}_r}\,\mathbf{x}_r(t), \end{aligned} \tag{1.7}$$

and which can uniquely be solved for the reduced state $\mathbf{x}_r(t)$. In case of an orthogonal projection, i. e. $\mathbf{W} = \mathbf{V}$, one refers to $\boldsymbol{\epsilon}(t) \perp \mathbf{V}$ as *Galerkin condition.* The following lemma shows that only the subspaces span($\mathbf{V}$) and span($\mathbf{W}$) determine the transfer function of the reduced model and that the chosen bases are irrelevant.

Lemma 1.2 ([79])**.** *Let* $\mathbf{T}_1, \mathbf{T}_2 \in \mathbb{R}^{n\times n}$ *be non-singular matrices, then the reduced transfer function* $\boldsymbol{G}_r(s) = \mathbf{C}_r\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r$ *is unchanged if we replace* $\mathbf{V}$ *and* $\mathbf{W}$ *by* $\mathbf{VT}_1$ *and* $\mathbf{WT}_2$*, respectively, because they span the same subspace.*

The projective approach has led to various methods for MOR of LTI systems, which cannot be reviewed here in their entirety. The interested reader is rather referred to the monograph [8], the classical surveys [16, 73, 90] or the more recent ones [19, 23].

One quite self-evident approach to MOR is *modal reduction.* There, the projection matrices $\mathbf{W}$ and $\mathbf{V}$ are chosen to span left and right invariant subspaces, in order to preserve some eigenvalues of the original model in the reduced one. If we aim at approximating only the transfer behaviour of $\boldsymbol{G}(s)$ by $\boldsymbol{G}_r(s)$ as good as possible, this, however, is often too restrictive. By letting the reduced model $\boldsymbol{G}_r(s)$ have poles, that are not poles of $\boldsymbol{G}(s)$, the transfer behaviour usually may be better approximated. Modal reduction is therefore not discussed here; nevertheless, it should be mentioned that in some cases, it is still a valuable tool, e. g. when dealing with certain mechanical structures [36] or by combining the method with other approaches [157].

1.4 Balanced Truncation

One of the most important methods for MOR is the so-called *truncated balanced realization* (TBR), or simply *balanced truncation*, as it achieves three of the four previously mentioned main goals: TBR generally yields a good approximation, it preserves stability (a variant also passivity), and an a priori error bound is available. Merely in a large-scale setting with $N > 5000$ (depending on the hardware), the numerical computations involved may be tough, which is not yet fully resolved. This, in fact, will be tackled in Part III of this work, but at first, the basic concept of TBR is reviewed in the following.

Balanced truncation consists of two main steps: find a state representation of the original system, such that each state variable is as well controllable as it is observable (the balancing step); and subsequently, eliminate those state variables that are least controllable/observable (the truncation step). This method permits the nice physical interpretation that those states are neglected which are difficult to reach and simultaneously difficult to observe, and consequently, which least contribute to the energy transfer from the input to the output. Therefore, the starting point of balanced truncation is to suitably quantify "observability" and "controllability" in a system, which is reviewed next. The outcome will be a method also denoted as *Lyapunov balancing*, in order to distinguish it from other types of balancing to-be-reviewed later on. The proofs of the coming results can be found in the book [8] or references therein.

For a linear system, a state $\mathbf{x}_e \in \mathbb{R}^N$ is called reachable, if there exists an input $\mathbf{u}(t)$, such that $\mathbf{x}(T) = \mathbf{x}_e$ after some finite time $T > 0$, starting from $\mathbf{x}(0) = \mathbf{0}$. Accordingly, if the minimum input energy that is required to reach the state $\mathbf{x}_e$ is small, this $\mathbf{x}_e$ is well controllable—and if much energy is necessary to reach $\mathbf{x}_e$, it is poorly controllable. In a dual way, observability of a state $\mathbf{x}_0$ can be quantified by measuring the output energy that results from setting $\mathbf{u}(t) = \mathbf{0}$, $\forall t \geq 0$, with initial state $\mathbf{x}(0) = \mathbf{x}_0$. As asymptotically stable systems are treated, both energies $J_c(\mathbf{x}_e)$ and $J_o(\mathbf{x}_0)$, measuring controllability of $\mathbf{x}_e$ and observability of $\mathbf{x}_0$, respectively, remain bounded, and are given by the functionals

$$J_c(\mathbf{x}_e) = \min_{\mathbf{x}(-\infty)=\mathbf{0},\, \mathbf{x}(0)=\mathbf{x}_e} \int_{-\infty}^{0} \mathbf{u}(t)^T \mathbf{u}(t)\,\mathrm{d}t, \qquad \text{and} \tag{1.8}$$

$$J_o(\mathbf{x}_0) = \int_{0}^{\infty} \mathbf{y}(t)^T \mathbf{y}(t)\,\mathrm{d}t, \quad \mathbf{x}(0) = \mathbf{x}_0, \quad \text{where } \mathbf{u}(t) = \mathbf{0},\ \forall t. \tag{1.9}$$

It can be shown that both energies (1.8) and (1.9) satisfy

$$J_c(\mathbf{x}_e) = \mathbf{x}_e^T \mathbf{P}^{-1} \mathbf{x}_e, \qquad \text{and} \qquad J_o(\mathbf{x}_0) = \mathbf{x}_0^T \mathbf{E}^T \mathbf{Q} \mathbf{E} \mathbf{x}_0, \tag{1.10}$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{N \times N}$ solve the two dual Lyapunov equations

$$\mathbf{A}\mathbf{P}\mathbf{E}^T + \mathbf{E}\mathbf{P}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0}, \tag{1.11}$$

$$\mathbf{A}^T\mathbf{Q}\mathbf{E} + \mathbf{E}^T\mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0}. \tag{1.12}$$

For an asymptotically stable system that is fully controllable and observable, both $\mathbf{P}$ and $\mathbf{Q}$ are symmetric positive definite. The matrices $\mathbf{P}$ and $\mathbf{E}^T\mathbf{QE}$ are then called the *Controllability* and *Observability Gramian* of the system, respectively. It follows from (1.10) that states $\mathbf{x}_e$ that require low energy to reach, and thus are well controllable, lie in the span of those eigenvectors that correspond to large eigenvalues of $\mathbf{P}$. Accordingly, states that lie in the span of those eigenvectors that correspond to large eigenvalues of $\mathbf{E}^T\mathbf{QE}$ yield much energy in the output, and thus are well observable.

For model order reduction, only those states may be neglected that are hard to reach and simultaneously difficult to observe. In order to identify these states, introduce a regular state transformation of the form $\mathbf{z}=\mathbf{Tx}$, with $\mathbf{T}\in\mathbb{R}^{N\times N}$ non-singular. Then it follows from the two Lyapunov equations (1.11) and (1.12) that the Gramians transform correspondingly: $\widetilde{\mathbf{P}}=\mathbf{TPT}^T$, $\widetilde{\mathbf{E}}^T\widetilde{\mathbf{Q}}\widetilde{\mathbf{E}}=\mathbf{T}^{-T}\mathbf{E}^T\mathbf{QET}^{-1}$. This implies that the eigenvalues of the product of the Gramians are invariant with respect to state transformations, $\widetilde{\mathbf{P}}\widetilde{\mathbf{E}}^T\widetilde{\mathbf{Q}}\widetilde{\mathbf{E}}=\mathbf{TPE}^T\mathbf{QET}^{-1}$, and thus are inherent properties of a system. The so-called *Hankel singular values* (HSV) $\sigma_i \geq 0$ of a system arise from these eigenvalues and are defined as

$$\sigma_i = \sqrt{\lambda_i\left(\mathbf{PE}^T\mathbf{QE}\right)}. \tag{1.13}$$

Definition 1.1. The state-space realization (1.1) of system $\boldsymbol{G}(s)$ is called *balanced* if

$$\mathbf{P} = \mathbf{E}^T\mathbf{QE} = \mathbf{\Sigma} = \operatorname{diag}\left(\sigma_1, \sigma_2, \ldots, \sigma_N\right), \tag{1.14}$$

with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0$.

It can be shown that there always exists a state transformation $\mathbf{T}$ that simultaneously diagonalizes both Gramians, and thereby balances the system. As discussed in [11], the Hankel singular values usually decay very rapidly. This fact motivates the truncation of those state variables that correspond to small Hankel singular values, as then the approximation error can be expected to be small. Let the large-scale system be balanced and the Gramians be partitioned as

$$\mathbf{P} = \mathbf{E}^T\mathbf{QE} = \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \end{bmatrix}, \tag{1.15}$$

with $\mathbf{\Sigma}_1=\operatorname{diag}\left(\sigma_1, \sigma_2, \ldots, \sigma_n\right)$, $\mathbf{\Sigma}_2=\operatorname{diag}\left(\sigma_{n+1}, \sigma_{n+2}, \ldots, \sigma_N\right)$, and accordingly partition

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \end{bmatrix}. \tag{1.16}$$

Then the reduced model $\boldsymbol{G}_r(s)$ of order n, obtained by balanced truncated, follows from truncating the second block of state variables and reads as

$$\begin{aligned} \mathbf{E}_{11}\dot{\mathbf{x}}_r(t) &= \mathbf{A}_{11}\mathbf{x}_r(t) + \mathbf{B}_1\mathbf{u}(t), \\ \mathbf{y}_r(t) &= \mathbf{C}_1\mathbf{x}_r(t). \end{aligned} \tag{1.17}$$

Lemma 1.3 ([8])**.** *Let the reduced system (1.17) be obtained by balanced truncation with $\sigma_i > \sigma_j,\ i = 1, \ldots, n,\ j = n+1, \ldots, N$, then it has the following properties: it is asymptotically stable, balanced, a minimal realization, and it satisfies*

$$\sigma_{n+1} \leq \|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_\infty} \leq 2\left(\sigma_{n+1} + \ldots + \sigma_N\right). \tag{1.18}$$

This reveals a nice property of balanced truncation: a rigorous upper bound on the $\mathcal{H}_\infty$ error is available—without having to compute the reduced system. Very recently, Minh et al. [141] presented a new lower bound on the $\mathcal{H}_\infty$ error. Without going into details, the new bound suggests that in some cases discarding other states than the ones that correspond to the smallest σ_j can lead to lower error in the $\mathcal{H}_\infty$ norm. This surprising result was also illustrated by a small example and it actually contradicts the above mentioned motivation for balanced truncation; however, it seems to be unlikely that this applies to large-scale models.

Quite the contrary, in a large-scale setting it is more likely that $\sigma_i \gg \sigma_j,\ i = 1, \ldots, n,\ j = n+1, \ldots, N$; and then the reduced model obtained by balanced truncation presumably is close to having locally minimal $\mathcal{H}_2$ error, cf. [104].

What is left to show is a numerically stable scheme for computing a balanced truncated system. There are different ways to achieve this, see e. g. [35]; in the next lemma the so-called *square root algorithm* is reviewed, for which a convenient implementation is available in Matlab by the function `balancmr`.

Lemma 1.4. *Given the two* Cholesky *decompositions $\mathbf{P} = \mathbf{R}\mathbf{R}^T$ and $\mathbf{Q} = \mathbf{U}\mathbf{U}^T$ and the singular value decomposition $\mathbf{U}^T\mathbf{E}\mathbf{R} = \mathbf{M}\boldsymbol{\Sigma}\mathbf{N}^T$, where $\mathbf{M}^T = \mathbf{M}^{-1}$ and $\mathbf{N}^T = \mathbf{N}^{-1}$ are orthogonal, and $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N)$, the state transformation $\mathbf{z} = \mathbf{T}\mathbf{x}$ that simultaneously diagonalizes both Gramians and thereby balances the system is given by*

$$\mathbf{T} = \boldsymbol{\Sigma}^{-1/2}\mathbf{M}^T\mathbf{U}^T\mathbf{E} \qquad \textit{and} \qquad \mathbf{T}^{-1} = \mathbf{R}\mathbf{N}\boldsymbol{\Sigma}^{-1/2}. \tag{1.19}$$

In order to obtain the reduced model, it is not necessary to balance the complete system. Instead, one may partition

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{M}_2 \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} \mathbf{N}_1 & \mathbf{N}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 \end{bmatrix} \tag{1.20}$$

with $\mathbf{M}_1 \in \mathbb{R}^{N\times n}$, $\mathbf{N}_1 \in \mathbb{R}^{N\times n}$ and $\mathbf{\Sigma}_1 \in \mathbb{R}^{n\times n}$, and define

$$\mathbf{W}^T = \mathbf{\Sigma}_1^{-1/2}\mathbf{M}_1^T\mathbf{U}^T \in \mathbb{R}^{n\times N} \qquad \text{and} \qquad \mathbf{V} = \mathbf{R}\mathbf{N}_1\mathbf{\Sigma}_1^{-1/2} \in \mathbb{R}^{N\times n}. \tag{1.21}$$

The matrices that realize the reduced system by balanced truncation then are

$$\mathbf{A}_r = \mathbf{W}^T\mathbf{A}\mathbf{V}, \quad \mathbf{E}_r = \mathbf{W}^T\mathbf{E}\mathbf{V}, \quad \mathbf{B}_r = \mathbf{W}^T\mathbf{B}, \quad \text{and} \quad \mathbf{C}_r = \mathbf{C}\mathbf{V}. \tag{1.22}$$

By construction, $\mathbf{E}_r$ is the identity matrix [209], and due to (1.22), $\mathbf{V}$ and $\mathbf{W}$ can be interpreted as projection matrices, defining the projector $\mathbf{\Pi} = \mathbf{E}\mathbf{V}\mathbf{W}^T$. This reveals the connection of balanced truncation to projective MOR and shows how the reduced system can be obtained in one step. In a large-scale setting it is advisable to drop $\mathbf{\Sigma}_1^{-1/2}$ in (1.21), leading to lower condition numbers of the projection matrices $\mathbf{W}^T$ and $\mathbf{V}$, cf. [8]. This yields a reduced system (1.22) with equal transfer function and which is balanced up to diagonal scaling.

A vast literature on balanced truncation is available. The fundamentals were derived by Mullis and Roberts [143] and by Moore [142], and the generalization to an $\mathbf{E}$ matrix is due to Hsu [103]. Many further generalizations of the method were treated by various authors, including the—by no means exhaustive—examples: non-minimal [185] and unstable [183, 221] systems, also systems with *differential algebraic equations* (DAE) [26, 136, 160, 180, 181] or inhomogeneous initial conditions [99], and furthermore, second-order [44, 45, 139, 159, 218], periodic [191], time-varying [173], bilinear [29], and infinite dimensional systems [158]. With regards to the numerical implementation, the square-root method [123, 185] was reviewed, but also other approaches are possible, such as the Schur method [171] or the balancing-free square-root method [190]. As aforementioned, the presented approach is referred to as Lyapunov balanced truncation, as the solutions of two dual Lyapunov equations are simultaneously diagonalized. For SISO and symmetric MIMO systems it also possible to diagonalize the solution of one Sylvester equation, which is denoted as Cross-Gramian, cf. [61, 62, 63, 124, 177]. Moreover, other types of balancing solve Riccati equations instead of Lyapunov equations, such that the bounded realness or the positive realness (i. e. passivity) of the original system is preserved in the reduced one. As these approaches go beyond the focus of this work, the interested reader is rather referred to the survey [87] and references therein, where also stochastic and frequency weighted balancing is discussed. The same holds for other related model reduction methods, such as optimal Hankel norm approximation and singular perturbation approximation, for which references can be found in the survey [19].

1.5 Numerical Solution of Lyapunov Equations

All above mentioned methods for balanced truncation require the solution of large matrix equations. As this work is concerned with Lyapunov balancing, the equations to-be-solved are two dual Lyapunov equations. From a numerical point of view, their solution is the bottleneck of the method, which is reviewed below. As solving (1.12) for $\mathbf{Q}$ is dual to solving (1.11) for $\mathbf{P}$, merely the latter is treated in the following, and reference to (1.12) is made only if absolutely necessary.

1.5.1 Direct Solution

The first type of methods for solving the Lyapunov equation (1.11) are direct or also called dense methods [178]: the Bartels-Stewart algorithm [18] and Hammarling's method [98]. Both algorithms were originally stated without an $\mathbf{E}$ matrix; the generalization was due to Penzl [152]. The original methods start with a Schur decomposition of $\mathbf{A}=\mathbf{TDT}^*$, where $\mathbf{T}^*=\mathbf{T}^{-1}$ is unitary and $\mathbf{D}$ is upper triangular, and transform the Lyapunov equation (1.11) (for $\mathbf{E}$ identity) with $\mathbf{T}$,

$$\mathbf{D}\widetilde{\mathbf{P}} + \widetilde{\mathbf{P}}\mathbf{D}^* + \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^* = \mathbf{0}, \tag{1.23}$$

such that $\widetilde{\mathbf{B}} = \mathbf{T}^*\mathbf{B}$. Once (1.23) is solved for $\widetilde{\mathbf{P}}$, then $\mathbf{P}$ can be obtained by $\mathbf{P} = \mathbf{T}\widetilde{\mathbf{P}}\mathbf{T}^*$.

In the Bartels-Stewart algorithm, (1.23) is solved for the columns of $\widetilde{\mathbf{P}}$ by forward substitutions, which is possible owing to the triangularity of $\mathbf{D}$. Then one complex linear system solve is required for each column of $\widetilde{\mathbf{P}}$. This method is available in Matlab in the function `lyap`, but it may also be formulated with a real Schur decomposition [178].

Hammarling's method directly solves for the Cholesky factor $\widetilde{\mathbf{R}}$ of the solution $\widetilde{\mathbf{P}}=\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^T$. This is done by recursively reducing the order of (1.23) by one and thereby preserving the triangular structure. Assuming the right-hand side in factored form $\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^*$, $\widetilde{\mathbf{R}}$ can be directly computed and $\mathbf{R}$ may be obtained from $\mathbf{R}=\mathbf{T}\widetilde{\mathbf{R}}$. Hammarling's method is therefore well suited for square-root balanced truncation, and it is implemented in Matlab in the function `lyapchol`.

1.5.2 Approximate Solution

Large-scale models often result from some kind of (often spatial) discretization. In such a case, the generated matrices are typically sparse. This is the key to reducing storage requirements in large-scale settings, as only non-zero entries have to be stored. In order to also keep computational efforts manageable it is essential to employ numerical operations that can exploit sparsity. The Schur decomposition, which is mandatory for both the Bartels-Stewart algorithm and Hammarling's method, is a dense matrix factorization whose storage requirements are $\mathcal{O}(N^2)$ and whose computational complexity is $\mathcal{O}(N^3)$. As sparsity is lost in the Schur decomposition, direct methods are feasible on a typical modern hardware only for systems with N up to a few thousands. In a large-scale setting, however, computing the full rank solution $\mathbf{P}$ is an ill-conditioned problem anyway, because of the typically rapid decay of its eigenvalues [11, 154]. Owing to its low numerical rank, $\mathbf{P}$ may then be well approximated by a non-negative definite $\widehat{\mathbf{P}} \in \mathbb{R}^{N\times N}$ of $\mathrm{rank}(\widehat{\mathbf{P}}) = q$, $q \ll N$, which may be factorized as $\widehat{\mathbf{P}} = \mathbf{Z}\mathbf{Z}^T$, with $\mathbf{Z} \in \mathbb{R}^{N\times q}$. The matrix $\mathbf{Z}$ is denoted as low-rank (Cholesky) factor (of $\widehat{\mathbf{P}}$), regardless of the fact that $\mathbf{Z}$ itself might have full column rank. To perform the square root algorithm, the Cholesky factor $\mathbf{R}$ then has to be substituted by $\mathbf{Z}$, whereas the subsequent steps remain unchanged.

The benefit of using low-rank factors is that storage requirements scale down to $\mathcal{O}(Nq)$, and—depending on the method—computational complexity additionally may be reduced significantly. The drawbacks are that the a priori error bound (1.18) is strictly speaking lost, and that stability in the reduced system cannot be guaranteed. In practice, however, it seems like the latter is not a problem and that the reduced system is close to the one obtained by direct methods [93].

Various approaches have been derived to compute low-rank Cholesky factors $\mathbf{Z}$. The first type of methods is based on the integral form of the Gramian [8],

$$\mathbf{P} = \int_0^\infty e^{\widetilde{\mathbf{A}}\tau}\widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T e^{\widetilde{\mathbf{A}}^T\tau}\,\mathrm{d}\tau = \frac{1}{2\pi}\int_{-\infty}^{\infty} (\imath\omega\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}\mathbf{B}^T\left(-\imath\omega\mathbf{E}-\mathbf{A}^T\right)^{-1}\mathrm{d}\omega, \qquad (1.24)$$

with $\widetilde{\mathbf{A}} = \mathbf{E}^{-1}\mathbf{A}$ and $\widetilde{\mathbf{B}} = \mathbf{E}^{-1}\mathbf{B}$. An approximation $\widehat{\mathbf{P}}$ can be computed by integrating (1.24) in time domain [166], sometimes denoted as balanced *proper orthogonal decomposition* (POD) [146], whereas numerical quadrature in the frequency domain is usually referred to as *poor man's TBR* [155].

The second type of methods project the system matrices $\mathbf{E}$, $\mathbf{A}$ and $\mathbf{B}$ and then solve the reduced Lyapunov equation of order q by direct methods. One typically projects onto Krylov subspaces, who will be defined in Section 1.6.1. Different types of Krylov subspaces are available. Probably the easiest choice is to use their original formulation, which could also be called *classical* Krylov subspaces; their application to solve Lyapunov equations is discussed in [48, 102, 110, 111, 113, 166, 176]. The numerical efficiency of this approach, however, suffers from possibly high dimension of the Krylov subspace that is required for good approximation. To reduce the order of the Lyapunov equation that has to be solved by direct methods, an extended Krylov subspace [100, 118, 174] may be applied. This approach is usually referred to as *extended Krylov subspace method* (EKSM), although in the original work [174] it was denoted as *Krylov-plus-inverted-Krylov* (K-PIK). More recently also rational Krylov subspaces are used for the projection [25, 53, 54, 214], then referred to as *rational Krylov subspace method* (RKSM). Even though the numerical effort to compute a *rational* Krylov subspace is higher than in both other cases, this might be compensated by the lower order of the reduced Lyapunov equation that is required for sufficient approximation. It should be noted, that there are also further approaches that use Krylov subspaces, which e. g. try to minimize the residual [129], or work for DAE systems [182].

The third type of methods for computing $\mathbf{Z}$ is based on the *low-rank alternating direction implicit* (LR-ADI) iteration [27, 32, 33, 93, 127, 128, 153], which is also denoted as *Cholesky factor ADI* (CF-ADI) iteration or *low-rank Cholesky factor ADI* (LRCF-ADI) iteration. Wachspress [200] was the first to consider the Lyapunov equation as an ADI model problem. In his original formulation, the ADI iteration has storage requirements of $\mathcal{O}(N^2)$. In a large-scale setting, however, it is mandatory to use the low rank formulation with storage requirements of $\mathcal{O}(Nn)$, which is due to Penzl [153] and Li and White [128]. In this work, therefore always the low-rank formulation is meant when just referring to "the ADI iteration". The performance of this method heavily depends on a good shift selection, which is typically achieved by solving a minimax problem [199, 200] or by a heuristic approach [153]. LR-ADI is also denoted as *low-rank Smith* (LR-Smith) iteration, which stems from an alternative derivation. If a given set of l shifts is cyclically reused, then one also refers to the method as LR-Smith(l) iteration, and a modification of this [95, 170] prevents the low-rank factor $\mathbf{Z}$ from having linearly dependent columns. Recent results [34, 210] suggest adaptively chosen shifts instead of a priori solving the minimax problem.

Methods that combine ideas of LR-ADI and Krylov projections can be found in [17, 38, 112], and it was proven that there is a strong connection of RKSM and LR-ADI [55, 67, 213], which will also be discussed in Section 5.4. For further reading, the surveys [28, 175] are recommended.

The numerical solution of (1.11) with RKSM and LR-ADI is treated in more detail in Part III. All three aforementioned types of methods for the approximate solution of (1.11) are either based on projections onto Krylov subspaces (EKSM, RKSM), may be interpreted as such (ADI [213]), or are at least connected to them (poor man's TBR [155], balanced POD [146]). Krylov subspaces, however, provide a family of methods for MOR in their own right, which is why they will be reviewed in the next section.

1.6 Moment Matching: Model Reduction Via Krylov Subspaces

Consider system (1.1) and define

$$\boldsymbol{X}(s) = (s\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B}, \tag{1.25}$$

which describes the transfer function from the input to the states in frequency domain. To get local information about the system, $\boldsymbol{X}(s)$ may be evaluated for a certain frequency $s_0 \in \mathbb{C}$, leading to the m-dimensional block $(s_0\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}$. If appropriate frequencies s_i, $i=1,\ldots,k$ are used, then the union of the blocks $(s_i\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}$ contains the most dominant directions in state space. This is the basic idea behind rational Krylov subspaces and is reviewed in this section.

1.6.1 Krylov Subspaces

Given the state-space realization (1.1) of a system $\boldsymbol{G}(s)$, and using the notation $\mathbf{A}_{s_0}= (\mathbf{A}-s_0\mathbf{E})$, the rational input Krylov subspace $\mathcal{K}(s_0,q_0)$ of order q_0 at s_0 is defined as

$$\mathcal{K}(s_0,q_0) = \operatorname{span}\left\{\mathbf{A}_{s_0}^{-1}\mathbf{B},\ \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{q_0-1}\mathbf{A}_{s_0}^{-1}\mathbf{B}\right\}, \tag{1.26}$$

and s_0 is denoted as *shift* or *expansion point*. A rational input Krylov subspace $\mathcal{K}_b$ generally merges different expansion points s_i, $i=1,\ldots,k$ with respective orders q_i,

$$\mathcal{K}_b = \mathcal{K}(s_1,q_1)\ \cup\ \mathcal{K}(s_2,q_2)\ \cup\ \ldots\ \cup\ \mathcal{K}(s_k,q_k). \tag{1.27}$$

An important property of a Krylov subspace $\mathcal{K}_b$ is that it can be nested, see e. g. [128, Theorem 5.4]. To demonstrate this property, assume e. g. k different expansion points $s_i, i=1,\ldots,k$, all of order $q_1=\ldots=q_k=1$, then

$$\mathcal{K}_b = \operatorname{span}\left\{\mathbf{A}_{s_1}^{-1}\mathbf{B},\ \mathbf{A}_{s_2}^{-1}\mathbf{B},\ \ldots,\ \mathbf{A}_{s_k}^{-1}\mathbf{B}\right\}, \tag{1.28}$$

$$= \operatorname{span}\left\{\mathbf{A}_{s_1}^{-1}\mathbf{B},\ \mathbf{A}_{s_2}^{-1}\mathbf{E}\mathbf{A}_{s_1}^{-1}\mathbf{B},\ \ldots,\ \mathbf{A}_{s_k}^{-1}\mathbf{E}\ldots\mathbf{A}_{s_2}^{-1}\mathbf{E}\mathbf{A}_{s_1}^{-1}\mathbf{B}\right\}. \tag{1.29}$$

If $m=1$, (1.27) will be denoted as *single-input rational Krylov* subspace $\mathcal{K}_s$, whereas if $m>1$, it is referred to as *multi-input rational Krylov* or *block-input rational Krylov* subspace $\mathcal{K}_b$. For every new shift or additional order, the dimension of the block-input subspace grows by (actually at most) m, which might be undesirable in some settings. A remedy is to introduce tangential directions $\mathbf{l}_i \in \mathbb{C}^m$ whereby the *tangential-input rational Krylov* subspace $\mathcal{K}_t$ is defined as

$$\mathcal{K}_t = \operatorname{span}\left\{\mathbf{A}_{s_1}^{-1}\mathbf{B}\mathbf{l}_1,\ \mathbf{A}_{s_2}^{-1}\mathbf{B}\mathbf{l}_2,\ \ldots,\ \mathbf{A}_{s_k}^{-1}\mathbf{B}\mathbf{l}_k\right\}. \tag{1.30}$$

Note that, in general, the nested property cannot be preserved for the tangential Krylov subspace (1.30); the property in fact is dependent on the choice of $\mathbf{l}_i$. Further note that the tangential-input Krylov subspace can be considered as a generalization: it directly includes single-input Krylov subspaces because then the tangential directions $\mathbf{l}_i$ become scalars and do not alter the subspace; and if m tangential directions $\mathbf{l}_i$, that form a basis of $\mathbb{R}^m$, are used for each shift, then $\mathcal{K}_t$ essentially is a block-input Krylov subspace.

Krylov subspaces originate from the classical $\mathcal{K}_\infty$, which is defined as

$$\mathcal{K}_\infty = \operatorname{span}\left\{\mathbf{E}^{-1}\mathbf{B},\ \mathbf{E}^{-1}\mathbf{A}\mathbf{E}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{E}^{-1}\mathbf{A}\right)^{q_\infty-1}\mathbf{E}^{-1}\mathbf{B}\right\}, \tag{1.31}$$

and which can be shown to be related to rational Krylov subspaces by letting $s_0 \to \infty$. Another important subspace is the already mentioned extended Krylov subspace $\mathcal{K}_e$; it combines $\mathcal{K}_\infty$ with (1.26) for $s_0 = 0$:

$$\mathcal{K}_e = \operatorname{span}\left\{\mathbf{E}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{E}^{-1}\mathbf{A}\right)^{q_e-1}\mathbf{E}^{-1}\mathbf{B},\ \mathbf{A}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{A}^{-1}\mathbf{E}\right)^{q_e-1}\mathbf{A}^{-1}\mathbf{B}\right\}. \tag{1.32}$$

Both (1.31) and (1.32), however, are of minor interest in this work, as we will see in Chapter 4, that the main contribution—i. e. pseudo-optimality—requires rational Krylov subspaces with $s_0 \neq 0$.

For all mentioned *input* Krylov subspaces, also dual *output* Krylov subspaces can be defined by simply substituting $\mathbf{A}^T$ for $\mathbf{A}$, $\mathbf{E}^T$ for $\mathbf{E}$, $\mathbf{C}^T$ for $\mathbf{B}$ and $\mathbf{r}_i \in \mathbb{C}^p$ for $\mathbf{l}_i$. Output Krylov subspaces are of minor interest in this work, as well, since for pseudo-optimality only one type of Krylov subspaces suffices.

In the remainder of this work, $\mathbf{V} \in \mathbb{R}^{N\times n}$ will exclusively denote a matrix whose columns form a basis of any of the above mentioned input Krylov subspaces, whereas $\mathbf{W} \in \mathbb{R}^{N\times n}$ may—but is not restricted to—denote a matrix whose columns form a basis of an output Krylov subspace.

1.6.2 Moment Matching

Of course it is no coincidence that $\mathbf{V}$ and $\mathbf{W}$ denote not only projection matrices in Section 1.3 but also bases of Krylov subspaces in Section 1.6.1. The reason is that the transfer function $\boldsymbol{G}(s)$ may be locally approximated by using Krylov subspaces for projecting (1.1); this is referred to as *moment matching*.

Definition 1.2. Given the transfer function $\boldsymbol{G}(s)=\mathbf{C}\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}$ and an expansion point s_0, the Taylor series expansion of $\boldsymbol{G}(s)$ is defined as

$$\boldsymbol{G}(s)=\sum_{i=0}^{\infty}\mathbf{M}_i^{s_0}\left(s-s_0\right)^i, \tag{1.33}$$

where $\mathbf{M}_i^{s_0}$ are called the *moments* of $\boldsymbol{G}(s)$ at s_0; they satisfy

$$\mathbf{M}_i^{s_0}=-\mathbf{C}\left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^i\mathbf{A}_{s_0}^{-1}\mathbf{B}. \tag{1.34}$$

Expanding $\boldsymbol{G}(s)$ at $s_0\to\infty$, the Taylor series is given by

$$\boldsymbol{G}(s)=\sum_{i=1}^{\infty}\mathbf{M}_i^{\infty}s^{-i}. \tag{1.35}$$

The $\mathbf{M}_i^{\infty}$ are called the *Markov parameters* of $\boldsymbol{G}(s)$, and they satisfy

$$\mathbf{M}_i^{\infty}=\mathbf{C}\left(\mathbf{E}^{-1}\mathbf{A}\right)^{i-1}\mathbf{E}^{-1}\mathbf{B}. \tag{1.36}$$

With these definitions, the main theorem of Krylov-based MOR can be stated.

Theorem 1.5 (Moment matching [85])**.** *Given the projection matrices* $\mathbf{V},\mathbf{W}\in\mathbb{R}^{N\times n}$*, let the reduced model* $\boldsymbol{G}_r(s)=\mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r$ *with Taylor series expansion* $\boldsymbol{G}_r(s)=\sum_{i=0}^{\infty}\widehat{\mathbf{M}}_i^{s_0}\left(s-s_0\right)^i=\sum_{i=1}^{\infty}\widehat{\mathbf{M}}_i^{\infty}s^{-i}$ *be given by projection:* $\mathbf{A}_r=\mathbf{W}^T\mathbf{A}\mathbf{V}$*,* $\mathbf{E}_r=\mathbf{W}^T\mathbf{E}\mathbf{V}$*,* $\mathbf{B}_r=\mathbf{W}^T\mathbf{B}$*, and* $\mathbf{C}_r=\mathbf{C}\mathbf{V}$*. Assume that* s_0 *is neither an eigenvalue of* $\mathbf{E}^{-1}\mathbf{A}$*, nor an eigenvalue of* $\mathbf{E}_r^{-1}\mathbf{A}_r$*. If one of the following,*

$$\operatorname{span}\left\{\mathbf{A}_{s_0}^{-1}\mathbf{B},\ \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{q_0-1}\mathbf{A}_{s_0}^{-1}\mathbf{B}\right\}\subseteq\operatorname{span}(\mathbf{V}), \tag{1.37}$$

$$\operatorname{span}\left\{\mathbf{A}_{s_0}^{-T}\mathbf{C}^T,\ \mathbf{A}_{s_0}^{-T}\mathbf{E}^T\mathbf{A}_{s_0}^{-T}\mathbf{C}^T,\ \ldots,\ \left(\mathbf{A}_{s_0}^{-T}\mathbf{E}^T\right)^{q_0-1}\mathbf{A}_{s_0}^{-T}\mathbf{C}^T\right\}\subseteq\operatorname{span}(\mathbf{W}), \tag{1.38}$$

is satisfied, then $\mathbf{M}_i^{s_0}=\widehat{\mathbf{M}}_i^{s_0}$*,* $i=0,\ldots,q_0-1$*. If both (1.37) and (1.38) are satisfied, then* $\mathbf{M}_i^{s_0}=\widehat{\mathbf{M}}_i^{s_0}$*,* $i=0,\ldots,2q_0-1$*. Accordingly, if one of the following,*

$$\operatorname{span}\left\{\mathbf{E}^{-1}\mathbf{B},\ \mathbf{E}^{-1}\mathbf{A}\mathbf{E}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{E}^{-1}\mathbf{A}\right)^{q_\infty-1}\mathbf{E}^{-1}\mathbf{B}\right\}\subseteq\operatorname{span}(\mathbf{V}), \tag{1.39}$$

$$\operatorname{span}\left\{\mathbf{E}^{-T}\mathbf{C}^T,\ \mathbf{E}^{-T}\mathbf{A}^T\mathbf{E}^{-T}\mathbf{C}^T,\ \ldots,\ \left(\mathbf{E}^{-T}\mathbf{A}^T\right)^{q_\infty-1}\mathbf{E}^{-T}\mathbf{C}^T\right\}\subseteq\operatorname{span}(\mathbf{W}), \tag{1.40}$$

is satisfied, then $\mathbf{M}_i^{\infty}=\widehat{\mathbf{M}}_i^{\infty}$*,* $i=0,\ldots,q_\infty-1$*. If both (1.39) and (1.40) are satisfied, then* $\mathbf{M}_i^{\infty}=\widehat{\mathbf{M}}_i^{\infty}$*,* $i=0,\ldots,2q_\infty-1$*.*

Theorem 1.5 states that if the sequence $\mathbf{A}_{s_0}^{-1}\mathbf{B}, \dots, \left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{q_0-1}\mathbf{A}_{s_0}^{-1}\mathbf{B}$ is contained in the subspace spanned by the columns of $\mathbf{V}$, then $\boldsymbol{G}_r(s)$ and its first q_0-1 derivatives with respect to s are equal to $\boldsymbol{G}(s)$ and its first q_0-1 derivatives at the point $s=s_0$. This is the fundamental relation between projections with Krylov subspaces and local interpolation. It follows from (1.25) that in order to preserve local information in the reduced system, it is sufficient to evaluate $\boldsymbol{X}(s)$ at a certain frequency and use this information for projection. Because of its local nature, the crucial question in the moment matching method has always been the choice of expansion points s_i and respective orders q_i for a good global approximation. Although various approaches towards this aim are available in the literature, there is still potential for improving their numerical efficiency. This question, however, will be treated later in this work; but it requires some basic remarks on the numerical implementation to compute bases of Krylov subspaces, as done in the following.

1.6.3 Numerical Considerations

Krylov subspace methods rely on evaluating (1.25) for a certain frequency s_0. To this end, a *linear system of equations* (LSE), $(s_0\mathbf{E}-\mathbf{A})\,\mathbf{V}_0=\mathbf{B}$, has to be solved for $\mathbf{V}_0$. The numerical solution of LSEs has been massively studied: there are direct and various iterative methods available. Direct methods are based on Gaussian elimination, but there are also modifications for sparse matrices—both available in Matlab by the backslash operator. If a rational Krylov subspace with higher order has to be computed, multiple LSEs with only varying right-hand sides have to be solved; then, it is advisable to compute an LU-decomposition of $\mathbf{A}_{s_0}$ a priori, because the LSEs may then be solved by forward/backward substitutions with low numerical effort for all right-hand sides.

Iterative methods start with an initial approximation of $\mathbf{V}_0$ and try to improve it in every step. The final $\mathbf{V}_0$ indeed is an "inexact solve", but powerful methods try to monitor or estimate the actual error. Methods of this kind are the *generalized minimal residual method* (GMRES), or the *biconjugate gradient method* (BiCG), together with its variants *biconjugate gradient stabilized method* (BiCGSTAB) and *conjugate gradient squared method* (CGS), to mention just the most popular of them; implementations are also available in Matlab. Convergence of these methods can be significantly accelerated by preconditioning and restarting; details on this, however, are out of the scope of this work—but it is crucial to note that LSEs can be solved efficiently. For details on iterative methods, and how inexact solves effect moment matching in Theorem 1.5, please refer to [4, 215] and references therein.

As the solution of an LSE still requires most of the numerical effort in rational Krylov subspace methods—irrespective of whether direct or iterative methods are used—, it is inevitable to keep the number of LSEs to be solved as low as possible; this will become important in Part III.

Assume that a $\mathbf{V}$ which spans a rational Krylov subspace was computed by solving multiple LSEs. As $\mathbf{V}$ is subsequently used for projection, it is advisable in finite precision to employ an orthonormal basis of the same subspace, cf. [77] (which does not change the reduced system, owing to Lemma 1.2). The orthogonalization of $\mathbf{V}$ can be achieved by a (modified) Gram-Schmidt process.

Another issue with rational Krylov subspaces (1.26) is that computing $\mathbf{V}$ as

$$\mathbf{V} = \left[\mathbf{A}_{s_0}^{-1}\mathbf{B},\ \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{B},\ \ldots,\ \left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{q_0-1}\mathbf{A}_{s_0}^{-1}\mathbf{B}\right], \tag{1.41}$$

has to be avoided. Instead, one should compute the blocks in $\mathbf{V}$ recursively by $\mathbf{V}_i = \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{V}_{i-1}$, $i = 2, \ldots, q_0$—with immediate orthonormalization. Such a numerical implementation is usually denoted as Arnoldi or Arnoldi-type algorithm; a generalization that simultaneously computes biorthogonal bases of input and output Krylov subspaces is referred to as Lanczos(-type) algorithm. From a theoretical point of view, it is regardless which basis of a Krylov subspace is employed; "Arnoldi algorithm" should therefore be understood in this work as an algorithm to compute any basis of a rational Krylov subspace—irrespective of the details in the implementation. We will see in Section 2.3 that this is in fact equivalent to solving a particular Sylvester equation.

Finally, it is noted that the Krylov subspaces are assumed in the following to have maximum rank. For minimal realizations of SISO systems, this can be shown to always hold for arbitrary combinations of shifts s_i and respective orders q_i, see e. g. [189]. This assumption, however, might not apply to block and tangential Krylov subspaces. Then deflation techniques should be employed to compute a $\mathbf{V}$ with full column rank. This is discussed in [72] and can be incorporated into the presented framework in a straightforward way. Nevertheless, rational Krylov subspaces have storage requirements of $\mathcal{O}(Nq)$, and the bottleneck of their computation is to solve LSEs. As this is possible with exploiting sparsity, rational Krylov subspaces are well suited for large-scale models.

1.6.4 Notes and References

The idea of locally approximating a rational function—i. e. *rational interpolation*—has a long history and contributions can be found under various names: matching the moments $\mathbf{M}_i^\infty$ at infinity is usually referred to as *partial realization*; matching moments

$\mathbf{M}_i^0$ at $s = 0$ is called *Padé approximation*; the generalization to $s = s_0$ is denoted *shifted Padé* and to multiple shifts s_i *multipoint Padé*. In the literature on reducing RC circuits, these methods are also known as *asymptotic waveform evaluation* (AWE) [156] and *complex frequency hopping* (CFH) [47].

The drawback of the original formulations of these methods is that they rely on the explicit computation of moments—which is numerically ill-conditioned. Following Theorem 1.5, moment matching yet may also be enforced implicitly. To this end, Krylov subspaces are computed by Arnoldi or Lanczos type algorithms, which were originally introduced to solve LSEs and eigevalue problems (the generalization to rational Krylov subspaces is due to Ruhe [165]). Villemagne and Skelton [50] were among the first ones to use Krylov subspaces for MOR, but the general framework for projective model reduction by Krlov subspaces was due to Grimme [85] and Freund [73] (see [85] also for a nice historical overview). Model order reduction in this sense is called *Padé via Lanczos* (PVL) [60], *Krylov subspace method*, or simply *rational Krylov* (RK).

To mention all contributions in Krylov-based model reduction since Grimme would go beyond the scope of this work. References that are relevant in some kind, will anyway be given in the subsequent chapters, where appropriate. As the case $\det(\mathbf{E}) = 0$ is not considered in this work, it is worth noting here that the projective framework was recently generalized to this case [96]. For more details, please refer to the recent surveys [19, 23].

It is indeed judicious to call RK a *framework*: it has been shown [75, 78, 81, 97], that any reduced model may be constructed through projections onto Krylov subspaces (at least in the SISO case; MIMO is little more involved). Consequently, it is misleading to assume that moment matching per se guarantees a good approximation; quite the contrary, arbitrarily bad approximations may be generated by moment matching. However, this also implies that even the best reduced model may be found—it is just a matter of choosing the right shifts. Another consequence is that any model of order $n < N$ matches (in fact more than) $2n$ moments of any model of order N. The important question therefore is not if moments *are* matched in the reduced model, but instead to know *where* the moments are matched, and if these locations are *optimal* in some sense.

1.7 Problem Formulation

Let us recap the main objectives of MOR as stated in Section 1.2: good approximation, numerical efficiency, preservation of structural properties (probably most importantly: stability), and quantitative information on the error (preferably an upper bound). If numerical efficiency is deemed to be of minor significance, then definitely balanced truncation together with direct methods for solving the Lyapunov equations should be used, because it fully accomplishes all remaining goals.

The situation changes with the reduction of large-scale models: then direct solvers are inappropriate and have to be replaced by approximate solutions, which, however, base upon projections onto Krylov subspaces (see Part III). This is actually not surprising, owing to the aforementioned generality of Krylov subspace methods: (almost) any reduced model can be generated. Consequently, Krylov subspaces may be understood as a parametrization of reduced models, and without loss of generality we thus may assume that the reduced model is constructed through a projection onto an input Krylov subspace. The remaining degrees of freedom are then a sequence of shifts s_i (in the MIMO case additionally a sequence of respective tangential directions $\mathbf{l}_i$) and the direction of projection by means of the matrix $\mathbf{W}$. Given any SISO reduced model, there is a combination of s_i and $\mathbf{W}$, which generates it; and therefore, these quantities may be used to parametrize all reduced models.

The question is now: how to choose s_i and $\mathbf{W}$? This dissertation makes a suggestion for $\mathbf{W}$, by introducing the concept of $\mathcal{H}_2$ pseudo-optimality, which is the main result of this thesis. It will turn out that this suggestion for $\mathbf{W}$ happens implicitly because the reduced matrices can be directly computed without having to set up the generating $\mathbf{W}$ explicitly. It will further be shown that $\mathcal{H}_2$ pseudo-optimality can be embedded in a slightly alternative approach to MOR, namely a cumulative framework using Krylov subspaces. The justification of combining the cumulative idea with $\mathcal{H}_2$ pseudo-optimality is that this offers a number of advantages: stability is preserved in the reduced model; the degree of freedom in $\mathbf{W}$ is fixed by the choice of the Krylov subspace $\mathbf{V}$, which in turn leads to its automatic determination and, as will be shown, which is in some sense optimal; the reduced order can be accumulated, which has not been possible before; the approximation error is guaranteed to decrease monotonically; and the approach can be regarded as numerically efficient, as the main numerical effort remains to compute rational Krylov subspaces.

The drawback of the proposed framework might be that fulfilling all remaining goals solely depends on the selection of shifts s_i. On that account, we will see that dependence of the reduced model on the shifts in fact is virtually "doubled" in $\mathcal{H}_2$ pseudo-optimal model reduction, which could be expressed as "all problems are shifted to the shifts". Admittedly, it is then still possible to generate very bad approximations in this framework. Nevertheless, a lot of structure will become apparent in it, which then may be exploited by algorithms. The hope is that this marks the beginning of powerful algorithms for choosing shifts s_i (and therefore, yielding a good approximation) and computing error bounds with acceptable tightness and computational effort—as then the mentioned objectives of model order reduction would entirely be satisfied.

Providing powerful algorithms for choosing shifts s_i is beyond the scope of this work, and even more, finding algorithms that entirely satisfy all goals might even be impossible at this point. The aim of this thesis is rather to provide the theoretical foundation of this framework, from which possible solutions hopefully may emanate. First and promising ideas, are indeed suggested in the doctoral thesis of Panzer [148], where an optimization technique skilfully exploits $\mathcal{H}_2$ pseudo-optimality to choose shifts adaptively and where rigorous and computationally efficient upper bounds for special system classes are presented.

$\mathcal{H}_2$ pseudo-optimality is presented in this work in the context of rational Krylov subspaces (because descriptions in other contexts, such as transfer functions, are already available in the literature). To this end, it is mandatory to characterize any $\mathbf{V}$ that spans a rational Krylov subspace as the solution of particular Sylvester equations. This connection is reviewed in Chapter 2 and extended with some new results—all of which will be used in the subsequent chapters.

The first conclusion that follows from these Sylvester equations is that the error system $\boldsymbol{G}(s)-\boldsymbol{G}_r(s)$ can be factorized if $\mathbf{V}$ spans a Krylov subspace. This is discussed in Chapter 3, which also paves the way for both the error analysis discussed in the thesis of Panzer [148], and an iterated reduction scheme using Krylov subspaces. The latter allows to accumulate a reduced model by any number of independently reduced ones; the resulting framework is denoted as *cumulative model order reduction* (CURE) and is discussed in [148] and Chapter 3.

The concept of $\mathcal{H}_2$ pseudo-optimality is presented in Chapter 4. Its relation to $\mathcal{H}_2$ optimal MOR is disclosed, together with its effect on the CURE framework, i. e. it is discussed how CURE benefits from $\mathcal{H}_2$ pseudo-optimality. The chapter provides the first general and detailed description of $\mathcal{H}_2$ pseudo-optimality in the context of Krylov subspaces.

The application of both the CURE framework and $\mathcal{H}_2$ pseudo-optimality to solve large-scale Lyapunov equations follows in Chapter 5, which also forms Part III of this thesis. The new results in this chapter are, firstly, the presentation of a numerically efficient low-rank formulation of the residual that results from approximate solutions, secondly, the disclosure of how to replicate the ADI iteration by rational Krylov subspace methods, and thirdly, the exploitation of this link in order to propose enhancements of the ADI iteration.

Part II

Theory: $\mathcal{H}_2$ Pseudo-Optimality

2 Duality of Sylvester Equations and Krylov Subspaces

Krylov subspaces have been studied mainly by numerical mathematicians, who developed practicable algorithms: the Arnoldi and Lanczos processes. The characterization of a Krylov subspace by its numerical instructions, however, is probably too detailed and non-constructive for system theoretical investigations. There, it is instead convenient (and as we will see, also sufficient) to use a more abstract level, namely Sylvester equations and a projection-based framework.

The main efforts in this direction are presumably the two theses by Grimme [85] and Vandendorpe [189]. Theorem 1.5 already contains the main result of Grimme in condensed form, whereas this chapter serves to review results due to Vandendorpe—among others—, and to present some new results already published in [211, 212].

The focus of this chapter is to uncover the strong connection of Krylov subspaces and particular Sylvester equations, which can actually be perceived as a duality. It should be stressed, that all discussions in the subsequent chapters rely on this duality, which turn Sylvester equations into the fundamental tool of this work. It is therefore important to acquaint oneself with the upcoming notation in order to facilitate examination of the subsequent chapters.

A certain type of Sylvester equation is of interest here, which is characterized in Section 2.1, together with its numerical solution. This type also yields the solution for the pole placement problem in control theory, which is reviewed in Section 2.2. The fundamental duality of rational Krylov subspaces and Sylvester equations is reviewed in Section 2.3 and an extension is presented in Section 2.4. The results of these sections not only provide elementary tools for the subsequent chapters, but also entail a convenient parametrization of all reduced models of order n that interpolate the original one at given n points, which is presented in Section 2.5. Finally, these results are discussed in Section 2.6, and it is analysed how the previously mentioned degrees of freedom s_i and $\mathbf{W}$ translate into the Sylvester framework.

2.1 Sparse-Dense Sylvester Equations

The Sylvester equation of interest in this work has a particular structure, which will be denoted hereafter as “sparse-dense”, and which is defined as follows.

Definition 2.1. Given the large and sparse matrices $\mathbf{A},\mathbf{E}\in\mathbb{C}^{N\times N}$ and $\mathbf{B}\in\mathbb{C}^{N\times m}$, let $\mathbf{S},\mathbf{R}\in\mathbb{C}^{n\times n}$ and $\mathbf{L}\in\mathbb{C}^{m\times n}$, with $n\ll N$, be small and dense, then

$$\mathbf{AVR}-\mathbf{EVS}=\mathbf{BL} \tag{2.1}$$

is called *sparse-dense Sylvester equation* for the solution $\mathbf{V}\in\mathbb{C}^{N\times n}$.

It should be noted, that the sign convention in (2.1) will facilitate the following discussions. The dimensions of the matrices in (2.1) and also the sparsity of $\mathbf{A}$, $\mathbf{E}$, and $\mathbf{B}$ are exemplified in Figure 2.1, where each “$\star$” denotes a non-zero entry. It is essential that the matrices on the right-hand sides of all products in (2.1) are small (because then they are allowed to be dense, too). This is the key for the solution of (2.1)—notwithstanding its large order. By contrast, the large-scale Lyapunov equations (1.11) and (1.12), or the Sylvester equation for the Cross-Gramian, share large and sparse matrices on both sides of all products (which in this context could be called *sparse-sparse*). They often require iterative solution techniques, which are separately discussed in Part III. Although the results therein can be generalized to solve (sparse-sparse) Sylvester equations as well, only Lyapunov equations are discussed in Part III. Consequently, Sylvester equations appear in this work only in the form (2.1), and for a concise presentation we may therefore drop the preceding term “sparse-dense”, i. e. unless explicitly announced, “Sylvester equation” will always refer to a sparse-dense one hereafter.

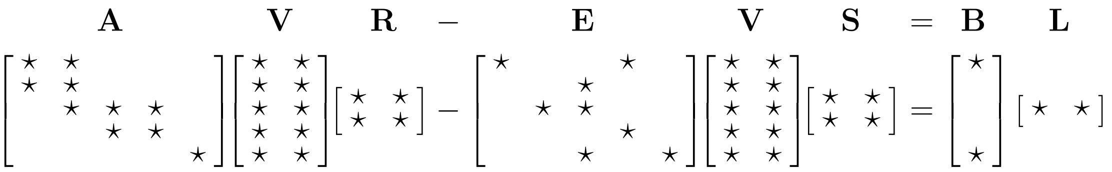


Figure 2.1: Dimensions and sparsity of matrices in sparse-dense Sylvester equations.

In the Sylvester equations of interest to this work, the matrix $\mathbf{R}$ in (2.1) is non-singular. Then $\mathbf{R}$ may be cancelled and (2.1) changes to

$$\mathbf{AV}-\mathbf{EVS}=\mathbf{BL}, \tag{2.2}$$

where $\mathbf{SR}^{-1} \to \mathbf{S}$ and $\mathbf{LR}^{-1} \to \mathbf{L}$. The benefit is that the direct solution techniques of Section 1.5.1 can be readily generalized to solve (2.2) by transforming the small and dense $\mathbf{S}$ to Schur form, and subsequently, compute the columns of the transformed $\widetilde{\mathbf{V}}$ by forward substitutions. The solution $\mathbf{V}$ of (2.2) is finally found by back transformation; a pseudo-code of this method is shown in Algorithm 2.1, where $\mathbf{M}_{ij}$ denotes the (i, j)-entry of a matrix $\mathbf{M}$, and where $\mathbf{M}_i$ denotes the ith column of $\mathbf{M}$.

Algorithm 2.1 Solution of sparse-dense Sylvester equations (2.2)

```
Input: $\mathbf{E}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{S}$, $\mathbf{L}$
Output: $\mathbf{V}$ such that $\mathbf{AV} - \mathbf{EVS} = \mathbf{BL}$
1: $\mathbf{S} = \mathbf{U}\widetilde{\mathbf{S}}\mathbf{U}^*$, with $\mathbf{U}^*\mathbf{U} = \mathbf{I}$ and $\widetilde{\mathbf{S}}$ upper triangular        // Schur decomposition
2: $\widetilde{\mathbf{L}} = \mathbf{LU}$
3: for $i = 1, \ldots, n$ do
4:     Solve $\left(\mathbf{A} - \widetilde{\mathbf{S}}_{ii}\mathbf{E}\right)\widetilde{\mathbf{V}}_i = \mathbf{B}\widetilde{\mathbf{L}}_i + \sum_{j=1}^{i-1} \widetilde{\mathbf{S}}_{ji}\widetilde{\mathbf{V}}_j$    for $\widetilde{\mathbf{V}}_i$
5: end for
6: $\widetilde{\mathbf{V}} = \left[\widetilde{\mathbf{V}}_1, \ldots, \widetilde{\mathbf{V}}_n\right]$
7: $\mathbf{V} = \widetilde{\mathbf{V}}\mathbf{U}^*$
```

Owing to $n \ll N$, the main numerical effort in Algorithm 2.1 are n solves for LSEs of order N in Step 4 (which is already a first hint at the duality of Krylov subspaces and Sylvester equations). With the same reasoning as in Section 1.6.3, we may therefore assume that (2.2) is solvable in admissible time—regardless of whether direct or iterative methods are used in Step 4.

If one uses an eigenvalue decomposition in Step 1 instead of the Schur decomposition, a closed formula for the solution $\mathbf{V}$ can be stated, which was shown in [8, 177]: let $\mathbf{S} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1}$ denote the eigen-decomposition, such that $\mathbf{t}_i$ are the columns of $\mathbf{T}$ and $\hat{\mathbf{t}}_i^*$ are the rows of $\mathbf{T}^{-1}$, and assume that $\boldsymbol{\Lambda}$ is diagonal with entries λ_i. Then,

$$\mathbf{V} = \sum_{i=1}^{n} (\mathbf{A} - \lambda_i\mathbf{E})^{-1} \mathbf{BLt}_i\hat{\mathbf{t}}_i^* \tag{2.3}$$

$$= \left[(\mathbf{A} - \lambda_1\mathbf{E})^{-1} \mathbf{BLt}_1, \ldots, (\mathbf{A} - \lambda_n\mathbf{E})^{-1} \mathbf{BLt}_n\right] \mathbf{T}^{-1}. \tag{2.4}$$

Further details on the numerical solution of (2.2) can be found e. g. in [39]. Owing to Lemma 1.2, we are mainly interested in the subspace that is spanned by the columns of $\mathbf{V}$; the actual basis becomes relevant only in numerical implementations. We therefore state the following invariance property.

Lemma 2.1. *Let $\mathbf{V}$ satisfy (2.2). Then every $\widetilde{\mathbf{V}}$, which spans the same subspace,*

$\operatorname{span}(\widetilde{\mathbf{V}}) = \operatorname{span}(\mathbf{V})$, *solves a Sylvester equation*

$$\mathbf{A}\widetilde{\mathbf{V}} - \mathbf{E}\widetilde{\mathbf{V}}\widetilde{\mathbf{S}} = \mathbf{B}\widetilde{\mathbf{L}}, \tag{2.5}$$

where $\mathbf{S}$ *and* $\widetilde{\mathbf{S}}$ *share equal Jordan canonical form.*

Proof. Because of $\operatorname{span}(\widetilde{\mathbf{V}}) = \operatorname{span}(\mathbf{V})$, there exists a non-singular $\mathbf{T} \in \mathbb{C}^{n\times n}$, such that $\widetilde{\mathbf{V}} = \mathbf{V}\mathbf{T}$. Substituting $\mathbf{V} = \widetilde{\mathbf{V}}\mathbf{T}^{-1}$ in (2.2), and multiplication with $\mathbf{T}$ from the right, yields (2.5), with $\widetilde{\mathbf{L}} = \mathbf{L}\mathbf{T}$ and $\widetilde{\mathbf{S}} = \mathbf{T}^{-1}\mathbf{S}\mathbf{T}$, which completes the proof. □

2.2 An Excursus on Sylvester Equations and Pole Placement

Before discussing the connection to rational Krylov subspaces, this section provides a short excursus on a method for static state-feedback in control theory known as the *pole-placement* problem: we are searching for a feedback $\mathbf{u}(t) = -\mathbf{R}\mathbf{x}(t)$ such that the closed-loop system $\mathbf{E}\dot{\mathbf{x}}(t) = (\mathbf{A} - \mathbf{B}\mathbf{R})\,\mathbf{x}(t)$ has its eigenvalues at prescribed locations. The following lemma reviews how Sylvester equations can be linked to this problem; it shows, how $n \leq N$ poles/eigenvalues can be assigned, and additionally, that the transfer function, which is obtained from replacing the output by the resulting feedback $\mathbf{R}$ and from adding a unit feed-through, has transmission zeros at the desired locations.

Lemma 2.2. *Given the Sylvester equation*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{S} = \mathbf{B}\mathbf{L}, \tag{2.6}$$

with $\mathbf{S} \in \mathbb{C}^{n\times n}$, $\mathbf{L} \in \mathbb{C}^{m\times n}$, *assume that the pair* $(\mathbf{L}, \mathbf{S})$ *is observable. Let* $\mathbf{\Lambda}(\mathbf{S})$ *denote the set of eigenvalues of* $\mathbf{S}$ *and further assume* $\mathbf{\Lambda}(\mathbf{S}) \cap \mathbf{\Lambda}(\mathbf{E}^{-1}\mathbf{A}) = \emptyset$. *If* $\mathbf{R} \in \mathbb{C}^{m\times N}$ *is such that* $\mathbf{R}\mathbf{V} = \mathbf{L}$, *then the following holds.*

- *Pole placement:* $\mathbf{\Lambda}(\mathbf{S})$ *become eigenvalues of* $\mathbf{E}^{-1}(\mathbf{A} - \mathbf{B}\mathbf{R})$,
- *Zero placement:* $\mathbf{\Lambda}(\mathbf{S})$ *become transmissions zeros of* $\mathbf{R}\,(s\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B} + \mathbf{I}$.

Proof. Observability of the pair $(\mathbf{L}, \mathbf{S})$ and $\mathbf{\Lambda}(\mathbf{S}) \cap \mathbf{\Lambda}(\mathbf{E}^{-1}\mathbf{A}) = \emptyset$ guarantee that $\mathbf{V}$ has full column rank. Then, substituting $\mathbf{R}\mathbf{V} = \mathbf{L}$ in the Sylvester equation (2.6) reads as $(\mathbf{A} - \mathbf{B}\mathbf{R})\,\mathbf{V} = \mathbf{E}\mathbf{V}\mathbf{S}$, which proves that the n eigenvalues of $\mathbf{S}$ are assigned in the closed loop, and furthermore, that $\mathbf{V}$ spans the corresponding invariant subspace in the closed-loop system.

To prove the "zero placement", let s_i denote an eigenvalue of $\mathbf{S}$. Then s_i is a transmission zero of the above systems if there exists an $\mathbf{l}_i$, such that $\left[\mathbf{R}\,(s_i\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B} + \mathbf{I}\right]\mathbf{l}_i = \mathbf{0}$, cf. [222], which is equivalent to $\mathbf{R}\,(s_i\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}\mathbf{l}_i = -\mathbf{l}_i$. Because of Lemma 2.1, we may assume without loss of generality that $\mathbf{S} = \operatorname{diag}(s_1, \ldots, s_n)$ and $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_n]$. Then using (2.4), the columns $\mathbf{v}_i$ of $\mathbf{V}$ are given by $\mathbf{v}_i = (\mathbf{A}-s_i\mathbf{E})^{-1}\,\mathbf{B}\mathbf{l}_i$, and therefore, $-\mathbf{R}\,(s_i\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}\mathbf{l}_i = \mathbf{R}\mathbf{v}_i$. From $\mathbf{R}\mathbf{V} = \mathbf{L}$ it follows that $\mathbf{R}\mathbf{v}_i = \mathbf{l}_i$, and thus, $\mathbf{R}\,(s_i\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}\mathbf{l}_i = -\mathbf{l}_i$, which is the above condition for a transmission zero. □

This lemma shows that the solution of a Sylvester equation can be used to place $n \leq N$ poles of a closed-loop system. If $n < N$, then this is also known as *partial pole placement*, which paves the way to an iterative procedure of assigning the desired poles. Algorithm 2.2 shows how all N eigenvalues can be assigned one after the other, by iteratively computing feedbacks $\mathbf{R}_i$ that place only one desired eigenvalue—without affecting previously assigned ones. The total feedback that assigns all N poles is then given by $\mathbf{R} = \sum_{i=1}^{N} \mathbf{R}_i$.

Algorithm 2.2 Iterative Pole Placement

```
Input: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_n]$, $\mathbf{S} = \operatorname{diag}(s_1, \ldots, s_n)$
Output: $\mathbf{R}$ such that $\boldsymbol{\Lambda}(\mathbf{A}-\mathbf{B}\mathbf{R}) = \boldsymbol{\Lambda}(\mathbf{S})$
1: $\mathbf{V} = []$
2: for $i = 1 \to N$ do
3:     $\mathbf{v}_i = (\mathbf{A} - s_i\mathbf{I})^{-1}\,\mathbf{B}\mathbf{l}_i$
4:     $\mathbf{V} = [\mathbf{V},\, \mathbf{v}_i]$
5:     $\widehat{\mathbf{L}} = [\mathbf{0},\, \ldots,\, \mathbf{0},\, \mathbf{l}_i]$
6:     find $\mathbf{R}_i$ such that $\mathbf{R}_i\mathbf{V} = \widehat{\mathbf{L}}$,          // e.g. $\mathbf{R}_i = \widehat{\mathbf{L}}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T$
7:     $\mathbf{A} \leftarrow \mathbf{A} - \mathbf{B}\mathbf{R}_i$
8: end for
9: $\mathbf{R} = \sum_{i=1}^{N} \mathbf{R}_i$
```

An elaborate discussion of the pole placement problem in the multi-variable case $m > 1$ is due to Roppenecker [163]. Up to the author's knowledge, Bhattacharyya and De Souza [37] were the first ones to discover the link between pole placement and Sylvester equations, which was then used in [43] to maximize the conditioning of **A**–**BR**. This is possible due to the (almost) free choice of $\mathbf{l}_i$ in Algorithm 2.2. Kautsky et. al. [116] used this degree of freedom to minimize sensitivity to perturbations in the system matrices, which is the basis of the function `place` in Matlab. It can be shown [144], that this also maximizes a stability margin with respect to disturbances of the system matrices.

If only $n<N$ poles are assigned, there is an additional degree of freedom in the $N-n$ remaining eigenvalues, which corresponds to the different solutions that are possible in Step 6 of Algorithm 2.2. The natural idea is to pick n eigenvalues of $\mathbf{E}^{-1}\mathbf{A}$, which then

are assigned at the desired locations, and leave the remaining N–n ones unchanged [163, 167]. It is however also possible to use the not explicitly assigned $N-n$ eigenvalues as an additional degree of freedom to minimize the conditioning of the closed-loop system [122], or to minimize the norm of the feedback $\mathbf{R}$, and thereby assuring that the remaining $N-n$ eigenvalues are assigned inside a pre-defined region in the complex plane [49].

An iterative pole placement like Algorithm 2.2 is also presented in [162]. Furthermore, a link between pole placement and the *linear quadratic regulator* (LQR) is discussed in [164]; it, however, can be shown with a simple example, cf. [57], that the poles resulting from LQR cannot be located at arbitrary positions.

As we have seen, the Sylvester equation provides the solution for the fundamental pole placement problem in control theory. As any method for determining a static state-feedback $\mathbf{R}$ in fact places the poles at some locations, the Sylvester equation (2.6) can be seen as a parametrization of all possible feedbacks. The following sections instead show that this Sylvester equation also serves as a parametrization of all possible reduced order models.

2.3 Sylvester Equation for the Interpolation Data

This section presents the duality of rational Krylov subspaces and Sylvester equations. It will be shown, that the interpolation data, i. e. the shifts s_i and (in the MIMO case also) the tangential directions $\mathbf{l}_i$, may conveniently be specified by the matrices $\mathbf{S}$ and $\mathbf{L}$ from (2.2). This turns the Sylvester equation into the fundamental tool in this work. The section comprises results from [76, 79, 189]; however, presentation is slightly different to better meet the needs of this work. As the statement of the most general case requires quite cumbersome notation, we first start with the case that the rational Krylov subspace contains only a single expansion point s_0.

Lemma 2.3. *Given the expansion point s_0 and the tangential directions $\mathbf{l}_1,\dots,\mathbf{l}_q$, assume that s_0 is not an eigenvalue of $\mathbf{E}^{-1}\mathbf{A}$. Then the columns of $\mathbf{V}\in\mathbb{C}^{N\times q}$ form a basis of the tangential-input rational Krylov subspace*

$$\operatorname{span}(\mathbf{V}) = \operatorname{span}\left\{\mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_1, \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_1 + \mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_2, \dots, \sum_{\nu=0}^{q-1}\left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{\nu}\mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_{q-\nu}\right\}, \quad (2.7)$$

if and only if there exists an observable pair $(\mathbf{L},\mathbf{S})$, $\mathbf{S}\in\mathbb{C}^{q\times q}$, $\mathbf{L}\in\mathbb{C}^{m\times q}$, which admits *the Jordan canonical form* $\mathbf{J}$,

$$\mathbf{T}^{-1}\mathbf{S}\mathbf{T} = \mathbf{J} = \begin{bmatrix} s_0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & s_0 \end{bmatrix}, \quad \textit{and} \quad \mathbf{L}\mathbf{T} = [\mathbf{l}_1, \ldots, \mathbf{l}_q], \tag{2.8}$$

for an appropriate transformation matrix $\mathbf{T} \in \mathbb{C}^{q \times q}$, *such that the Sylvester equation*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{S} = \mathbf{B}\mathbf{L} \tag{2.9}$$

is satisfied.

Moreover, the reduced model $\boldsymbol{G}_r(s) = \mathbf{C}_r(s\mathbf{E}_r – \mathbf{A}_r)^{-1}\mathbf{B}_r$ *from (1.7) fulfils the tangential interpolation*

$$\left(\mathbf{M}_0^{s_0} - \widehat{\mathbf{M}}_0^{s_0}\right)\mathbf{l}_1 = \mathbf{0}, \tag{2.10}$$

$$\left(\mathbf{M}_0^{s_0} - \widehat{\mathbf{M}}_0^{s_0}\right)\mathbf{l}_2 + \left(\mathbf{M}_1^{s_0} - \widehat{\mathbf{M}}_1^{s_0}\right)\mathbf{l}_1 = \mathbf{0}, \tag{2.11}$$

$$\vdots$$

$$\sum_{\nu=0}^{q-1}\left(\mathbf{M}_\nu^{s_0} - \widehat{\mathbf{M}}_\nu^{s_0}\right)\mathbf{l}_{q-\nu} = \mathbf{0}, \tag{2.12}$$

if s_0 *is not a pole of* $\boldsymbol{G}_r(s)$.

Proof. Use Lemma 2.1 to transform $\mathbf{S}$ in (2.9) by $\mathbf{T}$ to Jordan canonical form, then the proof is contained in [76, 79, 189]. An alternative proof is actually given in Theorem 2.15. □

The next theorem generalizes the above result to the most general case and thereby describes the duality of Krylov subspaces and Sylvester equations.

Theorem 2.4 (The "duality")**.** *Given* k *distinct expansion points* s_i, $i = 1, \ldots, k$, *assume that none of them is an eigenvalue of* $\mathbf{E}^{-1}\mathbf{A}$, *and assign to each shift of them* m_i *Jordan blocks* $\mathbf{J}_{ij}$, $j = 1, \ldots, m_i$, *of dimension* q_{ij}:

$$\mathbf{J}_{ij} = \begin{bmatrix} s_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & s_i \end{bmatrix} \in \mathbb{C}^{q_{ij} \times q_{ij}}. \tag{2.13}$$

Let there be q_{ij} *tangential directions for each Jordan block* $\mathbf{J}_{ij}$, *such that* $K = \sum_{i=1}^{k} m_i$ *tangential-input rational Krylov subspaces like (2.7) can be defined. Then the columns* *of* $\mathbf{V}$ *form a basis of the union of all* K *tangential-input rational Krylov subspaces if and only if there exists an observable pair* $(\mathbf{L},\mathbf{S})$ *of appropriate dimensions, which admits the Jordan canonical form* $\mathbf{J}$*,*

$$\mathbf{T}^{-1}\mathbf{S}\mathbf{T}=\mathbf{J}=\operatorname{diag}\left(\mathbf{J}_{11},\ldots,\mathbf{J}_{1m_1},\mathbf{J}_{21},\ldots,\mathbf{J}_{2m_2},\ldots,\mathbf{J}_{k1},\ldots,\mathbf{J}_{km_k}\right), \tag{2.14}$$

for an appropriate transformation matrix $\mathbf{T}$*, such that* $\mathbf{LT}$ *has the aforementioned tangential directions as columns, and such that the Sylvester equation*

$$\mathbf{AV}-\mathbf{EVS}=\mathbf{BL} \tag{2.15}$$

is satisfied. Then $1\leq m_i\leq m$ *holds true.*

Moreover, the reduced system $\boldsymbol{G}_r(s)=\mathbf{C}_r(s\mathbf{E}_r-\mathbf{A}_r)^{-1}\mathbf{B}_r$ *from (1.7) fulfils the tangential interpolation defined by (2.10)–(2.12) for all* K *Jordan blocks, if none of the* s_i *is a pole of* $\boldsymbol{G}_r(s)$*.*

Proof. Following the lemma of Hautus, the pair $(\mathbf{L},\mathbf{S})$ is observable, if and only if $\forall s_i$, $i=1,\ldots,k$, $\operatorname{rank}\left[s_i\mathbf{I}-\mathbf{S}^*,\ \mathbf{L}^*\right]=n$. Therefore, a necessary condition for observability is that the geometric multiplicity of each eigenvalue s_i is smaller or equal to m, which proves $1\leq m_i\leq m$. The rest of the proof follows from the fact that the K Jordan blocks are decoupled in the eigen-decomposition of $\mathbf{S}$, such that Lemma 2.3 may be applied to each of them independently. An alternative proof can also be found in [189]. □

Theorem 1.5 shows that there is some kind of duality between Krylov subspaces and solutions of Sylvester equations: any basis of a Krylov subspace solves a particular Sylvester equation with an observable pair $(\mathbf{L},\mathbf{S})$, where the shifts s_i correspond to the eigenvalues of $\mathbf{S}$ (including higher multiplicities), and the tangential directions correspond to the columns of $\mathbf{L}$ (after transforming $\mathbf{S}$ to Jordan canonical form). Conversely, any solution of a sparse-dense Sylvester equation (2.2), where the pair $(\mathbf{L},\mathbf{S})$ is observable, spans a rational Krylov subspace, where the expansion points and tangential directions are encoded in $\mathbf{S}$ and $\mathbf{L}$. The pair $(\mathbf{S},\mathbf{L})$ thus serves as a convenient specification of the interpolation data: eigenvalues of $\mathbf{S}$ correspond to expansion points, where higher multiplicities are reflected in Jordan blocks, and the tangential directions are determined by the columns of $\mathbf{L}$.

The pair $(\mathbf{L},\mathbf{S})$ will be used subsequently to derive various result, so it is important to get an idea of its structure. For the sake of generality, however, Theorem 2.4 requires quite cumbersome notation, whereas this can be significantly simplified in relevant cases for this work, such as single-input or block-input Krylov subspaces. Although this is actually redundant, subsequently three cases are examined in more detail, in order to clarify the structure of $(\mathbf{L}, \mathbf{S})$.

Corollary 2.5 (single-input, one shift). *Given the expansion point s_0, assume $m=1$ and that s_0 is not an eigenvalue of $\mathbf{E}^{-1}\mathbf{A}$. Then the columns of $\mathbf{V}\in\mathbb{C}^{N\times q}$ form a basis of the single-input rational Krylov subspace*

$$\operatorname{span}(\mathbf{V}) = \operatorname{span}\left\{\mathbf{A}_{s_0}^{-1}\mathbf{b},\ \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{b},\ \ldots,\ \left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{q-1}\mathbf{A}_{s_0}^{-1}\mathbf{b}\right\}, \tag{2.16}$$

if and only if there exists an observable pair $(\mathbf{l}, \mathbf{S})$, $\mathbf{S}\in\mathbb{C}^{q\times q}$, $\mathbf{l}\in\mathbb{C}^{1\times q}$, which admits the Jordan canonical form $\mathbf{J}$,

$$\mathbf{T}^{-1}\mathbf{S}\mathbf{T} = \mathbf{J} = \begin{bmatrix} s_0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & s_0 \end{bmatrix}, \tag{2.17}$$

for an appropriate transformation matrix $\mathbf{T}\in\mathbb{C}^{q\times q}$, such that the Sylvester equation

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{S} = \mathbf{b}\mathbf{l}, \tag{2.18}$$

is satisfied.

Moreover, the reduced model $\boldsymbol{G}_r(s)=\mathbf{C}_r(s\mathbf{E}_r-\mathbf{A}_r)^{-1}\mathbf{B}_r$ from (1.7) matches the $p\times 1$ moments $\mathbf{M}_i^{s_0}=\widehat{\mathbf{M}}_i^{s_0}$, $i=0,\ldots,q-1$ if s_0 is not a pole of $\boldsymbol{G}_r(s)$.

Corollary 2.5 is a direct consequence of Lemma 2.3, because scalar tangential directions do not alter the Krylov subspace, nor do they influence moment matching. To illustrate Corollary 2.5, consider the following example.

Example 2.1. Let $\mathbf{S}=\mathbf{J}$ be given by (2.17) and let $\mathbf{l}=[1,0,\ldots,0]\in\mathbb{R}^{1\times q}$. Then the Sylvester equation (2.18) is solved by

$$\mathbf{V} = \left[\mathbf{A}_{s_0}^{-1}\mathbf{b},\ \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{b},\ \ldots,\ \left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{q-1}\mathbf{A}_{s_0}^{-1}\mathbf{b}\right]. \tag{2.19}$$

Corollary 2.6 (block-input, multiple shifts). *Given k distinct expansion points s_i, $i=1,\ldots,k$, assume that none of them is an eigenvalue of $\mathbf{E}^{-1}\mathbf{A}$. Then the columns of $\mathbf{V}\in\mathbb{C}^{N\times km}$ form a basis of the block-input rational Krylov subspace*

$$\operatorname{span}(\mathbf{V}) = \operatorname{span}\left\{\mathbf{A}_{s_1}^{-1}\mathbf{B},\ \mathbf{A}_{s_2}^{-1}\mathbf{B},\ \ldots,\ \mathbf{A}_{s_k}^{-1}\mathbf{B}\right\}, \tag{2.20}$$

if and only if there exists an observable pair $(\mathbf{L},\mathbf{S})$, $\mathbf{S}\in\mathbb{C}^{km\times km}$, $\mathbf{L}\in\mathbb{C}^{m\times km}$, *which admits the Jordan canonical form* $\mathbf{J}$,

$$\mathbf{T}^{-1}\mathbf{S}\mathbf{T}=\mathbf{J}=\begin{bmatrix} s_1\mathbf{I} & & \\ & \ddots & \\ & & s_k\mathbf{I}\end{bmatrix}, \quad \textit{and} \quad \mathbf{L}\mathbf{T}=[\mathbf{I},\ldots,\mathbf{I}], \tag{2.21}$$

where $\mathbf{I}$ *is the* $m\times m$ *identity matrix, and for an appropriate transformation matrix* $\mathbf{T}\in\mathbb{C}^{km\times km}$, *such that the Sylvester equation*

$$\mathbf{A}\mathbf{V}-\mathbf{E}\mathbf{V}\mathbf{S}=\mathbf{B}\mathbf{L}, \tag{2.22}$$

is satisfied.

Moreover, the reduced model $\boldsymbol{G}_r(s)=\mathbf{C}_r(s\mathbf{E}_r-\mathbf{A}_r)^{-1}\mathbf{B}_r$ *from (1.7) matches the* $p\times m$ *block moments* $\mathbf{M}_0^{s_i}=\widehat{\mathbf{M}}_0^{s_i}$, $i=1,\ldots,k$ *if none of the* s_i *is a pole of* $\boldsymbol{G}_r(s)$.

Corollary 2.6 follows from Theorem 2.4, with $m_i=m$, $\forall i$, and $q_{ij}=1$, $\forall i,j$. Then there are m tangential directions to each eigenvalue, which have to span the whole $\mathbb{R}^m$, owing to observability. Then complete block moments are matched instead of individual tangential directions. To illustrate Corollary 2.6, again consider a short example.

Example 2.2. Let $\mathbf{I}$ denote the $m\times m$ identity matrix and $\mathbf{S}$, $\widetilde{\mathbf{S}}$ and $\mathbf{L}$, $\widetilde{\mathbf{L}}$ be given by

$$\mathbf{S}=\begin{bmatrix} s_1\mathbf{I} & & \\ & \ddots & \\ & & s_k\mathbf{I}\end{bmatrix}, \qquad \widetilde{\mathbf{S}}=\begin{bmatrix} s_1\mathbf{I} & \mathbf{I} & & \\ & \ddots & \ddots & \\ & & \ddots & \mathbf{I} \\ & & & s_k\mathbf{I}\end{bmatrix}, \quad \text{and} \tag{2.23}$$

$$\mathbf{L}=[\,\mathbf{I}\;\;\ldots\;\;\mathbf{I}\,], \qquad \widetilde{\mathbf{L}}=[\,\mathbf{I}\;\;\mathbf{0}\;\;\ldots\;\;\mathbf{0}\,], \tag{2.24}$$

Then the Sylvester equations $\mathbf{A}\mathbf{V}-\mathbf{E}\mathbf{V}\mathbf{S}=\mathbf{B}\mathbf{L}$, and $\mathbf{A}\widetilde{\mathbf{V}}-\mathbf{E}\widetilde{\mathbf{V}}\widetilde{\mathbf{S}}=\mathbf{B}\widetilde{\mathbf{L}}$ are solved by

$$\mathbf{V}=\left[\mathbf{A}_{s_1}^{-1}\mathbf{B},\;\ldots,\;\mathbf{A}_{s_k}^{-1}\mathbf{B}\right], \quad \text{and} \tag{2.25}$$

$$\widetilde{\mathbf{V}}=\left[\mathbf{A}_{s_1}^{-1}\mathbf{B},\;\mathbf{A}_{s_2}^{-1}\mathbf{E}\mathbf{A}_{s_1}^{-1}\mathbf{B},\;\ldots,\;\mathbf{A}_{s_k}^{-1}\mathbf{E}\ldots\mathbf{A}_{s_2}^{-1}\mathbf{E}\mathbf{A}_{s_1}^{-1}\mathbf{B}\right]. \tag{2.26}$$

This example shows, how switching to a nested basis (2.26) of the same Krylov subspace (2.25) affects the matrices $\mathbf{S}$ and $\mathbf{L}$ in the Sylvester equation.

Corollary 2.7 (tangential-input). *Given* n *distinct expansion points* s_i, $i=1,\ldots,n$, *and the tangential directions* $\mathbf{l}_1,\ldots,\mathbf{l}_n$, *assume that none of the* s_i *is an eigenvalue of* $\mathbf{E}^{-1}\mathbf{A}$. *Then the columns of* $\mathbf{V}\in\mathbb{C}^{N\times n}$ *form a basis of the tangential-input rational*

Krylov subspace

$$\mathrm{span}(\mathbf{V}) = \mathrm{span}\left\{\mathbf{A}_{s_1}^{-1}\mathbf{B}\mathbf{l}_1,\ \mathbf{A}_{s_2}^{-1}\mathbf{B}\mathbf{l}_2,\ \ldots,\ \mathbf{A}_{s_n}^{-1}\mathbf{B}\mathbf{l}_n\right\}, \tag{2.27}$$

if and only if there exists an observable pair $(\mathbf{L}, \mathbf{S})$*,* $\mathbf{S} \in \mathbb{C}^{n \times n}$*,* $\mathbf{L} \in \mathbb{C}^{m \times n}$*, which admits the Jordan canonical form* $\mathbf{J}$*,*

$$\mathbf{T}^{-1}\mathbf{S}\mathbf{T} = \mathbf{J} = \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_n \end{bmatrix}, \quad \textit{and} \quad \mathbf{L}\mathbf{T} = [\mathbf{l}_1, \ldots, \mathbf{l}_n], \tag{2.28}$$

for an appropriate transformation matrix $\mathbf{T} \in \mathbb{C}^{n \times n}$*, such that the Sylvester equation*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{S} = \mathbf{B}\mathbf{L}, \tag{2.29}$$

is satisfied.

Moreover, the reduced model $\boldsymbol{G}_r(s) = \mathbf{C}_r(s\mathbf{E}_r – \mathbf{A}_r)^{-1}\mathbf{B}_r$ *from (1.7) fulfils the tangential interpolation* $\left(\mathbf{M}_0^{s_i} - \widehat{\mathbf{M}}_0^{s_i}\right)\mathbf{l}_i = \mathbf{0}$*,* $i = 1, \ldots, n$ *if none of the* s_i *is a pole of* $\boldsymbol{G}_r(s)$*.*

Corollary 2.7 follows from Theorem 2.4, with $m_i = 1$, $\forall i$ and $q_{ij} = 1$, $\forall i, j$, and is illustrated in the following example.

Example 2.3. Let both $\mathbf{S}$ and $\mathbf{L}$ be given as in (2.28). Then the Sylvester equation (2.29) is solved by

$$\mathbf{V} = \left[\mathbf{A}_{s_1}^{-1}\mathbf{B}\mathbf{l}_1,\ \mathbf{A}_{s_2}^{-1}\mathbf{B}\mathbf{l}_2,\ \ldots,\ \mathbf{A}_{s_n}^{-1}\mathbf{B}\mathbf{l}_n\right]. \tag{2.30}$$

Remark 2.8. Theorem 2.4 and Corollaries 2.5–2.7 describe the duality of Krylov subspaces and Sylvester equations for moment matching/tangential interpolation. This duality, however, can be generalized in two ways. It is possible to incorporate invariant subspaces in $\mathbf{V}$, which would lead to modal approximation. The generalized eigenvectors in $\mathbf{V}$ would then correspond to the unobservable part in the pair $(\mathbf{L}, \mathbf{S})$, cf. [189]. It is further possible to include the subspace (1.31) in $\mathbf{V}$, which would lead to matching the Markov parameters. This case can be incorporated in the Sylvester equation by either introducing a singular $\mathbf{R}$ in (2.1), cf. [189], or by replacing $\mathbf{B}$ with $\mathbf{B}_{m_\infty} = \left(\mathbf{A}\mathbf{E}^{-1}\right)^{m_\infty}\mathbf{B}$, cf. [212]. For further details please refer to [76, 79, 189, 212], as these generalizations are irrelevant for this work.

Remark 2.9. Sylvester equations can not only be connected to Krylov subspaces, but also to the Loewner matrix, cf. [8, Remark 6.1.2]. The Loewner and shifted Loewner matrices are the main tool for an approach to model a system, when only given a set of measurements of its frequency response. This, however, is a different concept than the one pursued in this work, as it is based on given (tangential) interpolation data, instead of a given state-space realization $(\mathbf{A}, \mathbf{E}, \mathbf{B}, \mathbf{C})$. The interested reader is therefore referred to [125, 134] for further details.

To summarise, the matrix $\mathbf{V}$ may be equivalently interpreted in three different ways:

- the solution of a particular Sylvester equation,
- a matrix whose columns span the union of Krylov subspaces, or
- the outcome of a numerical procedure such as the Arnoldi or Lanczos processes.

It was already mentioned, that the first interpretation—that is, solution of a Sylvester equation—is most appropriate for system theoretical considerations. To this end, we have to find a triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ that composes the Sylvester equation (2.15), in order to turn this equation into a tool. There are three cases that may occur, when computing a triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$: the first one is, that we encode the desired interpolation data—i. e. shifts and tangential directions—in the matrices $\mathbf{S}$ and $\mathbf{L}$, and then solve the Sylvester equation (2.15). The solution could be based on the ideas presented in Section 2.1, and hence, this case is already completed. The second case is, that $\mathbf{V}$ is already given (analytically as in the examples above, or computed by a numerical procedure), and that we have to compute back to the corresponding $\mathbf{S}$ and $\mathbf{L}$. Finally, the third case is, that a compatible triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ is simultaneously computed by some numerical procedure. The two latter cases are detailed in the remainder of this section.

Assume that an Arnoldi-like process shall be adapted in order to not only compute $\mathbf{V}$, but also $\mathbf{S}$ and $\mathbf{L}$. The basic iterative procedure then is as follows: assume that a compatible triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ is already given and that we want to expand the Krylov subspace by the shift s_0. Then for computing the new column in $\mathbf{V}$, an LSE with the matrix $(\mathbf{A} - s_0\mathbf{E})$ has to be solved, whereas the matrix $\mathbf{S}$ has to be extended by a column and row, with s_0 on the diagonal and zeros elsewhere. If, on the one hand, the right-hand side in the LSE is $\mathbf{B}\mathbf{l}_0$, then $\mathbf{L}$ has to be extended by the additional column $\mathbf{l}_0$, and if, on the other hand, the right-hand side is a previous column in $\mathbf{V}$ this amounts to an extension of $\mathbf{L}$ by zeros, and an additional entry in $\mathbf{S}$ above the diagonal. Subsequently, all operations in the Gram-Schmidt process to orthogonalize the columns of $\mathbf{V}$ can be translated into appropriate modifications of $\mathbf{S}$ and $\mathbf{L}$. A pseudo-code of this approach can be found in [77] and a Matlab implementation in [148].

Now assume, that we want to compute back to the corresponding $\mathbf{S}$ and $\mathbf{L}$ for an already given $\mathbf{V}$. Towards this aim, two approaches are suggested in the following propositions. Both of them are based on projections, so they require the matrices

$\mathbf{A}_r = \mathbf{W}^*\mathbf{A}\mathbf{V}$, $\mathbf{E}_r = \mathbf{W}^*\mathbf{E}\mathbf{V}$ and $\mathbf{B}_r = \mathbf{W}^*\mathbf{B}$, where $\mathbf{W}$ is arbitrary, but such that each s_i is not an eigenvalue of $\mathbf{E}_r^{-1}\mathbf{A}_r$; as, apart from this, $\mathbf{W}$ may chosen arbitrarily, one may simply take $\mathbf{W} = \mathbf{V}$.

Proposition 2.10. *Given* $\mathbf{V}$ *whose columns span an input rational Krylov subspace, and an arbitrary* $\mathbf{W}$*, define* $\mathbf{B}_\perp = \mathbf{B} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$ *and assume that* $\mathbf{E}_r$ *is non-singular and that both* $\mathbf{B}_\perp$ *and* $[\mathbf{E}\mathbf{V}, \mathbf{B}]$ *have full column rank. Then there exists a unique* $\mathbf{S}$ *and a unique* $\mathbf{L}$*, such that the Sylvester equation*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{S} = \mathbf{B}\mathbf{L}, \tag{2.31}$$

is satisfied, and they are given by

$$\mathbf{L} = \left(\mathbf{B}_\perp^*\mathbf{B}_\perp\right)^{-1}\mathbf{B}_\perp^*\left(\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r\right), \tag{2.32}$$

$$\mathbf{S} = \mathbf{E}_r^{-1}\left(\mathbf{A}_r - \mathbf{B}_r\mathbf{L}\right). \tag{2.33}$$

Proof. Existence of $\mathbf{S}$ and $\mathbf{L}$ was proven in Theorem 2.4. Rewrite the Sylvester equation to

$$\begin{bmatrix} \mathbf{E}\mathbf{V} & \mathbf{B} \end{bmatrix}\begin{bmatrix} \mathbf{S} \\ \mathbf{L} \end{bmatrix} = \mathbf{A}\mathbf{V}. \tag{2.34}$$

As we assume that $[\mathbf{E}\mathbf{V}, \mathbf{B}]$ has full column rank, it follows that $\mathbf{S}$ and $\mathbf{L}$ are unique. Multiplying (2.31) from the left with the projector $\mathbf{I} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{W}^*$ yields

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{B}_\perp\mathbf{L}, \tag{2.35}$$

and multiplying (2.35) with $\left(\mathbf{B}_\perp^*\mathbf{B}_\perp\right)^{-1}\mathbf{B}_\perp^*$ from the left yields (2.32); (2.33) then follows from multiplying (2.31) with $\mathbf{W}^*$ from the right. □

A second approach to compute the corresponding $\mathbf{S}$ and $\mathbf{L}$ for a given $\mathbf{V}$ requires a priori knowledge of the expansions points s_i (including their multiplicities) and is presented next.

Proposition 2.11. *Given* $\mathbf{V}$ *whose columns span an input rational Krylov subspace with shifts* s_i *(and in the MIMO case also tangential directions* $\mathbf{l}_i$*), and an arbitrary* $\mathbf{W}$*, but such that* $\mathbf{E}_r$ *is non-singular and such that* $\mathbf{B}_r$ *has full column rank, define the matrix* $\mathbf{J}$ *in Jordan canonical form as in (2.13), (2.14), let* $\widetilde{\mathbf{L}}$ *have the tangential directions as columns, and let* $\mathbf{T}$ *satisfy the small (dense-dense) Sylvester equation*

$$\mathbf{E}_r^{-1}\mathbf{A}_r\,\mathbf{T} - \mathbf{T}\,\mathbf{J} - \mathbf{E}_r^{-1}\mathbf{B}_r\widetilde{\mathbf{L}} = \mathbf{0}. \tag{2.36}$$

Assume that $[\mathbf{EV},\mathbf{B}]$ *has full column rank, then there exists a unique* $\mathbf{S}$ *and a unique* $\mathbf{L}$*, such that the Sylvester equation*

$$\mathbf{AV} - \mathbf{EVS} = \mathbf{BL}, \tag{2.37}$$

is satisfied, and they are given by $\mathbf{S}=\mathbf{TJT}^{-1}$*, and* $\mathbf{L}=\widetilde{\mathbf{L}}\mathbf{T}^{-1}$.

Proof. Existence of $\mathbf{S}$ and $\mathbf{L}$ was proven in Theorem 2.4, and uniqueness in Proposition 2.10. Owing to Theorem 2.4, there is a $\widetilde{\mathbf{V}}$ that satisfies the Sylvester equation

$$\mathbf{A}\widetilde{\mathbf{V}} - \mathbf{E}\widetilde{\mathbf{V}}\mathbf{J} = \mathbf{B}\widetilde{\mathbf{L}}, \tag{2.38}$$

and from Lemma 2.1 it follows that $\widetilde{\mathbf{V}} = \mathbf{VT}$, $\mathbf{J} = \mathbf{T}^{-1}\mathbf{ST}$ and $\widetilde{\mathbf{L}} = \mathbf{LT}$, which proves $\mathbf{S}=\mathbf{TJT}^{-1}$, and $\mathbf{L}=\widetilde{\mathbf{L}}\mathbf{T}^{-1}$. Substituting $\widetilde{\mathbf{V}}=\mathbf{VT}$ in (2.38) and multiplying it with $\mathbf{W}^*$ from the left leads to (2.36), which completes the proof. □

Both approaches of Propositions 2.10 and 2.11 can be used to compute compatible $\mathbf{S}$ and $\mathbf{L}$ for a given $\mathbf{V}$. The former is based on a projection onto the orthogonal complement of span($\mathbf{W}$) and is capable of identifying the shifts that were used in the Krylov subspace. (It was already published in [212].) By contrast, the latter a priori requires the knowledge of the shifts and tangential directions and only computes the correct transformation matrix $\mathbf{T}$ that corresponds to the given $\mathbf{V}$. This approach is based on a projection onto span($\mathbf{EV}$) and is inspired by [14, 15]. The proofs of Propositions 2.10 and 2.11 already contain results that are discussed in the following two sections: the second type of Sylvester equation (2.35) (which is the basis of the error factorization in Chapter 3), and a parametrization of the family of reduced models that matches moments.

It should finally be noted that from now on we may assume that a triple $(\mathbf{V},\mathbf{S},\mathbf{L})$ which fulfils the Sylvester equation $\mathbf{AV}$–$\mathbf{EVS}=\mathbf{BL}$ is given—irrespective of which above mentioned approach is used to compute it. This triple provides all necessary data: $\mathbf{V}$ spans the rational Krylov subspace, the eigenvalues of $\mathbf{S}$ correspond to the expansion points (including higher multiplicities), and the columns of $\mathbf{L}$ encode the tangential directions. Instead of characterizing a rational Krylov subspace by its expansion points s_i with respective multiplicities and tangential directions $\mathbf{l}_i$, we thus will hereafter also use the pair $(\mathbf{S},\mathbf{L})$ to conveniently define the interpolation data of a Krylov subspace. For further details on the Sylvester equation (2.15), its solution and numerical stability please refer to [76, 77, 79, 189, 212].

2.4 Sylvester Equation for the Projection

As we have seen, the Sylvester equation (2.15) provides the desired interpolation data in terms of the eigenvalues of $\mathbf{S}$ and columns of $\mathbf{L}$. It should be stressed, that this interpolation data, still is independent from the remaining degree of freedom in projective MOR—which is the matrix $\mathbf{W}$—, and thus also from the matrices of the reduced model. The purpose of this section is to provide a second type of Sylvester equation for $\mathbf{V}$, which does not encode the interpolation data, but instead the reduced dynamics. Besides the Sylvester equation (2.15), this new Sylvester equation will be the second fundamental tool in the subsequent chapters.

Lemma 2.12. *Given* $\mathbf{V}$ *that solves the Sylvester equation (2.15), and an arbitrary* $\mathbf{W}$*, but such that* $\mathbf{E}_r = \mathbf{W}^*\mathbf{E}\mathbf{V}$ *is non-singular, define the projector* $\mathbf{\Pi} = \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{W}^*$ *and* $\mathbf{B}_\perp = (\mathbf{I}-\mathbf{\Pi})\,\mathbf{B} = \mathbf{B}-\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$*. Then* $\mathbf{V}$ *also satisfies a second type of Sylvester equations, namely*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{B}_\perp\mathbf{L}. \tag{2.39}$$

Proof. The proof simply follows by multiplying (2.15) with $(\mathbf{I}-\mathbf{\Pi})$ from the left. □

In order to distinguish both types of Sylvester equations, we will use the labels "$\mathbf{B}$-Sylvester equation" for the first type (2.15), and "$\mathbf{B}_\perp$-Sylvester equation" for the second one (2.39). The notation "$\mathbf{B}_\perp$" stems from the fact that $\mathbf{B}_\perp$ is orthogonal to the column span of $\mathbf{W}$: $\mathbf{W}^*\mathbf{B}_\perp = \mathbf{0}$. It follows that $\mathbf{B}_\perp$ closes the vector chain from the columns of $\mathbf{B}$ to its respective projections, which are the columns of $\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$: $\mathbf{B} = \mathbf{B}_\perp + \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$. This is illustrated in Figure 2.2.

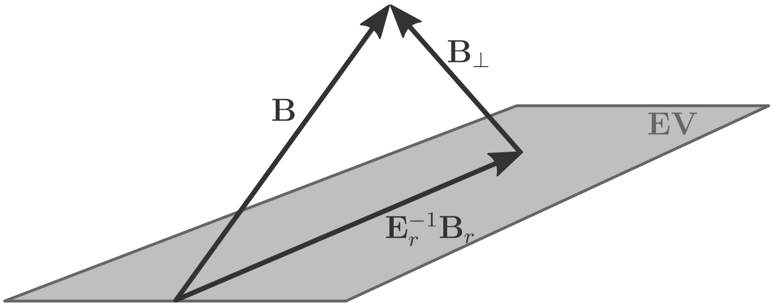


Figure 2.2: Vector chain of $\mathbf{B}$, $\mathbf{B}_\perp$, and $\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$

The result of Lemma 2.12 was published in preliminary form in [211] and in general form in [212]. Although the $\mathbf{B}_\perp$-Sylvester equation can be easily derived, it has been largely overlooked in the literature. This is remarkable, as (2.39) can be seen as the generalization of the Arnoldi equation to rational Krylov subspaces: let $\mathbf{E} = \mathbf{I}$, $m = 1$

and consider the (classical) Arnoldi method [12] that computes an orthogonal $\mathbf{V}_k$ whose columns span the subspace given by the sequence $\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^k\mathbf{B}$. This $\mathbf{V}_k$ then can be shown to satisfy

$$\mathbf{AV}_k = \mathbf{V}_k\mathbf{H}_k + \mathbf{r}_k\mathbf{e}_k^*, \tag{2.40}$$

where $\mathbf{H}_k = \mathbf{V}_k^*\mathbf{AV}_k$ is upper Hessenberg, $\mathbf{e}_k$ is the last column of the $k \times k$ identity matrix, and $\mathbf{r}$ is the nonzero residual, cf. [8]. Therefore, the $\mathbf{B}_\perp$-Sylvester equation (2.39) is the counterpart of (2.40) for rational Krylov subspaces as it connects $\mathbf{V}$ to the projection of $\mathbf{A}$ and the residual of $\mathbf{B}$. This is why (2.39) is also denoted as *Arnoldi-like equation* by Frangos and Jaimoukha, cf. [70, 71]. They derived the $\mathbf{B}_\perp$-Sylvester equation for the rational Arnoldi algorithm in [70], and for the *modified* rational Arnoldi algorithm in [71], both of which differ only in the right-hand sides of the LSEs that have to be solved. Nevertheless, both approaches yield the same subspace due to the "nested property" in rational Krylov subspaces; see also [68] for a discussion at full length. The results of Frangos and Jaimoukha are based on the particular course of action in numerical implementations, which leads to different formulations of the $\mathbf{B}_\perp$-Sylvester equation for the Arnoldi and Lanczos process, and which unfortunately buries the connection to the $\mathbf{B}$-Sylvester equation. Within the projective framework pursued in this work, the result instead may be proven in the most general form, with no constraints on the numerical implementation. The interested reader is also referred to [69], where the $\mathbf{B}_\perp$-Sylvester equation is exploited to further reduce an already reduced model that is too large (and thereby preserve the matching of certain moments), and also to [3] where sparse-sparse Sylvester equations are iteratively solved—to which we will come back in Part III.

Remark 2.13. The $\mathbf{B}_\perp$-Sylvester equation (2.39) is primarily used as a tool in the following chapters, but it nevertheless leads to remarkable insights by itself: as the $\mathbf{B}_\perp$-Sylvester equation is similar to the $\mathbf{B}$-Sylvester equation, Theorem 2.4 proves that the columns of $\mathbf{V}$ must span the rational Krylov subspace for the input $\mathbf{B}_\perp$ and eigenvalues of $\mathbf{E}_r^{-1}\mathbf{A}_r$. This statement may be even extended, as the matrix $\mathbf{W}$ in Lemma 2.12 is arbitrary (as long as $\mathbf{E}_r$ is non-singular). Therefore, as there are infinitely many admissible $\mathbf{W}$, there are also infinitely many $\mathbf{B}_\perp$-Sylvester equations (2.39). That means that as soon as the columns of $\mathbf{V}$ span one rational Krylov subspace, they in fact span infinitely many rational Krylov subspaces—all of which are connected through projections. In particular, $\mathbf{V}$ spans the Krylov subspace for the input $\mathbf{B}_\perp$, with the reduced eigenvalues as shifts and the columns of $\mathbf{L}$, after transforming $\mathbf{E}_r^{-1}\mathbf{A}_r$ to Jordan canonical form, as tangential directions.

2.5 Parametrized Family of Reduced Dynamics

After comprehensively characterizing bases of Krylov subspaces, given by $\mathbf{V}$ and the respective interpolation data, i. e. shifts and tangential directions in terms of $\mathbf{S}$ and $\mathbf{L}$, we are now ready to thoroughly examine the remaining degrees of freedom contained in the reduced model. This is carried out in this section and thereby, the results of Astolfi [13, 14, 15] and of [212] are slightly generalized and equipped with new proofs.

Lemma 2.14. *If* $\mathbf{V}$ *solves the Sylvester equation*

$$\mathbf{AV} - \mathbf{EVS} = \mathbf{BL}, \tag{2.41}$$

then for any $\mathbf{W} \in \mathbb{C}^{N\times n}$ *the matrices of the reduced system satisfy*

$$\mathbf{A}_r = \mathbf{E}_r\mathbf{S} + \mathbf{B}_r\mathbf{L}. \tag{2.42}$$

Proof. The proof directly follows by multiplying (2.41) with $\mathbf{W}^*$ from the left. □

Lemma 2.14 in fact delivers a family of reduced models that interpolate $\boldsymbol{G}(s)$: define $\mathbf{E}_r = \mathbf{I}$, then the family may be parametrized by the reduced input $\mathbf{B}_r$, because $\mathbf{A}_r$ is then defined by (2.42), and $\mathbf{C}_r$ is independent from $\mathbf{W}$; this is stated in the next theorem, which in addition gives a new proof of (tangential) interpolation by rational Krylov subspaces.

Theorem 2.15. *Given a triple* $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ *that satisfies the Sylvester equation*

$$\mathbf{AV} - \mathbf{EVS} = \mathbf{BL}, \tag{2.43}$$

with $\mathbf{S} \in \mathbb{C}^{n\times n}$ *and* $\mathbf{L} \in \mathbb{C}^{m\times n}$*, assume* $(\mathbf{L}, \mathbf{S})$ *observable and* $\boldsymbol{\Lambda}(\mathbf{S}) \cap \boldsymbol{\Lambda}(\mathbf{E}^{-1}\mathbf{A}) = \emptyset$*. Define the family of reduced models* $\boldsymbol{G}_{\mathbf{F}}(s)$*, parametrized in* $\mathbf{F} \in \mathbb{C}^{n\times m}$*, as follows:*

$$\begin{aligned} \dot{\mathbf{x}}_r(t) &= (\mathbf{S} + \mathbf{FL})\,\mathbf{x}_r(t) + \mathbf{Fu}(t), \\ \mathbf{y}_r(t) &= \mathbf{CVx}_r(t). \end{aligned} \tag{2.44}$$

If $\mathbf{F}$ *is such that* $\boldsymbol{\Lambda}(\mathbf{S}) \cap \boldsymbol{\Lambda}(\mathbf{S} + \mathbf{FL}) = \emptyset$*, then* $\boldsymbol{G}_{\mathbf{F}}(s)$ *(tangentially) interpolates* $\boldsymbol{G}(s)$ *as encoded in the pair* $(\mathbf{L}, \mathbf{S})$ *and as defined in Theorem 2.4.*

Proof. Due to Lemma 2.1, changing the basis of $\mathbf{V}$ amounts to a state transformation of (2.44), to which the transfer behaviour of $\boldsymbol{G}_{\mathbf{F}}(s)$ stays invariant. We may therefore assume without loss of generality that $\mathbf{S}$ is in Jordan canonical form and that $\mathbf{L}$ has the desired tangential directions $\mathbf{l}_i$ as columns. As the Jordan blocks are decoupled, it is sufficient to proof the theorem only for one Jordan block of dimension q. We may therefore assume the triple $(\mathbf{V},\mathbf{S},\mathbf{L})$ with $\mathbf{S}\in\mathbb{C}^{q\times q}$ and $\mathbf{L}\in\mathbb{C}^{m\times q}$ given by

$$\mathbf{V}=\left[\mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_1,\dots,\sum_{\nu=0}^{q-1}\left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{\nu}\mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_{q-\nu}\right],\ \mathbf{S}=\begin{bmatrix} s_0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & s_0 \end{bmatrix},\ \mathbf{L}=[\mathbf{l}_1,\dots,\mathbf{l}_q]\,, \tag{2.45}$$

and it is left to prove that $\boldsymbol{G}_{\mathbf{F}}(s)$ satisfies the tangential interpolation (2.10)–(2.12) for arbitrary choices of $\mathbf{F}$. To this end, let $\mathbf{e}_i$ denote the i-th column of the $q\times q$ identity matrix, then it follows from (2.45) that

$$(\mathbf{S}+\mathbf{F}\mathbf{L}-s_0\mathbf{I})\,\mathbf{e}_1 \;=\; \mathbf{F}\mathbf{l}_1, \tag{2.46}$$

$$(\mathbf{S}+\mathbf{F}\mathbf{L}-s_0\mathbf{I})\,\mathbf{e}_i \;=\; \mathbf{F}\mathbf{l}_i+\mathbf{e}_{i-1}, \quad i=2,\dots,q \tag{2.47}$$

Due to the assumption $\boldsymbol{\Lambda}(\mathbf{S})\cap\boldsymbol{\Lambda}(\mathbf{S}+\mathbf{F}\mathbf{L})=\emptyset$, we can solve this for $\mathbf{e}_i$, $i=1,\dots,q$, and by recursively using the results it follows that

$$\mathbf{e}_i \;=\; \sum_{\nu=0}^{i-1}(\mathbf{S}+\mathbf{F}\mathbf{L}-s_0\mathbf{I})^{-(\nu+1)}\,\mathbf{F}\mathbf{l}_{i-\nu}. \tag{2.48}$$

Then,

$$\sum_{\nu=0}^{i-1}\widehat{\mathbf{M}}_{\nu}^{s_0}\mathbf{l}_{i-\nu} \;=\; -\mathbf{C}\mathbf{V}\sum_{\nu=0}^{i-1}(\mathbf{S}+\mathbf{F}\mathbf{L}-s_0\mathbf{I})^{-(\nu+1)}\,\mathbf{F}\mathbf{l}_{i-\nu} \tag{2.49}$$

$$\overset{(2.48)}{=} -\mathbf{C}\mathbf{V}\mathbf{e}_i \tag{2.50}$$

$$\overset{(2.45)}{=} -\mathbf{C}\sum_{\nu=0}^{i-1}\left(\mathbf{A}_{s_0}^{-1}\mathbf{E}\right)^{\nu}\mathbf{A}_{s_0}^{-1}\mathbf{B}\mathbf{l}_{i-\nu} \tag{2.51}$$

$$= \sum_{\nu=0}^{i-1}\mathbf{M}_{\nu}^{s_0}\mathbf{l}_{i-\nu}, \tag{2.52}$$

holds for $i=1,\dots,q$, which proves tangential interpolation (2.10)–(2.12) of $\boldsymbol{G}_{\mathbf{F}}(s)$. □

It should be noted that the reduced models in the family $\boldsymbol{G}_{\mathbf{F}}(s)$ are *not* obtained by a projection of $\boldsymbol{G}(s)$; they are instead directly constructed such they interpolate $\boldsymbol{G}(s)$, and hence, Theorem 2.15 presents a new projection-independent proof of moment matching/tangential interpolation based on rational Krylov subspaces.

The benefit of the Theorem 2.15 is a parametrization of all reduced models that interpolate the original one: given $(\mathbf{V},\mathbf{S},\mathbf{L})$ such that the assumptions hold, one may select any $\mathbf{B}_r$ such that $\boldsymbol{\Lambda}(\mathbf{S})\cap\boldsymbol{\Lambda}(\mathbf{S}+\mathbf{B}_r\mathbf{L})=\emptyset$; then the reduced model $\boldsymbol{G}_r(s)$ that interpolates $\boldsymbol{G}(s)$ is given by $\mathbf{E}_r = \mathbf{I}$, $\mathbf{A}_r = \mathbf{S} + \mathbf{B}_r\mathbf{L}$ and $\mathbf{C}_r = \mathbf{CV}$.

A remaining question is if *every* reduced model that interpolates $\boldsymbol{G}(s)$ with the interpolation data $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ can be parametrized by $\mathbf{F}$, i. e. how general is the family $\boldsymbol{G}_{\mathbf{F}}(s)$ in Theorem 2.15. The next theorem shows that $\boldsymbol{G}_{\mathbf{F}}(s)$ is indeed more general than the projection framework using $\mathbf{W}$.

Theorem 2.16. *Given a triple* $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ *that satisfies the Sylvester equation*

$$\mathbf{AV} - \mathbf{EVS} = \mathbf{BL}, \tag{2.53}$$

with $\mathbf{S} \in \mathbb{C}^{n \times n}$ *and* $\mathbf{L} \in \mathbb{C}^{m \times n}$, *define the family of reduced systems* $\boldsymbol{G}_{\mathbf{F}}(s)$, *parametrized in* $\mathbf{F} \in \mathbb{C}^{n \times m}$,

$$\begin{aligned} \dot{\mathbf{x}}_r(t) &= (\mathbf{S} + \mathbf{FL})\,\mathbf{x}_r(t) + \mathbf{Fu}(t), \\ \mathbf{y}_r(t) &= \mathbf{CV}\mathbf{x}_r(t), \end{aligned} \tag{2.54}$$

and the family of reduced systems $\boldsymbol{G}_r(s)$, *parametrized in* $\mathbf{W}$,

$$\begin{aligned} \mathbf{E}_r\dot{\mathbf{x}}_r(t) &= \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{B}_r\mathbf{u}(t), \\ \mathbf{y}_r(t) &= \mathbf{C}_r\mathbf{x}_r(t). \end{aligned} \tag{2.55}$$

Then the following statements hold.

i) For any $\mathbf{W}$ *such that* $\mathbf{E}_r$ *is non-singular, there exists a unique* $\mathbf{F}$ *such that* $\mathbf{E}_r^{-1}\mathbf{B}_r = \mathbf{F}$ *and* $\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{S} + \mathbf{FL}$, *which means that the transfer functions* $\boldsymbol{G}_r(s) = \boldsymbol{G}_{\mathbf{F}}(s)$ *are equal.*

ii) If $[\mathbf{EV}, \mathbf{B}]$ *has full column rank* $n + m$, *then for any* $\mathbf{F}$ *there exists a* $\mathbf{W}$ *such that* $\mathbf{F} = \mathbf{E}_r^{-1}\mathbf{B}_r$ *and* $\mathbf{S} + \mathbf{FL} = \mathbf{E}_r^{-1}\mathbf{A}_r$, *which means that the transfer functions* $\boldsymbol{G}_r(s) = \boldsymbol{G}_{\mathbf{F}}(s)$ *are equal.*

Proof. The output $\mathbf{C}_r = \mathbf{CV}$ is independent from $\mathbf{W}$ and $\mathbf{F}$. To prove *i)*, note that for any $\mathbf{W}$ such that $\mathbf{E}_r$ is non-singular, $\boldsymbol{G}_r(s) = \mathbf{C}_r(s\mathbf{I} - \mathbf{E}_r^{-1}\mathbf{A}_r)^{-1}\mathbf{E}_r^{-1}\mathbf{B}_r$. Then choose the unique $\mathbf{F} = \mathbf{E}_r^{-1}\mathbf{B}_r$ and it follows from (2.42) that $\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{S} + \mathbf{FL}$ which completes the proof for this part.

To prove *ii)*, it is sufficient due to Lemma 2.14 to show existence of a $\mathbf{W}$ such that $\mathbf{E}_r = \mathbf{W}^*\mathbf{EV} = \mathbf{I}$ and $\mathbf{B}_r = \mathbf{W}^*\mathbf{B} = \mathbf{F}$, which is equivalent to $\mathbf{W}^*[\mathbf{EV}, \mathbf{B}] = [\mathbf{I}, \mathbf{F}]$. We may therefore restrict $\mathbf{W}$ to be contained in the subspace span$(\mathbf{W}) \subset$ span$[\mathbf{EV},\ \mathbf{B}]$, which shows non-uniqueness of $\mathbf{W}$. Construct $\mathbf{W} = [\mathbf{EV},\ \mathbf{B}]\mathbf{K}$, then $\mathbf{K} \in \mathbb{C}^{(n+m) \times n}$ such that the above mentioned statement holds has to satisfy $\mathbf{K}^*[\mathbf{EV}, \mathbf{B}]^*[\mathbf{EV}, \mathbf{B}] = [\mathbf{I}, \mathbf{F}]$, which exists for any choice of $\mathbf{F}$ only if $[\mathbf{EV}, \mathbf{B}]$ has full column rank. □

Remark 2.17. The statements of Theorems 2.15 and 2.16 may actually be generalized to reduced DAE systems. Following the proof of Theorem 2.15, a DAE family $\boldsymbol{G}_{\mathbf{F},\mathbf{H}} = \mathbf{CV}(s\mathbf{H} - (\mathbf{HS} + \mathbf{FL}))^{-1}\mathbf{F}$ with singular $\mathbf{H}$ can still fulfil tangential interpolation for the ODE parts of $\mathbf{S}$ and $\mathbf{L}$. The accurate conditions, however, would require further analysis, but this is omitted because it is arguable if it makes sense to construct a reduced DAE system for an original ODE model.

Remark 2.18. It should be stressed that Theorem 2.16 is not restricted to reduced models that match moments of $\boldsymbol{G}(s)$ (only DAE systems are excluded); it would also apply for reduced models that match Markov parameters or that preserve eigenvalues of the original model; these cases, however, are not of interest here.

Remark 2.19. A stronger statement of Theorem 2.16 follows for single inputs $m = 1$. As we consider matrices $\mathbf{V}$ whose columns span rational Krylov subspaces, and as we further assume that the pair $(\mathbf{E}^{-1}\mathbf{A}, \mathbf{E}^{-1}\mathbf{b})$ is controllable, then $[\mathbf{EV}, \mathbf{b}]$ is guaranteed to have full column rank $n+1$, and hence, the two families $\boldsymbol{G}_r(s)$ and $\boldsymbol{G}_{\mathbf{F}}(s)$ are equivalent; this was already proven by Astolfi [15]. Additionally taking into account the generality of projections by Krylov subspaces [75, 78], it follows that given a minimal system of order N, for any minimal model of order $n < N$, there exists a triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$, such that this reduced model is contained in both families $\boldsymbol{G}_r(s)$ and $\boldsymbol{G}_{\mathbf{F}}(s)$. The difference is that $\mathbf{F}$ provides a unique parametrization, whereas $\mathbf{W}$ does not.

It was shown that the reduced models in the families (2.54) and (2.55) are equivalent; only the matrix $\mathbf{W}$ is not unique. In order to remove this redundancy one could define the following parametrization.

Corollary 2.20. *Given a triple* $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ *that satisfies the Sylvester equation*

$$\mathbf{AV} - \mathbf{EVS} = \mathbf{BL}, \tag{2.56}$$

with $\mathbf{S} \in \mathbb{C}^{n\times n}$ *and* $\mathbf{L} \in \mathbb{C}^{m\times n}$, *define* $\widehat{\mathbf{W}} = [\mathbf{EV}, \mathbf{B}]$, *and assume that* $\widehat{\mathbf{W}}$ *has full column rank* $n+m$. *If* $\mathbf{W}$ *is constructed as* $\mathbf{W} = \widehat{\mathbf{W}}\mathbf{K}$, *where* $\mathbf{K}^* = [\mathbf{I}, \mathbf{F}](\widehat{\mathbf{W}}^*\widehat{\mathbf{W}})^{-1}$, *then* $\mathbf{W}$ *and* $\mathbf{F}$ *are in a one-to-one relation and consequently the families* $\boldsymbol{G}_r(s)$ *in (2.55) and* $\boldsymbol{G}_{\mathbf{F}}(s)$ *in (2.54) are equivalent.*

Proof. Note that $[\mathbf{A}_r, \mathbf{E}_r, \mathbf{B}_r] = \mathbf{W}^*[\mathbf{AV}, \mathbf{EV}, \mathbf{B}]$, and hence, only the subspace spanned by the columns of $[\mathbf{AV}, \mathbf{EV}, \mathbf{B}]$ is relevant for $\mathbf{W}$. It follows from the $\mathbf{B}$-Sylvester equation that $\operatorname{span}(\mathbf{AV}) \subseteq \operatorname{span}[\mathbf{EV}, \mathbf{B}]$ and therefore taking $\mathbf{W} = \widehat{\mathbf{W}}\mathbf{K}$ is sufficient. The rest of the proof is already contained in the proof of Theorem 2.16. □

Remark 2.21. Due to the $\mathbf{B}_\perp$-Sylvester equation, $\operatorname{span}[\mathbf{EV}, \mathbf{B}_\perp] = \operatorname{span}[\mathbf{EV}, \mathbf{B}]$, and therefore, it would also be possible to define the alternative parametrization of the projected models by $\widehat{\mathbf{W}} = [\mathbf{EV}, \mathbf{B}_\perp]$, and then proceeding similarly to Corollary 2.20.

Most of the results in this section were found by Astolfi: the family $\boldsymbol{G}_{\mathbf{F}}(s)$ is introduced in [14], where the free parameter $\mathbf{F}$ is used to assign eigenvalues and/or zeros, or to render the model passive, lossless, dissipative, or compartmental; a similar discussion can be found in [15], where also a generalization of the notion of moments for non-linear systems is presented; the equivalence of various families of reduced models is studied in [13]. The proof of moment matching/tangential interpolation in Theorem 2.15 appears to be new, whereas the generalizations of (2.42) to $\mathbf{E} \neq \mathbf{I}$ and multiple inputs, and its connection to the $\mathbf{B}$- and $\mathbf{B}_\perp$-Sylvester equations were first published in [212]. Ionescu and Astolfi studied the choice of appropriate shifts s_i such that the family $\boldsymbol{G}_{\mathbf{F}}(s)$ preserves either passivity, cf. [107], or a port-Hamiltonian structure, cf. [105]. The authors also deepened the understanding of moments in the non-linear case in [106], and furthermore, they chose $\mathbf{F}$ in [108] such that the reduced model becomes non-minimal, which results in a reduced model of smaller order than n, but which still achieves moment matching. This approach, however, seems cumbersome, because a reduced model of order $n/2$ that matches n moments could have been directly constructed through a two-sided projection. Finally, Ahmad et. al. [3] employed a closely related parametrization of the reduced dynamics, i. e. of Theorem 2.15.

2.6 Chapter Overview and Outlook

The aim of this chapter was to describe bases of rational Krylov subspaces—not only by its interpolation data, i. e. expansion points and tangential directions, but also by useful Sylvester equations. We found that any $\mathbf{V}$ in fact satisfies two types of Sylvester equations with specific structure, which in turn define the matrices $\mathbf{S}$, $\mathbf{L}$, and $\mathbf{B}_\perp$—all of which will become important in the subsequent chapters. The reason for this is, that the $\mathbf{B}$-Sylvester equation (2.15), the $\mathbf{B}_\perp$-Sylvester equation (2.39), and also (2.42), will be the main tools to derive the various results. It should be stressed, that the just mentioned matrices allow for concrete interpretations: $\mathbf{S}$ and $\mathbf{L}$ encode the expansion points and tangential directions, respectively, and $\mathbf{B}_\perp$ represents the residual of $\mathbf{B}$ after projection.

The aim of the subsequent chapters now can be depicted as follows: assume that a triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ is given, which satisfies the $\mathbf{B}$-Sylvester equation (2.15). This in fact implies that two issues in MOR have already been solved: firstly, the interpolation points s_i with multiplicities have been selected (and in the MIMO case additionally the tangential directions $\mathbf{l}_i$), and secondly, a numerical procedure for computing $\mathbf{V}$ has been implemented (either with direct or iterative solvers of the LSEs). Needless to say that both issues are far from trivial—their solution, however, is postponed for the moment. Then the remaining degrees of freedom in model order reduction can be parametrized in two ways: either with $\mathbf{W}$, leading to the family $\boldsymbol{G}_r(s)$ in (2.55), or with $\mathbf{F}$, leading to the family $\boldsymbol{G}_{\mathbf{F}}(s)$ in (2.54). The main contribution of this thesis is then to suggest a unique way, how fix this degree of freedom. This choice will coincide with the concept of $\mathcal{H}_2$ pseudo-optimality and has considerable advantages: stability is preserved in the reduced model; the degree of freedom is uniquely determined; the reduced model satisfies some kind of optimality; the reduced order can be accumulated, such that the approximation error is guaranteed to decrease monotonically; and the main numerical effort remains the computation of the triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$. The statement of the concept of $\mathcal{H}_2$ pseudo-optimality, however, requires the analysis of the approximation error, which is why this is first conducted in the next chapter.

What is left over, then are expedient methods to reveal suitable shifts and tangential directions. But as already mentioned, this is omitted in this work, and the interested reader is instead referred to the thesis of Panzer [148]. It should also be noted, that numerical issues in the computation of a triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ are independent from the theoretical concept of $\mathcal{H}_2$ pseudo-optimality and hence omitted; in this regard, the interested reader is referred to the thesis [215] and references therein.

Finally, it should be noted, that all results of this chapter can be formulated in a dual way for the output side: let the columns of $\mathbf{W}$ form a basis of an output rational Krylov subspace, then there exist $\mathbf{C}$- and $\mathbf{C}_\perp$-Sylvester equations and a similar parametrization of the reduced dynamics. The duality follows by replacing $\mathbf{A}$ with $\mathbf{A}^*$, $\mathbf{E}$ with $\mathbf{E}^*$, $\mathbf{B}$ with $\mathbf{C}^*$ and $\mathbf{V}$ with $\mathbf{W}$, cf. [212]. The details, however, are omitted, since the dual results are equivalent to the input side [13] and not essential for this work.

3 Error Analysis

This chapter discusses a factorization of the error model. It is based on the $\mathbf{B}_\perp$-Sylvester equation and therefore requires that the columns of $\mathbf{V}$ span a rational Krylov subspace. The basic factorization is presented in Section 3.1, whereas Section 3.2 discusses an incremental model order reduction, which emerges from iterative error factorizations in each step. The results are based on the publications [149, 211, 212].

3.1 Factorization of the Error System

Irrespective of whether we try to solve a large-scale Lyapunov equation in TBR or compute a reduced model by Krylov-based projections, we aim at approximating $\boldsymbol{X}(s)$, defined as

$$\boldsymbol{X}(s) = (s\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B}, \tag{3.1}$$

by $\boldsymbol{X}(s) \approx \mathbf{V}\boldsymbol{X}_r(s)$, where $\boldsymbol{X}_r(s)$ satisfies

$$\boldsymbol{X}_r(s) = (s\mathbf{E}_r - \mathbf{A}_r)^{-1}\,\mathbf{B}_r. \tag{3.2}$$

Then the error $\boldsymbol{E}(s)$ reads as

$$\boldsymbol{E}(s) = \boldsymbol{X}(s) - \mathbf{V}\boldsymbol{X}_r(s). \tag{3.3}$$

The error $\boldsymbol{E}(s)$ is typically described by the residual error $\boldsymbol{R}(s)$, which is defined next, cf. [85].

Definition 3.1. Given $\mathbf{V}$, the original model (1.1), and its reduction (1.4), the residual error $\boldsymbol{R}(s)$ is defined as

$$\boldsymbol{R}(s) = \mathbf{B} - (s\mathbf{E} - \mathbf{A})\,\mathbf{V}\,(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\,\mathbf{B}_r. \tag{3.4}$$

With this definition, it readily follows for the error that

$$\boldsymbol{E}(s) = (s\mathbf{E} - \mathbf{A})^{-1}\,\boldsymbol{R}(s), \tag{3.5}$$

and it is straightforward to show that the residual error $\boldsymbol{R}(s)$ satisfies the Petrov-Galerkin condition, $\mathbf{W}^*\boldsymbol{R}(s)=\mathbf{0}$, whereas the error $\boldsymbol{E}(s)$ generally does not, $\mathbf{W}^*\boldsymbol{E}(s)\neq\mathbf{0}$. It should be noted that the error model $\boldsymbol{G}_e(s) = \boldsymbol{G}(s) - \boldsymbol{G}_r(s)$ can be similarly described by the residual error, as it holds irrespectively of the output $\mathbf{C}$ that

$$\boldsymbol{G}_e(s) = \boldsymbol{G}(s) - \boldsymbol{G}_r(s) = \mathbf{C}\left[\boldsymbol{X}(s) - \mathbf{V}\boldsymbol{X}_r(s)\right] = \mathbf{C}\boldsymbol{E}(s) = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\boldsymbol{R}(s). \quad (3.6)$$

The next theorem investigates the residual error $\boldsymbol{R}(s)$, which is already the main result of this section. This important statement is the basis of a number of applications in this thesis. Although it was previously observed by Frangos and Jaimoukha in a slightly different form in [68], it has been largely overlooked in the literature.

Theorem 3.1. *Let* $\mathbf{V}$ *satisfy the Sylvester equation*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{B}_\perp\mathbf{L}, \quad (3.7)$$

where $\mathbf{B}_\perp=\mathbf{B}-\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$. *Define the feed-through model* $\boldsymbol{G}_f(s)=\mathbf{L}\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r+\mathbf{I}$ *of reduced order* n, *then the residual error* $\boldsymbol{R}(s)$ *may be factorized as*

$$\boldsymbol{R}(s) = \mathbf{B}_\perp\boldsymbol{G}_f(s). \quad (3.8)$$

Proof. It follows from the Sylvester equation (3.7) that

$$\left(s\mathbf{E} - \mathbf{A}\right)\mathbf{V} = s\mathbf{E}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r - \mathbf{B}_\perp\mathbf{L} = \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\left(s\mathbf{E}_r - \mathbf{A}_r\right) - \mathbf{B}_\perp\mathbf{L}. \quad (3.9)$$

Substituting this in the definition of the residual error (3.4) yields

$$\begin{aligned}\boldsymbol{R}(s) &= \mathbf{B} - \left[\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\left(s\mathbf{E}_r - \mathbf{A}_r\right) - \mathbf{B}_\perp\mathbf{L}\right]\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r && (3.10)\\ &= \mathbf{B} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{B}_\perp\mathbf{L}\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r && (3.11)\\ &= \mathbf{B}_\perp + \mathbf{B}_\perp\mathbf{L}\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r, && (3.12)\end{aligned}$$

which completes the proof. □

A consistent statement also holds for the error model $\boldsymbol{G}_e(s)$, which was also published in [211, 212].

Corollary 3.2. *Let* $\mathbf{V}$ *satisfy the Sylvester equation*

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{B}_\perp\mathbf{L}. \quad (3.13)$$

Then the error model $\boldsymbol{G}_e(s) = \boldsymbol{G}(s) - \boldsymbol{G}_r(s)$ *can be factorized by*

$$\boldsymbol{G}_e(s) = \boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s). \tag{3.14}$$

where $\boldsymbol{G}_\perp(s)$ *of order* N *and the feed-through model* $\boldsymbol{G}_f(s)$ *of order* n *are defined as*

$$\boldsymbol{G}_\perp = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B}_\perp \tag{3.15}$$

$$\boldsymbol{G}_f(s) = \mathbf{L}\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r + \mathbf{I} \tag{3.16}$$

Proof. The statement is a direct consequence of Theorem 3.1 and (3.6). □

It should be stressed that $\boldsymbol{G}_\perp(s)$ shares $\mathbf{E}$, $\mathbf{A}$ and $\mathbf{C}$ with $\boldsymbol{G}(s)$ and only differs from it in its input. Furthermore, the feed-through model $\boldsymbol{G}_f(s)$, which is introduced in Theorem 3.1 and its Corollary 3.2, is of small order n and shares $\mathbf{E}_r$, $\mathbf{A}_r$ and $\mathbf{B}_r$ with the reduced model $\boldsymbol{G}_r(s)$. The zeros of $\boldsymbol{G}_f(s)$ are investigated in the next lemma.

Lemma 3.3. *Given a triple* $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ *that satisfies the* $\mathbf{B}$*-Sylvester equation (2.15), let* s_0 *be an eigenvalue of* $\mathbf{S}$ *and assume that* $(\mathbf{L}, \mathbf{S})$ *is observable and that* s_0 *is not a pole of* $\boldsymbol{G}_f(s)$*. Then* s_0 *is a transmission zero of* $\boldsymbol{G}_f(s)$*.*

Proof. $\boldsymbol{G}_f(s)$ is an $m \times m$ proper transfer function, and due to its feed-through term, it has full column normal rank. For a proper definition of the concept of *normal rank* see e. g. [222]. It then follows from [222, Lemma 3.27], that s_0 is a transmission zero of $\boldsymbol{G}_f(s)$ if there exists a nonzero $\mathbf{u}_0 \in \mathbb{C}^m$ such that $\boldsymbol{G}_f(s_0)\mathbf{u}_0 = \mathbf{0}$. For the construction of a suitable $\mathbf{u}_0$, let the nonzero $\mathbf{x}_0$ denote an eigenvector to the eigenvalue s_0, i. e. $(s_0\mathbf{I} - \mathbf{S})\mathbf{x}_0 = \mathbf{0}$, and pick $\mathbf{u}_0 = -\mathbf{L}\mathbf{x}_0$. Note that this $\mathbf{u}_0$ is nonzero as $(\mathbf{L}, \mathbf{S})$ is assumed observable. Then it follows that

$$\mathbf{0} = \left(s_0\mathbf{E}_r - \mathbf{E}_r\mathbf{S}\right)\mathbf{x}_0 = \left(s_0\mathbf{E}_r - \left(\mathbf{E}_r\mathbf{S} + \mathbf{B}_r\mathbf{L}\right) + \mathbf{B}_r\mathbf{L}\right)\mathbf{x}_0 \tag{3.17}$$

$$\overset{(2.42)}{=} \left(s_0\mathbf{E}_r - \mathbf{A}_r + \mathbf{B}_r\mathbf{L}\right)\mathbf{x}_0 = \left(s_0\mathbf{E}_r - \mathbf{A}_r\right)\mathbf{x}_0 - \mathbf{B}_r\mathbf{u}_0 \tag{3.18}$$

Now taking $\mathbf{L}\mathbf{x}_0 + \mathbf{u}_0 = \mathbf{0}$ and replacing $\mathbf{x}_0$ by (3.18), yields $\mathbf{L}\left(s_0\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r\mathbf{u}_0 + \mathbf{u}_0 = \mathbf{0}$, where the inverse exists due to the assumption made. This proves that there exists a nonzero $\mathbf{u}_0$ such that $\boldsymbol{G}_f(s_0)\mathbf{u}_0 = \mathbf{0}$. □

Remark 3.4. It should be noted that the assumption that s_0 is not a pole of $\boldsymbol{G}_f(s)$ is not restrictive: it means that s_0 is not an eigenvalue of $\mathbf{E}_r^{-1}\mathbf{A}_r$, which is the same assumption that is required anyway for moment matching to hold. Another argument that compensation is very unlikely in practice is that we are searching for a stable reduced model, which means that the poles of $\boldsymbol{G}_f(s)$ should lie in the left half of the complex plane, whereas the shifts s_0—the zeros of $\boldsymbol{G}_f(s)$—are typically chosen in the right half of the complex plane. Finally, assuming that $(\mathbf{L}, \mathbf{S})$ is observable is also not restrictive, as this is required for moment matching, as well.

Remark 3.5. It should also be noted that the following generalizations of Lemma 3.3 are straightforward: if the eigenvalue s_0 has geometric multiplicity $1 < m_0 < m$, then there exist m_0 linearly independent $\mathbf{u}_i$ such that $\boldsymbol{G}_f(s_0)\mathbf{u}_i$, $i = 1, \dots, m_0$; if the eigenvalue s_0 has geometric multiplicity $m_0 = m$, then s_0 is a blocking zero, i. e. $\boldsymbol{G}_f(s_0) = \mathbf{0}$. A further generalization follows, if $\mathbf{S}$ contains a Jordan block of dimension q_0 to the eigenvalue s_0. Then it can be shown that s_0 is also a transmission zero of the first q_0-1 derivatives of $\boldsymbol{G}_f(s)$; as this case is of minor interest, the details are omitted for brevity; a proof in the SISO case can be found in [211].

The factorization of the error also brings forth another interesting question: is there a perturbation of the state-space realization of the original model $\boldsymbol{G}(s)$ such that the reduced model $\boldsymbol{G}_r(s)$ exactly approximates its transfer behaviour, i. e. such that $\boldsymbol{G}_r(s)$ is a minimal realization of the perturbed $\boldsymbol{G}(s)$? Although the answer is not directly related to the objective of this work, it is still presented in the next lemma, because it provides interesting insight into MOR based on rational Krylov subspaces. This is inspired from [68].

Lemma 3.6. *Given the perturbations* $\boldsymbol{\Delta}_{\mathbf{A}} = \mathbf{B}_\perp \mathbf{L}\mathbf{E}_r^{-1}\mathbf{W}^*\mathbf{E}$ *and* $\boldsymbol{\Delta}_{\mathbf{B}} = \mathbf{B}_\perp$*, define the perturbed dynamics* $\boldsymbol{X}_{\boldsymbol{\Delta}}(s) = [s\mathbf{E} - (\mathbf{A} - \boldsymbol{\Delta}_{\mathbf{A}})]^{-1}(\mathbf{B} - \boldsymbol{\Delta}_{\mathbf{B}})$ *and* $\boldsymbol{G}_{\boldsymbol{\Delta}}(s) = \mathbf{C}\boldsymbol{X}_{\boldsymbol{\Delta}}(s)$*, and assume that* $\mathbf{V}$ *satisfies the* $\mathbf{B}$*- and* $\mathbf{B}_\perp$*-Sylvester equations (2.15) and (3.7). Then* $\boldsymbol{X}_{\boldsymbol{\Delta}}(s) = \mathbf{V}\boldsymbol{X}_r(s)$ *for all* s *such that* $\det[s\mathbf{E} - (\mathbf{A} - \boldsymbol{\Delta}_{\mathbf{A}})] \neq \mathbf{0}$*. Furthermore, all eigenvalues of* $\mathbf{E}^{-1}(\mathbf{A} - \boldsymbol{\Delta}_{\mathbf{A}})$ *are either eigenvalues of* $\mathbf{E}_r^{-1}\mathbf{A}_r$ *or not controllable, i. e.* $\boldsymbol{G}_r(s)$ *is a minimal realization of* $\boldsymbol{G}_{\boldsymbol{\Delta}}(s)$*.*

Proof. It follows from the $\mathbf{B}_\perp$-Sylvester equation (3.7) that

$$\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{A}\mathbf{V} - \mathbf{B}_\perp\mathbf{L} = \left(\mathbf{A} - \mathbf{B}_\perp\mathbf{L}\mathbf{E}_r^{-1}\mathbf{W}^*\mathbf{E}\right)\mathbf{V} = (\mathbf{A} - \boldsymbol{\Delta}_{\mathbf{A}})\,\mathbf{V}, \tag{3.19}$$

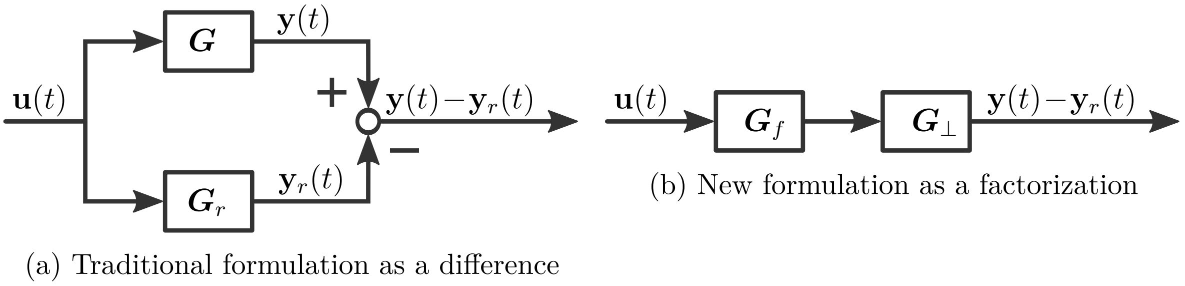
which proves that $\Lambda(\mathbf{E}_r^{-1}\mathbf{A}_r) \subset \Lambda(\mathbf{E}^{-1}(\mathbf{A} - \boldsymbol{\Delta}_{\mathbf{A}}))$. Now consider

$$\begin{aligned}
\mathbf{B} - \boldsymbol{\Delta}_{\mathbf{B}} \;=\; \mathbf{B} - \mathbf{B}_\perp \;&=\; \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r && (3.20)\\
&=\; \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}(s\mathbf{E}_r - \mathbf{A}_r)(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r && (3.21)\\
&=\; \left(s\mathbf{E}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r\right)(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r && (3.22)\\
&\overset{(3.19)}{=}\; [s\mathbf{E} - (\mathbf{A} - \boldsymbol{\Delta}_{\mathbf{A}})]\,\mathbf{V}(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r, && (3.23)
\end{aligned}$$

which proves that all remaining eigenvalues of $\mathbf{E}^{-1}(\mathbf{A}-\boldsymbol{\Delta}_{\mathbf{A}})$ are not controllable in $\boldsymbol{X}_{\boldsymbol{\Delta}}(s)$, and also that $\boldsymbol{X}_{\boldsymbol{\Delta}}(s)=\mathbf{V}\boldsymbol{X}_r(s)$ for all other values of s. $\boldsymbol{G}_{\boldsymbol{\Delta}}(s)=\boldsymbol{G}_r(s)$ then follows by multiplication with $\mathbf{C}$. □

Lemma 3.6 presents a neat observation in MOR by rational Krylov subspaces, but it is merely of theoretical interest. By contrast, Theorem 3.1 and its Corollary 3.2 have a strong impact on error analysis in MOR. Assume for example that $\boldsymbol{G}_r(s)$ approximates $\boldsymbol{G}(s)$ well, then the error $\boldsymbol{G}_e(s)$ should be small for all s. In light of the error model $\boldsymbol{G}_e(s)=\boldsymbol{G}(s)-\boldsymbol{G}_r(s)$, this means that the difference of two large and similar quantities would result in a small value. The traditional formulation of the error model therefore is numerically ill-conditioned, as the "small" error dynamics are overwhelmed by the "large" dynamics of original and reduced model, and thus, are hard to identify.

By using Theorem 3.1 and Corollary 3.2, the error model can instead be factorized into $\boldsymbol{G}_\perp(s)$ of original order N and $\boldsymbol{G}_f(s)$ of reduced order n. The difference between the traditional formulation, $\boldsymbol{G}_e(s)=\boldsymbol{G}(s)-\boldsymbol{G}_r(s)$, and the new one, $\boldsymbol{G}_e(s)=\boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$, may be illustrated as a parallel and a series connection, respectively, in a block diagram; this is shown in Figure 3.1, where Figure 3.1a corresponds to $\boldsymbol{G}_e(s)=\boldsymbol{G}(s)-\boldsymbol{G}_r(s)$ and where Figure 3.1b describes $\boldsymbol{G}_e(s)=\boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$. The benefit is that only the real error dynamics are triggered in the new formulation $\boldsymbol{G}_e(s)=\boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$, as no subtraction in the output occurs. The interpretation is that $\mathbf{B}_\perp$ must be worse controllable than $\mathbf{B}$ (which will be also verified by investigating the Lyapunov equation in Part III).



(a) Traditional formulation as a difference

(b) New formulation as a factorization

Figure 3.1: Block diagrams describing the error

The factorized formulation is particularly helpful in a large-scale setting, because only $\mathbf{B}_\perp$ and $\mathbf{L}$ are required, which in turn may be computed with negligible numerical effort. The benefit is that the factorization is described by state-space models that preserve sparsity of the original matrices. In conclusion, the factorized formulation is better suited for further analysis than the traditional formulation $\boldsymbol{G}_e(s)=\boldsymbol{G}(s)-\boldsymbol{G}_r(s)$; a first application is presented in the next section. Finally, it should be noted, that the above results could be generalized to cases where invariant subspaces and also subspaces like

(1.31) are contained in span($\mathbf{V}$), which, however, is omitted as these cases are irrelevant for this work.

3.2 Cumulative Approximation using Krylov Subspaces

We have seen that the error can be factorized by $\boldsymbol{E}(s)=\boldsymbol{X}_{\perp}(s)\boldsymbol{G}_f(s)$, with $\boldsymbol{X}_{\perp}(s)=(s\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}_{\perp}$. The second factor $\boldsymbol{G}_f(s)$ is of reduced order n, so it can be analysed without difficulty. The unknown remains $\boldsymbol{X}_{\perp}(s)$, which yet is equal to the original $\boldsymbol{X}(s)$—except for a different input. Then why not approximate $\boldsymbol{X}_{\perp}(s)$ in a second step? Just replace $\mathbf{B}$ by $\mathbf{B}_{\perp}$, then the reduction works the same way as in the first step. It then turns out that this can be repeated as often as desired, which yields an *iterative* or also called *cumulative* or *incremental* framework for model reduction by rational Krylov subspaces. The idea was first published in [149] and is discussed in this section.

3.2.1 The Cumulative Framework

The presentation unfortunately requires some inconvenient notation, which is clarified first. Let all quantities of the reduction and factorization performed so far have an additional index "1", to denote the first reduction. Accordingly, using the projection matrices $\mathbf{V}_1$ and $\mathbf{W}_1$, the original dynamics $\boldsymbol{X}(s)$ are approximated by

$$\boldsymbol{X}(s)=\mathbf{V}_1\boldsymbol{X}_{r,1}(s)+\boldsymbol{X}_{\perp,1}(s)\boldsymbol{G}_{f,1}(s) \tag{3.24}$$

with

$$\boldsymbol{X}_{r,1}(s)=\left(s\mathbf{E}_{r,1}-\mathbf{A}_{r,1}\right)^{-1}\mathbf{B}_{r,1}, \tag{3.25}$$

$$\boldsymbol{X}_{\perp,1}(s)=\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}_{\perp,1}, \tag{3.26}$$

$$\boldsymbol{G}_{f,1}(s)=\mathbf{L}_1\left(s\mathbf{E}_{r,1}-\mathbf{A}_{r,1}\right)^{-1}\mathbf{B}_{r,1}+\mathbf{I}, \tag{3.27}$$

where $\mathbf{E}_{r,1}=\mathbf{W}_1^*\mathbf{E}\mathbf{V}_1$, $\mathbf{A}_{r,1}=\mathbf{W}_1^*\mathbf{A}\mathbf{V}_1$, $\mathbf{B}_{r,1}=\mathbf{W}_1^*\mathbf{B}$, and where $\mathbf{B}_{\perp,1}$ and $\mathbf{L}_1$ are given by the $\mathbf{B}_{\perp}$-Sylvester equation (3.7). As already mentioned, the idea is now to consider $\boldsymbol{X}_{\perp,1}(s)$ as the original data, and reduce it in a second step, using the projection matrices $\mathbf{V}_2$ and $\mathbf{W}_2$, where $\mathbf{V}_2$ is assumed to span a rational Krylov subspace to the input $\mathbf{B}_{\perp,1}$, that is, $\mathbf{V}_2$ satisfies

$$\mathbf{A}\mathbf{V}_2-\mathbf{E}\mathbf{V}_2\mathbf{S}_2=\mathbf{B}_{\perp,1}\mathbf{L}_2. \tag{3.28}$$

The resulting reduced quantities are denoted with the index “2”, i. e. $\mathbf{E}_{r,2} = \mathbf{W}_2^*\mathbf{E}\mathbf{V}_2$, $\mathbf{A}_{r,2} = \mathbf{W}_2^*\mathbf{A}\mathbf{V}_2$ and $\mathbf{B}_{r,2} = \mathbf{W}_2^*\mathbf{B}_{\perp,1}$, and it follows that $\mathbf{V}_2$ also satisfies a corresponding $\mathbf{B}_\perp$-Sylvester equation

$$\mathbf{A}\mathbf{V}_2 - \mathbf{E}\mathbf{V}_2\mathbf{E}_{r,2}^{-1}\mathbf{A}_{r,2} = \mathbf{B}_{\perp,2}\mathbf{L}_2, \tag{3.29}$$

with $\mathbf{B}_{\perp,2} = \mathbf{B}_{\perp,1} – \mathbf{E}\mathbf{V}_2\mathbf{E}_{r,2}^{-1}\mathbf{B}_{r,2}$. A question that then arises is, if both reduced quantities can be conveniently merged into a single, accumulated model. The goal thus is to find quantities with index “tot”, that define the total reduction and error after two steps; the result is presented in the next theorem.

Theorem 3.7. *Let a reduction and factorization (3.24)–(3.27) be given, and assume that* $\boldsymbol{X}_{\perp,1}(s)$ *was reduced and factorized in a second step, where* $\mathbf{V}_2$ *satisfies (3.28). Then the error of the total reduction can be factorized as*

$$\boldsymbol{X}(s) = \mathbf{V}_{\text{tot}}\boldsymbol{X}_{r,\text{tot}}(s) + \boldsymbol{X}_{\perp,2}(s)\boldsymbol{G}_{f,\text{tot}}(s) \tag{3.30}$$

with

$$\boldsymbol{X}_{r,\text{tot}}(s) = \left(s\mathbf{E}_{r,\text{tot}} - \mathbf{A}_{r,\text{tot}}\right)^{-1}\mathbf{B}_{r,\text{tot}}, \tag{3.31}$$

$$\boldsymbol{X}_{\perp,2}(s) = \left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B}_{\perp,2}, \tag{3.32}$$

$$\boldsymbol{G}_{f,\text{tot}}(s) = \mathbf{L}_{\text{tot}}\left(s\mathbf{E}_{r,\text{tot}} - \mathbf{A}_{r,\text{tot}}\right)^{-1}\mathbf{B}_{r,\text{tot}} + \mathbf{I}, \tag{3.33}$$

and where the accumulated quantities are given by

$$\mathbf{B}_{\perp,2} = \mathbf{B}_{\perp,1} - \mathbf{E}\mathbf{V}_2\mathbf{E}_{r,2}^{-1}\mathbf{B}_{r,2}, \qquad \mathbf{V}_{\text{tot}} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix}, \qquad \mathbf{L}_{\text{tot}} = \begin{bmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{bmatrix}, \tag{3.34}$$

$$\mathbf{A}_{r,\text{tot}} = \begin{bmatrix} \mathbf{A}_{r,1} & \mathbf{0} \\ \mathbf{B}_{r,2}\mathbf{L}_1 & \mathbf{A}_{r,2} \end{bmatrix}, \qquad \mathbf{E}_{r,\text{tot}} = \begin{bmatrix} \mathbf{E}_{r,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{r,2} \end{bmatrix}, \quad \mathbf{B}_{r,\text{tot}} = \begin{bmatrix} \mathbf{B}_{r,1} \\ \mathbf{B}_{r,2} \end{bmatrix}. \tag{3.35}$$

Proof. Owing to the Sylvester equation (3.29), $\boldsymbol{X}_{\perp,1}(s)$ can be described by its reduction and factorized error,

$$\boldsymbol{X}_{\perp,1}(s) = \mathbf{V}_2\boldsymbol{X}_{r,2}(s) + \boldsymbol{X}_{\perp,2}(s)\boldsymbol{G}_{f,2}(s), \tag{3.36}$$

with $\mathbf{B}_{\perp,2} = \mathbf{B}_{\perp,1} - \mathbf{E}\mathbf{V}_2\mathbf{E}_{r,2}^{-1}\mathbf{B}_{r,2}$. Replacing $\boldsymbol{X}_{\perp,1}(s)$ in (3.24) by (3.36) yields

$$\boldsymbol{X}(s) = \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{r,1}(s) \\ \boldsymbol{X}_{r,2}(s)\boldsymbol{G}_{f,1}(s) \end{bmatrix} + \boldsymbol{X}_{\perp,2}(s)\boldsymbol{G}_{f,2}(s)\boldsymbol{G}_{f,1}(s), \tag{3.37}$$

which proves $\mathbf{V}_{\text{tot}} = [\mathbf{V}_1 \ \mathbf{V}_2]$. Consider $\boldsymbol{X}_{r,\text{tot}}(s)$ from (3.31) and use the definitions

(3.35), then

$$\boldsymbol{X}_{r,\mathrm{tot}}(s) = \begin{bmatrix} s\mathbf{E}_{r,1}-\mathbf{A}_{r,1} & \mathbf{0} \\ -\mathbf{B}_{r,2}\mathbf{L}_1 & s\mathbf{E}_{r,2}-\mathbf{A}_{r,2} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{B}_{r,1} \\ \mathbf{B}_{r,2} \end{bmatrix} \tag{3.38}$$

$$= \begin{bmatrix} (s\mathbf{E}_{r,1}-\mathbf{A}_{r,1})^{-1} & \mathbf{0} \\ (s\mathbf{E}_{r,2}-\mathbf{A}_{r,2})^{-1}\mathbf{B}_{r,2}\mathbf{L}_1(s\mathbf{E}_{r,1}-\mathbf{A}_{r,1})^{-1} & (s\mathbf{E}_{r,2}-\mathbf{A}_{r,2})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{r,1} \\ \mathbf{B}_{r,2} \end{bmatrix} \tag{3.39}$$

$$= \begin{bmatrix} \boldsymbol{X}_{r,1}(s) \\ \boldsymbol{X}_{r,2}(s)\boldsymbol{G}_{f,1}(s) \end{bmatrix}, \tag{3.40}$$

which proves that $\boldsymbol{X}_{r,\mathrm{tot}}(s)$ from (3.31) and with the definitions (3.35) indeed describes the dynamics required by (3.37). By comparing (3.37) with (3.30), it follows that it is left to prove that $\boldsymbol{G}_{f,\mathrm{tot}}(s) = \boldsymbol{G}_{f,2}(s)\boldsymbol{G}_{f,1}(s)$. Consider $\boldsymbol{G}_{f,\mathrm{tot}}(s)$ from (3.33) and use (3.34), then

$$\boldsymbol{G}_{f,\mathrm{tot}}(s) = \begin{bmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{bmatrix} \boldsymbol{X}_{r,\mathrm{tot}}(s) + \mathbf{I} \overset{(3.40)}{=} \begin{bmatrix} \mathbf{L}_1 & \mathbf{L}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{r,1}(s) \\ \boldsymbol{X}_{r,2}(s)\boldsymbol{G}_{f,1}(s) \end{bmatrix} + \mathbf{I} \tag{3.41}$$

$$= \mathbf{L}_2\boldsymbol{X}_{r,2}(s)\boldsymbol{G}_{f,1}(s) + \mathbf{L}_1\boldsymbol{X}_{r,1}(s) + \mathbf{I} \tag{3.42}$$

$$= \mathbf{L}_2\boldsymbol{X}_{r,2}(s)\boldsymbol{G}_{f,1}(s) + \boldsymbol{G}_{f,1}(s) \tag{3.43}$$

$$= (\mathbf{L}_2\boldsymbol{X}_{r,2}(s) + \mathbf{I})\,\boldsymbol{G}_{f,1}(s), \tag{3.44}$$

which completes the proof. □

Theorem 3.7 shows, that the error after the second reduction step (3.30), has again the identical structure that was already apparent after the first step (3.24). It is therefore possible to iterate the reduction and subsequent factorization, i. e. repetitively exploit Theorem 3.7 as often as desired; the next corollary describes how to recursively generate the matrices that describe this procedure—i. e. the cumulative framework.

Corollary 3.8. *Assume that the reduction and factorization of the error from Theorem 3.7 has been recursively performed for k steps, where the columns of each $\mathbf{V}_i$ form a basis of a rational Krylov subspace, i. e. it satisfies a corresponding Sylvester equation similar to (3.28). Then the error of the total reduction can be factorized as*

$$\boldsymbol{X}(s) = \mathbf{V}_{\mathrm{tot}}\boldsymbol{X}_{r,\mathrm{tot}}(s) + \boldsymbol{X}_{\perp,i}(s)\boldsymbol{G}_{f,\mathrm{tot}}(s) \tag{3.45}$$

with

$$\boldsymbol{X}_{r,\mathrm{tot}}(s) = (s\mathbf{E}_{r,\mathrm{tot}} - \mathbf{A}_{r,\mathrm{tot}})^{-1}\mathbf{B}_{r,\mathrm{tot}}, \tag{3.46}$$

$$\boldsymbol{X}_{\perp,i}(s) = (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}_{\perp,i}, \tag{3.47}$$

$$\boldsymbol{G}_{f,\mathrm{tot}}(s) = \mathbf{L}_{\mathrm{tot}}(s\mathbf{E}_{r,\mathrm{tot}} - \mathbf{A}_{r,\mathrm{tot}})^{-1}\mathbf{B}_{r,\mathrm{tot}} + \mathbf{I}, \tag{3.48}$$

where the final matrices can recursively be generated by

$$\mathbf{A}_{r,\mathrm{tot}} \leftarrow \begin{bmatrix} \mathbf{A}_{r,\mathrm{tot}} & \mathbf{0} \\ \mathbf{B}_{r,i}\mathbf{L}_{\mathrm{tot}} & \mathbf{A}_{r,i} \end{bmatrix}, \qquad \mathbf{E}_{r,\mathrm{tot}} \leftarrow \begin{bmatrix} \mathbf{E}_{r,\mathrm{tot}} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{r,i} \end{bmatrix}, \ \mathbf{B}_{r,\mathrm{tot}} \leftarrow \begin{bmatrix} \mathbf{B}_{r,\mathrm{tot}} \\ \mathbf{B}_{r,i} \end{bmatrix}, \tag{3.49}$$

$$\mathbf{B}_{\perp,i} = \mathbf{B}_{\perp,i-1} - \mathbf{E}\mathbf{V}_i\mathbf{E}_{r,i}^{-1}\mathbf{B}_{r,i}, \quad \mathbf{V}_{\mathrm{tot}} \leftarrow [\, \mathbf{V}_{\mathrm{tot}} \ \ \mathbf{V}_i \,], \qquad \mathbf{L}_{\mathrm{tot}} \leftarrow [\, \mathbf{L}_{\mathrm{tot}} \ \ \mathbf{L}_i \,], \tag{3.50}$$

for $i = 1, \dots, k$ with $\mathbf{B}_{\perp,0} = \mathbf{B}$ and where $\mathbf{V}_{\mathrm{tot}}$, $\mathbf{L}_{\mathrm{tot}}$, $\mathbf{A}_{r,\mathrm{tot}}$, $\mathbf{E}_{r,\mathrm{tot}}$ and $\mathbf{B}_{r,\mathrm{tot}}$ are all initialized as empty matrices.

Proof. The proof follows by recursively applying Theorem 3.7. □

According to [148], the procedure of Corollary 3.8 will be denoted as *cumulative reduction* (CURE) in the following. CURE is of course also valid if $\boldsymbol{G}(s)$, with arbitrary output $\mathbf{C}$, has to be approximated instead of $\boldsymbol{X}(s)$. The result is obvious with Corollary 3.8, but it is nevertheless presented for completeness.

Corollary 3.9. *Let all variables be as defined in Corollary 3.8 and assume that the columns of each $\mathbf{V}_i$ form a basis of a rational input Krylov subspace. Then the model $\boldsymbol{G}(s) = \mathbf{C}\,(s\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B}$ can be reduced in a cumulative procedure with k steps, such that the total reduction and error factorization is given by*

$$\boldsymbol{G}(s) = \boldsymbol{G}_{r,\mathrm{tot}}(s) + \boldsymbol{G}_{\perp,i}(s)\boldsymbol{G}_{f,\mathrm{tot}}(s) \tag{3.51}$$

with

$$\boldsymbol{G}_{r,\mathrm{tot}}(s) = \mathbf{C}_{r,\mathrm{tot}}\,(s\mathbf{E}_{r,\mathrm{tot}} - \mathbf{A}_{r,\mathrm{tot}})^{-1}\,\mathbf{B}_{r,\mathrm{tot}}, \tag{3.52}$$

$$\boldsymbol{G}_{\perp,i}(s) = \mathbf{C}\,(s\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B}_{\perp,i}, \tag{3.53}$$

and $\mathbf{C}_{r,\mathrm{tot}} = \mathbf{C}\mathbf{V}_{\mathrm{tot}}$.

Corollaries 3.8 and 3.9 define the cumulative framework CURE. The basic procedure of three steps in CURE is illustrated in Figure 3.2, where the dimensions of the matrices of the state-space realizations are described in the form

$$\boldsymbol{G}(s) = \mathbf{C}\,(s\mathbf{E} - \mathbf{A})^{-1}\,\mathbf{B} + \mathbf{D} = \left[\begin{array}{c|c} \mathbf{E}, \mathbf{A} & \mathbf{B} \\ \hline \mathbf{C} & \mathbf{D} \end{array}\right]. \tag{3.54}$$

The matrices of corresponding steps in CURE are printed with equal shades of grey, and the shade of grey becomes darker with consecutive steps. It should be stressed that $\mathbf{A}$, $\mathbf{E}$ and $\mathbf{C}$ remain unchanged in $\boldsymbol{G}_{\perp}(s)$ during all steps of CURE.

It is next shown that CURE is not restrictive, which means that the results of Chapter 2 also apply to the accumulated quantities in CURE.

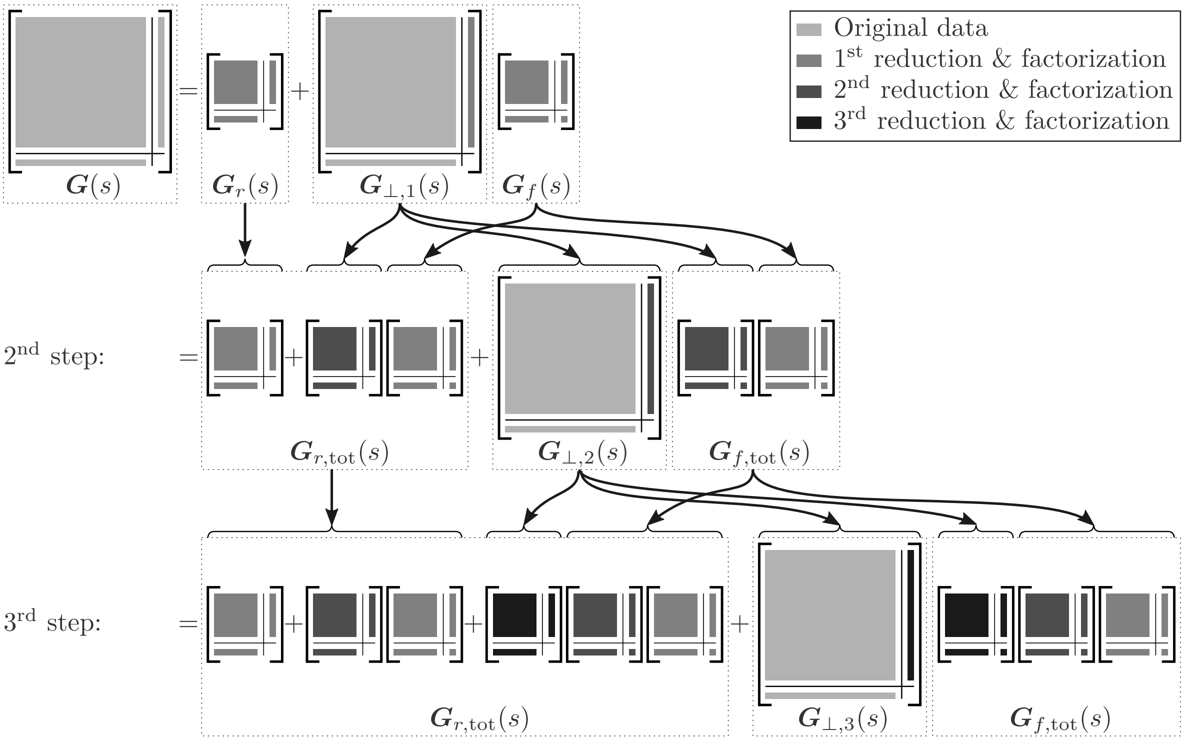

Original data
1st reduction & factorization
2nd reduction & factorization
3rd reduction & factorization
$\boldsymbol{G}(s)$
=
$\boldsymbol{G}_r(s)$
+
$\boldsymbol{G}_{\perp,1}(s)$
$\boldsymbol{G}_f(s)$
2nd step:
$\boldsymbol{G}_{r,\mathrm{tot}}(s)$
$\boldsymbol{G}_{\perp,2}(s)$
$\boldsymbol{G}_{f,\mathrm{tot}}(s)$
3rd step:
$\boldsymbol{G}_{r,\mathrm{tot}}(s)$
$\boldsymbol{G}_{\perp,3}(s)$
$\boldsymbol{G}_{f,\mathrm{tot}}(s)$

Figure 3.2: Illustration of three steps of the cumulative framework CURE.

3.2.2 Properties of the Cumulative Framework

The first question is whether $\mathbf{V}_{\mathrm{tot}}$ satisfies a $\mathbf{B}$-Sylvester equation, and consequently, whether its columns form a basis of a rational input Krylov subspace. The answer is given in the next lemma.

Lemma 3.10. *Let all variables be as defined in Corollary 3.8 and assume that the columns of each $\mathbf{V}_i$ form a basis of a recursively computed input rational Krylov subspace, i. e. each $\mathbf{V}_i$ satisfies due to Theorem 2.4*

$$\mathbf{A}\mathbf{V}_i - \mathbf{E}\mathbf{V}_i\mathbf{S}_i = \mathbf{B}_{\perp,i-1}\mathbf{L}_i, \tag{3.55}$$

for $i=1,\dots,k$ and $\mathbf{B}_{\perp,0}=\mathbf{B}$. Then the accumulated matrix $\mathbf{V}_{\mathrm{tot}}=[\mathbf{V}_1,\dots,\mathbf{V}_k]$ satisfies

$$\mathbf{A}\mathbf{V}_{\mathrm{tot}} - \mathbf{E}\mathbf{V}_{\mathrm{tot}}\mathbf{S}_{\mathrm{tot}} = \mathbf{B}\mathbf{L}_{\mathrm{tot}}, \tag{3.56}$$

where $\mathbf{S}_{\mathrm{tot}}$ and $\mathbf{L}_{\mathrm{tot}}$ can be generated by the recursive procedure

$$\mathbf{S}_{\mathrm{tot}} \leftarrow \begin{bmatrix} \mathbf{S}_{\mathrm{tot}} & -\mathbf{E}_{r,\mathrm{tot}}^{-1}\mathbf{B}_{r,\mathrm{tot}}\mathbf{L}_i \\ \mathbf{0} & \mathbf{S}_i \end{bmatrix}, \qquad \mathbf{L}_{\mathrm{tot}} \leftarrow \begin{bmatrix} \mathbf{L}_{\mathrm{tot}} & \mathbf{L}_i \end{bmatrix}, \tag{3.57}$$

in which $\mathbf{V}_{\text{tot}}$*,* $\mathbf{S}_{\text{tot}}$*, and* $\mathbf{L}_{\text{tot}}$ *are initialized as empty matrices.*

Proof. The proof is done by induction. The case $i=1$ is true due to Theorem 2.4. Then assume that (3.56) holds at step i–1 and let $\mathbf{V}_i$ satisfy (3.55). Arranging both Sylvester equations (3.55) and (3.56) in one equation reads as

$$\mathbf{A}\left[\begin{array}{cc}\mathbf{V}_{\text{tot}} & \mathbf{V}_i\end{array}\right]-\mathbf{E}\left[\begin{array}{cc}\mathbf{V}_{\text{tot}} & \mathbf{V}_i\end{array}\right]\left[\begin{array}{cc}\mathbf{S}_{\text{tot}} & \mathbf{0}\\ \mathbf{0} & \mathbf{S}_i\end{array}\right]=\left[\begin{array}{cc}\mathbf{B}\mathbf{L}_{\text{tot}} & \mathbf{B}_{\perp,i-1}\mathbf{L}_i\end{array}\right]. \tag{3.58}$$

Replacing $\mathbf{B}_{\perp,i-1}=\mathbf{B}-\sum_{\nu=1}^{i-1}\mathbf{E}\mathbf{V}_\nu\mathbf{E}_{r,\nu}^{-1}\mathbf{B}_{r,\nu}=\mathbf{B}-\mathbf{E}\mathbf{V}_{\text{tot}}\mathbf{E}_{r,\text{tot}}^{-1}\mathbf{B}_{r,\text{tot}}$ in (3.58) yields

$$\mathbf{A}\left[\begin{array}{cc}\mathbf{V}_{\text{tot}} & \mathbf{V}_i\end{array}\right]-\mathbf{E}\left[\begin{array}{cc}\mathbf{V}_{\text{tot}} & \mathbf{V}_i\end{array}\right]\left[\begin{array}{cc}\mathbf{S}_{\text{tot}} & -\mathbf{E}_{r,\text{tot}}^{-1}\mathbf{B}_{r,\text{tot}}\mathbf{L}_i\\ \mathbf{0} & \mathbf{S}_i\end{array}\right]=\mathbf{B}\left[\begin{array}{cc}\mathbf{L}_{\text{tot}} & \mathbf{L}_i\end{array}\right], \tag{3.59}$$

which completes the proof. □

Remark 3.11. It should be noted that due to its upper-triangular block structure, the eigenvalues of $\mathbf{S}_{\text{tot}}$ are the union of the eigenvalues of $\mathbf{S}_i$, $i=1,\ldots,k$. The columns of $\mathbf{V}_{\text{tot}}$ thus form a basis of the rational Krylov subspace with input $\mathbf{B}$, and where the shifts and tangential directions are encoded in $\mathbf{S}_{\text{tot}}$ and $\mathbf{L}_{\text{tot}}$. It is therefore irrespective of whether $\mathbf{B}$ or $\mathbf{B}_{\perp,i}$ is used to compute $\mathbf{V}_{i+1}$, as this does not change the subspace that is spanned by the columns of $\mathbf{V}_{\text{tot}}$. Lemma 3.10 thus actually gives an alternative proof of the “nested property” of rational Krylov subspace, which was already mentioned.

Owing to the Sylvester equation (3.56), one could define a parametrized family of reduced models like in Section 2.5 by using $\mathbf{S}_{\text{tot}}$ and $\mathbf{L}_{\text{tot}}$; yet this is senseless as it would destroy the cumulative idea. Nevertheless, it might be interesting to note that Lemma 2.14 can be maintained in CURE.

Lemma 3.12. *Let all variables be as defined in Lemma 3.10, then*

$$\mathbf{A}_{r,\text{tot}}=\mathbf{E}_{r,\text{tot}}\mathbf{S}_{\text{tot}}+\mathbf{B}_{r,\text{tot}}\mathbf{L}_{\text{tot}}. \tag{3.60}$$

Proof. The proof is straightforward by inserting the definitions (3.57) and (3.49). □

It follows from this result, that also the $\mathbf{B}_\perp$-Sylvester equation holds for the cumulated quantities in CURE.

Lemma 3.13. *Let all variables be as defined in Lemma 3.10, then* $\mathbf{V}_{\text{tot}}$ *satisfies*

$$\mathbf{A}\mathbf{V}_{\text{tot}}-\mathbf{E}\mathbf{V}_{\text{tot}}\mathbf{E}_{r,\text{tot}}^{-1}\mathbf{A}_{r,\text{tot}}=\mathbf{B}_{\perp,i}\mathbf{L}_{\text{tot}}. \tag{3.61}$$

Proof. The proof follows by replacing $\mathbf{S}_{\text{tot}}$ in (3.56) by (3.60) and noting that $\mathbf{B}_{\perp,i} = \mathbf{B} - \mathbf{E}\mathbf{V}_{\text{tot}}\mathbf{E}_{r,\text{tot}}^{-1}\mathbf{B}_{r,\text{tot}}$. □

Finally, it can also be shown that the transmission zeros of $\boldsymbol{G}_{f,\text{tot}}$ are the union of the transmission zeros of $\boldsymbol{G}_{f,i}$, $i=1,\ldots,k$.

Lemma 3.14. *Let s_0 be an eigenvalue of $\mathbf{S}_{\text{tot}}$, and assume that s_0 is not a pole of $\boldsymbol{G}_{f,\text{tot}}(s)$, and that the pair $(\mathbf{L}_{\text{tot}}, \mathbf{S}_{\text{tot}})$ is observable. Then s_0 is a transmission zero of $\boldsymbol{G}_{f,\text{tot}}(s)$.*

Proof. The proof is analogous to the one of Lemma 3.3 as the only requirement is that (3.60) holds. □

The original goal of MOR was to (tangentially) interpolate $\boldsymbol{X}(s)$, or equivalently $\boldsymbol{G}(s)$. At first sight, it seems like this property should be lost in CURE, because at steps $i = 2,\ldots,k$ we actually interpolate $\boldsymbol{X}_{\perp,i-1}(s)$ instead of $\boldsymbol{X}(s)$. Nevertheless, Lemma 3.14 shows that interpolation at the primary expansion points indeed is preserved in CURE: it follows from (3.57), that $\mathbf{S}_{\text{tot}}$ unites the eigenvalues of all $\mathbf{S}_i$, $i = 1,\ldots,k$, which precisely are the expansion points; as these eigenvalues are transmission zeros of $\boldsymbol{G}_{f,\text{tot}}(s)$, they are also transmission zeros of the error, $\mathbf{E}_{\text{tot}}(s) = \boldsymbol{X}_{\perp,k}(s)\boldsymbol{G}_{f,\text{tot}}(s)$, and hence, interpolation is preserved. The remaining question is, what are the tangential directions? The answer to this is quite involved in the most general case, so for a concise presentation, we consider only the case of $k=2$ steps of CURE in the next lemma. As we work with the Sylvester equation, we call the pair $(\mathbf{S},\mathbf{L})$ *interpolation data*, knowing that the expansion points s_i are encoded as eigenvalues of $\mathbf{S}$ and that the respective tangential directions $\mathbf{l}_i$ are encoded as columns of $\mathbf{L}$, after transforming $\mathbf{S}$ to Jordan canonical form, cf. Lemma 2.3.

Lemma 3.15. *Given two steps of CURE, where $\mathbf{V}_1$ and $\mathbf{V}_2$ satisfy*

$$\mathbf{A}\mathbf{V}_1 - \mathbf{E}\mathbf{V}_1\mathbf{S}_1 = \mathbf{B}\mathbf{L}_1, \tag{3.62}$$

$$\mathbf{A}\mathbf{V}_2 - \mathbf{E}\mathbf{V}_2\mathbf{S}_2 = \mathbf{B}_{\perp,1}\mathbf{L}_2. \tag{3.63}$$

Then $\mathbf{V}_{\text{tot}}\boldsymbol{X}_{r,\text{tot}}(s)$ interpolates $\boldsymbol{X}(s)$ at the data $(\mathbf{S}_1,\mathbf{L}_1)$ and $(\mathbf{S}_2,\mathbf{L}_2{+}\mathbf{L}_1\mathbf{M})$, where $\mathbf{M}$ satisfies

$$\mathbf{S}_1\mathbf{M} - \mathbf{M}\mathbf{S}_2 = \mathbf{E}_{r,1}^{-1}\mathbf{B}_{r,1}\mathbf{L}_2. \tag{3.64}$$

Proof. Due to Lemma 3.10,

$$\mathbf{A}\left[\begin{array}{cc}\mathbf{V}_1 & \mathbf{V}_2\end{array}\right] - \mathbf{E}\left[\begin{array}{cc}\mathbf{V}_1 & \mathbf{V}_2\end{array}\right]\left[\begin{array}{cc}\mathbf{S}_1 & -\mathbf{E}_{r,1}^{-1}\mathbf{B}_{r,1}\mathbf{L}_2 \\ \mathbf{0} & \mathbf{S}_2\end{array}\right] = \mathbf{B}\left[\begin{array}{cc}\mathbf{L}_1 & \mathbf{L}_2\end{array}\right]. \tag{3.65}$$

Introduce the matrix $\mathbf{T}$ to transform $\mathbf{S}_{\mathrm{tot}}$:

$$\mathbf{T} = \begin{bmatrix} \mathbf{I} & \mathbf{M} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathbf{T}^{-1}\mathbf{S}_{\mathrm{tot}}\mathbf{T} = \begin{bmatrix} \mathbf{S}_1 & \mathbf{S}_1\mathbf{M} - \mathbf{M}\mathbf{S}_2 - \mathbf{E}_{r,1}^{-1}\mathbf{B}_{r,1}\mathbf{L}_2 \\ \mathbf{0} & \mathbf{S}_2 \end{bmatrix} \tag{3.66}$$

The matrix $\mathbf{T}$ thus transforms $\mathbf{S}_{\mathrm{tot}}$ to block-diagonal structure if and only if (3.64) holds. In order to transform (3.65), multiply it with $\mathbf{T}$ from the right hand side, which leads to $\mathbf{L}_{\mathrm{tot}}\mathbf{T} = [\mathbf{L}_1, \mathbf{L}_2 + \mathbf{L}_1\mathbf{M}]$ and which completes the proof. □

Lemma 3.15 shows how the tangential directions that are encoded in $\mathbf{L}_2$, and which are applied to the input $\mathbf{B}_{\perp,1}$, trace back to the original input $\mathbf{B}$, namely by $\mathbf{L}_2 + \mathbf{L}_1\mathbf{M}$. This shows that the proper choice of tangential directions in the CURE framework might not be transparent. The situation, however, changes if either there is a single-input, $m = 1$, or only block-input rational Krylov subspaces are employed. Then the interpolation is irrespective of the choice of the $\mathbf{L}_i$ and every eigenvalue s_i of $\mathbf{S}_{\mathrm{tot}}$ becomes a blocking zero in the error: $\boldsymbol{G}_{f,\mathrm{tot}}(s_i) = \mathbf{0}$. There is a final remark in order, concerning the CURE framework.

Remark 3.16. An important property of CURE is that the individual reduction steps are decoupled: we can choose the interpolation points freely, as this choice is neither depending on previous steps, nor does it affect the subsequent ones; and even more, once an expansion point is incorporated in the accumulated quantities, the interpolation property will not be destroyed (if no singularities occur). Furthermore, it follows from the structure of $\mathbf{E}_{r,\mathrm{tot}}$ and $\mathbf{A}_{r,\mathrm{tot}}$, that also the reduced eigenvalues of the individual reduction steps are accumulated in the total reduced model. In addition, for the recursive construction of the accumulated quantities, there is no need to individually access the quantities of previous steps; it is instead sufficient to use only the accumulated quantities so far (index “tot”), cf. (3.49) and (3.57). It is therefore indeed justified to call CURE a cumulative framework, as all previous data is kept, while new one can be added independently. Finally, it should be noted that the additional numerical effort of CURE is negligible: the recursive factorization only requires the computation of $\mathbf{L}_i$ and $\mathbf{B}_{\perp,i}$, which are easy to calculate, cf. Chapter 2.

CURE was first published in [149], where it is denoted as “iterative model order reduction”. Therein, $\boldsymbol{X}_{\perp,i}(s)$ was investigated in every step, in order to adaptively select interpolation points. This idea is elaborated by Panzer in his thesis [148], where also the notation “CURE” is introduced. It is interesting to note that a quite similar approach was presented by Ahmad et. al. [3], however, only in the context of solving large-scale Sylvester and Lyapunov equations. Also related to CURE is the work of

Lefteriu and Antoulas [125], who introduced the idea of recursive interpolation in the Loewner framework.

All in all, CURE is a framework that allows to recursively accumulate a reduced model, or equivalently, an approximation of a large-scale Lyapunov equation. Thereby, the reduced order can be adaptively chosen. This is probably the main improvement that CURE contributes: recursively perform independent reduction steps, until at some point the desired accuracy is reached. Even more, this functionality is independent from the degrees of freedom in Krylov-based projection methods, which are the selection of interpolation points and tangential directions, and also the choice of the projection matrices $\mathbf{W}_i$. To conclude, the CURE framework preserves all the flexibility that is available in Krylov-based projection methods.

It remains to determine interpolation points, tangential directions, and the projection matrix $\mathbf{W}$ that induce a good approximation. The following chapter addresses the latter by introducing the concept of $\mathcal{H}_2$ pseudo-optimality. And as it will turn out, this concept perfectly suits into the CURE framework, as it will assure nice properties.

4 Analysis of $\mathcal{H}_2$ Pseudo-Optimality

pseudo-: false, not genuine, fake

(wiktionary.org)

The statement “something is optimal” is not absolute; instead, it certainly relates to some measure. Consequently, in the context of model order reduction one first of all has to suitably quantify the error with respect to which optimality is desired. To this end, there are various adequate ways, all of which result in well-defined system norms; see e. g. [8] for a nice overview. A common system norm is the $\mathcal{H}_\infty$ norm, as it is an induced norm, which provides a concrete interpretation: the maximum amplification possible. Measuring the error in the $\mathcal{H}_\infty$ norm hence provides valuable information, and what is more, having a reduced model that minimizes the error in the $\mathcal{H}_\infty$ norm indeed might be the ultimate solution in many practical applications of MOR. A severe drawback of the $\mathcal{H}_\infty$ norm, however, is that it is hard to handle, both analytically and numerically, and especially in a large-scale setting.

This is the main reason, why in this thesis the error is quantified in the $\mathcal{H}_2$ norm, as it is not only numerically accessible even in a large-scale setting, but also “analytically convenient” [161]. This allows to derive convenient necessary conditions for optimality, which also led to an algorithmic implementation in the famous IRKA algorithm [92]. One drawback of the $\mathcal{H}_2$ norm is that interpretation is not as simple as for the $\mathcal{H}_\infty$ norm: it is “the expected root-mean-square (rms) value of the output when the input is a unit variance white noise process” [219]. Another drawback is that an $\mathcal{H}_2$ optimal reduced model might simply not be one’s desire; one may think of many practical applications where an $\mathcal{H}_\infty$ optimal reduced model would instead be the benchmark solution. In addition, the $\mathcal{H}_2$ norm is induced either in the single-input or single-output case, but not in general, cf. [8, p. 144] and [46]. Nevertheless, an $\mathcal{H}_2$ optimal reduced model typically also yields small error in the $\mathcal{H}_\infty$ norm, and it certainly is better to achieve at least $\mathcal{H}_2$ optimality than none.

The aim of this chapter is twofold. On the one hand, $\mathcal{H}_2$ optimal MOR is reviewed. On the other hand, the concept of $\mathcal{H}_2$ pseudo-optimality is comprehensively described, and it is discussed how the state-of-the-art in $\mathcal{H}_2$ optimal MOR may be improved by exploiting $\mathcal{H}_2$ pseudo-optimality. To this end, Section 4.1 reviews the $\mathcal{H}_2$ inner product and norm, and discusses different ways for their computation. There are convenient necessary conditions for $\mathcal{H}_2$ optimality, which can be stated in different forms. This is reviewed in Section 4.2, together with an iterative algorithm that, upon convergence, yields locally optimal reduced models. Section 4.3 presents the main result of this work: the concept of $\mathcal{H}_2$ pseudo-optimality is introduced; various equivalent necessary and sufficient conditions for $\mathcal{H}_2$ pseudo-optimality are defined; and a numerically efficient algorithm for the direct computation of an $\mathcal{H}_2$ pseudo-optimal reduced models is stated. Finally, the results are discussed and possible applications are suggested. The contributions of this chapter were partly published in preliminary form in [208].

4.1 The Hilbert Space $\mathcal{H}_2$

Let $\mathcal{H}_2^{(p,m)}$ denote the set of all asymptotically stable systems, which have m inputs and p outputs and a strictly proper transfer function. We will very much make use of the notation $\mathcal{H}_2^{(p,m)}$ in the following, however, in order to improve readability, we will drop "(p, m)" hereafter and use only $\mathcal{H}_2$ instead. This should not lead to confusion, because all models are generally assumed to have m inputs and p outputs, whereas deviations should become clear from the context.

4.1.1 $\mathcal{H}_2$ Inner Product

Now define the $\mathcal{H}_2$ inner product (which will induce a respective norm) as follows.

Definition 4.1 ([21])**.** Let $\boldsymbol{G}(s)$ and $\boldsymbol{H}(s)$ be $\mathcal{H}_2$ functions, then their $\mathcal{H}_2$ inner product is defined as

$$\begin{aligned} \langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \operatorname{trace}\left[\boldsymbol{G}^*(\imath\omega)\boldsymbol{H}(\imath\omega)\right] \mathrm{d}\omega \qquad (4.1) \\ &= \int_{0}^{\infty} \operatorname{trace}\left[\boldsymbol{G}^*(t)\boldsymbol{H}(t)\right] \mathrm{d}t, \qquad (4.2) \end{aligned}$$

where the second equation is a consequence of Parseval's theorem.

With this inner product, the set $\mathcal{H}_2$ becomes a Hilbert space, cf. [24]. There is a convenient way to compute the $\mathcal{H}_2$ inner product based on large-scale (sparse-sparse) Sylvester equations, which is presented in the next lemma.

Lemma 4.1. *Let* $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B}$ *and* $\boldsymbol{H}(s) = \mathbf{C}_{\mathrm{H}}\left(s\mathbf{E}_{\mathrm{H}} - \mathbf{A}_{\mathrm{H}}\right)^{-1}\mathbf{B}_{\mathrm{H}}$ *be* $\mathcal{H}_2$ *functions, then the* $\mathcal{H}_2$ *inner product is given by*

$$\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \mathrm{trace}\left(\mathbf{B}^*\mathbf{Y}\mathbf{B}_{\mathrm{H}}\right) = \mathrm{trace}\left(\mathbf{C}\mathbf{X}\mathbf{C}_{\mathrm{H}}^*\right), \tag{4.3}$$

where $\mathbf{X}$ *and* $\mathbf{Y}$ *satisfy*

$$\mathbf{A}\mathbf{X}\mathbf{E}_{\mathrm{H}}^* + \mathbf{E}\mathbf{X}\mathbf{A}_{\mathrm{H}}^* + \mathbf{B}\mathbf{B}_{\mathrm{H}}^* = \mathbf{0}, \tag{4.4}$$

$$\mathbf{A}^*\mathbf{Y}\mathbf{E}_{\mathrm{H}} + \mathbf{E}^*\mathbf{Y}\mathbf{A}_{\mathrm{H}} + \mathbf{C}^*\mathbf{C}_{\mathrm{H}} = \mathbf{0}. \tag{4.5}$$

Proof. Define

$$\mathbf{Y} = \mathbf{E}^{-*}\int_0^\infty \left(\mathrm{e}^{\mathbf{A}^*\mathbf{E}^{-*}t}\mathbf{C}^*\mathbf{C}_{\mathrm{H}}\mathrm{e}^{\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{A}_{\mathrm{H}}t}\right)\mathrm{d}t\;\mathbf{E}_{\mathrm{H}}^{-1}, \tag{4.6}$$

then

$$\begin{aligned}\mathbf{A}^*\mathbf{Y}\mathbf{E}_{\mathrm{H}} + \mathbf{E}^*\mathbf{Y}\mathbf{A}_{\mathrm{H}} &= \int_0^\infty \Big(\mathbf{A}^*\mathbf{E}^{-*}\mathrm{e}^{\mathbf{A}^*\mathbf{E}^{-*}t}\mathbf{C}^*\mathbf{C}_{\mathrm{H}}\mathrm{e}^{\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{A}_{\mathrm{H}}t} + \\ &\qquad + \mathrm{e}^{\mathbf{A}^*\mathbf{E}^{-*}t}\mathbf{C}^*\mathbf{C}_{\mathrm{H}}\mathrm{e}^{\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{A}_{\mathrm{H}}t}\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{A}_{\mathrm{H}}\Big)\,\mathrm{d}t \qquad (4.7)\\ &= \left[\mathrm{e}^{\mathbf{A}^*\mathbf{E}^{-*}t}\mathbf{C}^*\mathbf{C}_{\mathrm{H}}\mathrm{e}^{\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{A}_{\mathrm{H}}t}\right]_0^\infty = \mathbf{0} - \mathbf{C}^*\mathbf{C}_{\mathrm{H}}, \qquad (4.8)\end{aligned}$$

which proves that $\mathbf{Y}$ satisfies (4.5). From the time-domain formulation of the $\mathcal{H}_2$ inner product (4.2) it follows that

$$\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \mathrm{trace}\left[\mathbf{B}^*\mathbf{E}^{-*}\int_0^\infty \left(\mathrm{e}^{\mathbf{A}^*\mathbf{E}^{-*}t}\mathbf{C}^*\mathbf{C}_{\mathrm{H}}\mathrm{e}^{\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{A}_{\mathrm{H}}t}\right)\mathrm{d}t\;\mathbf{E}_{\mathrm{H}}^{-1}\mathbf{B}_{\mathrm{H}}\right], \tag{4.9}$$

which proves $\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \mathrm{trace}(\mathbf{B}^*\mathbf{Y}\mathbf{B}_{\mathrm{H}})$. By using $\mathrm{trace}[\boldsymbol{G}^*(t)\boldsymbol{H}(t)] = \mathrm{trace}[\boldsymbol{H}(t)\boldsymbol{G}^*(t)]$, the proof of $\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \mathrm{trace}(\mathbf{C}\mathbf{X}\mathbf{C}_{\mathrm{H}}^*)$ can be derived analogously to the above one. □

An alternative way to compute the $\mathcal{H}_2$ inner product is based on the *pole-residue representation* of $\boldsymbol{G}(s)$. For the ease of presentation, assume for the moment that $\mathbf{E}^{-1}\mathbf{A}$ is diagonalizable, then the pole-residue representation is defined as follows.

Definition 4.2. Let $\mathbf{A} = \mathrm{diag}[\lambda_1\mathbf{I}_1, \ldots, \lambda_k\mathbf{I}_k]$, $\mathbf{B} = [\mathbf{B}_1^*, \ldots, \mathbf{B}_k^*]^*$ and $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_k]$, with $\mathbf{B}_i \in \mathbb{C}^{m_i \times m}$ and $\mathbf{C}_i \in \mathbb{C}^{p \times m_i}$ and where $\mathbf{I}_i$ is the $m_i \times m_i$ identity matrix. Then

$$\boldsymbol{G}(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i\mathbf{B}_i}{s - \lambda_i}, \tag{4.10}$$

is called the pole-residue representation of $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{B}$; the $\mathbf{B}_i$ and $\mathbf{C}_i$ are called input and output residues (residue directions), respectively.

Remark 4.2. It should be noted that the pole-residue representation is not unique. Consider e. g. non-singular $\mathbf{T}_i \in \mathbb{C}^{m_i \times m_i}$ and define the transformed residues $\widetilde{\mathbf{B}}_i = \mathbf{T}_i \mathbf{B}_i$ and $\widetilde{\mathbf{C}}_i = \mathbf{C}_i \mathbf{T}_i^{-1}$, then obviously, $\boldsymbol{G}(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i \mathbf{B}_i}{s-\lambda_i} = \sum_{i=1}^{k} \frac{\widetilde{\mathbf{C}}_i \widetilde{\mathbf{B}}_i}{s-\lambda_i}$. The row span of $\mathbf{B}_i$ and the column span of $\mathbf{C}_i$, however, is invariant under this transformation. Then the subspace spanned by the columns of $\mathbf{B}_i^*$ and $\mathbf{C}_i$ is unique, and it is indeed judicious to also call $\mathbf{B}_i$ and $\mathbf{C}_i$ residue *directions.*

With the above definition, we may formulate the $\mathcal{H}_2$ inner product based on poles and residues as follows.

Lemma 4.3. *Let $\boldsymbol{G}(s)$ and $\boldsymbol{H}(s)$ be $\mathcal{H}_2$ functions, and assume that $\boldsymbol{H}(s)$ is given in pole-residue representation: $\boldsymbol{H}(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i \mathbf{B}_i}{s-\lambda_i}$. Then the $\mathcal{H}_2$ inner product is given by*

$$\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \sum_{i=1}^{k} \operatorname{trace} \left[\mathbf{C}_i^* \boldsymbol{G}(-\overline{\lambda}_i) \mathbf{B}_i^* \right]^* . \tag{4.11}$$

Proof. An admissible state-space realization of $\boldsymbol{H}(s)$ is given in Definition 4.2 together with $\mathbf{E}_\mathrm{H} = \mathbf{I}$ identity. Now consider (4.5) and divide $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_k]$ into the blocks $\mathbf{Y}_1 \in \mathbb{C}^{N \times m_i}$. It then follows due to the block-diagonal structure of $\mathbf{A}_\mathrm{H}$ that (this is actually the direct application of (2.4))

$$\mathbf{Y}_i = -\left(\mathbf{A}^* + \lambda_i \mathbf{E}^* \right)^{-1} \mathbf{C}^* \mathbf{C}_i. \tag{4.12}$$

Now using Lemma 4.1 to compute the $\mathcal{H}_2$ inner product yields

$$\begin{aligned} \langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} &= \operatorname{trace} \left[-\mathbf{B}^* \sum_{i=1}^{k} \left(\mathbf{A}^* + \lambda_i \mathbf{E}^* \right)^{-1} \mathbf{C}^* \mathbf{C}_i \mathbf{B}_i \right] = \operatorname{trace} \left[\sum_{i=1}^{k} \boldsymbol{G}^*(-\overline{\lambda}_i) \mathbf{C}_i \mathbf{B}_i \right] && (4.13) \\ &= \sum_{i=1}^{k} \operatorname{trace} \left[\mathbf{B}_i \boldsymbol{G}^*(-\overline{\lambda}_i) \mathbf{C}_i \right], && (4.14) \end{aligned}$$

then transposition with complex conjugation completes the proof. □

Remark 4.4. It should be noted, that if $\boldsymbol{G}(s)$ admits a state-space realization with real matrices then obviously, $\overline{\boldsymbol{G}}(s) = \boldsymbol{G}(\overline{s})$. It then directly follows, that the $\mathcal{H}_2$ inner product could also be formulated as $\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \sum_{i=1}^{k} \operatorname{trace} \left[\mathbf{C}_i^T \boldsymbol{G}(-\lambda_i) \mathbf{B}_i^T \right]$. If both models $\boldsymbol{G}(s)$ and $\boldsymbol{H}(s)$ admit a real valued realization in state-space, then it follows from Lemma 4.1 that the $\mathcal{H}_2$ inner product is also real valued, and it could be formulated as $\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \sum_{i=1}^{k} \operatorname{trace} \left[\mathbf{C}_i^* \boldsymbol{G}(-\overline{\lambda}_i) \mathbf{B}_i^* \right]$. As it will be easier to state the subsequent results if these assumptions are avoided, we, however, still use the slightly more inconvenient formulation in (4.11).

We will also make use of the next lemma, which introduces a slight modification of the above one, and which appears to be new.

Lemma 4.5. *Let $\boldsymbol{G}(s)$ and $\boldsymbol{H}(s)$ be $\mathcal{H}_2$ functions, and assume that $\boldsymbol{H}(s)$ is given in pole-residue representation: $\boldsymbol{H}(s)=\sum_{i=1}^{k}\frac{\mathbf{C}_i\mathbf{B}_i}{s-\lambda_i}$. Let $\boldsymbol{H}'(s)$ denote the derivative of $\boldsymbol{H}(s)$ with respect to s, then*

$$\langle \boldsymbol{G}, \boldsymbol{H}'\rangle_{\mathcal{H}_2} = \sum_{i=1}^{k} \operatorname{trace}\left[\mathbf{C}_i^* \boldsymbol{G}'(-\overline{\lambda}_i)\mathbf{B}_i^*\right]^*. \tag{4.15}$$

Proof. We first of all seek for an appropriate state-space realization of $\boldsymbol{H}'(s)$. To this end, define $\mathbf{E}_{\mathrm{H}'}=\mathbf{I}$ and

$$\mathbf{A}_{\mathrm{H}'}=\begin{bmatrix} \widetilde{\mathbf{A}}_1 & & \\ & \ddots & \\ & & \widetilde{\mathbf{A}}_k \end{bmatrix}, \quad \mathbf{B}_{\mathrm{H}'}=\begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \vdots \\ \widetilde{\mathbf{B}}_k \end{bmatrix}, \quad \mathbf{C}_{\mathrm{H}'}=\begin{bmatrix} \widetilde{\mathbf{C}}_1 & \dots & \widetilde{\mathbf{C}}_k \end{bmatrix}, \tag{4.16}$$

where

$$\widetilde{\mathbf{A}}_i=\begin{bmatrix} \lambda_i\mathbf{I} & \mathbf{I} \\ \mathbf{0} & \lambda_i\mathbf{I} \end{bmatrix}, \qquad \widetilde{\mathbf{B}}_i=\begin{bmatrix} \mathbf{0} \\ \mathbf{B}_i \end{bmatrix}, \qquad \widetilde{\mathbf{C}}_i=[\ -\mathbf{C}_i \quad \mathbf{0}\ ]. \tag{4.17}$$

This defines the desired state-space realization by $\boldsymbol{H}'(s)=\mathbf{C}_{\mathrm{H}'}\left(s\mathbf{E}_{\mathrm{H}'}-\mathbf{A}_{\mathrm{H}'}\right)^{-1}\mathbf{B}_{\mathrm{H}'}$, because of

$$\mathbf{C}_{\mathrm{H}'}\left(s\mathbf{E}_{\mathrm{H}'}-\mathbf{A}_{\mathrm{H}'}\right)^{-1}\mathbf{B}_{\mathrm{H}'} = \sum_{i=1}^{k}\widetilde{\mathbf{C}}_i\begin{bmatrix} (s-\lambda_i)\mathbf{I} & -\mathbf{I} \\ \mathbf{0} & (s-\lambda_i)\mathbf{I} \end{bmatrix}^{-1}\widetilde{\mathbf{B}}_i \tag{4.18}$$

$$= \sum_{i=1}^{k}\frac{1}{(s-\lambda_i)^2}\widetilde{\mathbf{C}}_i\begin{bmatrix} (s-\lambda_i)\mathbf{I} & \mathbf{I} \\ \mathbf{0} & (s-\lambda_i)\mathbf{I} \end{bmatrix}\widetilde{\mathbf{B}}_i \tag{4.19}$$

$$= \sum_{i=1}^{k}\frac{-\mathbf{C}_i\mathbf{B}_i}{(s-\lambda_i)^2} = \boldsymbol{H}'(s). \tag{4.20}$$

Now consider (4.5) and divide $\mathbf{Y}=[\widetilde{\mathbf{Y}}_1,\dots,\widetilde{\mathbf{Y}}_k]$ into the blocks $\widetilde{\mathbf{Y}}_i=[\mathbf{Y}_{i,1}\,\mathbf{Y}_{i,2}]$, because then it holds for each block that

$$\mathbf{A}^*\begin{bmatrix} \mathbf{Y}_{i,1} & \mathbf{Y}_{i,2} \end{bmatrix}+\mathbf{E}^*\begin{bmatrix} \mathbf{Y}_{i,1} & \mathbf{Y}_{i,2} \end{bmatrix}\begin{bmatrix} \lambda_i\mathbf{I} & \mathbf{I} \\ \mathbf{0} & \lambda_i\mathbf{I} \end{bmatrix}=\mathbf{C}^*\begin{bmatrix} \mathbf{C}_i & \mathbf{0} \end{bmatrix}, \tag{4.21}$$

and hence,

$$\mathbf{Y}_{i,1} = \left(\mathbf{A}^*+\lambda_i\mathbf{E}^*\right)^{-1}\mathbf{C}^*\mathbf{C}_i, \tag{4.22}$$

$$\mathbf{Y}_{i,2} = -\left(\mathbf{A}^*+\lambda_i\mathbf{E}^*\right)^{-1}\mathbf{E}^*\left(\mathbf{A}^*+\lambda_i\mathbf{E}^*\right)^{-1}\mathbf{C}^*\mathbf{C}_i. \tag{4.23}$$

Now we use Lemma 4.1 to compute the $\mathcal{H}_2$ inner product by

$$\langle \boldsymbol{G}, \boldsymbol{H}' \rangle_{\mathcal{H}_2} = \operatorname{trace}\left(\mathbf{B}^*\mathbf{Y}\mathbf{B}_{\mathrm{H}'}\right) = \sum_{i=1}^{k} \operatorname{trace}\left(\mathbf{B}^*\widetilde{\mathbf{Y}}_i\widetilde{\mathbf{B}}_i\right) = \sum_{i=1}^{k} \operatorname{trace}\left(\mathbf{B}^*\mathbf{Y}_{i,2}\mathbf{B}_i\right) \tag{4.24}$$

$$= \sum_{i=1}^{k} \operatorname{trace}\left[-\mathbf{B}^*\left(\mathbf{A}^*+\lambda_i\mathbf{E}^*\right)^{-1}\mathbf{E}^*\left(\mathbf{A}^*+\lambda_i\mathbf{E}^*\right)^{-1}\mathbf{C}^*\mathbf{C}_i\mathbf{B}_i\right] \tag{4.25}$$

$$= \sum_{i=1}^{k} \operatorname{trace}\left[\mathbf{B}_i^*\mathbf{C}_i^*\boldsymbol{G}'(-\overline{\lambda}_i)\right]^*, \tag{4.26}$$

from which the statement can be concluded. □

It is interesting to note that the derivative with respect to s “jumps” in Lemma 4.5 from one model to the other. This result also induces the next statement.

Corollary 4.6. *Let $\boldsymbol{G}(s)$ and $\boldsymbol{H}(s)$ be $\mathcal{H}_2$ functions, and assume that $\boldsymbol{H}(s)$ is given in pole-residue representation, $\boldsymbol{H}(s)=\sum_{i=1}^{k}\frac{\mathbf{C}_i\mathbf{B}_i}{s-\lambda_i}$, then*

$$\langle \boldsymbol{G}, \boldsymbol{H}' \rangle_{\mathcal{H}_2} = \langle \boldsymbol{G}', \boldsymbol{H} \rangle_{\mathcal{H}_2}. \tag{4.27}$$

Proof. Consider Lemma 4.3 and replace $\boldsymbol{G}(s)$ by $\boldsymbol{G}'(s)$, then obviously $\langle \boldsymbol{G}', \boldsymbol{H} \rangle_{\mathcal{H}_2} = \sum_{i=1}^{k} \operatorname{trace}\left[\mathbf{C}_i^*\boldsymbol{G}'(-\overline{\lambda}_i)\mathbf{B}_i^*\right]^*$, which equals $\langle \boldsymbol{G}, \boldsymbol{H}' \rangle_{\mathcal{H}_2}$ due to Lemma 4.5. □

The formula for the $\mathcal{H}_2$ inner product based on poles and residues, as in Lemma 4.3, requires that at least one of the systems can be transformed to diagonal form. If this is not the case, Lemma 4.3 can be generalized, which however requires quite cumbersome notation. For the sake of a concise presentation, we only present the case of $\widetilde{m}$ Jordan blocks of equal dimensions to one eigenvalue λ in the following. The most general case can then be deduced from combining this result with Lemma 4.3.

Lemma 4.7. *Let $\boldsymbol{G}(s)$ and $\boldsymbol{H}(s)$ be $\mathcal{H}_2$ functions, and assume $1 \leq \widetilde{m} \leq m$ and that $\boldsymbol{H}(s)$ is given by $\boldsymbol{H}(s)=\mathbf{C}_{\mathrm{H}}\left(s\mathbf{I}-\mathbf{A}_{\mathrm{H}}\right)^{-1}\mathbf{B}_{\mathrm{H}}$, with*

$$\mathbf{A}_{\mathrm{H}} = \begin{bmatrix} \lambda\mathbf{I} & & & \\ -\mathbf{I} & \ddots & & \\ & \ddots & \ddots & \\ & & -\mathbf{I} & \lambda\mathbf{I} \end{bmatrix}, \quad \mathbf{B}_{\mathrm{H}} = \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_k \end{bmatrix}, \quad \mathbf{C}_{\mathrm{H}} = \begin{bmatrix} \mathbf{C}_1 & \ldots & \mathbf{C}_k \end{bmatrix}, \tag{4.28}$$

and where $\mathbf{I}$ denotes the $\widetilde{m}\times\widetilde{m}$ identity matrix and $\mathbf{B}_i \in \mathbb{C}^{\widetilde{m}\times m}$ and $\mathbf{C}_i \in \mathbb{C}^{p\times\widetilde{m}}$. Then the $\mathcal{H}_2$ inner product is given by

$$\langle \boldsymbol{G}, \boldsymbol{H} \rangle_{\mathcal{H}_2} = \sum_{i=1}^{k}\sum_{j=i}^{k} \operatorname{trace}\left[\mathbf{C}_j^* \boldsymbol{G}^{(j-i)}(-\overline{\lambda})\mathbf{B}_i^*\right]^* \frac{1}{(j-i)!} \tag{4.29}$$

$$= \sum_{i=1}^{k}\sum_{j=i}^{k} \operatorname{trace}\left[\mathbf{C}_j^* \mathbf{M}_{(j-i)}^{-\overline{\lambda}} \mathbf{B}_i^*\right]^*, \tag{4.30}$$

where $\boldsymbol{G}^{(j-i)}(s)$ denotes the $(j-i)$th derivative of $\boldsymbol{G}(s)$ with respect to s, and where the moments $\mathbf{M}_{(j-i)}^{-\overline{\lambda}}$ are defined in Section 1.6.2.

Proof. Consider (4.5) and divide $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_k]$ into the blocks $\mathbf{Y}_1 \in \mathbb{C}^{N\times\widetilde{m}}$. It then follows that

$$\mathbf{Y}_i = \sum_{j=i}^{k} -\left[\left(\mathbf{A}^* + \lambda\mathbf{E}^*\right)^{-1}\mathbf{E}^*\right]^{(j-i)}\left(\mathbf{A}^* + \lambda\mathbf{E}^*\right)^{-1}\mathbf{C}^*\mathbf{C}_j. \tag{4.31}$$

Using $\mathbf{M}_{(j-i)}^{-\overline{\lambda}} = -\mathbf{C}\left[\left(\mathbf{A}+\overline{\lambda}\mathbf{E}\right)^{-1}\mathbf{E}\right]^{(j-i)}\left(\mathbf{A}+\overline{\lambda}\mathbf{E}\right)^{-1}\mathbf{B}$ and $\boldsymbol{G}^{(j-i)}(-\overline{\lambda}) = \mathbf{M}_{(j-i)}^{-\overline{\lambda}}(j-i)!$, then the proof is analogous to the one of Lemma 4.3, hence omitted. □

Remark 4.8. The above lemma describes the case that the state-space representation of $\boldsymbol{H}(s)$ contains one eigenvalue λ of algebraic multiplicity $\widetilde{m}\,k$, and that there are $\widetilde{m}$ Jordan blocks, all of dimension k. Although $\mathbf{A}_{\mathrm{H}}$ in (4.28) is not in Jordan canonical form, it will become apparent in Section 4.3, why this unusual definition was used in the lemma.

Using the $\mathcal{H}_2$ inner product, the respective $\mathcal{H}_2$ norm is defined as follows.

Definition 4.3. Let $\boldsymbol{G}(s)$ be an $\mathcal{H}_2$ function, then its $\mathcal{H}_2$ norm is defined as

$$\|\boldsymbol{G}\|_{\mathcal{H}_2} = \sqrt{\langle \boldsymbol{G}, \boldsymbol{G} \rangle_{\mathcal{H}_2}} = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \operatorname{trace}\left[\boldsymbol{G}^*(\imath\omega)\boldsymbol{G}(\imath\omega)\right] \mathrm{d}\omega\right)^{\frac{1}{2}} \tag{4.32}$$

$$= \left(\int_0^{\infty} \operatorname{trace}\left[\boldsymbol{G}^*(t)\boldsymbol{G}(t)\right] \mathrm{d}t\right)^{\frac{1}{2}}. \tag{4.33}$$

It is obvious by Lemmata 4.1, 4.3 and 4.7, that also the norm of a system can be computed by its Gramian or based on its pole-residue formulation. This is stated next for completeness.

Corollary 4.9. *Let $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}$ be an $\mathcal{H}_2$ function, then its $\mathcal{H}_2$ norm is given by*

$$\|\boldsymbol{G}\|_{\mathcal{H}_2}^2 = \operatorname{trace}\left(\mathbf{B}^*\mathbf{Q}\mathbf{B}\right) = \operatorname{trace}\left(\mathbf{C}\mathbf{P}\mathbf{C}^*\right), \tag{4.34}$$

where $\mathbf{P}$ and $\mathbf{Q}$ satisfy the two dual Lyapunov equations (1.11) and (1.12). Furthermore, if $\mathbf{E}^{-1}\mathbf{A}$ is diagonalizable, let $\boldsymbol{G}(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i\mathbf{B}_i}{s-\lambda_i}$ denote the pole-residue representation, *then the $\mathcal{H}_2$ norm is also given by*

$$\|\boldsymbol{G}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^{k} \operatorname{trace}\left[\mathbf{C}_i^* \boldsymbol{G}(-\overline{\lambda}_i)\mathbf{B}_i^*\right]. \tag{4.35}$$

If otherwise $\boldsymbol{G}(s)$ admits the state-space realization given by $\mathbf{E}=\mathbf{I}$ identity, and by

$$\mathbf{A} = \begin{bmatrix} \lambda\mathbf{I} & & & \\ -\mathbf{I} & \ddots & & \\ & \ddots & \ddots & \\ & & -\mathbf{I} & \lambda\mathbf{I} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_k \end{bmatrix}, \quad \mathbf{C} = [\ \mathbf{C}_1 \ \ \dots \ \ \mathbf{C}_k\ ], \tag{4.36}$$

where $\mathbf{I}$ denotes the $\widetilde{m}\times\widetilde{m}$ identity matrix and $\mathbf{B}_i \in \mathbb{C}^{\widetilde{m}\times m}$ and $\mathbf{C}_i \in \mathbb{C}^{p\times\widetilde{m}}$, then its $\mathcal{H}_2$ norm is given by

$$\|\boldsymbol{G}\|_{\mathcal{H}_2}^2 = \sum_{i=1}^{k}\sum_{j=i}^{k} \operatorname{trace}\left[\mathbf{C}_j^* \boldsymbol{G}^{(j-i)}(-\overline{\lambda})\mathbf{B}_i^*\right] \frac{1}{(j-i)!}. \tag{4.37}$$

Proof. The proof is a direct consequence of Lemmata 4.1, 4.3 and 4.7, by noting that the system norm is real valued and hence the final complex conjugation in (4.11) and (4.29) can be omitted. □

4.1.2 $\mathcal{H}_2$ Error in Model Order Reduction

The previous section discusses the $\mathcal{H}_2$ inner product on a general level, i. e. for two anonymous models. Indeed, the $\mathcal{H}_2$ norm is of course used in this work to measure *the approximation error* in model order reduction. In order to clarify this, the formulations of Section 4.1.1 are now reviewed for all relevant quantities. To this end, first of all some notation is introduced, which will be frequently used in the remainder: given the original model $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}$ and its reduction $\boldsymbol{G}_r(s) = \mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r$, then $\mathbf{X}$ and $\mathbf{Y}$ are defined as the solutions of

$$\mathbf{A}\mathbf{X}\mathbf{E}_r^* + \mathbf{E}\mathbf{X}\mathbf{A}_r^* + \mathbf{B}\mathbf{B}_r^* = \mathbf{0}, \tag{4.38}$$

$$\mathbf{A}^*\mathbf{Y}\mathbf{E}_r + \mathbf{E}^*\mathbf{Y}\mathbf{A}_r + \mathbf{C}^*\mathbf{C}_r = \mathbf{0}. \tag{4.39}$$

Furthermore, suppose

$$\mathbf{A}_r\mathbf{P}_r\mathbf{E}_r^* + \mathbf{E}_r\mathbf{P}_r\mathbf{A}_r^* + \mathbf{B}_r\mathbf{B}_r^* = \mathbf{0}, \tag{4.40}$$

$$\mathbf{A}_r^*\mathbf{Q}_r\mathbf{E}_r + \mathbf{E}_r^*\mathbf{Q}_r\mathbf{A}_r + \mathbf{C}_r^*\mathbf{C}_r = \mathbf{0}, \tag{4.41}$$

then $\mathbf{P}_r$ and $\mathbf{E}_r^T\mathbf{Q}_r\mathbf{E}_r$ denote the Controllability and Observability Gramians, respectively, of the reduced model. We are now ready to describe the $\mathcal{H}_2$ error in model order reduction.

Corollary 4.10. *Given the original model $\boldsymbol{G}(s)=\mathbf{C}\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}$, assume a reduced model $\boldsymbol{G}_r(s)=\mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r$ that admits a state-space realization with exclusively real matrices. Then the $\mathcal{H}_2$ norm $\|\boldsymbol{G}_e\|_{\mathcal{H}_2}$ of the error $\boldsymbol{G}_e(s)=\boldsymbol{G}-\boldsymbol{G}_r$ is given by*

$$\|\boldsymbol{G}_e\|_{\mathcal{H}_2}^2 = \|\boldsymbol{G}\|_{\mathcal{H}_2}^2 + \|\boldsymbol{G}_r\|_{\mathcal{H}_2}^2 - 2\left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2}, \tag{4.42}$$

where $\left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2}$ can be computed by

$$\left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2} = \mathrm{trace}\left(\mathbf{B}^*\mathbf{Y}\mathbf{B}_r\right) = \mathrm{trace}\left(\mathbf{C}\mathbf{X}\mathbf{C}_r^*\right), \tag{4.43}$$

for which $\mathbf{X}$ and $\mathbf{Y}$ satisfy (4.38) and (4.39). Assume that $\boldsymbol{G}_r(s)$ admits the pole-residue representation $\boldsymbol{G}_r(s)=\sum_{i=1}^k \frac{\mathbf{C}_i\mathbf{B}_i}{s-\lambda_i}$, then $\left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2}$ can also be computed by

$$\left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2} = \sum_{i=1}^k \mathrm{trace}\left[\mathbf{C}_i^*\boldsymbol{G}(-\overline{\lambda}_i)\mathbf{B}_i^*\right]. \tag{4.44}$$

Proof. Note that $\left\langle \boldsymbol{G}_r,\boldsymbol{G}\right\rangle_{\mathcal{H}_2} = \left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2}^*$, which equals $\left\langle \boldsymbol{G},\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2}$ if both $\boldsymbol{G}(s)$ and $\boldsymbol{G}_r(s)$ admit state-space realizations with real matrices. Equation (4.42) then follows from evaluating $\|\boldsymbol{G}_e\|_{\mathcal{H}_2}^2 = \left\langle \boldsymbol{G}-\boldsymbol{G}_r,\boldsymbol{G}-\boldsymbol{G}_r\right\rangle_{\mathcal{H}_2}$ and the rest is a direct application of Lemmata 4.1 and 4.3. □

Remark 4.11. It should be noted that (4.44) is a convenient representation of the $\mathcal{H}_2$ inner product of original and reduced model, as the original model has to be evaluated solely at the mirror images of the reduced poles—which is often feasible even in a large-scale setting.

Not only the definition of the $\mathcal{H}_2$ norm based on Gramians (4.34), but also its application to the error model in Corollary 4.10 can be found in many textbooks, see e. g. [8]. The expression of the $\mathcal{H}_2$ norm based on the pole-residue formulation is due to Antoulas, and a proof in the SISO case can be found e. g. in the book [8]. It should be noted that the MIMO case is stated on p. 145 therein without proof, however with a small typo: Equation (4.35) is printed with $\boldsymbol{G}^*$, instead of the correct $\boldsymbol{G}$, which could yield complex values for system norms. See also [197], where an alternative formulation of this result is presented, which is based on an element-wise access of the residual matrices $\mathbf{B}_i$ and $\mathbf{C}_i$. An alternative proof of the $\mathcal{H}_2$ inner product based on poles and residues, i. e. of

Lemmata 4.3 and 4.7, can be found in [92], but for the SISO case only; the MIMO case is used e. g. in [21]. By contrast, Lemma 4.5 and all the proofs given here (which are based on explicit formulae for the solution of Sylvester equations) appear to be new.

4.2 $\mathcal{H}_2$ Optimal Model Order Reduction

The results of the previous section provide convenient analytical and numerical access to the $\mathcal{H}_2$ norm, which paves the way to various contributions on minimizing the $\mathcal{H}_2$ error norm in MOR. This fact is reflected in the large literature that is available on this subject, and which this section tries to review.

4.2.1 The Problem

For the statement of the problem, we require additional notation: let $\dim(\boldsymbol{G}_r)$ denote the McMillan degree of $\boldsymbol{G}_r(s)$, or equivalently, the order of any minimal realization of $\boldsymbol{G}_r(s)$ in state-space; for details on this and a definition of the McMillan degree, please refer to e. g. [222]. The general problem of $\mathcal{H}_2$ optimal MOR then may be defined as follows.

Problem 4.1. Given the original model $\boldsymbol{G}(s) = \mathbf{C}\,(s\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}$ and a reduced order n, we are searching for the $\mathcal{H}_2$ optimal reduced model $\boldsymbol{G}_r(s)$ with $\dim(\boldsymbol{G}_r) = n$, which satisfies

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2} = \min_{\dim(\boldsymbol{H}_r)=n} \|\boldsymbol{G} - \boldsymbol{H}_r\|_{\mathcal{H}_2}\,. \tag{4.45}$$

Let J denote the approximation error measured in the $\mathcal{H}_2$ norm,

$$J = \|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2}\,, \tag{4.46}$$

then the problem is to minimize J in the set of all asymptotically stable systems of fixed order. This, in fact, is a non-convex optimization problem, which has “no apparent explicit closed-form solution” [203]. Only recently, Problem 4.1 was tackled for the reduced orders $n = 1$ in [2], and $n = 2$ in [1]. It however seems like the computations involved are not suitable in a large-scale setting, such that these contributions unfortunately are merely of theoretical interest.

As the global solution of Problem 4.1 is generally not possible, one instead considers first-order necessary conditions, which $\boldsymbol{G}_r(s)$ has to fulfil in order to minimize J. These conditions can then be used to construct at least locally $\mathcal{H}_2$ optimal reduced models. The next subsection first of all reviews different formulations of these conditions.

4.2.2 First-Order Necessary Conditions

Assume that we have some valid parametrization of the reduced model, i. e. a set of quantities that uniquely defines the transfer behaviour of $\boldsymbol{G}_r(s)$. Then it is obvious that the gradient of J with respect to these parameters must vanish in the optimum—and that this is irrespective of the kind of parametrization we had chosen. The analysis of the vanishing gradient then leads to so-called *first-order necessary conditions* that $\boldsymbol{G}_r(s)$ has to satisfy, and which accordingly may be exploited to generate optimal reduced models. But it should be stressed that, due to the nature of gradient based optimization, one certainly can enforce merely local optimality instead global one.

The remaining question is still how to parametrize the reduced model, for which one possibility would be to take e. g. the poles and residues—or another one to directly use the matrices $\mathbf{E}_r$, $\mathbf{A}_r$, $\mathbf{B}_r$ and $\mathbf{C}_r$. The derivation of J with respect to these different parameters thus triggers also diverse formulations of the first-order necessary conditions; the first one, which is based on poles and residues, is presented in the next theorem.

Theorem 4.12 ([92])**.** *Given $\boldsymbol{G}(s)$, let $\boldsymbol{G}_r(s)$ be a local minimizer of J, and suppose that $\boldsymbol{G}_r(s)$ admits the pole-residue formulation $\boldsymbol{G}_r(s) = \sum_{i=1}^{n} \frac{\mathbf{c}_i \mathbf{b}_i}{s-\lambda_{r,i}}$, with the residue directions $\mathbf{b}_i^* \in \mathbb{C}^m$ and $\mathbf{c}_i \in \mathbb{C}^p$. Then, for $i = 1, \ldots, n$,*

$$\boldsymbol{G}(-\overline{\lambda}_{r,i})\mathbf{b}_i^* = \boldsymbol{G}_r(-\overline{\lambda}_{r,i})\mathbf{b}_i^*, \tag{4.47}$$

$$\mathbf{c}_i^* \boldsymbol{G}(-\overline{\lambda}_{r,i}) = \mathbf{c}_i^* \boldsymbol{G}_r(-\overline{\lambda}_{r,i}), \tag{4.48}$$

$$\mathbf{c}_i^* \boldsymbol{G}'(-\overline{\lambda}_{r,i})\mathbf{b}_i^* = \mathbf{c}_i^* \boldsymbol{G}_r'(-\overline{\lambda}_{r,i})\mathbf{b}_i^*, \tag{4.49}$$

where $\boldsymbol{G}'(s)$ denotes the first derivative of $\boldsymbol{G}(s)$ with respect to s.

Remark 4.13. Theorem 4.12 is valid only if $\mathbf{E}_r^{-1}\mathbf{A}_r$ is diagonalizable. If this is not the case, the above conditions can be generalized by considering also higher derivatives of $\boldsymbol{G}(s)$ and $\boldsymbol{G}_r(s)$. However, this would require cumbersome notation, which is omitted for brevity; the interested reader is instead referred to [197].

The tangential directions $\mathbf{b}_i$ and $\mathbf{c}_i$ degenerate to scalars in the SISO case, and hence can be omitted. This is clarified in the next corollary.

Corollary 4.14. *Given the SISO model $\boldsymbol{G}(s)$, let $\boldsymbol{G}_r(s) = \mathbf{c}_r \left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1} \mathbf{b}_r$ be a local minimizer of J, and suppose that $\mathbf{E}_r^{-1}\mathbf{A}_r$ can be diagonalized to* $\operatorname{diag}[\lambda_1, \ldots, \lambda_n]$*. Then* *for* $i=1,\dots,n$,

$$\boldsymbol{G}(-\overline{\lambda}_{r,i}) = \boldsymbol{G}_r(-\overline{\lambda}_{r,i}) \tag{4.50}$$
$$\boldsymbol{G}'(-\overline{\lambda}_{r,i}) = \boldsymbol{G}'_r(-\overline{\lambda}_{r,i}). \tag{4.51}$$

Theorem 4.12 and Corollary 4.14 show that an $\mathcal{H}_2$ optimal reduced model has to interpolate the original model at the mirror images of its poles with respect to the imaginary axis. In the MIMO case, additionally, the tangential directions for interpolation are defined by the residue directions of the reduced model. Equations (4.47)–(4.49) and (4.50)–(4.51) are usually referred to as "interpolatory conditions" or "Meier-Luenberger conditions". The interpolation property renders these conditions appealing in the context of MOR by Krylov subspaces, which will be discussed in Section 4.2.3.

Owing to the duality of Krylov subspaces and Sylvester equations, it naturally follows that there must also be a version of the first order necessary conditions that is based on Sylvester equations. This formulation indeed arises by the derivation of J with respect to $\mathbf{E}_r$, $\mathbf{A}_r$, $\mathbf{B}_r$ and $\mathbf{C}_r$, and is usually denoted as the "Wilson conditions"; it is reviewed in the following theorem.

Theorem 4.15. *Given* $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}$, *let* $\boldsymbol{G}_r(s) = \mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r$ *be a local minimizer of* J*, where all matrices in both state-space realizations are assumed to be real, then*

$$\mathbf{Q}_r\mathbf{E}_r\mathbf{P}_r - \mathbf{Y}^T\mathbf{E}\mathbf{X} = \mathbf{0}, \tag{4.52}$$
$$\mathbf{Q}_r\mathbf{B}_r - \mathbf{Y}^T\mathbf{B} = \mathbf{0}, \tag{4.53}$$
$$\mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X} = \mathbf{0}. \tag{4.54}$$

where $\mathbf{X}$*,* $\mathbf{Y}$*,* $\mathbf{P}_r$ *and* $\mathbf{Q}_r$ *satisfy (4.38)–(4.41).*

Proof. The proof for $\mathbf{E}=\mathbf{I}$ can be found in the original work of Wilson [203]. The general case is straightforward to show, by noting that $\mathbf{Y}$ changes to $\mathbf{E}^T\mathbf{Y}\mathbf{E}_r$ and that $\mathbf{Q}_r$ changes to $\mathbf{E}_r^T\mathbf{Q}_r\mathbf{E}_r$. □

It should be noted that one can deduce from (4.52)–(4.54) appropriate projection matrices by $\mathbf{V}=\mathbf{X}\mathbf{P}_r^{-1}$ and $\mathbf{W}=\mathbf{Y}\mathbf{Q}_r^{-1}$, that would indeed yield the corresponding reduced model, however, only if the Wilson conditions are already satisfied. These projection matrices $\mathbf{V}$ and $\mathbf{W}$ actually span rational Krylov subspaces. To uncover this fact, substitute $\mathbf{X}=\mathbf{V}\mathbf{P}_r$ and $\mathbf{Y}=\mathbf{W}\mathbf{Q}_r$ in (4.38) and (4.39), then it follows that

$$\mathbf{A}\mathbf{V}-\mathbf{E}\mathbf{V}\left(-\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\right)=-\mathbf{B}\left(\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\right), \tag{4.55}$$

$$\mathbf{A}^*\mathbf{W}-\mathbf{E}^*\mathbf{W}\left(-\mathbf{Q}_r\mathbf{A}_r\mathbf{E}_r^{-1}\mathbf{Q}_r^{-1}\right)=-\mathbf{C}^*\left(\mathbf{C}_r\mathbf{E}_r^{-1}\mathbf{Q}_r^{-1}\right). \tag{4.56}$$

Using Theorem 2.4, this shows that the projection matrices $\mathbf{V}$ and $\mathbf{W}$ that satisfy the Wilson conditions in fact span rational Krylov subspaces, which guarantee interpolation at the mirror images of the reduced poles. This indicates that the Wilson conditions and the Meier-Luenberger conditions in Theorem 4.12 are actually equivalent.

Before going into this, a third formulation of the first-order necessary conditions is reviewed, which is denoted as the "Hyland-Bernstein conditions". The following theorem, however, presents a slightly different form of the original formulation in [104], in order to better suit the notation of this work. To the best of the author's knowledge, this has not been presented before.

Theorem 4.16. *Given $\boldsymbol{G}(s)=\mathbf{C}\left(s\mathbf{E}-\mathbf{A}\right)^{-1}\mathbf{B}$, let $\boldsymbol{G}_r(s)=\mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r$ be a local minimizer of J, then there exist projection matrices $\mathbf{V}$, $\mathbf{W}\in\mathbb{R}^{N\times n}$, such that $\boldsymbol{G}_r(s)$ is given by projection as in (1.7), and such that*

$$\mathbf{A}\widehat{\mathbf{P}}\mathbf{E}^T+\mathbf{E}\widehat{\mathbf{P}}\mathbf{A}^T+\mathbf{B}\mathbf{B}^T=\mathbf{B}_\perp\mathbf{B}_\perp^T, \tag{4.57}$$

$$\mathbf{A}^T\widehat{\mathbf{Q}}\mathbf{E}+\mathbf{E}^T\widehat{\mathbf{Q}}\mathbf{A}+\mathbf{C}^T\mathbf{C}=\mathbf{C}_\perp^T\mathbf{C}_\perp, \tag{4.58}$$

where $\widehat{\mathbf{P}}=\mathbf{V}\mathbf{P}_r\mathbf{V}^T$, $\widehat{\mathbf{Q}}=\mathbf{W}\mathbf{Q}_r\mathbf{W}^T$, $\mathbf{B}_\perp=\mathbf{B}-\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$ and $\mathbf{C}_\perp=\mathbf{C}-\mathbf{C}_r\mathbf{E}_r^{-1}\mathbf{W}^*\mathbf{E}$, and where $\mathbf{P}_r$ and $\mathbf{Q}_r$ are defined by (4.40) and (4.41).

Proof. The proof is derived from the main theorem in [104]. It should be noted that the notation of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{Q}}$ is interchanged in [104]. Further note that $\mathbf{Q}$ changes to $\mathbf{E}^T\mathbf{Q}\mathbf{E}$ for $\mathbf{E}\neq\mathbf{I}$ and hence, $\widehat{\mathbf{Q}}$ changes to $\mathbf{E}^T\widehat{\mathbf{Q}}\mathbf{E}$. Consider $\widehat{\mathbf{P}}\mathbf{E}^T\widehat{\mathbf{Q}}\mathbf{E}=\boldsymbol{G}^T\mathbf{M}\boldsymbol{\Gamma}$ with $\boldsymbol{G}^T=\mathbf{V}$, $\mathbf{M}=\mathbf{P}_r\mathbf{E}_r^T\mathbf{Q}_r\mathbf{E}_r$ and $\boldsymbol{\Gamma}=\mathbf{E}_r^{-1}\mathbf{W}^T\mathbf{E}$. The eigenvalues of $\mathbf{M}$ are the squared Hankel singular values of the reduced model and hence positive. Therefore, $\boldsymbol{G}^T\mathbf{M}\boldsymbol{\Gamma}$ is an admissible factorization required in [104] and clearly, $\boldsymbol{\Gamma}\boldsymbol{G}^T=\mathbf{I}$, and $\boldsymbol{\tau}=\boldsymbol{G}^T\boldsymbol{\Gamma}$ defines a projector. Then $\mathbf{I}=\mathbf{E}_r$, $\boldsymbol{\Gamma}\mathbf{E}^{-1}\mathbf{A}\boldsymbol{G}^T=\mathbf{A}_r$, $\boldsymbol{\Gamma}\mathbf{E}^{-1}\mathbf{B}=\mathbf{B}_r$, and $\mathbf{C}\boldsymbol{G}^T=\mathbf{C}_r$ define the reduced model. Obviously, $\mathrm{rank}(\widehat{\mathbf{P}})=\mathrm{rank}(\widehat{\mathbf{Q}})=\mathrm{rank}(\widehat{\mathbf{P}}\widehat{\mathbf{Q}})=n$ and finally, using Proposition 2.4 from [104], the statement follows with $\mathbf{B}_\perp=\mathbf{E}(\mathbf{I}-\boldsymbol{\tau})\mathbf{E}^{-1}\mathbf{B}$ and $\mathbf{C}_\perp=\mathbf{C}(\mathbf{I}-\boldsymbol{\tau})$. □

We have now reviewed three different formulations of necessary conditions for local $\mathcal{H}_2$ optimality, so it is important to analyse their difference; i. e. is there a formulation that is more powerful than the other? As all of them are derived from setting some gradients to zero, this is not the case, and in fact, they are all equivalent to each other; this is reviewed in the next proposition.

Proposition 4.17. *The first order necessary conditions of Theorem 4.12, of Theorem 4.15 and of Theorem 4.16 are equivalent to each other.*

Proof. The proof in the SISO case for simple poles is given in [92], the MIMO case with multiple poles can be found in [41]. See also [188], where the discrete time case is discussed. □

Both the Wilson and the Hyland-Bernstein conditions are theoretically more convenient, as neither do they alter for multiple inputs, nor do they change for reduced eigenvalues with higher orders; see also the discussion in [188]. By contrast, the Meier-Luenberger conditions pave the way to a fixed point iteration, which is based on rational Krylov subspaces, and which can generate locally optimal reduced models. This is reviewed next. Finally, it should be stressed that the conditions are of first-order type and hence not sufficient for locally minimal error; Kammler [115] presented examples, where the conditions are indeed satisfied also by local maxima.

4.2.3 Iterative Rational Krylov Algorithm (IRKA)

The Meier-Luenberger conditions suggest to use the mirror images of the reduced eigenvalues as shifts for rational Krylov subspaces, and then project again. Recursively applying this idea yields the fixed point iteration due to Gugercin et al. [92], which is known as the *iterative rational Krylov algorithm* (IRKA). The famous IRKA algorithm can still be regarded as state-of-the-art in $\mathcal{H}_2$ optimal MOR, which is not only due to its striking simplicity. Its basic procedure is shown in Algorithm 4.1.

It should be noted that Algorithm 4.1 merely displays the basic concept of IRKA, instead of a numerical implementation. It is e. g. left undefined, how to measure convergence of IRKA in Step 8, as there are different possibilities. One would be to compute the difference in the vector $[\lambda_1, \ldots, \lambda_n]$ between the current and the preceding iteration step. Then, however, the eigenvalues λ_i of two successive iterations have to be matched appropriately.

We know from Lemma 1.2 that the bases of $\mathbf{V}$ and $\mathbf{W}$ are irrelevant and only the subspaces spanned by their columns define the reduced model. Therefore, one would usually compute real bases of $\mathbf{V}$ and $\mathbf{W}$ in Steps 3 and 4, respectively, which is always

Algorithm 4.1 Iterative rational Krylov algorithm (IRKA)

```
Input: $\mathbf{E}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and reduced order $n$
Output: locally $\mathcal{H}_2$ optimal reduced model $\boldsymbol{G}_r(s) = \mathbf{C}_r\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r$ of order $n$
1: Make initial choice of the set $\{s_1, \ldots, s_n\}$, that is closed under conjugation; select
   $\mathbf{b}_i^* \in \mathbb{C}^m$ and $\mathbf{c}_i \in \mathbb{C}^p$, that satisfy $\mathbf{b}_i = \overline{\mathbf{b}}_j$ and $\mathbf{c}_i = \overline{\mathbf{c}}_j$ if $s_i = \overline{s}_j$.
2: repeat
3:     $\mathbf{V} = \left[(\mathbf{A} - s_1\mathbf{E})^{-1}\mathbf{B}\mathbf{b}_1^*,\ \ldots,\ (\mathbf{A} - s_n\mathbf{E})^{-1}\mathbf{B}\mathbf{b}_n^*\right]$
4:     $\mathbf{W} = \left[\left(\mathbf{A}^T - \overline{s}_1\mathbf{E}^T\right)^{-1}\mathbf{C}^T\mathbf{c}_1,\ \ldots,\ \left(\mathbf{A}^T - \overline{s}_n\mathbf{E}^T\right)^{-1}\mathbf{C}^T\mathbf{c}_n\right]$
5:     $\mathbf{E}_r = \mathbf{W}^*\mathbf{E}\mathbf{V}$, $\mathbf{A}_r = \mathbf{W}^*\mathbf{A}\mathbf{V}$, $\mathbf{B}_r = \mathbf{W}^*\mathbf{B}$ and $\mathbf{C}_r = \mathbf{C}\mathbf{V}$
6:     Compute eigenvalue decomposition $\mathbf{E}_r^{-1}\mathbf{A}_r = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$, with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$
7:     Assign $s_i = -\overline{\lambda}_i$, $[\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*]^* = \mathbf{U}^{-1}\mathbf{E}_r^{-1}\mathbf{B}_r$ and $[\mathbf{c}_1, \ldots, \mathbf{c}_n] = \mathbf{C}_r\mathbf{U}$
8: until converged
```

possible as the sets s_i, $\mathbf{b}_i$ and $\mathbf{c}_i$ are closed under conjugation. This reduces complex arithmetic and hence improves numerical conditioning.

From Theorem 2.4 it is clear that the computation of bases of rational Krylov subspaces in Steps 3 and 4 of Algorithm 4.1 is equivalent to the solution of particular Sylvester equations. To this end, consider the Sylvester equations

$$\mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\left(-\mathbf{A}_r^*\mathbf{E}_r^{-*}\right) = \mathbf{B}\left(\mathbf{B}_r^*\mathbf{E}_r^{-*}\right), \tag{4.59}$$

$$\mathbf{A}^T\mathbf{W} - \mathbf{E}^T\mathbf{W}\left(-\mathbf{E}_r^{-1}\mathbf{A}_r\right) = \mathbf{C}^T\mathbf{C}_r, \tag{4.60}$$

whose solutions span the same subspaces as $\mathbf{V}$ and $\mathbf{W}$ in Algorithm 4.1. Equations (4.59) and (4.60) thus serve as a substitute for Steps 3, 4, 6 and 7 in IRKA. It should be stressed, that this substitution would not change the basic concept of IRKA—it only represents a different numerical implementation. Such an approach was denoted in [217] as *two-sided iteration algorithm* (TSIA), and the idea has also appeared in [187]. Owing to its conceptual equivalence, the acronym TSIA will not be used in the remainder. Nevertheless, using (4.59) and (4.60) instead of Steps 3, 4, 6 and 7 in IRKA may have numerical advantages, especially if reduced eigenvalues of higher orders occur; see e. g. the discussion in [188].

There is an important property of the IRKA algorithm: assume that IRKA has converged, then the resulting reduced model is guaranteed to have locally minimal $\mathcal{H}_2$ error—even though IRKA is based on the first-order necessary conditions which also local maxima satisfy. This property follows from the particular fixed point iteration of IRKA, see e. g. [23]. Convergence of IRKA, however, cannot be guaranteed in general. Only for certain system classes convergence of IRKA can be ensured, cf. [65]. In general, it may happen that IRKA does not converge, but which is not a severe problem in practice. However, IRKA sometimes requires unacceptably many iterations for convergence, which gets worse the higher the reduced order is. In such a case, the CURE framework seems to be appropriate, by dividing the problem into smaller parts with better convergence behaviour. This approach is denoted as “CUREd IRKA” in [148], where also numerical examples are given.

4.2.4 Overview on $\mathcal{H}_2$ Optimal Model Order Reduction

The interpolatory conditions for $\mathcal{H}_2$ optimality were originally derived in the context of network synthesis by Aigrain and Williams [5], and then rediscovered in the control literature by Meier and Luenberger [137]. Since then, $\mathcal{H}_2$ optimal approximation was addressed by many researchers. Miller [140] formulated the Aigrain-Williams conditions also for sampled data, whereas Riggs and Edgar [161] stated the Meier-Luenberger conditions corresponding to local optimality in a finite time interval of the impulse response. They were able to also treat multiple inputs, but only single outputs. Maybe the first reference of the interpolatory conditions in the MIMO case was by Krajewski et al. [119], who presented a block version. Maybe the most famous work on that topic is due to Gugercin et al. [92] who not only introduced the IRKA algorithm, but also presented a new proof of the Meier-Luenberger conditions based on structured optimality. The proof in the MIMO case can be found in the work of Van Dooren et al. [187], whereas the generalization to multiple poles was presented by Vossen et al. [197] and Van Dooren et al. [188]. An IRKA-like algorithm that minimizes the $\mathcal{H}_{2,\alpha}$ norm of the error, where $\alpha \in \mathbb{R}$ is an additional stability margin, can be found in the work of Vossen [196].

Wilson [203] derived his version of the optimality conditions from setting to zero the gradient of J with respect to $\mathbf{A}_r$, $\mathbf{B}_r$ and $\mathbf{C}_r$. Hirzinger and Kreisselmeier [101] computed these gradients also for different input functions, such as a unit step or the impulse response of a linear shaping filter. Wilson and Mishra [205] generalized the conditions to inputs of the form $u(t) = \frac{t^k}{k!}$. A more recent proof of the Wilson conditions was presented by Van Dooren et al. [187].

Hyland and Bernstein [104] showed the equivalence of their version of the necessary optimality conditions to the Wilson conditions. The equivalence of all three formulations was proven by Gugercin et al. [92] in the SISO case, and in the most general case by Bunse-Gerstner et al. [41] and Van Dooren et al. [188].

The idea of an IRKA-like fixed point iteration was independently introduced by

Lepschy et al. [126] and Lucas [130, 131] for SISO systems. A block-MIMO version was suggested by Krajewski et al. [120], which then was slightly modified by Ferrante et al. [64] in order to improve convergence properties—but with higher numerical effort. All of these references, however, are based on data of the transfer function $\boldsymbol{G}(s)$ (i. e. poles, residues and characteristic polynomials), which is not desirable in a large-scale setting. The "breakthrough" for large-scale systems was due to Gugercin et al. [92] who formulated the idea of a fixed point iteration in a state-space setting using projections onto Krylov subspaces. Van Dooren et al. [188] discussed the occurrence of higher-order poles, which would in fact lead to a slight modification of the IRKA algorithm. A discrete version of IRKA, the *MIMO Iterative Rational Interpolation Algorithm* (MIRIAm), was introduced by Bunse-Gerstner et al. [42]. Although IRKA might not converge in general, Flagg et al. [65] proved that IRKA is guaranteed to locally converge at least for "state-space-symmetric systems". To overcome convergence problems, Krajewski and Viaro [121] suggested a modification of IRKA with guaranteed convergence. If the original IRKA, however, converges—which is not known a priori—the proposed modification seems to lead to a higher number of iterations for convergence. The modified IRKA of Krajewski and Viaro therefore appears to be a reasonable choice, only if IRKA does not converge in the first place. A quite similar modification of IRKA was suggested by Wolf et al. [208], which is based on the conditions for $\mathcal{H}_2$ pseudo-optimality in Section 4.3.3, but without rigorous convergence analysis. Various derivatives of the IRKA algorithm are available from different authors: Gugercin [89] suggested an *iterative SVD-rational Krylov based model reduction method* (ISRK) that, on the one hand, guarantees stability preservation, but on the other hand, requires a Gramian; in a similar way, Gugercin et al. [94] showed how IRKA can preserve a port-Hamiltonian structure in the reduced model; a frequency weighted version of IRKA was presented by Anic et al. [7]; Poussot-Vassal and Vuillemin [157] combined IRKA with the preservation of some user-defined eigenvalues, which can be desirable in the aeronautic field; an IRKA-like algorithm based on measurements in the Loewner framework is due to Beattie and Gugercin [24], and finally, an extension to time-varying discrete-time systems over finite horizons was suggested by Melchior et al. [138], which, however, is not yet applicable for large-scale models.

Instead of recursively mirroring the reduced eigenvalues in the IRKA algorithm, it is also possible to employ gradients and Hessians for optimization. Bryson and Carrier [40] proposed a Newton-Raphson algorithm, which is based on data of the transfer function $\boldsymbol{G}(s)$, but, as already mentioned above, this is not applicable in a large-scale setting. This also seems to be the case for the optimization on a Stiefel manifold suggested by

Yan and Lam [219]. Xu and Zeng [216] transformed this idea into an optimization on a Grassmann Manifold (and thereby facilitating the computation of the gradient), which was enhanced by Zeng [220] with oblique projections. However, it remains unclear if these approaches are suitable for large-scale problems. Beattie and Gugercin [22] proposed a Newton algorithm which they later generalized for MIMO models and in addition, which they also equipped with a trust region algorithm, cf. [21]. Panzer et al. [149] then addressed numerical drawbacks of this approach by employing the concept of $\mathcal{H}_2$ pseudo-optimality and the CURE framework, all of which was improved by Panzer in his thesis [148].

A gradient-based approach has some advantages over the IRKA algorithm. Probably the main drawbacks of IRKA are that, on the one hand, it might not converge, and that, on the other hand, the whole algorithm must be restarted with a higher reduced order n, if the approximation was not sufficient in the first place. By contrast, the cumulative approach in [148, 149] can not only guarantee convergence, but also adaptively choose the reduced order—without the need of entirely restarting the algorithm. An important ingredient in this approach is the concept of "$\mathcal{H}_2$ pseudo-optimality". The proper definition and detailed analysis of this concept now follows and it is the main result of this thesis.

4.3 $\mathcal{H}_2$ Pseudo-Optimality

The basic idea behind $\mathcal{H}_2$ pseudo-optimality is to divide the set of reduced models with fixed order into disjoint subsets. The respective global minimizer in any of these subsets will be marked with the prefix "pseudo". As any locally $\mathcal{H}_2$ optimal reduced model necessarily is also the $\mathcal{H}_2$ pseudo-optimal reduced model in its respective subset, the concept of $\mathcal{H}_2$ pseudo-optimality may be exploited for optimization. This is discussed in the end of this section. First of all, the notation "pseudo" is defined.

Definition 4.4. Given a subset $\mathcal{G} \subset \mathcal{H}_2$ of reduced models with fixed order n. Then the reduced model $\boldsymbol{G}_r(s)$ that satisfies

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2} = \min_{\boldsymbol{H}_r \in \mathcal{G}} \|\boldsymbol{G} - \boldsymbol{H}_r\|_{\mathcal{H}_2} \tag{4.61}$$

is called "$\mathcal{H}_2$ pseudo-optimal" (with respect to $\mathcal{G}$).

It should be stressed, that the $\mathcal{H}_2$ pseudo-optimal reduced model is the *global* minimizer of the $\mathcal{H}_2$ error norm in its respective subset $\mathcal{G}$, and that basically no assumptions are made on the subset $\mathcal{G}$. It is therefore possible to let an $\mathcal{H}_2$ pseudo-optimal reduced model be an arbitrarily bad approximation of $\boldsymbol{G}(s)$, by just selecting an arbitrary bad subset $\mathcal{G}$. The resulting $\mathcal{H}_2$ pseudo-optimal reduced model would then indeed be far from local $\mathcal{H}_2$ optimality. This is the reason, why alternative labels, like e. g. "suboptimality" would be inappropriate as notation, and "pseudo-optimality" is chosen instead.

4.3.1 The Problem

The nature of $\mathcal{H}_2$ pseudo-optimality is illustrated in Figure 4.1. The grey square on the left-hand side represents the set of all reduced models of fixed order n. After slicing the square into infinitesimal strips, we get disjoint subsets $\mathcal{G}$. In each of theses subsets we can identify a unique minimizer of the $\mathcal{H}_2$ error norm: the $\mathcal{H}_2$ pseudo-optimal reduced model, denoted as "×". Following the trace of $\mathcal{H}_2$ pseudo-optimal reduced models in adjoining subsets $\mathcal{G}$, we will eventually come across a locally $\mathcal{H}_2$ optimal reduced model, denoted as "⊗". Due to non-linearity, there may be several local minima, and the best of them is the global minimum, which is denoted as "⊚".

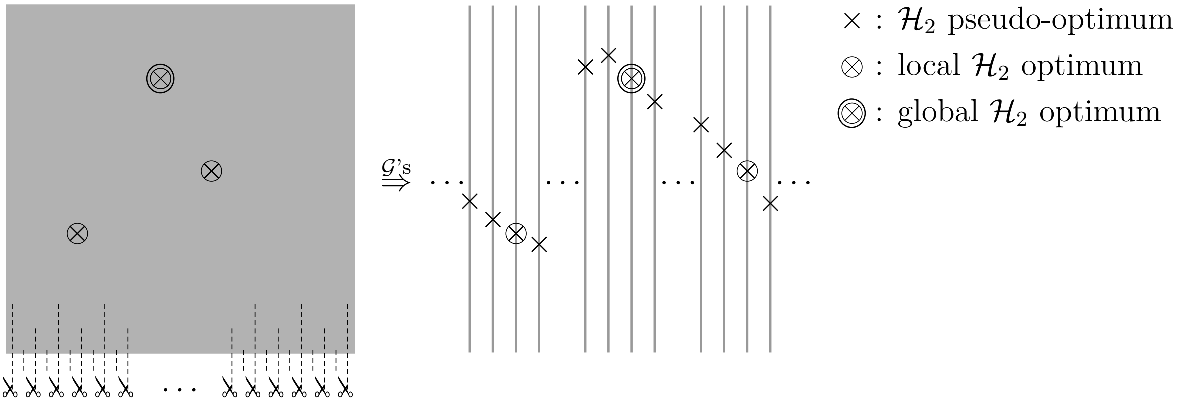

× : $\mathcal{H}_2$ pseudo-optimum
⊗ : local $\mathcal{H}_2$ optimum
⊚ : global $\mathcal{H}_2$ optimum
$\mathcal{G}$'s
. . .

Figure 4.1: Illustration of $\mathcal{H}_2$ pseudo-optimality.

The objective now is to translate Figure 4.1 into a mathematical language. To this end, we first of all have to suitably define the subsets $\mathcal{G}$. For SISO models, fixing the reduced eigenvalues and varying only the residues is a proper choice; this is reviewed in the next lemma.

Lemma 4.18 ([137])**.** *Let $\mathcal{G}(\mathcal{L})$ be the set of all SISO models with fixed eigenvalues $\mathcal{L}=\{\lambda_1, \ldots \lambda_n\}$, where $\lambda_i \neq \lambda_j$, for $i \neq j$, and $\mathrm{Re}(\lambda_i)<0$, $i=1, \ldots n$. Then $\boldsymbol{G}_r(s)$ is the $\mathcal{H}_2$ pseudo-optimal reduced model, i. e. it satisfies*

$$\|\boldsymbol{G}-\boldsymbol{G}_r\|_{\mathcal{H}_2} = \min_{\boldsymbol{H}_r \in \mathcal{G}(\mathcal{L})} \|\boldsymbol{G}-\boldsymbol{H}_r\|_{\mathcal{H}_2}\,, \tag{4.62}$$

if and only if

$$\boldsymbol{G}(-\overline{\lambda}_i) = \boldsymbol{G}_r(-\overline{\lambda}_i), \quad i = 1, \ldots n. \tag{4.63}$$

Lemma 4.18 is sometimes referred to as Gaier's result [74, Theorem 3, p. 86], but it seems like this finding was already known to Walsh [201, Theorem 1, p. 224] in the 1920's. Since then, many researches made use of $\mathcal{H}_2$ pseudo-optimal reduced models (see the discussion in Section 4.3.9), which were often denoted as the "least-squares solution". We do not use this term here, in order to avoid confusion with the independent least-squares model reduction by Gugercin and Antoulas [91]. The available literature on $\mathcal{H}_2$ pseudo-optimality provides various numerical methods to compute the $\mathcal{H}_2$ pseudo-optimal reduced model for a given set $\mathcal{G}(\mathcal{L})$, but all of which directly construct the reduced model, i. e. without determining a connecting projection to the original model. The knowledge of the corresponding $\mathbf{V}$ that spans a suitable Krylov subspace is however essential in the CURE framework. This is the reason, why the available literature is not applicable in the CURE framework, and why the discussion here is required.

The results that are next presented are the following: first of all, the necessary and sufficient interpolatory condition for SISO $\mathcal{H}_2$ pseudo-optimality, as shown in Lemma 4.18, is generalized to multiple inputs and outputs. In order to make the result available in the CURE framework, $\mathcal{H}_2$ pseudo-optimality is subsequently embedded in a projective framework using rational Krylov subspaces. In addition, a numerically efficient algorithm is proposed to compute an $\mathcal{H}_2$ pseudo-optimal reduced model that results from projection. Finally, various new conditions for $\mathcal{H}_2$ pseudo-optimality are presented, which are proven to be equivalent to each other, and which also include counterparts of the Meier-Luenberger, Wilson, and Hyland-Bernstein conditions. These easy-to-evaluate conditions can then be either used a priori, for the construction of $\mathcal{H}_2$ pseudo-optimal reduced models, or a posteriori, to analyse (the distance to) $\mathcal{H}_2$ pseudo-optimality. Preliminary versions of these results are published in [208, 213]. The benefits and possible applications of $\mathcal{H}_2$ pseudo-optimality are finally discussed in Section 4.3.8.

4.3.2 Interpolatory Conditions for $\mathcal{H}_2$ Pseudo-Optimality

The necessary and sufficient interpolatory condition for $\mathcal{H}_2$ pseudo-optimality in the MIMO case is presented in the next theorem; the statement is almost identical to the result of Beattie and Gugercin [24], but it introduces a new and shorter proof, which in turn is inspired by the work of Gugercin et al. [92] for the SISO case.

Theorem 4.19. *Given a fixed set* $\mathcal{L}_\mathcal{B} = \{(\lambda_1, \mathbf{B}_1), \ldots, (\lambda_k, \mathbf{B}_k)\}$ *of pairs* $(\lambda_i, \mathbf{B}_i)$*, where* λ_i *with* $\lambda_i \neq \lambda_j$*,* $i \neq j$*, and* $\operatorname{Re}(\lambda_i) < 0$*,* $i = 1, \ldots k$*, denote eigenvalues, and where* $\mathbf{B}_i \in \mathbb{C}^{m_i \times m}$ *with* $1 \leq m_i \leq m$*,* $i = 1, \ldots k$*, denote input residues, define the set* $\mathcal{G}(\mathcal{L}_\mathcal{B})$ *of all reduced models having the pairs* $(\lambda_i, \mathbf{B}_i)$ *of poles and input residues as follows:*

$$\mathcal{G}(\mathcal{L}_\mathcal{B}) = \left\{ \boldsymbol{H}_r(s) \,\middle|\, \exists\, \mathbf{C}_i \in \mathbb{C}^{p \times m_i} : \boldsymbol{H}_r(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i \mathbf{B}_i}{s - \lambda_i} \right\}. \tag{4.64}$$

Then $\boldsymbol{G}_r(s)$ *is the unique* $\mathcal{H}_2$ *pseudo-optimal reduced model, i. e. it satisfies*

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2} = \min_{\boldsymbol{H}_r \in \mathcal{G}(\mathcal{L}_\mathcal{B})} \|\boldsymbol{G} - \boldsymbol{H}_r\|_{\mathcal{H}_2}, \tag{4.65}$$

if and only if

$$\boldsymbol{G}(-\overline{\lambda}_i)\mathbf{B}_i^* = \boldsymbol{G}_r(-\overline{\lambda}_i)\mathbf{B}_i^*, \quad i = 1, \ldots k. \tag{4.66}$$

Proof. It can be readily verified, that the sum of two models from $\mathcal{G}(\mathcal{L}_\mathcal{B})$ stays in $\mathcal{G}(\mathcal{L}_\mathcal{B})$. Therefore, $\mathcal{G}(\mathcal{L}_\mathcal{B})$ is a closed subspace of $\mathcal{H}_2$, which essentially is the key to the proof of the statement, because we then may apply the Hilbert projection theorem: it states that $\boldsymbol{G}_r(s)$ is the unique minimizer of the $\mathcal{H}_2$ error norm in the subspace $\mathcal{G}(\mathcal{L}_\mathcal{B})$, if and only if

$$\langle \boldsymbol{G} - \boldsymbol{G}_r, \boldsymbol{H}_r \rangle_{\mathcal{H}_2} = 0, \tag{4.67}$$

for all $\boldsymbol{H}_r(s) \in \mathcal{G}(\mathcal{L}_\mathcal{B})$. Obviously, $\boldsymbol{H}_r(s)$ has the pole residue representation $\boldsymbol{H}_r(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i \mathbf{B}_i}{s - \lambda_i}$, where the λ_i's and $\mathbf{B}_i$'s are fixed and the $\mathbf{C}_i$'s are arbitrary. Using this and Lemma 4.3 for the $\mathcal{H}_2$ inner product in (4.67), results in

$$\sum_{i=1}^{k} \operatorname{trace} \left[\mathbf{C}_i^* \left(\boldsymbol{G}(-\overline{\lambda}_i) - \boldsymbol{G}_r(-\overline{\lambda}_i) \right) \mathbf{B}_i^* \right]^* = 0. \tag{4.68}$$

As (4.68) has to hold for any $\mathbf{C}_i$, this is equivalent to (4.66). □

It is obvious, that there exists also a dual version of the above theorem.

Theorem 4.20. *Given a fixed set* $\mathcal{L}_\mathcal{C} = \{(\lambda_1, \mathbf{C}_1), \ldots, (\lambda_k, \mathbf{C}_k)\}$ *of pairs* $(\lambda_i, \mathbf{C}_i)$*, where* λ_i *with* $\lambda_i \neq \lambda_j$*,* $i \neq j$*, and* $\operatorname{Re}(\lambda_i) < 0$*,* $i = 1, \ldots k$*, denote eigenvalues, and where* $\mathbf{C}_i \in \mathbb{C}^{p \times m_i}$ *with* $1 \leq m_i \leq p$*,* $i = 1, \ldots k$*, denote output residues, define the set* $\mathcal{G}(\mathcal{L}_\mathcal{C})$ *of all reduced models having the pairs* $(\lambda_i, \mathbf{C}_i)$ *of poles and output residues as follows:*

$$\mathcal{G}(\mathcal{L}_\mathcal{C}) = \left\{ \boldsymbol{H}_r(s) \,\middle|\, \exists\, \mathbf{B}_i \in \mathbb{C}^{m_i \times m} : \boldsymbol{H}_r(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i \mathbf{B}_i}{s - \lambda_i} \right\}. \tag{4.69}$$

Then $\boldsymbol{G}_r(s)$ is the unique $\mathcal{H}_2$ pseudo-optimal reduced model, i. e. it satisfies

$$\|\boldsymbol{G}-\boldsymbol{G}_r\|_{\mathcal{H}_2}=\min_{\boldsymbol{H}_r\in\mathcal{G}(\mathcal{L}_\mathcal{C})}\|\boldsymbol{G}-\boldsymbol{H}_r\|_{\mathcal{H}_2}\,, \tag{4.70}$$

if and only if

$$\mathbf{C}_i^*\boldsymbol{G}(-\overline{\lambda}_i)=\mathbf{C}_i^*\boldsymbol{G}_r(-\overline{\lambda}_i),\quad i=1,\,\dots\,k. \tag{4.71}$$

Proof. The proof is dual to the one of Theorem 4.19, hence omitted. □

To distinguish both cases of pseudo-optimality, we will denote a reduced model $\boldsymbol{G}_r(s)$, that satisfies Theorem 4.19 as "input $\mathcal{H}_2$ pseudo-optimal", and a reduced model that satisfies Theorem 4.20 as "output $\mathcal{H}_2$ pseudo-optimal".

Remark 4.21. It should be noted that both subsets $\mathcal{G}(\mathcal{L}_\mathcal{B})$ and $\mathcal{G}(\mathcal{L}_\mathcal{C})$ are not uniquely defined by the input residues $\mathbf{B}_i$ and output residues $\mathbf{C}_i$, respectively. Introduce e. g. non-singular transformation matrices $\mathbf{T}_i\in\mathbb{C}^{m_i\times m_i}$, and define the set $\mathcal{G}(\mathcal{L}_{\widetilde{\mathcal{B}}})$, with $\mathcal{L}_{\widetilde{\mathcal{B}}}=\{(\lambda_1,\widetilde{\mathbf{B}}_1),\dots,(\lambda_k,\widetilde{\mathbf{B}}_k)\}$, and where $\widetilde{\mathbf{B}}_i=\mathbf{T}_i\mathbf{B}_i$. Then it readily follows that $\mathcal{G}(\mathcal{L}_{\widetilde{\mathcal{B}}})=\mathcal{G}(\mathcal{L}_\mathcal{B})$, as the $\mathbf{T}_i$'s may be shifted into the $\mathbf{C}_i$'s, see also Remark 4.2. As a consequence, the subset $\mathcal{G}(\mathcal{L}_\mathcal{B})$ and the condition for $\mathcal{H}_2$ pseudo-optimality (4.66) is determined only by the row span of the $\mathbf{B}_i$'s.

Both Theorems 4.19 and 4.20 can be generalized to reduced models that contain poles of higher multiplicities. As the most general case would require cumbersome notation, we only consider the case of one eigenvalue, from which then the most general case can be deduced. This is introduced in the next theorem, which appears to be new.

Theorem 4.22. *Given a fixed eigenvalue λ, with $\operatorname{Re}(\lambda)<0$, and a fixed set of input residues $\mathcal{B}=\{\mathbf{B}_1,\,\dots\,\mathbf{B}_k\}$, where $\mathbf{B}_i\in\mathbb{C}^{\widetilde{m}\times m}$ and $1\leq\widetilde{m}\leq m$, define*

$$\mathbf{A}_\mathrm{H}=\begin{bmatrix}\lambda\mathbf{I} & & & \\ -\mathbf{I} & \ddots & & \\ & \ddots & \ddots & \\ & & -\mathbf{I} & \lambda\mathbf{I}\end{bmatrix},\quad \mathbf{B}_\mathrm{H}=\begin{bmatrix}\mathbf{B}_1\\ \vdots\\ \mathbf{B}_k\end{bmatrix}, \tag{4.72}$$

where $\mathbf{I}$ denotes the $\widetilde{m}\times\widetilde{m}$ identity matrix. Further define the set $\mathcal{G}(\lambda,\mathcal{B})$ of all reduced models having one eigenvalue λ with $\widetilde{m}$ Jordan blocks of dimensions k and with the input residues $\mathcal{B}$ as follows:

$$\mathcal{G}(\lambda,\mathcal{B})=\left\{\boldsymbol{H}_r(s)\,\middle|\,\exists\,\mathbf{C}_\mathrm{H}\in\mathbb{C}^{p\times k\widetilde{m}}:\boldsymbol{H}_r(s)=\mathbf{C}_\mathrm{H}\left(s\mathbf{I}-\mathbf{A}_\mathrm{H}\right)^{-1}\mathbf{B}_\mathrm{H}\right\}. \tag{4.73}$$

Then $\boldsymbol{G}_r(s)$ *is the unique* $\mathcal{H}_2$ *pseudo-optimal reduced model, i. e. it satisfies*

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2} = \min_{\boldsymbol{H}_r \in \mathcal{G}(\lambda,\mathcal{B})} \|\boldsymbol{G} - \boldsymbol{H}_r\|_{\mathcal{H}_2}\,, \tag{4.74}$$

if and only if

$$\left(\mathbf{M}_0^{-\overline{\lambda}} - \widehat{\mathbf{M}}_0^{-\overline{\lambda}}\right)\mathbf{B}_1^* = \mathbf{0} \tag{4.75}$$

$$\left(\mathbf{M}_0^{-\overline{\lambda}} - \widehat{\mathbf{M}}_0^{-\overline{\lambda}}\right)\mathbf{B}_2^* + \left(\mathbf{M}_1^{-\overline{\lambda}} - \widehat{\mathbf{M}}_1^{-\overline{\lambda}}\right)\mathbf{B}_1^* = \mathbf{0} \tag{4.76}$$

$$\vdots$$

$$\sum_{i=0}^{k-1}\left(\mathbf{M}_i^{-\overline{\lambda}} - \widehat{\mathbf{M}}_i^{-\overline{\lambda}}\right)\mathbf{B}_{k-i}^* = \mathbf{0} \tag{4.77}$$

Proof. It can be readily verified, that also $\mathcal{G}(\lambda,\mathcal{B})$ is a closed subspace of $\mathcal{H}_2$, and hence, we may again apply the Hilbert projection theorem, which states that $\boldsymbol{G}_r(s)$ is the unique minimizer of the $\mathcal{H}_2$ error norm in the subspace $\mathcal{G}(\lambda,\mathcal{B})$, if and only if

$$\langle \boldsymbol{G} - \boldsymbol{G}_r, \boldsymbol{H}_r\rangle_{\mathcal{H}_2} = 0, \tag{4.78}$$

for all $\boldsymbol{H}_r(s) \in \mathcal{G}(\lambda,\mathcal{B})$. Now subdivide $\mathbf{C}_{\mathrm{H}}$ into the blocks $\mathbf{C}_{\mathrm{H}} = [\mathbf{C}_1, \ldots, \mathbf{C}_k]$ with $\mathbf{C}_i \in \mathbb{C}^{p\times\widetilde{m}}$, and use Lemma 4.7 for the $\mathcal{H}_2$ inner product, then (4.78) reads as

$$\sum_{i=1}^{k}\sum_{j=i}^{k} \operatorname{trace}\left[\mathbf{C}_j^*\left(\mathbf{M}_{(j-i)}^{-\overline{\lambda}} - \widehat{\mathbf{M}}_{(j-i)}^{-\overline{\lambda}}\right)\mathbf{B}_i^*\right]^* = 0. \tag{4.79}$$

As (4.79) has to hold for arbitrary $\mathbf{C}_i$ this is equivalent to (4.75)–(4.77), which completes the proof. □

Remark 4.23. There would of course also be a dual version of Theorem 4.22 for output $\mathcal{H}_2$ pseudo-optimality. However, the statement should be obvious with Theorem 4.20, hence it is omitted for brevity.

It should be noted, that Theorem 4.19 generalizes three results that are available in the literature: the SISO case as stated in Lemma 4.18 follows from the fact that scalar "tangential directions" can be cancelled in (4.66). This statement is in fact more general, as it also applies to *single* inputs and *multiple* outputs. The MIMO case with single tangential directions, i. e. $m_i = 1$, $i = 1, \ldots, k$, was already presented in [24] and is directly included in Theorem 4.19. If block Krylov subspaces are employed, then $m_i = m$, $i = 1, \ldots, k$, and the conditions for $\mathcal{H}_2$ pseudo-optimality simplify. This has been published in [213] and is clarified in the following.

Corollary 4.24. *Given a fixed set* $\mathcal{L}_{\mathcal{B}} = \{(\lambda_1, \mathbf{B}_1), \ldots, (\lambda_k, \mathbf{B}_k)\}$ *of pairs* $(\lambda_i, \mathbf{B}_i)$*, where* λ_i *with* $\lambda_i \neq \lambda_j$*,* $i \neq j$*, and* $\mathrm{Re}(\lambda_i) < 0$*,* $i = 1, \ldots k$*, denote eigenvalues, and where* $\mathbf{B}_i \in \mathbb{C}^{m \times m}$*,* $i = 1, \ldots k$*, denote input residues, and are assumed non-singular, define the set* $\mathcal{G}(\mathcal{L}_{\mathcal{B}})$ *of all reduced models having the pairs* $(\lambda_i, \mathbf{B}_i)$ *of poles and input residues as follows:*

$$\mathcal{G}(\mathcal{L}_{\mathcal{B}}) = \left\{ \boldsymbol{H}_r(s) \,\middle|\, \exists\, \mathbf{C}_i \in \mathbb{C}^{p \times m} : \boldsymbol{H}_r(s) = \sum_{i=1}^{k} \frac{\mathbf{C}_i \mathbf{B}_i}{s - \lambda_i} \right\}. \tag{4.80}$$

Then $\boldsymbol{G}_r(s)$ *is the unique* $\mathcal{H}_2$ *pseudo-optimal reduced model, i. e. it satisfies*

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2} = \min_{\boldsymbol{H}_r \in \mathcal{G}(\mathcal{L}_{\mathcal{B}})} \|\boldsymbol{G} - \boldsymbol{H}_r\|_{\mathcal{H}_2}, \tag{4.81}$$

if and only if

$$\boldsymbol{G}(-\overline{\lambda}_i) = \boldsymbol{G}_r(-\overline{\lambda}_i), \quad i = 1, \ldots k. \tag{4.82}$$

Proof. The proof follows from Theorem 4.19 and by noting that non-singular $\mathbf{B}_i$ can be cancelled in (4.66). □

We have now discussed all possible cases for input $\mathcal{H}_2$ pseudo-optimality: SISO models, MIMO models—both in terms of tangential interpolation and matching complete block moments—, and also higher order poles. In this respect, if we just say that a reduced model is "input $\mathcal{H}_2$ pseudo-optimal", we actually mean that it is the input $\mathcal{H}_2$ pseudo-optimal reduced model in its respective subset, which then could be $\mathcal{G}(\mathcal{L})$ by Lemma 4.18, $\mathcal{G}(\mathcal{L}_{\mathcal{B}})$—either by Theorem 4.19 or by Corollary 4.24—, or also $\mathcal{G}(\lambda, \mathcal{B})$ by Theorem 4.22. Before we proceed, an important remark on $\mathcal{H}_2$ pseudo-optimality is in order, which applies to all mentioned cases.

Remark 4.25. It should be stressed that an important property of $\mathcal{H}_2$ pseudo-optimality is that once the poles and the input residues (or equivalently the output residues) are fixed, the above conditions become necessary *and* sufficient for $\mathcal{H}_2$ optimality. It is therefore reasonable to employ $\mathcal{H}_2$ pseudo-optimality for optimizing reduced models; this is discussed in Section 4.3.8.

The necessary and sufficient conditions for $\mathcal{H}_2$ pseudo-optimality are in the form of interpolatory Meier-Luenberger conditions. The next section presents equivalent conditions that include also counterparts of the Wilson and Hyland-Bernstein conditions.

4.3.3 New and Equivalent Conditions for $\mathcal{H}_2$ Pseudo-Optimality

The next two theorems provide new and elegant conditions for $\mathcal{H}_2$ pseudo-optimality in the context of projective MOR based on rational Krylov subspaces. These condition are “almost” (this will be well defined in Theorem 4.27) necessary and sufficient for $\mathcal{H}_2$ pseudo-optimality and they are in the form of easy-to-evaluate matrix equations. They thus serve as a valuable tool not only for the a priori construction but also for the a posteriori analysis of $\mathcal{H}_2$ pseudo-optimality, and additionally, they make it possible for the first time, to use $\mathcal{H}_2$ pseudo-optimality in a large-scale setting in the CURE framework. In this respect, these conditions are the most important result of this thesis. A preliminary version in the SISO case was published in [208].

Theorem 4.26. *Given $\boldsymbol{G}(s)$ and a $\mathbf{V}$ whose columns form a basis of a rational input Krylov subspace, define the reduced model $\boldsymbol{G}_r(s) = \mathbf{C}_r\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r$ by projection as in (1.7). This particularly means that $\boldsymbol{G}_r(s)$ is contained in the family $\boldsymbol{G}_{\mathbf{F}}(s)$ from Section 2.5, and that $\mathbf{V}$, $\mathbf{S}$, $\mathbf{L}$, $\mathbf{X}$, and $\mathbf{P}_r$ are defined by the equations (2.15), (4.38), and (4.40), and that $\widehat{\mathbf{P}} = \mathbf{V}\mathbf{P}_r\mathbf{V}^*$, and $\mathbf{B}_\perp = \mathbf{B} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$. Furthermore, let $\mathbf{E}_r^*\mathbf{Q}_f\mathbf{E}_r$ be the Observability Gramian of the system $\boldsymbol{G}_f(s)$, i. e. $\mathbf{Q}_f$ satisfies*

$$\mathbf{A}_r^*\mathbf{Q}_f\mathbf{E}_r + \mathbf{E}_r^*\mathbf{Q}_f\mathbf{A}_r + \mathbf{L}^*\mathbf{L} = \mathbf{0}. \tag{4.83}$$

Assume that both $\mathbf{B}_r$ and $\mathbf{B}_\perp$ have full column rank and that $\mathbf{P}_r$, the solution of (4.40), exists and is unique. Then, the following statements are equivalent:

i) $\mathbf{S} = -\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}$

ii) $\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^* = \mathbf{0}$

iii) $\mathbf{S}\mathbf{P}_r + \mathbf{P}_r\mathbf{S}^* - \mathbf{P}_r\mathbf{L}^*\mathbf{L}\mathbf{P}_r = \mathbf{0}$

iv) $\mathbf{X} = \mathbf{V}\mathbf{P}_r$

v) $\mathbf{A}\widehat{\mathbf{P}}\mathbf{E}^T + \mathbf{E}\widehat{\mathbf{P}}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{B}_\perp\mathbf{B}_\perp^*$

vi) $\mathbf{P}_r^{-1} = \mathbf{E}_r^*\mathbf{Q}_f\mathbf{E}_r$

vii) *$\boldsymbol{G}_f(s)$ is all-pass, i. e. it satisfies $\boldsymbol{G}_f(s)\boldsymbol{G}_f^*(-\overline{s}) = \mathbf{I}$*

Additionally, the statement

viii) $\mathbf{S}^*$ and $-\mathbf{E}_r^{-1}\mathbf{A}_r$ share the same Jordan canonical form

is necessary for statements i)–vii) and sufficient only if the columns of $\mathbf{V}$ form a basis of either a single-input or a block-input Krylov subspace.

Proof. The proof can be found in Appendix A. □

It should be highlighted that Theorem 4.26 does not assume stability of the reduced model. The next result connects the above conditions to $\mathcal{H}_2$ pseudo-optimality.

Theorem 4.27. *Let the conditions of Theorem 4.26 be satisfied, and assume that all interpolation points s_i of the Krylov subspace that is spanned by the columns of $\mathbf{V}$ are contained in the open right half of the complex plane, then*

i) $\boldsymbol{G}_r(s)$ *is input $\mathcal{H}_2$ pseudo-optimal.*
ii) The gradient of $J = \|\boldsymbol{G}-\boldsymbol{G}_r\|_{\mathcal{H}_2}^2$ *with respect to* $\mathbf{C}_r$ *vanishes.*

Conversely, let the reduced model $\boldsymbol{G}_r(s)$ be input $\mathcal{H}_2$ pseudo-optimal, then there exists a $\mathbf{V}$ whose columns form a basis of a rational input Krylov subspace, such that $\boldsymbol{G}_r(s)$ is contained in the family $\boldsymbol{G}_{\mathbf{F}}(s)$, and for which the conditions of Theorem 4.26 are satisfied.

Proof. The proof can be found in Appendix B. □

Remark 4.28. The above theorem shows that the conditions of Theorem 4.26 are sufficient for $\mathcal{H}_2$ pseudo-optimality, if the expansion points s_i are contained in the open right half of the complex plane. Conversely, they are not necessary, because the very same reduced model may be constructed through projections with different Krylov subspaces. As a consequence, there might even exist more than one $\mathcal{H}_2$ pseudo-optimal reduced model in the family $\boldsymbol{G}_{\mathbf{F}}(s)$. Nevertheless, if the reduced model is $\mathcal{H}_2$ pseudo-optimal, at least there always exists a Krylov projection that satisfies the conditions of Theorem 4.26. This is the reason, why the conditions can be exploited to derive an algorithm for the direct construction of (in fact, all possible) $\mathcal{H}_2$ pseudo-optimal reduced models; this is done in the Section 4.3.4.

Remark 4.29. There is a concrete interpretation of condition *vii)* of Theorem 4.26 in the SISO case: $\mathcal{H}_2$ pseudo-optimality requires that the interpolation points s_i and the reduced poles are mirror images of each other. $\boldsymbol{G}_f(s)$ obviously has the same poles as the reduced model and due to Lemma 3.3, its transmission zeros are exactly the interpolation points s_i. In the $\mathcal{H}_2$ pseudo-optimal case, $\boldsymbol{G}_f(s)$ hence takes the form $\boldsymbol{G}_f(s) = \frac{(s-s_1)\cdots(s-s_n)}{(s+\bar{s}_1)\cdots(s+\bar{s}_n)}$, which depicts that $\boldsymbol{G}_f(s)$ is all-pass.

Remark 4.30. We may assume that we choose the shifts in the open right half of the complex plane, then, unlike the interpolatory (kind of Meier-Luenberger) conditions of Section 4.3.2, the conditions of Theorem 4.26 are only sufficient but not necessary for $\mathcal{H}_2$ pseudo-optimality. Nevertheless, a kind of Wilson condition, *iv)*, and a kind of Hyland-Bernstein condition, *v)*, for $\mathcal{H}_2$ pseudo-optimality are included. The advantage of conditions *i)–viii)* is that neither do they alter for higher order poles nor do they change for multiple inputs, all of which is the case for the interpolatory (Meier-Luenberger) conditions. This is similar to the necessary conditions for local $\mathcal{H}_2$ optimality, which was already discussed by Van Dooren et al. [188], who also showed that the generality of the Wilson conditions may have numerical advantages.

It is obvious that there exist also dual versions of the above two theorems which provide easy-to-evaluate conditions for *output* $\mathcal{H}_2$ pseudo-optimality; this is presented next for completeness.

Theorem 4.31. *Given $\boldsymbol{G}(s)$ and a $\mathbf{W}$ whose columns form a basis of a rational output Krylov subspace, define the reduced model $\boldsymbol{G}_r(s) = \mathbf{C}_r\,(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\,\mathbf{B}_r$ by projection as in (1.7). This particularly means that there exist dual $\mathbf{S}_\mathrm{W}$ and $\mathbf{L}_\mathrm{W}$ such that*

$$\mathbf{A}^T\mathbf{W} - \mathbf{E}^T\mathbf{W}\mathbf{S}_\mathrm{W}^* = \mathbf{C}^T\mathbf{L}_\mathrm{W}^*, \tag{4.84}$$

$$\mathbf{A}^T\mathbf{W} - \mathbf{E}^T\mathbf{W}\mathbf{E}_r^{-*}\mathbf{A}_r^* = \mathbf{C}_\perp^*\mathbf{L}_\mathrm{W}^*, \tag{4.85}$$

with $\mathbf{C}_\perp = \mathbf{C} - \mathbf{C}_r\mathbf{E}_r^{-1}\mathbf{W}^*\mathbf{E}$. Let $\mathbf{Y}$ and $\mathbf{Q}_r$ be given by the equations (4.39) and (4.41), and define $\widehat{\mathbf{Q}} = \mathbf{W}\mathbf{Q}_r\mathbf{W}^*$, and the Controllability Gramian $\mathbf{P}_f$ of the feed-through model $\boldsymbol{G}_{f,\mathrm{W}}(s) = \mathbf{C}_r\,(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\,\mathbf{L}_\mathrm{W} + \mathbf{I}$, i. e. $\mathbf{P}_f$ satisfies

$$\mathbf{A}_r\mathbf{P}_f\mathbf{E}_r^* + \mathbf{E}_r\mathbf{P}_f\mathbf{A}_r^* + \mathbf{L}_\mathrm{W}\mathbf{L}_\mathrm{W}^* = \mathbf{0}. \tag{4.86}$$

Assume that both $\mathbf{C}_r$ and $\mathbf{C}_\perp$ have full row rank and that $\mathbf{Q}_r$, the solution of (4.41), exists and is unique. Then, the following statements are equivalent:

i) $\mathbf{S}_\mathrm{W}^* = -\mathbf{Q}_r\mathbf{A}_r\mathbf{E}_r^{-1}\mathbf{Q}_r^{-1}$

ii) $\mathbf{E}_r^{-*}\mathbf{C}_r^* + \mathbf{Q}_r\mathbf{L}_\mathrm{W} = \mathbf{0}$

iii) $\mathbf{S}_\mathrm{W}^*\mathbf{Q}_r + \mathbf{Q}_r\mathbf{S}_\mathrm{W} - \mathbf{Q}_r\mathbf{L}_\mathrm{W}\mathbf{L}_\mathrm{W}^*\mathbf{Q}_r = \mathbf{0}$

iv) $\mathbf{Y} = \mathbf{W}\mathbf{Q}_r$

v) $\mathbf{A}^T\widehat{\mathbf{Q}}\mathbf{E} + \mathbf{E}^T\widehat{\mathbf{Q}}\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{C}_\perp^*\mathbf{C}_\perp$

vi) $\mathbf{Q}_r^{-1} = \mathbf{E}_r\mathbf{P}_f\mathbf{E}_r^*$

vii) $\boldsymbol{G}_{f,\mathrm{W}}(s)$ is all-pass, i. e. it satisfies $\boldsymbol{G}_{f,\mathrm{W}}(s)\boldsymbol{G}_{f,\mathrm{W}}^*(-\overline{s}) = \mathbf{I}$

Additionally, the statement

viii) $\mathbf{S}_\mathrm{W}^*$ and $-\mathbf{E}_r^{-1}\mathbf{A}_r$ share the same Jordan canonical form

is necessary for statements i)–viii) and sufficient only if the columns of $\mathbf{W}$ form a basis of either a single-output or a block-output Krylov subspace.

Moreover, assume that the above conditions i)–vii) are satisfied, and that all interpolation points s_i of the Krylov subspace that is spanned by the columns of $\mathbf{W}$—i. e. the eigenvalues of $\mathbf{S}_\mathrm{W}$—are contained in the open right half of the complex plane, then *i)* $\boldsymbol{G}_r(s)$ *is output* $\mathcal{H}_2$ *pseudo-optimal.*

ii) The gradient of $J=\|\boldsymbol{G}-\boldsymbol{G}_r\|_{\mathcal{H}_2}^2$ *with respect to* $\mathbf{B}_r$ *vanishes.*

Conversely, let the reduced model $\boldsymbol{G}_r(s)$ *be output* $\mathcal{H}_2$ *pseudo-optimal, then there exists a* $\mathbf{W}$ *whose columns form a basis of a rational output Krylov subspace, such that* $\boldsymbol{G}_r(s)$ *is contained in the dual family to* $\boldsymbol{G}_{\mathbf{F}}(s)$*, and for which the above conditions i)–viii) are satisfied.*

Proof. The proof follows from duality and is hence omitted. □

As we have defined both conditions for input and output $\mathcal{H}_2$ pseudo-optimality, we are now ready to relate them to the first order necessary conditions for $\mathcal{H}_2$ optimality.

Lemma 4.32. *If both conditions for input and output* $\mathcal{H}_2$ *pseudo-optimality from Theorems 4.26 and 4.31 are satisfied, then the first order necessary conditions for* $\mathcal{H}_2$ *optimality from Theorems 4.12, 4.15 and 4.16 are satisfied.*

Proof. The proof readily follows from the equivalence of conditions *v)* from Theorems 4.26 and 4.31 to the Hyland-Bernstein conditions in (4.57), (4.58). □

If MOR is based on projections onto rational Krylov subspaces, the equivalent conditions of both Theorems 4.26 and 4.31 are a valuable tool for the a posteriori analysis of a given reduced model. This has considerable advantages over the necessary and sufficient interpolatory conditions of Section 4.3.2: the evaluation of the interpolatory conditions requires large-scale operations, because moments of the original model have to be computed; the conditions of Theorems 4.26 and 4.31 may be instead evaluated with small-scale operations only, e. g. by condition *ii)*, because the nature of rational Krylov subspaces is exploited. Another application of Theorems 4.26 and 4.31 is the direct construction of $\mathcal{H}_2$ pseudo-optimal model reduced models, which is discussed in the next section.

4.3.4 $\mathcal{H}_2$ Pseudo-Optimal Rational Krylov (PORK) Algorithm

Assume that a sequence of interpolation points s_i with respective tangential directions $\mathbf{L}_i$ is given. As discussed in Section 2.3, then it is possible to compute a triple $(\mathbf{V},\mathbf{S},\mathbf{L})$, such that the $\mathbf{B}$-Sylvester equation (2.15) is satisfied, where the columns of $\mathbf{V}$ form a basis of the rational input Krylov subspace, and where the interpolation points s_i are the eigenvalues of $\mathbf{S}$, and where the tangential directions $\mathbf{L}_i$ are incorporated in $\mathbf{L}$. Given such a triple $(\mathbf{V},\mathbf{S},\mathbf{L})$ that satisfies the $\mathbf{B}$-Sylvester equation (2.15), the exploitation of condition *iii)* of Theorem 4.26, and subsequently of condition *ii)*, and of equation

(2.42), is sufficient for the construction of an input $\mathcal{H}_2$ pseudo-optimal reduced model. (The sufficiency is in fact proven in the end of Appendix B.) The basic procedure is depicted in Algorithm 4.2—denoted as the *pseudo-optimal rational Krylov* (PORK) algorithm—which has been published for the SISO case in [208].

Algorithm 4.2 Pseudo-optimal rational Krylov (PORK)

```
Input: $\mathbf{V}$, $\mathbf{S}$, $\mathbf{L}$, $\mathbf{C}$, such that $\mathbf{AV}-\mathbf{EVS}=\mathbf{BL}$ is satisfied (see Section 2.3)
Output: input $\mathcal{H}_2$ pseudo-optimal reduced model $\boldsymbol{G}_r(s)=\mathbf{C}_r\,(s\mathbf{E}_r-\mathbf{A}_r)^{-1}\,\mathbf{B}_r$
1: $\mathbf{P}_r^{-1}=\text{lyap}(\mathbf{S}^*,-\mathbf{L}^*\mathbf{L})$    // direct solver for $\mathbf{S}^*\mathbf{P}_r^{-1}+\mathbf{P}_r^{-1}\mathbf{S}-\mathbf{L}^*\mathbf{L}=\mathbf{0}$, condition iii)
2: $\mathbf{B}_r=-\left(\mathbf{P}_r^{-1}\right)^{-1}\mathbf{L}^*$    // condition ii)
3: $\mathbf{A}_r=\mathbf{S}+\mathbf{B}_r\mathbf{L}$,    $\mathbf{E}_r=\mathbf{I}$,    $\mathbf{C}_r=\mathbf{CV}$
```

It should be noted that PORK requires any $\mathbf{V}$ that spans an input rational Krylov subspace, together with corresponding $\mathbf{S}$ and $\mathbf{L}$. (An admissible triple $(\mathbf{V},\mathbf{S},\mathbf{L})$ may be computed by a modified Arnoldi algorithm; for details please refer to Section 2.3.) This is already the main numerical effort to achieve the $\mathcal{H}_2$ pseudo-optimal reduced model, because it then remains in PORK to solve a Lyapunov equation and an LSE—but both of reduced order n. It should be stressed, that these remaining steps may be conducted irrespectively of the type of Krylov subspaces, namely single-, block- or tangential-input. This allows for a convenient implementation of PORK, as there is no need to distinguish the different cases. The reason for this is that the shifts s_i, the tangential directions $\mathbf{L}_i$, and even more, higher multiplicities of s_i and $\mathbf{L}_i$, are all encoded in the Jordan canonical form of the pair $(\mathbf{L},\mathbf{S})$, such that one does not have to bother with these details any more.

Another interesting aspect should be highlighted, which is not obvious from Algorithm 4.2: as the outcome of PORK is an $\mathcal{H}_2$ pseudo-optimal reduced model, it satisfies two particular properties. Firstly, the eigenvalues of $\mathbf{S}$—i. e. the shifts s_i—will become the mirror images of the eigenvalues of $\mathbf{A}_r$—i. e. the poles of the reduced model—, with corresponding multiplicities. Secondly, the tangential directions $\mathbf{L}_i$ will become the input residues $\mathbf{B}_i^*$ of the reduced model. Although this fact is not apparent from Algorithm 4.2, it is a consequence of Theorem 4.26. To illustrate the nature of PORK, a simple example is presented next.

Example 4.1. Assume a single-input, $m=1$, and that we have an expansion point $s_0\in\mathbb{R}$, at which we want to match the first two moments. Define $\mathbf{A}_{s_0}=(\mathbf{A}-s_0\mathbf{E})$, then an admissible triple $(\mathbf{V},\mathbf{S},\mathbf{L})$ is given by

$$\mathbf{V}=\left[\begin{array}{cc}\mathbf{A}_{s_0}^{-1}\mathbf{b} & \mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{b}\end{array}\right],\quad \mathbf{S}=\left[\begin{array}{cc}s_0 & 1\\ 0 & s_0\end{array}\right],\quad \mathbf{l}=\left[\begin{array}{cc}1 & 0\end{array}\right]. \tag{4.87}$$

Owing to condition *iii)* of Theorem 4.26, we solve at Step 1 of PORK the Lyapunov equation $\mathbf{P}_r^{-1}\mathbf{S}+\mathbf{S}^T\mathbf{P}_r^{-1}-\mathbf{l}^T\mathbf{l}=\mathbf{0}$ for

$$\mathbf{P}_r^{-1}=\begin{bmatrix} \frac{1}{2s_0} & -\frac{1}{4s_0^2} \\ -\frac{1}{4s_0^2} & \frac{1}{4s_0^3} \end{bmatrix}, \quad \text{and hence,} \quad \mathbf{P}_r=4s_0\begin{bmatrix} 1 & s_0 \\ s_0 & 2s_0^2 \end{bmatrix}. \tag{4.88}$$

The reduced matrices then follow from Steps 2 and 3 of PORK, and read as

$$\mathbf{b}_r=\begin{bmatrix} -4s_0 \\ -4s_0^2 \end{bmatrix},\ \mathbf{A}_r=\begin{bmatrix} -3s_0 & 1 \\ -4s_0^2 & s_0 \end{bmatrix},\ \mathbf{E}_r=\mathbf{I},\ \mathbf{C}_r=\begin{bmatrix} \mathbf{C}\mathbf{A}_{s_0}^{-1}\mathbf{b} & \mathbf{C}\mathbf{A}_{s_0}^{-1}\mathbf{E}\mathbf{A}_{s_0}^{-1}\mathbf{b} \end{bmatrix}. \tag{4.89}$$

Then $\boldsymbol{G}_r(s)=\mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{b}_r$ satisfies $\boldsymbol{G}(s_0)=\boldsymbol{G}_r(s_0)$ and $\boldsymbol{G}'(s_0)=\boldsymbol{G}_r'(s_0)$, where $\boldsymbol{G}'(s)$ denotes the derivative with respect to s. Moreover, $\mathbf{A}_r$ has one eigenvalue $-s_0$, with algebraic multiplicity 2 and geometric multiplicity 1. It is noteworthy, that $\mathbf{E}_r$, $\mathbf{A}_r$ and $\mathbf{b}_r$ are all independent from the original model.

The dual output PORK algorithm shall also be presented for completeness. The prerequisite is that the columns of $\mathbf{W}$ form a basis of a rational output Krylov subspace (may be single-output, block-output or tangential-output), and corresponding $\mathbf{S}_{\mathrm{W}}$ and $\mathbf{L}_{\mathrm{W}}$, such that $\mathbf{A}^T\mathbf{W}-\mathbf{E}^T\mathbf{W}\mathbf{S}_{\mathrm{W}}^*=\mathbf{C}^T\mathbf{L}_{\mathrm{W}}^*$ is satisfied, have to be given. Then the *output pseudo-optimal rational Krylov* (O-PORK) algorithm, that computes the output $\mathcal{H}_2$ pseudo-optimal reduced model, is presented in Algorithm 4.3.

Algorithm 4.3 Output pseudo-optimal rational Krylov (O-PORK)

```
Input: $\mathbf{W}$, $\mathbf{S}_{\mathrm{W}}$, $\mathbf{L}_{\mathrm{W}}$, $\mathbf{B}$, such that $\mathbf{A}^T\mathbf{W}-\mathbf{E}^T\mathbf{W}\mathbf{S}_{\mathrm{W}}^*=\mathbf{C}^T\mathbf{L}_{\mathrm{W}}^*$ is satisfied
Output: output $\mathcal{H}_2$ pseudo-optimal reduced model $\boldsymbol{G}_r(s)=\mathbf{C}_r\left(s\mathbf{E}_r-\mathbf{A}_r\right)^{-1}\mathbf{B}_r$
$\mathbf{Q}_r^{-1}=$ lyap($\mathbf{S}_{\mathrm{W}}, -\mathbf{L}_{\mathrm{W}}\mathbf{L}_{\mathrm{W}}^*$)          // condition iii) of Theorem 4.31
$\mathbf{C}_r=-\mathbf{L}_{\mathrm{W}}^*\left(\mathbf{Q}_r^{-1}\right)^{-1}$          // condition ii) of Theorem 4.31
$\mathbf{A}_r=\mathbf{S}_{\mathrm{W}}+\mathbf{L}_{\mathrm{W}}\mathbf{C}_r$,    $\mathbf{E}_r=\mathbf{I}$,    $\mathbf{B}_r=\mathbf{W}^*\mathbf{B}$
```

The kind of optimality, which the outcome of both PORK and O-PORK algorithms fulfil, might require some clarification. To this end, let a $\mathbf{V}$ be given whose columns form a basis of a rational input Krylov subspace, then PORK as in Algorithm 4.2 directly constructs a reduced model that is $\mathcal{H}_2$ pseudo-optimal. Bearing in mind the family $\boldsymbol{G}_{\mathbf{F}}(s)$ from Section 2.5, this means that PORK automatically picks an $\mathcal{H}_2$ pseudo-optimal reduced model out of the family $\boldsymbol{G}_{\mathbf{F}}(s)$. However, it should be highlighted that there is no guarantee, that the outcome of PORK minimizes the $\mathcal{H}_2$ error in the family $\boldsymbol{G}_{\mathbf{F}}(s)$; it is rather like there always exists (at least) one $\mathcal{H}_2$ pseudo-optimal reduced model—with respect to the subset $\mathcal{G}(\mathcal{L}_{\mathcal{B}})$—in the family $\boldsymbol{G}_{\mathbf{F}}(s)$, and which PORK automatically picks. PORK may thus be interpreted as a deliberate choice of the direction of projection, represented by the matrix $\mathbf{W}$ (which has been the remaining degree of freedom after fixing $\mathbf{V}$). Please note that PORK does not explicitly build up an appropriate matrix $\mathbf{W}$, it is instead implicitly determined by the outcome of PORK. If desired, it would be possible to subsequently construct a suitable $\mathbf{W}$ that goes with the reduced model; this, however, seems to be needless, as no circumstance is known to the author, where the explicit knowledge of $\mathbf{W}$ is beneficial.

By contrast, the explicit knowledge of $\mathbf{V}$ is essential because it is required in the CURE framework. To be precise, the factorization of the error model discussed in Section 3.1 is feasible only if $\mathbf{V}$ is known. This fact justifies the need for the PORK algorithm, because most existing ways to compute $\mathcal{H}_2$ pseudo-optimal reduced models rely on the direct construction of $\boldsymbol{G}_r(s)$, and hence, the connection to the Krylov subspace $\mathbf{V}$ would be lost. Another possibility to compute an $\mathcal{H}_2$ pseudo-optimal reduced model—at least in the SISO case—is the pole-placement approach due to Antoulas [9]. Although this way could preserve the connection to $\mathbf{V}$, it would require additional large-scale operations, which can be avoided in PORK: once that $\mathbf{V}$ is computed, all remaining steps in PORK are n-dimensional operations. The advantage of PORK over other approaches to compute $\mathcal{H}_2$ pseudo-optimal reduced models was also discussed in [208], to which the interested reader is referred to for further details.

4.3.5 Orthogonality in $\mathcal{H}_2$ Optimal Reduction

Another interesting point to study is orthogonality in $\mathcal{H}_2$ optimal MOR, which is presented in the next lemma.

Lemma 4.33. *If the reduced model $\boldsymbol{G}_r(s)$ is input or output $\mathcal{H}_2$ pseudo-optimal, then*

$$\langle \boldsymbol{G} - \boldsymbol{G}_r, \boldsymbol{G}_r \rangle_{\mathcal{H}_2} = \mathbf{0}. \tag{4.90}$$

Moreover, if the reduced model $\boldsymbol{G}_r(s)$ is locally $\mathcal{H}_2$ optimal, then additionally

$$\langle \boldsymbol{G} - \boldsymbol{G}_r, \boldsymbol{G}_r' \rangle_{\mathcal{H}_2} = \mathbf{0}, \tag{4.91}$$

where $\boldsymbol{G}_r'(s)$ denotes the derivative of $\boldsymbol{G}_r(s)$ with respect to s.

Proof. We prove only the case that $\boldsymbol{G}_r(s)$ is input $\mathcal{H}_2$ pseudo-optimal, as the proof for output $\mathcal{H}_2$ pseudo-optimality directly follows from duality. Use Lemmata 4.3 and 4.7 for the $\mathcal{H}_2$ inner product (4.90), then the first result follows from the necessary and sufficient conditions of Theorems 4.19 and 4.22. Use Lemma 4.5 for the $\mathcal{H}_2$ inner product (4.91), then the second result follows from the necessary conditions for local $\mathcal{H}_2$ optimality stated in Theorem 4.12. To prove (4.91) in case of higher multiplicities of the reduced poles would require cumbersome notation and is omitted for brevity. □

Lemma 4.33 presents the geometric interpretation that the error is orthogonal to the reduced model in the $\mathcal{H}_2$ pseudo-optimal case and, additionally, orthogonal to the derivative of the reduced model in the local $\mathcal{H}_2$ optimal case. A similar statement in terms of the error factorization from Section 3.1 is presented in the next lemma.

Lemma 4.34. *If the reduced model $\boldsymbol{G}_r(s)$ is input $\mathcal{H}_2$ pseudo-optimal and satisfies the conditions of Theorem 4.26, then*

$$\langle \boldsymbol{G}_\perp \boldsymbol{G}_f, \boldsymbol{G}_r \rangle_{\mathcal{H}_2} = \mathbf{0}. \tag{4.92}$$

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2} = \|\boldsymbol{G}_\perp\|_{\mathcal{H}_2}. \tag{4.93}$$

Moreover, if the reduced model $\boldsymbol{G}_r(s)$ is locally $\mathcal{H}_2$ optimal, then additionally

$$\langle \boldsymbol{G}_\perp, \boldsymbol{G}_r \rangle_{\mathcal{H}_2} = \mathbf{0}. \tag{4.94}$$

Proof. As $\boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s) = \boldsymbol{G}(s) - \boldsymbol{G}_r(s)$, the proof of (4.92) is already contained in Lemma 4.33. To prove (4.93), consider $\|\boldsymbol{G}_e\|_{\mathcal{H}_2}^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{trace}\,[\boldsymbol{G}_e^*(\imath\omega)\boldsymbol{G}_e(\imath\omega)]\,\mathrm{d}\omega$. Substituting $\boldsymbol{G}_e(s) = \boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$, and using $\boldsymbol{G}_f(s)\boldsymbol{G}_f^*(-\overline{s}) = \mathbf{I}$ (and hence, $\boldsymbol{G}_f(\imath\omega)\boldsymbol{G}_f^*(\imath\omega) = \mathbf{I}$), due to Theorem 4.26, the proof can be concluded. It is left to prove (4.94): $\boldsymbol{G}_\perp(s)$ shares $\mathbf{E}$, $\mathbf{A}$ and $\mathbf{C}$ with $\boldsymbol{G}(s)$, and we can use Lemma 4.1 to compute the $\mathcal{H}_2$ inner product: $\langle \boldsymbol{G}_\perp, \boldsymbol{G}_r \rangle_{\mathcal{H}_2} = \mathbf{B}_\perp^* \mathbf{Y} \mathbf{B}_r$. Owing to Theorem 4.26, the reduced model is a projection onto a rational input Krylov subspace, and due to Wilson's necessary conditions for local $\mathcal{H}_2$ optimality, there exists a $\mathbf{W}$ such that $\mathbf{Y} = \mathbf{W}\mathbf{Q}_r$. By definition, $\mathbf{W} \perp \mathbf{B}_\perp$, which completes the proof. □

The orthogonality of the error and reduced model has an important consequence, which is presented in the next theorem. It presents one of the main advantages of $\mathcal{H}_2$ pseudo-optimal reductions, which is the basis for the optimization procedures discussed in [148].

Theorem 4.35. *If the reduced model $\boldsymbol{G}_r(s)$ is input or output $\mathcal{H}_2$ pseudo-optimal, then*

$$\|\boldsymbol{G} - \boldsymbol{G}_r\|_{\mathcal{H}_2}^2 = \|\boldsymbol{G}\|_{\mathcal{H}_2}^2 - \|\boldsymbol{G}_r\|_{\mathcal{H}_2}^2, \tag{4.95}$$

and consequently, $\|\boldsymbol{G}\|_{\mathcal{H}_2} \geq \|\boldsymbol{G}_r\|_{\mathcal{H}_2}$.

Proof. Due to Lemma 4.33, $\langle \boldsymbol{G}-\boldsymbol{G}_r,\boldsymbol{G}_r\rangle_{\mathcal{H}_2}=\boldsymbol{0}$, which is equivalent to $\langle \boldsymbol{G},\boldsymbol{G}_r\rangle_{\mathcal{H}_2}=\langle \boldsymbol{G}_r,\boldsymbol{G}_r\rangle_{\mathcal{H}_2}=\|\boldsymbol{G}_r\|^2_{\mathcal{H}_2}$. Using this in (4.42), $\|\boldsymbol{G}_e\|^2_{\mathcal{H}_2}=\|\boldsymbol{G}\|^2_{\mathcal{H}_2}+\|\boldsymbol{G}_r\|^2_{\mathcal{H}_2}-2\langle \boldsymbol{G},\boldsymbol{G}_r\rangle_{\mathcal{H}_2}$, proves (4.95), which also yields $\|\boldsymbol{G}\|^2_{\mathcal{H}_2}=\|\boldsymbol{G}_r\|^2_{\mathcal{H}_2}+\|\boldsymbol{G}_e\|^2_{\mathcal{H}_2}\geq\|\boldsymbol{G}_r\|^2_{\mathcal{H}_2}$. □

Theorem 4.35 shows that the $\mathcal{H}_2$ norm of the reduced model cannot be larger than the $\mathcal{H}_2$ norm of the original model—if the reduced model is $\mathcal{H}_2$ pseudo-optimal. Equation (4.95) also has a nice interpretation, as it depicts some kind of Pythagorean equation, which might be illustrated by Thales' theorem like in Figure 4.2.

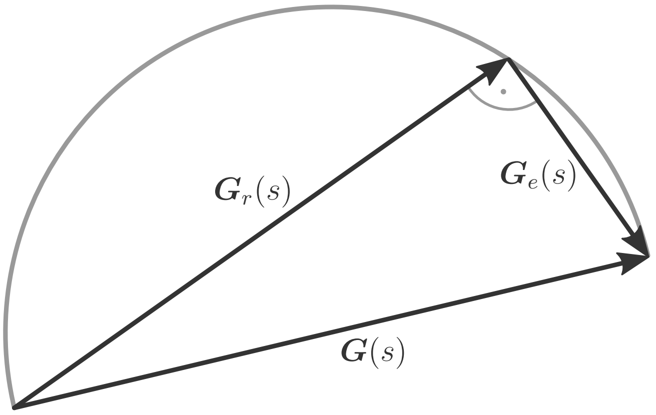


Figure 4.2: Original, $\mathcal{H}_2$ pseudo-optimal reduced, and corresponding error model.

The figure clearly shows that $\|\boldsymbol{G}_r\|_{\mathcal{H}_2}$ cannot grow larger than $\|\boldsymbol{G}\|_{\mathcal{H}_2}$, and that a larger $\|\boldsymbol{G}_r\|_{\mathcal{H}_2}$ forces the error, i. e. $\|\boldsymbol{G}_e\|_{\mathcal{H}_2}$, to decrease. This is a crucial consequence, as it may be exploited for optimization: instead of minimizing the $\mathcal{H}_2$ norm of the error—which is usually not accessible in a large-scale setting—one can equivalently maximize the $\mathcal{H}_2$ norm of the reduced model, as long as one reduces $\mathcal{H}_2$ pseudo-optimally (e. g. through PORK). The justification for this statement is that an $\mathcal{H}_2$ pseudo-optimal reduced model will always rest on the semi-circle drawn in Figure 4.2, and hence, maximizing the $\mathcal{H}_2$ norm of the reduced model forces the $\mathcal{H}_2$ norm of the error to decrease. This paradigm shift in the optimization procedure may be depicted as follows:

$$\min\|\boldsymbol{G}-\boldsymbol{G}_r\|_{\mathcal{H}_2} \quad \overset{\mathcal{H}_2\text{ pseudo-optimality}}{\Longrightarrow} \quad \max\|\boldsymbol{G}_r\|_{\mathcal{H}_2}\,. \tag{4.96}$$

This is one of the ideas behind the trust region optimization algorithm denoted as *stability-preserving, adaptive rational Krylov* (SPARK), which was introduced by Panzer et al. [149], and then improved by Panzer in his thesis [148]. The approach also makes use of the CURE framework, and as will be discussed in Section 4.3.7, Theorem 4.35 is the reason, why $\mathcal{H}_2$ pseudo-optimality is particularly beneficial in the CURE framework.

4.3.6 Nested Inner IRKA Loop for Multivariable Systems

Generally, the convergence of IRKA slows down with higher numbers of inputs and outputs. This can be improved by exploiting $\mathcal{H}_2$ pseudo-optimality in the MIMO case; which is discussed in this section, and which does not apply to SISO models.

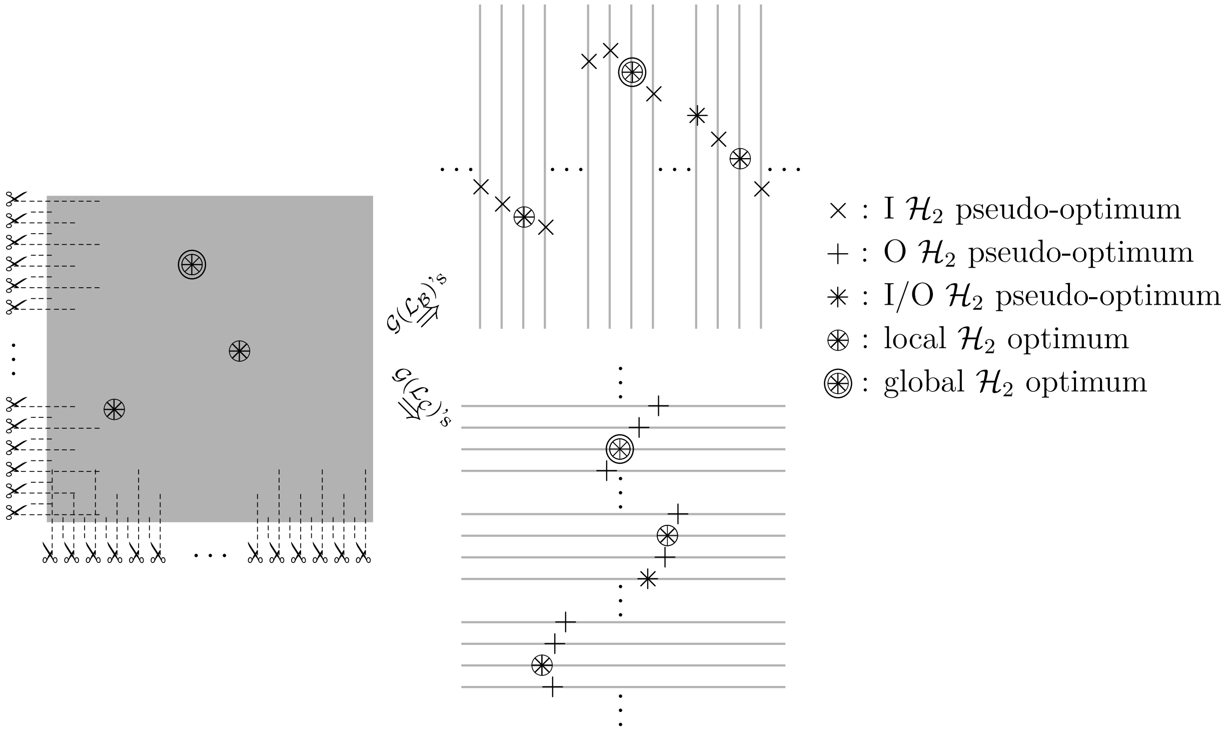
To identify the difference between SISO and MIMO, assume for the moment a SISO model, because then $\mathcal{H}_2$ pseudo-optimality is defined as the global minimizer in the subset $\mathcal{G}(\mathcal{L})$ as in Lemma 4.18. That means, that a set of reduced eigenvalues $\mathcal{L} = \{\lambda_1, \dots \lambda_n\}$ uniquely defines the respective subset. By contrast, in the MIMO case the reduced model may not only be input $\mathcal{H}_2$ pseudo-optimal, but also output $\mathcal{H}_2$ pseudo-optimal. That means, that the reduced model is the global minimizer in the subset $\mathcal{G}(\mathcal{L}_\mathcal{B})$ as in Theorem 4.19, or in the subset $\mathcal{G}(\mathcal{L}_\mathcal{C})$ as in Theorem 4.20, where to each reduced eigenvalue either an input residue direction or an output residue direction is associated. Consequently, any SISO reduced model is contained in only *one* subset $\mathcal{G}(\mathcal{L})$, whereas any MIMO reduced model is contained in *two* intersecting subsets $\mathcal{G}(\mathcal{L}_\mathcal{B})$ and $\mathcal{G}(\mathcal{L}_\mathcal{C})$. In order to clarify all possible cases that might occur with MIMO models, please consider Table 4.1.

Table 4.1: Possible cases of $\mathcal{H}_2$ pseudo-optimality for multivariable models

	input $\mathcal{H}_2$ pseudo-optimal	output $\mathcal{H}_2$ pseudo-optimal	locally $\mathcal{H}_2$ optimal
Case 1:	✓	—	—
Case 2:	—	✓	—
Case 3:	✓	✓	—
Case 4:	✓	✓	✓

Cases 1 and 2 are obvious, as they describe either input or output $\mathcal{H}_2$ pseudo-optimality; Case 4 is also clear, as local $\mathcal{H}_2$ optimality requires both input and output $\mathcal{H}_2$ pseudo-optimality; the interesting one is Case 3: a reduced model may be both input *and* output $\mathcal{H}_2$ pseudo-optimal, *without* being locally $\mathcal{H}_2$ optimal. Owing to Theorems 4.26 and 4.31 this means that both gradients of J, with respect to $\mathbf{B}_r$ *and* $\mathbf{C}_r$ vanish—but *not* the gradient of J with respect to $\mathbf{A}_r$. If one thinks of the Meier-Luenberger condition for local $\mathcal{H}_2$ optimality as in Theorem 4.12, this means that $\boldsymbol{G}(-\overline{\lambda}_{r,i})\mathbf{b}_i^* = \boldsymbol{G}_r(-\overline{\lambda}_{r,i})\mathbf{b}_i^*$, cf. (4.47), and $\mathbf{c}_i^*\boldsymbol{G}(-\overline{\lambda}_{r,i}) = \mathbf{c}_i^*\boldsymbol{G}_r(-\overline{\lambda}_{r,i})$, cf. (4.48), are satisfied, but not $\mathbf{c}_i^*\boldsymbol{G}'(-\overline{\lambda}_{r,i})\mathbf{b}_i^* = \mathbf{c}_i^*\boldsymbol{G}_r'(-\overline{\lambda}_{r,i})\mathbf{b}_i^*$, cf. (4.49).

Owing to the existence of simultaneous input and output $\mathcal{H}_2$ pseudo-optimality, Figure 4.1 does in fact not reflect all aspects of $\mathcal{H}_2$ pseudo-optimality in the MIMO case. This is generalized in Figure 4.3: it is illustrated that in the MIMO case there are two distinct ways to divide the set of all reduced models of fixed order. Either way, the global minimizer in the subsets are input (I) $\mathcal{H}_2$ pseudo-optimal, denoted as “×”, or output (O) $\mathcal{H}_2$ pseudo-optimal, denoted as “+”. If two subsets $\mathcal{G}(\mathcal{L}_\mathcal{B})$ and $\mathcal{G}(\mathcal{L}_\mathcal{C})$ overlap at an $\mathcal{H}_2$ pseudo-optimum, the reduced model is both input and output (I/O) $\mathcal{H}_2$ pseudo-optimal, denoted as “$\ast$”. Then every locally $\mathcal{H}_2$ optimal reduced model, denoted as “$\circledast$”, necessarily is both input and output $\mathcal{H}_2$ pseudo-optimal, and the global optimum is denoted as “$\textcircled{\circledast}$”. Figure 4.3 of course oversimplifies $\mathcal{H}_2$ pseudo-optimality in the MIMO case, but it still reflects its basic nature.



$\mathcal{G}(\mathcal{L}_\mathcal{B})$'s
$\mathcal{G}(\mathcal{L}_\mathcal{C})$'s
× : I $\mathcal{H}_2$ pseudo-optimum
+ : O $\mathcal{H}_2$ pseudo-optimum
$\ast$: I/O $\mathcal{H}_2$ pseudo-optimum
$\circledast$: local $\mathcal{H}_2$ optimum
$\textcircled{\circledast}$: global $\mathcal{H}_2$ optimum

Figure 4.3: Illustration of $\mathcal{H}_2$ pseudo-optimality in the MIMO case.

The occurrence of reduced models that are simultaneously input and output $\mathcal{H}_2$ pseudo-optimal may be exploited in IRKA. The original statement of IRKA, as in Algorithm 4.1, depicts a single loop, in which all data—i. e. interpolation points, input and output tangential directions—is updated at the same time. The idea now is to establish two nested loops: an inner loop in which the reduced eigenvalues are fixed, and which only corrects the tangential directions, and an outer loop, which performs an update of the interpolation points. The basic idea was proposed by Beattie and Gugercin [24].

The principle of the inner loop then is to alternate between input and output $\mathcal{H}_2$ pseudo-optimal reduced models for a given set of reduced eigenvalues, until at some point the residues converge to a set for which the reduced model is simultaneously input and output $\mathcal{H}_2$ pseudo-optimal. The task in the inner loop thus is to find the $\mathcal{H}_2$ pseudo-optimal output (input) residue directions for given poles and input (output) residue directions. This may basically be solved by the PORK and O-PORK algorithms. However, it will be beneficial in this setting to slightly modify these algorithms, which is discussed next.

Assume that an admissible triple $(\mathbf{V}, \mathbf{S}, \mathbf{L})$ is given and that the input $\mathcal{H}_2$ pseudo-optimal reduced model has been computed by PORK as in Algorithm 4.2. Then introduce the state transformation $\mathbf{z} = -\mathbf{P}_r^{-1}\mathbf{x}$, such that the reduced model reads as $\boldsymbol{G}_r(s) = -\mathbf{C}_r\mathbf{P}_r\left(s\mathbf{I} - \mathbf{P}_r^{-1}\mathbf{A}_r\mathbf{P}_r\right)^{-1}(-\mathbf{P}_r^{-1}\mathbf{B}_r)$. Owing to Step 2 of PORK, $\mathbf{B}_r = -\mathbf{P}_r\mathbf{L}^*$, then the input of the state-transformed (i. e. in $\mathbf{z}$ co-ordinates) model becomes $\mathbf{L}^*$. Furthermore, the dynamic matrix becomes

$$\begin{aligned} \mathbf{P}_r^{-1}\mathbf{A}_r\mathbf{P}_r &\overset{\text{Step 3}}{=} \mathbf{P}_r^{-1}\mathbf{S}\mathbf{P}_r - \mathbf{L}^*\mathbf{L}\mathbf{P}_r && (4.97)\\ &\overset{\text{Step 1}}{=} -\mathbf{S}^*\mathbf{P}_r^{-1}\mathbf{P}_r + \mathbf{L}^*\mathbf{L}\mathbf{P}_r - \mathbf{L}^*\mathbf{L}\mathbf{P}_r \quad = \quad -\mathbf{S}^*, && (4.98) \end{aligned}$$

and hence, the reduced model $\boldsymbol{G}_r(s)$ has the state-space realization

$$\begin{aligned} \dot{\mathbf{z}}(t) &= -\mathbf{S}^*\mathbf{z}(t) + \mathbf{L}^*\mathbf{u}(t),\\ \mathbf{y}(t) &= -\mathbf{C}\mathbf{V}\mathbf{P}_r\mathbf{z}(t). \end{aligned} \tag{4.99}$$

To summarise these findings: let $\mathbf{S}$ be diagonal with mirrored eigenvalues $-\overline{\lambda}_i$ on the diagonal, and $\mathbf{L} = [\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*]$ with input residue directions as columns, then it follows from (4.99), that the output residue directions for input $\mathcal{H}_2$ pseudo-optimality are given by the columns of $-\mathbf{C}\mathbf{V}\mathbf{P}_r = [\mathbf{c}_1, \ldots, \mathbf{c}_n]$. In a dual way, if $\mathbf{S}_\mathrm{W} = \mathbf{S}$ is given as above, together with output residue directions $\mathbf{L}_\mathrm{W}^* = [\mathbf{c}_1, \ldots, \mathbf{c}_n]$, then the input residue directions for output $\mathcal{H}_2$ pseudo-optimality are the rows of $-\mathbf{Q}_r\mathbf{W}^*\mathbf{B} = [\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*]^*$, and the reduced model reads as

$$\begin{aligned} \dot{\mathbf{z}}(t) &= -\mathbf{S}_\mathrm{W}^*\mathbf{z}(t) - \mathbf{Q}_r\mathbf{W}^*\mathbf{B}\mathbf{u}(t),\\ \mathbf{y}(t) &= \mathbf{L}_\mathrm{W}^*\mathbf{z}(t). \end{aligned} \tag{4.100}$$

Both reduced models (4.99) and (4.100) are the basis of the inner loop of IRKA we are about to derive. The benefit is that we may choose $\mathbf{S} = \mathbf{S}_\mathrm{W}$, such that it is possible to alternate between (4.99) and (4.100), in order to compute $\mathcal{H}_2$ pseudo-optimal output residue directions by (4.99) on the one hand, and on the other hand $\mathcal{H}_2$ pseudo-optimal input residue directions by (4.100).

The basic procedure of the modified IRKA algorithm with inner and outer loop

Algorithm 4.4 IRKA with inner loop for residue correction

```
Input: $\mathbf{E}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and reduced order $n$
Output: locally $\mathcal{H}_2$ optimal reduced model $\boldsymbol{G}_r(s)=\mathbf{C}_r\,(s\mathbf{E}_r-\mathbf{A}_r)^{-1}\,\mathbf{B}_r$ of order $n$
 1: Make initial choice of the set $\{s_1,\dots,s_n\}$, that is closed under conjugation; select
    $\mathbf{b}_i^*\in\mathbb{C}^m$ and $\mathbf{c}_i\in\mathbb{C}^p$, that satisfy $\mathbf{b}_i=\overline{\mathbf{b}}_j$ and $\mathbf{c}_i=\overline{\mathbf{c}}_j$ if $s_i=\overline{s}_j$.
 2: $\mathbf{V}=\left[(\mathbf{A}-s_1\mathbf{E})^{-1}\,\mathbf{B}\mathbf{b}_1^*,\ \dots,\ (\mathbf{A}-s_n\mathbf{E})^{-1}\,\mathbf{B}\mathbf{b}_n^*\right]$
 3: $\mathbf{W}=\left[\left(\mathbf{A}^T-\overline{s}_1\mathbf{E}^T\right)^{-1}\mathbf{C}^T\mathbf{c}_1,\ \dots,\ \left(\mathbf{A}^T-\overline{s}_n\mathbf{E}^T\right)^{-1}\mathbf{C}^T\mathbf{c}_n\right]$
 4: $\mathbf{E}_r=\mathbf{W}^*\mathbf{E}\mathbf{V}$, $\mathbf{A}_r=\mathbf{W}^*\mathbf{A}\mathbf{V}$, $\mathbf{B}_r=\mathbf{W}^*\mathbf{B}$ and $\mathbf{C}_r=\mathbf{C}\mathbf{V}$
 5: repeat
 6:     Compute eigenvalue decomposition $\mathbf{E}_r^{-1}\mathbf{A}_r=\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$, with $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\dots,\lambda_n)$
 7:     Assign $s_i=-\overline{\lambda}_i$, $[\mathbf{b}_1^*,\dots,\mathbf{b}_n^*]^*=\mathbf{U}^{-1}\mathbf{E}_r^{-1}\mathbf{B}_r$ and $[\mathbf{c}_1,\dots,\mathbf{c}_n]=\mathbf{C}_r\mathbf{U}$
 8:     $\mathbf{S}=\mathrm{diag}(s_1,\dots,s_n)$
 9:     for $1$ to $n$ do
10:         $\mathbf{V}_i=(\mathbf{A}-s_i\mathbf{E})^{-1}\,\mathbf{B}$,    $\mathbf{W}_i=\left(\mathbf{A}^T-\overline{s}_i\mathbf{E}^T\right)^{-1}\mathbf{C}^T$
11:     end for
12:     repeat
13:         for $1$ to $n$ do
14:             $\mathbf{b}_i=\frac{\mathbf{b}_i}{\|\mathbf{b}_i\|_2}$
15:         end for
16:         $\mathbf{V}=[\mathbf{V}_1\mathbf{b}_1^*,\ \dots,\ \mathbf{V}_n\mathbf{b}_n^*]$,    $\mathbf{L}=[\mathbf{b}_1^*,\ \dots,\ \mathbf{b}_n^*]$,
17:         $\mathbf{P}_r^{-1}=\mathrm{lyap}(\mathbf{S}^*,-\mathbf{L}^*\mathbf{L})$
18:         $[\mathbf{c}_1,\dots,\mathbf{c}_n]=-\mathbf{C}\mathbf{V}\,(\mathbf{P}_r^{-1})^{-1}$
19:         for $1$ to $n$ do
20:             $\mathbf{c}_i=\frac{\mathbf{c}_i}{\|\mathbf{c}_i\|_2}$
21:         end for
22:         $\mathbf{W}=[\mathbf{W}_1\mathbf{c}_1,\ \dots,\ \mathbf{W}_n\mathbf{c}_n]$,    $\mathbf{L}_{\mathrm{W}}^*=[\mathbf{c}_1,\ \dots,\ \mathbf{c}_n]$,
23:         $\mathbf{Q}_r^{-1}=\mathrm{lyap}(\mathbf{S},-\mathbf{L}_{\mathrm{W}}\mathbf{L}_{\mathrm{W}}^*)$
24:         $[\mathbf{b}_1^*,\dots,\mathbf{b}_n^*]^*=-\,(\mathbf{Q}_r^{-1})^{-1}\,\mathbf{W}^*\mathbf{B}$
25:     until converged
26:     $\mathbf{E}_r=\mathbf{W}^*\mathbf{E}\mathbf{V}$, $\mathbf{A}_r=\mathbf{W}^*\mathbf{A}\mathbf{V}$, $\mathbf{B}_r=\mathbf{W}^*\mathbf{B}$ and $\mathbf{C}_r=\mathbf{C}\mathbf{V}$
27: until converged
```

is shown in Algorithm 4.4. It should be highlighted, that the reduced model only depends on the *directions* of the input and output residues, and hence, we are allowed to normalize them in Steps 14 and 20 of the algorithm for a better numerical conditioning. The main numerical effort in Algorithm 4.4 remains the computation of the Krylov blocks in Step 10, whereas the inner loop, described by Steps 12–25, is based on mainly small-scale operations. The motivation for Algorithm 4.4 is as follows: although a single iteration of the outer loop requires higher numerical effort compared to one iteration of standard IRKA as in Algorithm 4.1, one hopes that Algorithm 4.4 requires less iterations of the outer loop for convergence, due to the optimized residue directions in the inner loop. Then it may happen, that the total effort for convergence is decreased by the modified IRKA in Algorithm 4.4. Finally, it should be stressed, that Algorithm 4.4 only describes the principle course of action, and it is not intended for direct numerical implementation; one would e. g. use real bases of $\mathbf{V}$ and $\mathbf{W}$ for complex conjugated interpolation points and tangential directions.

Wilson [204] was probably the first one, who proposed the mentioned inner loop, to optimize residue directions of MIMO models for given reduced eigenvalues. His algorithm, however, was not yet applicable in a large-scale setting. Only recently, Beattie and Gugercin [24] reinvented the idea. Although their inner loop is stated in the context of the Loewner framework, it is directly applicable to original models in state-space representations of the form (1.1). The algorithm in [24], however, uses Cauchy matrices, which may be "poorly conditioned". In this respect, Algorithm 4.4 may be seen as the state-space counterpart of the algorithm presented in [24], which in turn is based on transfer functions. It was also Beattie and Gugercin who named the inner loop the "residue correction" step, and furthermore, who presented promising numerical examples, that verified that the modified IRKA in Algorithm 4.4 can indeed outperform a standard implementation like in Algorithm 4.1. As the approach in [24] is conceptually equal to the one pursued here, we refrain from also presenting numerical examples.

4.3.7 $\mathcal{H}_2$ Pseudo-Optimality and the Cumulative Framework

The concept of $\mathcal{H}_2$ pseudo-optimality is particularly beneficial in the cumulative framework CURE. This is due to condition *vii)* of Theorem 4.26: the feed-through model $\boldsymbol{G}_f(s)$ becomes an all-pass in the $\mathcal{H}_2$ pseudo-optimal case. Consider e. g. a SISO model, then $\boldsymbol{G}_f(s)$ generates a 0 dB line in the magnitude plot, and hence, all dynamics are shifted to $\boldsymbol{G}_\perp(s)$. This is essential, as then $\boldsymbol{G}_\perp(s)$ may be subsequently reduced in the next iteration of CURE, so that all dynamics of the error are available for reduction. Instead, if $\boldsymbol{G}_f(s)$ is not all-pass, it might contain dynamics that are important for the reduced model. But since the reduced model is the cumulated reduction of the $\boldsymbol{G}_{\perp,i}(s)$, it proves hard to generate these dynamics of $\boldsymbol{G}_f(s)$ in the reduced model in subsequent iterations of CURE.

It is therefore reasonable to reduce $\mathcal{H}_2$ pseudo-optimally in each iteration of CURE. But in order to guarantee $\mathcal{H}_2$ pseudo-optimality in each iteration, we yet still have to prove that the cumulated reduced model stays $\mathcal{H}_2$ pseudo-optimal, if each individually reduced models have been $\mathcal{H}_2$ pseudo-optimal. This is clarified in the next lemma.

Lemma 4.36. *Let all variables be as defined in Corollary 3.9, and assume that the columns of each* $\mathbf{V}_i$ *form the basis of recursively computed rational input Krylov subspaces, i. e. each* $\mathbf{V}_i$ *satisfies (3.55). Further assume that each reduced model* $\boldsymbol{G}_{r,i}(s)$ *is input* $\mathcal{H}_2$ *pseudo-optimal, such that the conditions of Theorem 4.26 hold. Then the accumulated reduced model* $\boldsymbol{G}_{r,\text{tot}}(s)$ *is also input* $\mathcal{H}_2$ *pseudo-optimal, and the conditions of Theorem 4.26 also hold for the accumulated data.*

Proof. The proof is done by induction. The case $i = 1$ is trivial, as $\boldsymbol{G}_{r,1}(s)$ is input $\mathcal{H}_2$ pseudo-optimal due to the assumptions made. Then assume that the total reduced model $\boldsymbol{G}_{r,\text{tot}}(s)$ at step $i-1$ is input $\mathcal{H}_2$ pseudo-optimal and that the conditions of Theorem 4.26 hold for the cumulated data. It follows from Lemma 5.9 that $\mathbf{P}_{r,\text{tot}}$ can be recursively computed by

$$\mathbf{P}_{r,\text{tot}} \leftarrow \begin{bmatrix} \mathbf{P}_{r,\text{tot}} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{bmatrix} \tag{4.101}$$

where $\mathbf{P}_{12}$ and $\mathbf{P}_{22}$ are obtained from (5.29) and (5.30). Due to condition *ii)* of Theorem 4.26, we can substitute $\mathbf{E}_{r,\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{L}_{\text{tot}}^* = -\mathbf{B}_{r,\text{tot}}$ in (5.29), which yields

$$\mathbf{A}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{E}_{r,i}^* + \mathbf{E}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{A}_{r,i}^* = \mathbf{0}. \tag{4.102}$$

As both the total reduced model and the reduced model at the subsequent step i are assumed $\mathcal{H}_2$ pseudo-optimal, they are asymptotically stable and hence, the solution $\mathbf{P}_{12}$ of the above Sylvester equation exists and is unique, and we can identify $\mathbf{P}_{12} = \mathbf{0}$. Substituting this in (5.30) directly yields $\mathbf{P}_{22} = \mathbf{P}_{r,i}$. Then we have the following recursive formula for $\mathbf{P}_{r,\text{tot}}$, if each reduced model in CURE is input $\mathcal{H}_2$ pseudo-optimal:

$$\mathbf{P}_{r,\text{tot}} \leftarrow \begin{bmatrix} \mathbf{P}_{r,\text{tot}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{r,i} \end{bmatrix}. \tag{4.103}$$

It then can be readily verified with equations (4.103), (3.49) and (3.50) that condition *ii)* of Theorem 4.26 is satisfied also for the cumulated reduced model after step i, and consequently, that this model is input $\mathcal{H}_2$ pseudo-optimal due to Theorem 4.27. □

We are now ready to state the probably most important benefit of $\mathcal{H}_2$ pseudo-optimal MOR in the CURE framework; this was also proven by Panzer in his thesis [148].

Theorem 4.37. *Let all variables be as defined in Corollary 3.9, and assume that the columns of each* $\mathbf{V}_i$ *form the basis of recursively computed rational input Krylov subspaces, i. e. each* $\mathbf{V}_i$ *satisfies (3.55). Further assume that each reduced model* $\boldsymbol{G}_{r,i}(s)$ *is input* $\mathcal{H}_2$ *pseudo-optimal, such that the conditions of Theorem 4.26 hold. Then, the*

$\mathcal{H}_2$ norm of the error, $\|\boldsymbol{G}-\boldsymbol{G}_{r,\text{tot}}\|_{\mathcal{H}_2}$, decreases monotonically with each iteration of CURE. Moreover, if $\|\boldsymbol{G}_{r,i}\|_{\mathcal{H}_2} \neq 0$, $\forall i$, then the $\mathcal{H}_2$ norm of the error decreases strictly monotonically in each iteration of CURE.

Proof. Due to Lemma 4.36, the total reduced model $\boldsymbol{G}_{r,\text{tot}}(s)$ stays $\mathcal{H}_2$ pseudo-optimal, if each each reduced model $\boldsymbol{G}_{r,i}(s)$ already is. Then it follows from Theorem 4.35, that $\|\boldsymbol{G}_{r,\text{tot}}\|_{\mathcal{H}_2} \leq \|\boldsymbol{G}\|_{\mathcal{H}_2}$. It is therefore left to show, that $\|\boldsymbol{G}_{r,\text{tot}}\|_{\mathcal{H}_2}$ cannot decrease in each iteration of CURE, because this would ensure that the error monotonically decreases. To this end, consider $\mathbf{C}_{r,\text{tot}} = \mathbf{C}\,[\mathbf{V}_1, \ldots, \mathbf{V}_k]$, which can be recursively formulated as $\mathbf{C}_{r,\text{tot}} \leftarrow [\mathbf{C}_{r,\text{tot}},\ \mathbf{C}_{r,i}]$. Together with the recursive definition of $\mathbf{P}_{r,\text{tot}}$ in (4.103), and the computation of $\|\boldsymbol{G}_{r,\text{tot}}\|_{\mathcal{H}_2}$ by (4.34), i. e. $\|\boldsymbol{G}_{r,\text{tot}}\|^2_{\mathcal{H}_2} = \text{trace}\left(\mathbf{C}_{r,\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{C}^*_{r,\text{tot}}\right)$, the statement can be concluded. □

Remark 4.38. It should be noted that the statement of Theorem 4.37 is irrespective of the choice of interpolation points s_i. That means that, if the reduced model is computed by PORK in each iteration of CURE, the $\mathcal{H}_2$ error is guaranteed to decrease—no matter which interpolation points s_i and tangential directions $\mathbf{L}_i$ are plugged into PORK (to be precise: in fact poles and transmission zeros of the original model have to be excluded, which can be avoided in practical applications). In conclusion, although CURE combined with PORK does not at all ensure that one "does the right thing"—as this heavily depends of the choice of s_i and $\mathbf{L}_i$—, one at least "cannot do wrong" within this framework.

4.3.8 Discussion of $\mathcal{H}_2$ Pseudo-Optimality

To recap the findings thus far, if the original model is large-scale, then the projection onto rational Krylov subspaces is one of the major tools for MOR. Instead of directly generating the reduced model in a single projection step, the recursive error factorization discussed in Chapter 3 permits any desired number of individual, decoupled reduction steps and, furthermore, provides easy-to-implement recursive formulae to gather these reduced models in an accumulated one. The idea of this framework, denoted as CURE, may be illustrated as "salami slicing" or "divide and conquer", and its mathematical description reads as:

$$\boldsymbol{G}(s) = \boldsymbol{G}_{r,\text{tot}}(s) + \boldsymbol{G}_{\perp,i}(s)\boldsymbol{G}_{f,\text{tot}}(s) \tag{4.104}$$

Subsequently, it was shown that it is advisable to reduce $\mathcal{H}_2$ pseudo-optimally in each iteration (which also includes local $\mathcal{H}_2$ optimal reduction). This has several benefits: on the one hand, $\boldsymbol{G}_{f,\text{tot}}(s)$ becomes all-pass, and due to (4.93), the remaining dynamics of the error stay in $\boldsymbol{G}_{\perp,i}(s)$—which in turn may be reduced in a successive iteration of CURE. On the other hand, an $\mathcal{H}_2$ pseudo-optimal reduction in each iteration ensures that the error measured in the $\mathcal{H}_2$ norm decreases. It is notable that $\mathcal{H}_2$ pseudo-optimal reductions in CURE (with non-trivial reduced models) guarantee strictly monotonically decreasing error without making any assumption on the choice of interpolation points. This whole framework is available for general MIMO models, and the only large-scale operation is the calculation of bases of Krylov subspaces, and therefore, it can be considered numerically efficient.

In this respect, CURE combined with PORK solves many problems in large-scale model order reduction: stability of the reduced model is guaranteed, a proper choice of $\mathbf{W}$ is made, the feed-through model $\boldsymbol{G}_f(s)$ becomes all-pass, the reduced order can be adaptively selected, and a strictly monotonically decrease in the error may be secured. The only drawback is that the reduced model now depends twice as much on the proper choice of interpolation points (and in the MIMO case also tangential directions), because these also become the mirror images of the reduced poles (and the reduced input or output residues). This fact may be illustrated as "all problems are shifted to the shifts". Hence, it is indispensable in $\mathcal{H}_2$ pseudo-optimal reduction to be sure of having properly selected interpolation points. This issue would most appropriately be tackled with some kind of optimization, which, however, is out of the scope of this thesis; the interested reader is instead referred to the thesis of Panzer [148]. Nevertheless, the benefits of $\mathcal{H}_2$ pseudo-optimality for this optimization will be briefly pointed out in the next paragraph.

On the one hand, $\mathcal{H}_2$ pseudo-optimality can be exploited a posteriori: an obvious approach would be to run IRKA in each iteration of CURE. The drawback of this procedure is that local $\mathcal{H}_2$ optimality is generally not preserved in the accumulated reduced model; only $\mathcal{H}_2$ pseudo-optimality is maintained. The advantage, however, is that in combination with CURE, IRKA may be executed with arbitrary small reduced orders. It is to say that convergence of IRKA typically is faster the smaller the reduced order is. Furthermore, an entire restart would be required in IRKA without CURE, if the reduced order was too small for sufficient approximation. Nevertheless, combining IRKA with CURE, may cause unexpected trouble. This is due to the fixed point iteration in IRKA: the algorithm requires a large number of iterations to yield an analytically exact local minimizer of the $\mathcal{H}_2$ error; in practice, one always has to abort the iteration after a certain convergence tolerance is achieved. This is even deteriorated by numerical round-off errors. As a consequence, the outcome of IRKA can only be close to local $\mathcal{H}_2$ optimality and hence, is not even $\mathcal{H}_2$ pseudo-optimal. This would not be problematic if the outcome of IRKA was the final reduced model; in combination with

CURE, however, this means that monotonic decrease of the error is strictly speaking lost. Furthermore, $\boldsymbol{G}_f(s)$ is not all-pass, and hence, potentially important dynamics withdraw from being available for subsequent iterations of CURE. Although these errors might be small at first sight, by performing many iterations of CURE, they sum up and may considerably deteriorate the quality of the total reduced model. To avoid this, one can take the eigenvalues λ_i of the reduced model (and in the MIMO case also input or output residues) after IRKA was aborted, and plug a $\mathbf{V}$ whose columns form a basis of the rational Krylov subspace with their mirror images, $-\overline{\lambda}_i$, as expansion points into PORK. This would only marginally alter the reduced dynamics, but $\mathcal{H}_2$ pseudo-optimality is guaranteed, which increases robustness against the convergence tolerance in IRKA. Gilbert confirmed in his thesis [83, p. 61] that small perturbations on locally $\mathcal{H}_2$ optimal reduced models are not critical:

> *"Does an approximate solution of the pole optimization problem appreciably alter approximation accuracy? For numerous practical examples the answer is fortunately no. Pole position can be changed rather drastically in certain directions with little effect on system response (provided, of course, that the approximation coefficients are always chosen for minimum error [Editor: corresponds to $\mathcal{H}_2$ pseudo-optimality]). However, in some cases the tolerance to pole position shifts may be poor or the approximate poles may be poorly chosen. In these cases, the additional error can always be corrected by the addition of more terms (i. e., more poles) [Editor: this would correspond to the next iteration of CURE] in the approximating series."*

Although the CURE framework has not been available to Gilbert in the presented form, he in fact promotes its basic idea in combination with $\mathcal{H}_2$ pseudo-optimality and his statement may perhaps be translated into the framework here as follows: it is suggested to undertake a final step of PORK, after IRKA has been aborted in each iteration of CURE, because the advantages of this approach outweigh the drawbacks. Finally it should be noted, that the final execution of PORK does not even increase the numerical effort considerably, as the $\mathbf{V}$, which is required in PORK, is already available from the last iteration of IRKA.

On the other hand, $\mathcal{H}_2$ pseudo-optimality can also be exploited a priori: the idea is to search for locally $\mathcal{H}_2$ optimal reduced models only in the subset of $\mathcal{H}_2$ pseudo-optimal ones. Due to Lemma 4.33 and Thales' theorem, this may be illustrated as the search "on a semi-circle instead of in the whole plane". To be precise: once we have a set of reduced poles (and in the MIMO case also input or output residues) it is suggested to compute the $\mathcal{H}_2$ pseudo-optimal reduced model with PORK; subsequently, an optimization is required that finds an improved set of reduced poles (and in the MIMO case also input or output residues). This optimization could be based on gradients and maybe also Hessians. The very idea of this approach was already suggested by Wilson [203]. His evaluation of the gradient, however, required the knowledge of the Gramian of the original model, which is unfavourable in large-scale settings. Various authors revisited the idea of gradient based optimization since Wilson, which was already addressed right before Section 4.3. However, it seems like only very recently Panzer et al. [149] brought back $\mathcal{H}_2$ pseudo-optimality in a Krylov-based projection framework for large-scale models. As this is a complex subject, we will not dig deeper into it, and refer the interested reader to the thesis of Panzer [148], who fruitfully exploited in his approach, CURE, PORK, and also the maximization of $\|\boldsymbol{G}_r\|_{\mathcal{H}_2}$ as the optimization objective.

Apart from the just mentioned optimization, also smaller applications of $\mathcal{H}_2$ pseudo-optimality are imaginable, which shall be briefly reviewed here. Any of the conditions of Theorem 4.26 may be used as a convergence criterion in IRKA. Especially condition *iii)* suggests itself by providing with $\|\mathbf{E}_r^{-1}\mathbf{B}_r+\mathbf{P}_r\mathbf{L}^*\|_2$ a distance to a locally $\mathcal{H}_2$ optimal reduced model. This idea was already presented in [208], where also an alternative update scheme for IRKA was suggested. This modification can improve convergence of IRKA, and furthermore, can cause IRKA to yield better local minima after restart, both of which were verified by numerical examples. Although the suggested modification was motivated only heuristically, maybe the work of Krajewski and Viaro [121] can give hints for its rigorous convergence analysis.

Another application is the work of Panzer et al. [150], who introduced rigorous upper bounds on the approximation error, for which in turn $\mathcal{H}_2$ pseudo-optimality is also beneficial: two bounds—one on the $\mathcal{H}_2$ norm and one on the $\mathcal{H}_\infty$ norm—were suggested, both of which include the factor $\|\boldsymbol{G}_f\|_{\mathcal{H}_\infty}$; if the reduced model is computed by PORK, then $\boldsymbol{G}_f(s)$ is all-pass, and hence, no additional overestimation due to $\|\boldsymbol{G}_f\|_{\mathcal{H}_\infty}$ is introduced.

4.3.9 Overview on $\mathcal{H}_2$ Pseudo-Optimality

The basic idea of $\mathcal{H}_2$ pseudo-optimality is to find the optimal residues for fixed reduced poles and was proven and used by many researchers. In the context of approximating rational functions, the result was stated by Walsh [201] and Gaier [74]. Gilbert [83] was probably the first one to employ the result in the context of dynamical systems. He also presented a constructive approach to compute $\mathcal{H}_2$ pseudo-optimal reduced models, that required the solution of one LSE. Meier and Luenberger [137] already knew that the conditions for $\mathcal{H}_2$ pseudo-optimality are necessary and sufficient and could also derive (4.95). $\mathcal{H}_2$ pseudo-optimality (and local $\mathcal{H}_2$ optimality) in the context of signal processing was covered by McDonough and Huggins [135]. They elaborately exploited orthogonality and it seems appealing to investigate, if their ideas may be transferred to the presented large-scale setting. Wilson [203] suggested a constructive way to compute $\mathcal{H}_2$ pseudo-optimal reduced models, which relies on either the controllability or the observability canonical form of the reduced model. Later in [204], he proposed an iterative procedure to compute MIMO reduced models that are simultaneously input *and* output $\mathcal{H}_2$ pseudo-optimal, similar to what was presented in Section 4.3.6. Riggs and Edgar [161] generalized the conditions for $\mathcal{H}_2$ pseudo-optimality to, amongst others, finite time intervals, time delay systems, or higher order poles. Kimura [117] observed that in the $\mathcal{H}_2$ pseudo-optimal case one may equivalently maximize the $\mathcal{H}_2$ norm of the reduced model, instead of minimizing the $\mathcal{H}_2$ norm of the error. He, however, could not make a suggestion, how this can be exploited to optimize pole locations of the reduced model. The computation of $\mathcal{H}_2$ pseudo-optimal reduced models within given linear constraints was treated by Vilbé and Calvez [193]. This e. g. allows to find the $\mathcal{H}_2$ pseudo-optimal reduced model with a prescribed steady-state gain. The numerical implementation of this approach was improved first by Therapos [184], and then again by Vilbé et al. [195], through avoiding matrix inversion. $\mathcal{H}_2$ pseudo-optimal approximation was employed by Vilbé et al. [194] for suboptimal reduced poles, that result from computing time derivatives (and integrals) of the impulse response $\boldsymbol{G}(t)$. Spanos et al. [179] suggested an optimization algorithm that combined $\mathcal{H}_2$ pseudo-optimality with a line search for the poles, which is based on gradients. Easy-to-implement constructive algorithms to compute $\mathcal{H}_2$ pseudo-optimal reduced models were proposed by Lucas for continuous time [133] and discrete time models [132]. Gugercin [89] proved $\mathcal{H}_2$ pseudo-optimality of the IRKA-like ISRK algorithm, which was adapted by Gugercin et al. [94], in order to preserve a port-Hamiltonian structure in the reduced model. The first work to propose the necessary and sufficient interpolatory condition for MIMO, i. e. either input or output, $\mathcal{H}_2$ pseudo-optimality, as in Theorems 4.19 and 4.20, was due to Beattie and Gugercin [24]. They also suggested the inner IRKA loop, but only in the context of the Loewner framework, which seems to be numerically ill-conditioned in the Krylov-based projective model order reduction. For the block-case, i. e. that each reduced eigenvalue has geometric multiplicity m, the conditions for MIMO $\mathcal{H}_2$ pseudo-optimality were presented by Wolf and Panzer [213]. As already mentioned, none of the available literature suggested a constructive way to compute $\mathcal{H}_2$ pseudo-optimal reduced models while preserving the projection onto a $\mathbf{V}$, whose columns span the corresponding rational Krylov subspace, like it is possible with PORK. This was introduced by Wolf et al. [208], whereas this thesis represents its comprehensive analysis.

Finally, it is to say that there are two methods to solve large-scale Lyapunov (or Sylvester) equations, which, in fact, compute $\mathcal{H}_2$ pseudo-optimal approximations: the *alternating direction implicit* (ADI) iteration, see e. g. [27, 33, 128], and the approach by Ahmad et al. [3]. Although both of these methods were motivated completely differently, it can be shown that they can actually be interpreted as $\mathcal{H}_2$ pseudo-optimal approximations. As their inherent optimality property has not been recognized so far, this will be covered in the final part of this thesis.

Part III

Application: Large-Scale Lyapunov Equations

5 Approximate Solutions based on Rational Krylov Subspaces

The results obtained in Part II will be applied in this chapter to approximate the solution of large-scale Lyapunov equations in the form

$$\mathbf{APE}^T + \mathbf{EPA}^T + \mathbf{BB}^T = \mathbf{0}. \tag{5.1}$$

Because we assume that $\mathbf{E}$ is non-singular and that the eigenvalues of $\mathbf{E}^{-1}\mathbf{A}$ are contained in the open left half of the complex plane, the solution $\mathbf{P} = \mathbf{P}^T$ of (5.1) exists and is unique. We further assume that the pair $(\mathbf{E}^{-1}\mathbf{A}, \mathbf{E}^{-1}\mathbf{B})$ is controllable, which ensures that $\mathbf{P}$ is positive definite. For details on existence and uniqueness of the solution of (5.1), please refer to e. g. [10].

It was already mentioned in Section 1.5.2, that the main effort in balanced truncation is to find a low-rank Cholesky factor $\mathbf{Z}$, such that the low-rank approximation $\widehat{\mathbf{P}} = \mathbf{ZZ}^*$ satisfies $\widehat{\mathbf{P}} \approx \mathbf{P}$. However, in the following we employ the more general form $\widehat{\mathbf{P}} = \mathbf{VP}_r\mathbf{V}^*$, where the columns of $\mathbf{V} \in \mathbb{C}^{N \times n}$ span an appropriate subspace, and where $\mathbf{P}_r$ is the reduced solution. If $\mathbf{P}_r$ is symmetric, positive definite, then the low-rank Cholesky factor $\mathbf{Z}$ can be readily deduced from computing the Cholesky factorization, $\mathbf{P}_r = \mathbf{R}_r\mathbf{R}_r^*$, because then obviously $\mathbf{Z} = \mathbf{VR}_r$. The advantage of the formulation $\widehat{\mathbf{P}} = \mathbf{VP}_r\mathbf{V}^*$ is that the search for an approximate solution $\widehat{\mathbf{P}}$ can be split into two independent parts:

- the search for a subspace span($\mathbf{V}$),
- and the search for a reduced Lyapunov solution $\mathbf{P}_r$.

Concerning the first part, apparently bases of rational Krylov subspaces will be used, whereas the latter part will be extensively discussed in the following sections.

Numerous approaches to compute suitable $\widehat{\mathbf{P}}$ have been proposed by various researchers over the past decades, see e. g. the surveys [28, 48, 175]. In what follows, we try to adapt the ideas of the previous chapters to find a suitable $\widehat{\mathbf{P}}$. In Section 5.1, the basic idea of approximate solutions by projections with rational Krylov subspaces is reviewed. This approach unfortunately has some drawbacks, which are discussed in Section 5.2. Nevertheless, it seems reasonable to still pursue this approach in a large-scale setting. Section 5.3 not only presents how the method can be extended with the cumulative idea, but also how one can benefit from $\mathcal{H}_2$ pseudo-optimality in this context. In Section 5.4, the resulting approach will be shown to be equivalent to both the alternating directions implicit (ADI) iteration and the method of Ahmad et al. [3]. Preliminary versions of these contributions have been published in [207, 210, 213, 214].

5.1 Rational Krylov Subspace Method (RKSM)

First of all, the *rational Krylov subspace method* (RKSM) for computing $\widehat{\mathbf{P}}$ is reviewed.

5.1.1 Approximate Solution by RKSM

Let the columns of $\mathbf{V}$ form a basis of a rational input Krylov subspace, and let $\mathbf{W}$ be arbitrary but such the solution $\mathbf{P}_r$ of

$$\mathbf{A}_r\mathbf{P}_r\mathbf{E}_r^T + \mathbf{E}_r\mathbf{P}_r\mathbf{A}_r^T + \mathbf{B}_r\mathbf{B}_r^T = \mathbf{0} \tag{5.2}$$

exists and is unique; then $\widehat{\mathbf{P}} = \mathbf{V}\mathbf{P}_r\mathbf{V}^*$ is called the approximate solution of (5.1) by RKSM. It should be noted, that RKSM was introduced in [53] with a Galerkin projection $\mathbf{W}=\mathbf{V}$, but as the basic procedure is left unchanged, we still refer to the generalized method $\mathbf{W}\neq\mathbf{V}$ as RKSM. It should be further noted, that a change of basis does not change the approximation, which is stated in the next lemma.

Lemma 5.1. *Let the columns of* $\mathbf{V}_1$ *and of* $\mathbf{V}_2$ *form two different bases of the same subspace,* $\operatorname{span}(\mathbf{V}_1) = \operatorname{span}(\mathbf{V}_2)$*, and correspondingly let* $\mathbf{W}_1$ *and* $\mathbf{W}_2$ *be such that* $\operatorname{span}(\mathbf{W}_1)=\operatorname{span}(\mathbf{W}_2)$*. Given that* $\mathbf{P}_{r,1}$ *and* $\mathbf{P}_{r,2}$ *satisfy the respective reduced Lyapunov equations (5.2), the resulting low-rank approximations are equal:* $\widehat{\mathbf{P}}_1 = \mathbf{V}_1\mathbf{P}_{r,1}\mathbf{V}_1^* = \mathbf{V}_2\mathbf{P}_{r,2}\mathbf{V}_2^*=\widehat{\mathbf{P}}_2$*.*

Proof. Because of $\operatorname{span}(\mathbf{V}_1)=\operatorname{span}(\mathbf{V}_2)$, there exists a non-singular matrix $\mathbf{T}_\mathbf{V}\in\mathbb{C}^{n\times n}$, such that $\mathbf{V}_2=\mathbf{V}_1\mathbf{T}_\mathbf{V}$, and consequently, there is also a $\mathbf{T}_\mathbf{W}\in\mathbb{C}^{n\times n}$, such that $\mathbf{W}_2=\mathbf{W}_1\mathbf{T}_\mathbf{W}$. Substituting this in $\mathbf{A}_{r,2}\mathbf{P}_{r,2}\mathbf{E}_{r,2}^T+\mathbf{E}_{r,2}\mathbf{P}_{r,2}\mathbf{A}_{r,2}^T+\mathbf{B}_{r,2}\mathbf{B}_{r,2}^T=\mathbf{0}$ yields

$$\mathbf{T}_\mathbf{W}^T\mathbf{A}_{r,1}\mathbf{T}_\mathbf{V}\mathbf{P}_{r,2}\mathbf{T}_\mathbf{V}^T\mathbf{E}_{r,1}^T\mathbf{T}_\mathbf{W} + \mathbf{T}_\mathbf{W}^T\mathbf{E}_{r,1}\mathbf{T}_\mathbf{V}\mathbf{P}_{r,2}\mathbf{T}_\mathbf{V}^T\mathbf{A}_{r,1}^T\mathbf{T}_\mathbf{W} + \mathbf{T}_\mathbf{W}^T\mathbf{B}_{r,1}\mathbf{B}_{r,1}^T\mathbf{T}_\mathbf{W} = \mathbf{0}. \tag{5.3}$$

Then $\mathbf{T}_\mathbf{W}$ can be cancelled and due to uniqueness, $\mathbf{P}_{r,1}=\mathbf{T}_\mathbf{V}\mathbf{P}_{r,2}\mathbf{T}_\mathbf{V}^T$ can be identified, which is equivalent to $\mathbf{P}_{r,2} = \mathbf{T}_{\mathbf{V}}^{-1}\mathbf{P}_{r,1}\mathbf{T}_{\mathbf{V}}^{-T}$. Then, $\widehat{\mathbf{P}}_1 = \mathbf{V}_1\mathbf{T}_{\mathbf{V}}\mathbf{T}_{\mathbf{V}}^{-1}\mathbf{P}_{r,1}\mathbf{T}_{\mathbf{V}}^{-T}\mathbf{T}_{\mathbf{V}}^{T}\mathbf{V}_1^T = \mathbf{V}_2\mathbf{P}_{r,2}\mathbf{V}_2^T = \widehat{\mathbf{P}}_2$, which completes the proof. □

Lemma 5.1 shows that only the subspaces spanned by the columns of $\mathbf{V}$ and $\mathbf{W}$ affect the approximate solution $\widehat{\mathbf{P}}$, whereas the chosen bases are irrelevant. For a given $\mathbf{V}$ that spans a rational input Krylov subspace, the remaining degree of freedom in $\widehat{\mathbf{P}}$ is the choice of the subspace spanned by the columns of $\mathbf{W}$—or equivalently the selection of one reduced model out of the family $\boldsymbol{G}_{\mathbf{F}}(s)$. Before we dig deeper into this, let us first analyse the residual in the original Lyapunov equation.

5.1.2 The Residual in RKSM

The residual $\mathbf{R} \in \mathbb{C}^{N\times N}$ that follows from an arbitrary approximate solution $\widehat{\mathbf{P}} \in \mathbb{C}^{N\times N}$ is generally defined as

$$\mathbf{R} = \mathbf{A}\widehat{\mathbf{P}}\mathbf{E}^T + \mathbf{E}\widehat{\mathbf{P}}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T. \tag{5.4}$$

Proposition 5.2. *Let $\widehat{\mathbf{P}} = \mathbf{V}\mathbf{P}_r\mathbf{V}^*$, where $\mathbf{P}_r$ solves (5.2), then the residual satisfies the Petrov-Galerkin condition: $\mathbf{W}^*\mathbf{R}\mathbf{W} = \mathbf{0}$.*

Proof. The proof readily follows by multiplying (5.4) with $\mathbf{W}^*$ and $\mathbf{W}$ from the left and right, respectively. □

Even if the matrices $\mathbf{A}$, $\mathbf{E}$ and $\mathbf{B}$ are sparse, the residual $\mathbf{R}$ is generally dense, which makes it actually difficult to store and analyse $\mathbf{R}$. However, if $\widehat{\mathbf{P}}$ is computed by RKSM, then a convenient low-rank formulation of the residual can be computed with low numerical effort. This is stated in the next theorem.

Theorem 5.3. *Let the columns of $\mathbf{V}$ form a basis of a rational input Krylov subspace, and let $\mathbf{W}$ be arbitrary but such that the solution $\mathbf{P}_r$ of (5.2) exists and is unique. This particularly means that there exist $\mathbf{S}$ and $\mathbf{L}$ such that (2.15) is satisfied, and that $\mathbf{B}_\perp = \mathbf{B} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$. Define*

$$\mathbf{F} = \mathbf{E}\mathbf{V}\left(\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^*\right) \in \mathbb{C}^{N\times m}, \tag{5.5}$$

then the approximate solution $\widehat{\mathbf{P}} = \mathbf{V}\mathbf{P}_r\mathbf{V}^*$ yields the following residual:

$$\mathbf{R} = \begin{bmatrix} \mathbf{B}_\perp & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{B}_\perp^* \\ \mathbf{F}^* \end{bmatrix}. \tag{5.6}$$

Proof. Consider $\mathbf{B}=\mathbf{B}_\perp+\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$, then

$$\mathbf{B}\mathbf{B}^* = \mathbf{B}_\perp\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{V}^*\mathbf{E}^* + \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_\perp^* + \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{V}^*\mathbf{E}^*. \tag{5.7}$$

The residual is defined as

$$\mathbf{R} = \mathbf{A}\mathbf{V}\mathbf{P}_r\mathbf{V}^*\mathbf{E}^* + \mathbf{E}\mathbf{V}\mathbf{P}_r\mathbf{V}^*\mathbf{A}^* + \mathbf{B}\mathbf{B}^*. \tag{5.8}$$

Replacing $\mathbf{AV}$ by the Sylvester equation (2.39) leads to

$$\mathbf{R} = \mathbf{E}\mathbf{V}\left(\mathbf{E}_r^{-1}\mathbf{A}_r\mathbf{P}_r + \mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\right)\mathbf{V}^*\mathbf{E}^* + \mathbf{B}_\perp\mathbf{L}\mathbf{P}_r\mathbf{V}^*\mathbf{E}^* + \mathbf{E}\mathbf{V}\mathbf{P}_r\mathbf{L}^*\mathbf{B}_\perp^* + \mathbf{B}\mathbf{B}^*. \tag{5.9}$$

Substituting $\mathbf{E}_r^{-1}\mathbf{A}_r\mathbf{P}_r+\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}=-\mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*}$ and replacing $\mathbf{B}\mathbf{B}^*$ by (5.7) yields

$$\begin{aligned}\mathbf{R} &= \mathbf{B}_\perp\mathbf{L}\mathbf{P}_r\mathbf{V}^*\mathbf{E}^* + \mathbf{E}\mathbf{V}\mathbf{P}_r\mathbf{L}^*\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{V}^*\mathbf{E}^* + \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_\perp^* && (5.10)\\ &= \mathbf{B}_\perp\mathbf{B}_\perp^* + \mathbf{F}\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{F}^*. && (5.11)\end{aligned}$$

Expanding the formulation (5.6), leads to (5.11) and completes the proof. □

Theorem 5.3 shows that the rank of the residual $\mathbf{R}$ cannot exceed $2m$—irrespective of the reduce order n. The formulation (5.6) therefore massively reduces storage requirements compared to the traditional dense matrix that follows from (5.4). Additionally, common matrix norms of the residual $\mathbf{R}$—which are often used as a convergence criterion when iteratively approximating $\mathbf{P}$—can be easily computed with the low-rank formulation. The case of computing the Euclidean norm is presented in the next lemma.

Lemma 5.4. *The Euclidean norm $\|\mathbf{R}\|_2$ of the residual (5.6) can be calculated from the eigenvalues of a $2m\times 2m$ matrix,*

$$\|\mathbf{R}\|_2 = \max\left|\mathbf{\Lambda}\left(\begin{bmatrix}\mathbf{B}_\perp^T\mathbf{B}_\perp & \mathbf{B}_\perp^T\mathbf{F}\\ \mathbf{F}^T\mathbf{B}_\perp & \mathbf{F}^T\mathbf{F}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\right)\right|, \tag{5.12}$$

where $\mathbf{\Lambda}(\cdot)$ denotes the set of eigenvalues of a matrix.

Proof. Because $\mathbf{R}$ is symmetric, $\|\mathbf{R}\|_2 = \sqrt{\max\lambda\left(\mathbf{R}^2\right)} = \max\left|\lambda\left(\mathbf{R}\right)\right|$. For arbitrary matrices $\mathbf{M}$ and $\mathbf{N}$ of appropriate dimensions, $\mathbf{\Lambda}(\mathbf{MN})=\mathbf{\Lambda}(\mathbf{NM})$, and hence,

$$\mathbf{\Lambda}\left(\mathbf{R}\right) = \mathbf{\Lambda}\left(\begin{bmatrix}\mathbf{B}_\perp & \mathbf{F}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{B}_\perp^T\\ \mathbf{F}^T\end{bmatrix}\right) = \mathbf{\Lambda}\left(\begin{bmatrix}\mathbf{B}_\perp^T\\ \mathbf{F}^T\end{bmatrix}\begin{bmatrix}\mathbf{B}_\perp & \mathbf{F}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{bmatrix}\right), \tag{5.13}$$

which completes the proof. □

Lemma 5.4 shows that the Euclidean norm of $\mathbf{R}$ can be calculated even in a large-scale setting with marginal numerical effort: both $\mathbf{B}_\perp$ and $\mathbf{F}$ are found by mainly matrix vector products, and as we assume only a few inputs, the eigenvalue problem of dimension $2m{\times}2m$ may be easily solved. The low-rank formulation (5.6) of the residual is hence not only analytically convenient, but also numerically. Finally, the eigenvalues of the residual are studied in the next lemma.

Lemma 5.5. *Let the residual* $\mathbf{R}$ *be given by Theorem 5.3, and assume that* $\mathbf{R}$ *is real and that* $[\mathbf{B}_\perp,\ \mathbf{F}]$ *has full column rank. Then* $\mathbf{R}$ *has* m *positive and* m *negative eigenvalues.*

Proof. As $[\mathbf{B}_\perp,\ \mathbf{F}]$ has full column rank, let $\mathbf{X}_\perp \in \mathbb{R}^{N\times(N-2m)}$ denote its orthogonal complement and $\mathbf{I}$ the $m{\times}m$ identity matrix, then

$$\mathbf{R} = \mathbf{B}_\perp\mathbf{B}_\perp^* + \mathbf{F}\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{F}^* = [\mathbf{F},\ \ \mathbf{B}_\perp + \mathbf{F},\ \ \mathbf{X}_\perp\,] \begin{bmatrix} -\mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{F}^* \\ \mathbf{B}_\perp^* + \mathbf{F}^* \\ \mathbf{X}_\perp^* \end{bmatrix}. \tag{5.14}$$

Due to the assumptions, $[\mathbf{F},\ \mathbf{B}_\perp{+}\mathbf{F},\ \mathbf{X}_\perp]$ is real and non-singular, and hence, the proof follows from Sylvester's law of inertia, cf. [8, p. 189]. □

5.1.3 Iterative Procedure with RKSM

When applying RKSM to approximate $\mathbf{P}$, one typically performs an iterative procedure instead of computing $\widehat{\mathbf{P}}$ at once. The basic steps of such an approach can be pictured as follows:

- Compute $k \in \mathbb{N}^+$ new directions of a Krylov subspace: $\mathbf{V} \in \mathbb{R}^{N\times n} \rightarrow \mathbf{V} \in \mathbb{R}^{N\times(n+k)}$.
- Project equation (5.1) onto the subspace span($\mathbf{EV}$) and compute the solution $\mathbf{P}_r$ of the projected Lyapunov equation (5.2) using direct solvers.
- (Usually compute some norm of the residual in order to) evaluate an appropriate stopping criterion.
- If the desired accuracy is achieved, stop the algorithm; otherwise restart by calculating k additional directions.

We will see that the numerical effort of such an iterative approach may be reduced by the cumulative idea. Before going into this in Section 5.3, we will first of all review some literature and then also discuss the drawbacks of RKSM in Section 5.2.

5.1.4 Notes and References

Not only classical Krylov subspaces (1.31) have long been used to approximate the solution of (5.1), see e. g. [48, 102, 110, 111, 113, 166, 175, 176], but also extended

Krylov subspaces (1.32), cf. [100, 118, 174]. As the basic machinery has been available throughout these years, it is surprising that [53] seems to be the first reference that explicitly uses rational Krylov subspaces for projection, where also the label "RKSM" was introduced. An error analysis of RKSM is available in [25], and generalizations to the MIMO case can be found in [54]. For further reading, the recent surveys [28, 175] and references therein are recommended.

The residual $\mathbf{R}$ is often used as an indicator for the approximation of $\mathbf{P}$ by $\widehat{\mathbf{P}}$. As $\mathbf{R}$ is a dense matrix, one was typically computing its Frobenius norm or was using power methods to approximate the Euclidean norm, see e. g. [170, Section 4]. As stated above, the low-rank formulation of the residual from Theorem 5.3 eases this computation, which has been published in [214]. There it was also shown, how the low-rank formulation reads for the classical (1.31) and extended Krylov subspaces (1.32), and how $\|\mathbf{R}\|_2$ can be employed to upper bound $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_2$, by slightly generalizing the result of Hodel and Tenison [102]. In addition, it was discussed that a small Euclidean norm of the residual is neither necessary nor sufficient for a good approximation. This finding has been observed in numerical examples by Saak et al. [169]. The conclusion in [169] was then to suggest a particular "goal-oriented" convergence criterion if $\widehat{\mathbf{P}}$ is to be used for balanced truncation. The above mentioned bound on $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_2$ was improved in [150] in order to state a bound on the $\mathcal{H}_2$ error in model order reduction. This bound is also based on the error factorization discussed in Chapter 3, and additionally, a bound on the $\mathcal{H}_\infty$ error was proposed in [150].

5.2 The Dilemma of RKSM

This section discusses a drawback of RKSM, as the resulting approximation $\widehat{\mathbf{P}}$ of RKSM can to some extent not be as one might hope for. Basically, our objective is to approximate $\mathbf{P}$, such that the error $\mathbf{P}-\widehat{\mathbf{P}}$ is as small as possible. Similar to what was discussed in Chapter 4 for model order reduction, here again, one first has to define a suitable measure of the error—which in this case obviously should be a matrix norm. A self-evident choice would be the Euclidean norm $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_2$, but as $\mathbf{P}$ is generally dense, its analysis and computation are hard tasks in a large-scale setting. We will instead consider the Frobenius norm $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}$, because we will see that this is analytically more convenient, and also because a minimization of the Frobenius norm likewise minimizes the Euclidean norm due to $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}} \geq \left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_2$. The problem is accurately stated as follows.

Problem 5.1. Given the Lyapunov equation (5.1), we are searching for the approximate solution $\widehat{\mathbf{P}}$ of given rank n, which satisfies

$$\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}=\min_{\operatorname{rank}\left(\widetilde{\mathbf{P}}\right)=n}\left\|\mathbf{P}-\widetilde{\mathbf{P}}\right\|_{\mathrm{F}}. \tag{5.15}$$

The well-known solution of Problem 5.1 is obtained by retaining the n largest singular values of $\mathbf{P}$. This is often referred to as Eckart-Young-Mirsky theorem [56]. We of course do not know $\mathbf{P}$, and hence, it is impossible to compute its singular value decomposition. In practice one would instead only know an approximating subspace, which in the case here is a rational Krylov subspace $\operatorname{span}(\mathbf{V})$. As mentioned above, the approximate solution then reads as $\widehat{\mathbf{P}}=\mathbf{V}\mathbf{P}_r\mathbf{V}^*$, and consequently, we are merely searching for the optimal $\mathbf{P}_r$. This leads us again to some kind of pseudo-optimality, defined as follows.

Definition 5.1. Given a $\mathbf{V}\in\mathbb{C}^{N\times n}$ with full column rank, let $\mathbf{P}$ solve (5.1) and define the subset $\mathcal{P}$ of all symmetric approximations $\widehat{\mathbf{P}}$ that satisfy $\operatorname{span}(\widehat{\mathbf{P}})=\operatorname{span}(\mathbf{V})$. Then $\widehat{\mathbf{P}}$ that satisfies

$$\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}=\min_{\widetilde{\mathbf{P}}\in\mathcal{P}}\left\|\mathbf{P}-\widetilde{\mathbf{P}}\right\|_{\mathrm{F}} \tag{5.16}$$

is called “Frobenius pseudo-optimal” (with respect to $\mathcal{P}$).

The Frobenius pseudo-optimality that minimizes $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}$ in a certain subset, is similarly defined as $\mathcal{H}_2$ pseudo-optimality in minimizing $\|\boldsymbol{G}-\boldsymbol{G}_r\|_{\mathcal{H}_2}$. In order to find the Frobenius pseudo-optimal approximation $\widehat{\mathbf{P}}$, we require the following definition of the Frobenius inner product.

Definition 5.2. Let $\mathbf{A}$ and $\mathbf{B}$ be two matrices of appropriate dimensions, then the Frobenius inner product is defined as

$$\langle\mathbf{A},\mathbf{B}\rangle_{\mathrm{F}}=\operatorname{trace}\left(\mathbf{A}^*\mathbf{B}\right), \tag{5.17}$$

and it induces the Frobenius norm: $\|\mathbf{A}\|_{\mathrm{F}}=\sqrt{\langle\mathbf{A},\mathbf{A}\rangle_{\mathrm{F}}}$.

We are now ready to show that Frobenius pseudo-optimality is likewise uniquely defined, which is presented in the next theorem.

Theorem 5.6. *Given* $\mathbf{V}\in\mathbb{C}^{N\times n}$ *with full column rank, let* $\mathbf{P}$ *solve (5.1) and define the subset* $\mathcal{P}$ *of all symmetric approximations* $\widehat{\mathbf{P}}$ *that satisfy* $\operatorname{span}(\widehat{\mathbf{P}})=\operatorname{span}(\mathbf{V})$*. Assume that* $\mathbf{V}$ *is orthogonal,* $\mathbf{V}^*\mathbf{V}=\mathbf{I}$*, then* $\widehat{\mathbf{P}}=\mathbf{V}\mathbf{P}_r\mathbf{V}^*$ *is the Frobenius pseudo-optimal* *approximation with respect to $\mathcal{P}$, i. e. it satisfies*

$$\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}=\min_{\widetilde{\mathbf{P}}\in\mathcal{P}}\left\|\mathbf{P}-\widetilde{\mathbf{P}}\right\|_{\mathrm{F}}, \tag{5.18}$$

if and only if $\mathbf{P}_r=\mathbf{V}^*\mathbf{P}\mathbf{V}$.

Proof. The space of all $N\times N$ symmetric matrices is a Hilbert space with the Frobenius inner product $\langle\mathbf{A},\mathbf{B}\rangle_{\mathrm{F}}$. All symmetric approximations $\widehat{\mathbf{P}}$ that satisfy $\operatorname{span}(\widehat{\mathbf{P}})=\operatorname{span}(\mathbf{V})$ may be formulated as $\widehat{\mathbf{P}}=\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*$, with some $\widetilde{\mathbf{P}}_r$. Then it can be readily verified that the sum of two $\widehat{\mathbf{P}}=\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*$ stay in $\mathcal{P}$, such that that $\mathcal{P}$ is a closed subspace. We thus can apply the Hilbert projection theorem, which states that $\widehat{\mathbf{P}}$ is the unique minimizer of $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}$ in the subspace $\mathcal{P}$, if and only if

$$\left\langle\mathbf{P}-\widehat{\mathbf{P}},\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*\right\rangle_{\mathrm{F}}=0, \tag{5.19}$$

for all $\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*\in\mathcal{P}$. Then it follows

$$\begin{aligned}0&=\left\langle\mathbf{P},\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*\right\rangle_{\mathrm{F}}-\left\langle\widehat{\mathbf{P}},\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*\right\rangle_{\mathrm{F}}=\operatorname{trace}\left(\mathbf{P}\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*\right)-\operatorname{trace}\left(\mathbf{V}\mathbf{P}_r\widetilde{\mathbf{P}}_r\mathbf{V}^*\right) &(5.20)\\&=\operatorname{trace}\left(\mathbf{V}^*\mathbf{P}\mathbf{V}\widetilde{\mathbf{P}}_r\right)-\operatorname{trace}\left(\mathbf{P}_r\widetilde{\mathbf{P}}_r\right)=\operatorname{trace}\left[\left(\mathbf{V}^*\mathbf{P}\mathbf{V}-\mathbf{P}_r\right)\widetilde{\mathbf{P}}_r\right]. &(5.21)\end{aligned}$$

As $\operatorname{trace}\left[\left(\mathbf{V}^*\mathbf{P}\mathbf{V}-\mathbf{P}_r\right)\widetilde{\mathbf{P}}_r\right]=\mathbf{0}$ has to be satisfied for arbitrary $\widetilde{\mathbf{P}}_r=\widetilde{\mathbf{P}}_r^*$, this is equivalent to $\mathbf{V}^*\mathbf{P}\mathbf{V}-\mathbf{P}_r=\mathbf{0}$, which completes the proof. □

It is possible to state similar orthogonality conditions for Frobenius pseudo-optimality, as was already the case for $\mathcal{H}_2$ pseudo-optimality. This is presented in the next lemma.

Lemma 5.7. *If the approximate solution* $\widehat{\mathbf{P}}$ *is Frobenius pseudo-optimal, then*

$$\left\langle\mathbf{P}-\widehat{\mathbf{P}},\widehat{\mathbf{P}}\right\rangle_{\mathrm{F}}=0, \tag{5.22}$$

and also

$$\left\langle\mathbf{P},\widehat{\mathbf{P}}\right\rangle_{\mathrm{F}}=\left\langle\widehat{\mathbf{P}},\widehat{\mathbf{P}}\right\rangle_{\mathrm{F}}, \tag{5.23}$$

$$\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}^2=\|\mathbf{P}\|_{\mathrm{F}}^2-\left\|\widehat{\mathbf{P}}\right\|_{\mathrm{F}}^2, \tag{5.24}$$

$$\|\mathbf{P}\|_{\mathrm{F}}\geq\left\|\widehat{\mathbf{P}}\right\|_{\mathrm{F}}. \tag{5.25}$$

Proof. The proof of (5.22) is actually already contained in the proof of Theorem 5.6, because $\mathbf{P}-\widehat{\mathbf{P}}$ is orthogonal to all $\mathbf{V}\widetilde{\mathbf{P}}_r\mathbf{V}^*$. Equation (5.23) is a direct consequence of (5.22). Then consider $\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_{\mathrm{F}}^2=\left\langle\mathbf{P}-\widehat{\mathbf{P}},\mathbf{P}-\widehat{\mathbf{P}}\right\rangle_{\mathrm{F}}=\left\langle\mathbf{P}-\widehat{\mathbf{P}},\mathbf{P}\right\rangle_{\mathrm{F}}-\left\langle\mathbf{P}-\widehat{\mathbf{P}},\widehat{\mathbf{P}}\right\rangle_{\mathrm{F}}\overset{(5.22)}{=}$

$\langle \mathbf{P},\mathbf{P}\rangle_{\mathrm{F}} - \left\langle \widehat{\mathbf{P}},\mathbf{P}\right\rangle_{\mathrm{F}} \overset{(5.23)}{=} \|\mathbf{P}\|_{\mathrm{F}}^2 - \left\|\widehat{\mathbf{P}}\right\|_{\mathrm{F}}^2$, which proves (5.24), and from which (5.25) obviously follows. □

Lemma 5.7 shows that if we would compute the Frobenius pseudo-optimal approximation $\widehat{\mathbf{P}}$, the error $\mathbf{P}-\widehat{\mathbf{P}}$ would be perpendicular to $\widehat{\mathbf{P}}$, and in a cumulative framework we would achieve strictly monotonically convergence towards the real solution. Consequently, Frobenius pseudo-optimality features the same nice properties as $\mathcal{H}_2$ pseudo-optimality, which would pave the way for an advantageous framework to approximate $\mathbf{P}$. The drawback, however, is that the Frobenius pseudo-optimal approximation satisfies $\mathbf{P}_r = \mathbf{V}^*\mathbf{P}\mathbf{V}$, which is difficult to achieve without the knowledge of $\mathbf{P}$. To be precise, the following lemma states that the Frobenius pseudo-optimal approximation given by $\mathbf{P}_r = \mathbf{V}^*\mathbf{P}\mathbf{V}$ can in general *not* be achieved by RKSM, which could be called a “dilemma of RKSM”.

Lemma 5.8. *Given an orthogonal* $\mathbf{V}$ *whose columns span a rational input Krylov subspace, and let* $\mathbf{P}$ *solve (5.1), then there need not exist a* $\mathbf{W}$*, such that* $\mathbf{P}_r$ *solves (5.2) and the resulting approximation* $\widehat{\mathbf{P}} = \mathbf{V}\mathbf{P}_r\mathbf{V}^*$ *is the Frobenius pseudo-optimal one, given by* $\mathbf{P}_r = \mathbf{V}^*\mathbf{P}\mathbf{V}$. *In other words, it occurs that the Frobenius pseudo-optimal approximation* $\widehat{\mathbf{P}} = \mathbf{V}\mathbf{V}^*\mathbf{P}\mathbf{V}\mathbf{V}^*$ *cannot be generated by RKSM.*

Proof. For a given $\mathbf{V}$, the matrices $\mathbf{S}$ and $\mathbf{L}$ are fixed, and for any choice of $\mathbf{W}$, the reduced solution $\mathbf{P}_r$ by RKSM satisfies (5.2), and it also has to hold that $\mathbf{A}_r = \mathbf{E}_r\mathbf{S}+\mathbf{B}_r\mathbf{L}$. Using these equations yields

$$\mathbf{S}\mathbf{P}_r + \mathbf{P}_r\mathbf{S}^* = \mathbf{E}_r^{-1}\mathbf{A}_r\mathbf{P}_r + \mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*} - \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L}\mathbf{P}_r - \mathbf{P}_r\mathbf{L}^*\mathbf{B}_r^*\mathbf{E}_r^{-*} \tag{5.26}$$

$$= -\left(\mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*} - \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L}\mathbf{P}_r - \mathbf{P}_r\mathbf{L}^*\mathbf{B}_r^*\mathbf{E}_r^{-*}\right). \tag{5.27}$$

The right hand side of (5.27) obviously has maximum rank $2m$. Consequently, if $\mathbf{P}_r = \mathbf{V}^*\mathbf{P}\mathbf{V}$ is the Frobenius pseudo-optimal solution, then $\operatorname{rank}(\mathbf{S}\mathbf{V}^*\mathbf{P}\mathbf{V}+\mathbf{V}^*\mathbf{P}\mathbf{V}\mathbf{S}^*) \leq 2m$ has to be satisfied. As this condition is independent from $\mathbf{W}$, it is a necessary condition that $\mathbf{V}$ has to satisfy, such that RKSM can yield the Frobenius pseudo-optimal solution. However, this condition need not be satisfied for arbitrary Krylov subspaces. Moreover, simple numerical examples confirm that this indeed generally does not hold. □

In conclusion, Frobenius pseudo-optimal approximations $\widehat{\mathbf{P}}$ would on the one hand be desirable, due to nice properties, such as e. g. guaranteed convergence. On the other hand, the Frobenius pseudo-optimal approximation not necessarily may be attained by RKSM—one instead requires knowledge of $\mathbf{P}$ in order to identify the Frobenius pseudo-optimal solution. Needless to say, that this is improper in a large-scale setting. Consequently, it is indeed judicious to employ the concept of $\mathcal{H}_2$ pseudo-optimality not only for model order reduction, but also for the solution of Lyapunov equations—even in the absence of an output matrix $\mathbf{C}$. This is discussed in the next section.

5.3 A New Type of RKSM

The objective of this section is to exploit the concept of $\mathcal{H}_2$ pseudo-optimality also for the approximation of $\mathbf{P}$. The task thus is to translate the findings of Part II, made for model order reduction, into a new approach to approximately solve Lyapunov equations. We start with the presentation of the cumulative idea from Chapter 3.

5.3.1 Cumulative Framework for RKSM

The error due to approximate solutions $\widehat{\mathbf{P}}$ was already analysed in Theorem 5.3, by stating the low-rank formulation of the residual. This can be seen as a counterpart of Section 3.1, where the factorization of the error in model order reduction was presented. The next lemma hence directly proposes a cumulative framework for Lyapunov equations, similar to Section 3.2.

Lemma 5.9. *Let all variables be as defined in Corollary 3.8, and assume that the columns of each $\mathbf{V}_i$ form a basis of recursively computed rational input Krylov subspaces, i. e. each $\mathbf{V}_i$ satisfies (3.55). Then the total approximate solution $\widehat{\mathbf{P}}_{\text{tot}} = \mathbf{V}_{\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{V}_{\text{tot}}^*$ of the Lyapunov equation can be recursively obtained by*

$$\mathbf{P}_{r,\text{tot}} \leftarrow \begin{bmatrix} \mathbf{P}_{r,\text{tot}} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{bmatrix}, \tag{5.28}$$

where $\mathbf{P}_{12}$ and $\mathbf{P}_{22}$ satisfy

$$\mathbf{A}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{E}_{r,i}^* + \mathbf{E}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{A}_{r,i}^* + \mathbf{E}_{r,\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{L}_{\text{tot}}^*\mathbf{B}_{r,i}^* + \mathbf{B}_{r,\text{tot}}\mathbf{B}_{r,i}^* = \mathbf{0}, \tag{5.29}$$

$$\mathbf{A}_{r,i}\mathbf{P}_{22}\mathbf{E}_{r,i}^* + \mathbf{E}_{r,i}\mathbf{P}_{22}\mathbf{A}_{r,i}^* + \mathbf{B}_{r,i}\mathbf{B}_{r,i}^* + \mathbf{B}_{r,i}\mathbf{L}_{\text{tot}}\mathbf{P}_{12}\mathbf{E}_{r,i}^* + \mathbf{E}_{r,i}\mathbf{P}_{12}^*\mathbf{L}_{\text{tot}}^*\mathbf{B}_{r,i}^* = \mathbf{0}, \tag{5.30}$$

and where $\mathbf{V}_{\text{tot}}$, $\mathbf{L}_{\text{tot}}$, $\mathbf{A}_{r,\text{tot}}$, $\mathbf{E}_{r,\text{tot}}$, $\mathbf{P}_{r,\text{tot}}$, and $\mathbf{B}_{r,\text{tot}}$ are all initialized as empty matrices. Furthermore, let $\mathbf{B}_{\perp,i} = \mathbf{B}_{\perp,i-1} - \mathbf{E}\mathbf{V}_i\mathbf{E}_{r,i}^{-1}\mathbf{B}_{r,i}$, with $\mathbf{B}_{\perp,0} = \mathbf{B}$, and define

$$\mathbf{F}_{\text{tot}} = \mathbf{E}\mathbf{V}_{\text{tot}}\left(\mathbf{E}_{r,\text{tot}}^{-1}\mathbf{B}_{r,\text{tot}} + \mathbf{P}_{r,\text{tot}}\mathbf{L}_{\text{tot}}^*\right), \tag{5.31}$$

then the total approximate solution $\widehat{\mathbf{P}}_{\text{tot}}=\mathbf{V}_{\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{V}_{\text{tot}}^{*}$ *yields the following residual:*

$$\mathbf{R}_{\text{tot}} = \left[\begin{array}{cc} \mathbf{B}_{\perp,i} & \mathbf{F}_{\text{tot}} \end{array}\right] \left[\begin{array}{cc} \mathbf{I} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{B}_{\perp,i}^{*} \\ \mathbf{F}_{\text{tot}}^{*} \end{array}\right]. \tag{5.32}$$

Proof. The proof is done by induction. The case $i = 1$ is trivial, as then $\mathbf{P}_{r,\text{tot}}$ and $\mathbf{P}_{12}$ are empty matrices. Then assume that the cumulated reduced matrices satisfy the reduced Lyapunov equation (5.2) at step $i-1$, and define

$$\mathbf{A}_{+}=\left[\begin{array}{cc} \mathbf{A}_{r,\text{tot}} & \mathbf{0} \\ \mathbf{B}_{r,i}\mathbf{L}_{\text{tot}} & \mathbf{A}_{r,i} \end{array}\right], \mathbf{E}_{+}=\left[\begin{array}{cc} \mathbf{E}_{r,\text{tot}} & \mathbf{0} \\ \mathbf{0} & \mathbf{E}_{r,i} \end{array}\right], \mathbf{P}_{+}=\left[\begin{array}{cc} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^{T} & \mathbf{P}_{22} \end{array}\right], \mathbf{B}_{+}=\left[\begin{array}{c} \mathbf{B}_{r,\text{tot}} \\ \mathbf{B}_{r,i} \end{array}\right], \tag{5.33}$$

then $\mathbf{P}_{r,\text{tot}}$ at step i, i. e. $\mathbf{P}_{+}$, is determined by

$$\mathbf{A}_{+}\mathbf{P}_{+}\mathbf{E}_{+}^{*} + \mathbf{E}_{+}\mathbf{P}_{+}\mathbf{A}_{+}^{*} + \mathbf{B}_{+}\mathbf{B}_{+}^{*} = \mathbf{0}. \tag{5.34}$$

Executing simple matrix products shows

$$\mathbf{A}_{+}\mathbf{P}_{+}\mathbf{E}_{+}^{*}=\left[\begin{array}{cc} \mathbf{A}_{r,\text{tot}}\mathbf{P}_{11}\mathbf{E}_{r,\text{tot}}^{*} & \mathbf{A}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{E}_{r,i}^{*} \\ \mathbf{B}_{r,i}\mathbf{L}_{\text{tot}}\mathbf{P}_{11}\mathbf{E}_{r,\text{tot}}^{*} + \mathbf{A}_{r,i}\mathbf{P}_{12}^{*}\mathbf{E}_{r,\text{tot}}^{*} & \mathbf{B}_{r,i}\mathbf{L}_{\text{tot}}\mathbf{P}_{12}\mathbf{E}_{r,i}^{*} + \mathbf{A}_{r,i}\mathbf{P}_{22}\mathbf{E}_{r,i}^{*} \end{array}\right]. \tag{5.35}$$

Then it follows from the left upper block of (5.34) that $\mathbf{P}_{11} = \mathbf{P}_{r,\text{tot}}$, whereas from the right upper and left lower block (5.29) follows. The right lower block is equal to (5.30). It is left to prove the formulation of the residual, which was already done in Theorem 5.3 for a single reduction. The proof of Theorem 5.3 required that the $\mathbf{B}_{\perp}$-Sylvester equation (2.39) holds, which was proven in Lemma 3.13 to hold also for the cumulated data. This proves that the formulation of the residual of Theorem 5.3 also holds for the cumulated data. □

The cumulative framework suggested in Lemma 5.9 has considerable numerical advantages over standard iterative procedures as depicted in Section 5.1.3. In the standard framework actually only the matrix $\mathbf{V}$ is accumulated, whereas reduced Lyapunov equations of increasing orders have to be solved anew by direct methods. By contrast, within the framework of Lemma 5.9 it is possible to additionally accumulate the reduced Lyapunov solution $\mathbf{P}_{r}$. This has the advantage, that now the Lyapunov equation to-be-solved is (5.30), which has only the reduced order chosen in the current step of each iteration, whereas the additional Sylvester equation (5.29) can efficiently be solved with the ideas presented in Section 2.1. Consider for example a Lyapunov equation (5.1), and assume that at the current step we have $\mathbf{P}_{r,\text{tot}} \in \mathbb{R}^{n\times n}$, and that the matrix $\mathbf{V}$ is augmented by $\widehat{n}$ columns for the next iteration. In the standard iterative procedure, one would then have to solve a reduced Lyapunov equation of order $(n+\widehat{n})$—without a chance to recycle the previous reduced solution $\mathbf{P}_{r,\text{tot}}$. In the cumulative framework suggested above, one instead only has to solve a reduced Lyapunov equation of order $\widehat{n}$ and a Sylvester equation of dimensions $(n+\widehat{n})\times\widehat{n}$, both of which together considerably require less numerical effort.

The drawback of the cumulative approach is that the total approximation $\widehat{\mathbf{P}}_{\text{tot}} = \mathbf{V}_{\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{V}_{\text{tot}}^*$ differs from the one of standard RKSM. This is due to the fact, that the residual after each iteration contains both $\mathbf{B}_\perp$ and $\mathbf{F}$, whereas only $\mathbf{B}_\perp$ is incorporated in the computation of the subsequent $\mathbf{V}_i$. This can be resolved by employing $\mathcal{H}_2$ pseudo-optimality, which is presented in the next section.

5.3.2 $\mathcal{H}_2$ Pseudo-Optimal and Cumulative RKSM

The cumulative framework for the solution of large-scale Lyapunov equations with $\mathcal{H}_2$ pseudo-optimal approximations in each iteration is presented in the next lemma.

Lemma 5.10. *Let all variables be as defined in Corollary 3.8, and assume that the columns of each $\mathbf{V}_i$ form a basis of recursively computed rational input Krylov subspaces, i. e. each $\mathbf{V}_i$ satisfies (3.55). Further assume that the reduced data in each iteration satisfies the conditions of Theorem 4.26 for input $\mathcal{H}_2$ pseudo-optimality. Then the total approximate solution $\widehat{\mathbf{P}}_{\text{tot}} = \mathbf{V}_{\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{V}_{\text{tot}}^*$ of the Lyapunov equation can be recursively computed by*

$$\mathbf{P}_{r,\text{tot}} \leftarrow \begin{bmatrix} \mathbf{P}_{r,\text{tot}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{r,i} \end{bmatrix}, \tag{5.36}$$

where $\mathbf{P}_{r,i}$ satisfies

$$\mathbf{A}_{r,i}\mathbf{P}_{r,i}\mathbf{E}_{r,i}^* + \mathbf{E}_{r,i}\mathbf{P}_{r,i}\mathbf{A}_{r,i}^* + \mathbf{B}_{r,i}\mathbf{B}_{r,i}^* = \mathbf{0}, \tag{5.37}$$

and where $\mathbf{V}_{\text{tot}}$ and $\mathbf{P}_{r,\text{tot}}$ are initialized as empty matrices. Moreover, the total approximate solution $\widehat{\mathbf{P}}_{\text{tot}} = \mathbf{V}_{\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{V}_{\text{tot}}^*$ yields the residual $\mathbf{R}_{\text{tot}} = \mathbf{B}_{\perp,i}\mathbf{B}_{\perp,i}^*$, which can be recursively computed by $\mathbf{B}_{\perp,i} = \mathbf{B}_{\perp,i-1} - \mathbf{E}\mathbf{V}_i\mathbf{E}_{r,i}^{-1}\mathbf{B}_{r,i}$, with $\mathbf{B}_{\perp,0} = \mathbf{B}$. Finally, the accumulated reduced data, i, e. $\mathbf{V}_{\text{tot}}$, $\mathbf{S}_{\text{tot}}$, $\mathbf{L}_{\text{tot}}$, $\mathbf{A}_{r,\text{tot}}$, $\mathbf{E}_{r,\text{tot}}$, $\mathbf{P}_{r,\text{tot}}$, and $\mathbf{B}_{r,\text{tot}}$, also satisfy the conditions of Theorem 4.26 for input $\mathcal{H}_2$ pseudo-optimality.

Proof. The proof is done by induction. The case $i = 1$ is trivial, as then $\mathbf{P}_{r,\text{tot}}$ is an empty matrix. Then assume that the reduced Lyapunov equation holds at step i–1, and we can compute $\mathbf{P}_{r,\text{tot}}$ recursively by (5.28)–(5.30). As we assume that the conditions of Theorem 4.26 for input $\mathcal{H}_2$ pseudo-optimality are satisfied for each iteration, they are also satisfied for the cumulated data due to Lemma 4.36, which already proves the last part of the above statement. Substituting $\mathbf{E}_{r,\text{tot}}\mathbf{P}_{r,\text{tot}}\mathbf{L}_{\text{tot}}^{*}\mathbf{B}_{r,i}^{*}=-\mathbf{B}_{r,\text{tot}}$ due to condition *ii)* of Theorem 4.26 in (5.29), leads to

$$\mathbf{A}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{E}_{r,i}^{*}+\mathbf{E}_{r,\text{tot}}\mathbf{P}_{12}\mathbf{A}_{r,i}^{*}=\mathbf{0}. \tag{5.38}$$

We can identify the unique solution $\mathbf{P}_{12}=\mathbf{0}$, which, inserted in (5.30) proves (5.37). The proof of the residual follows from (5.31) and (5.32): using again condition *ii)* of Theorem 4.26 yields $\mathbf{F}=\mathbf{0}$ which completes the proof. □

The benefit of using $\mathcal{H}_2$ pseudo-optimal approximations in each iteration of the cumulative framework is that the individual iterations are decoupled due to the diagonal structure of $\mathbf{P}_{r,\text{tot}}$ in (5.36). This means that the total approximation is actually the sum of all individual iterations: $\widehat{\mathbf{P}}_{\text{tot}}=\sum_{i=1}^{k}\mathbf{V}_i\mathbf{P}_{r,i}\mathbf{V}_i^{*}$. Another benefit is that after each iteration we have

$$\mathbf{A}\widehat{\mathbf{P}}_{\text{tot}}\mathbf{E}^{*}+\mathbf{E}\widehat{\mathbf{P}}_{\text{tot}}\mathbf{A}^{*}+\mathbf{B}\mathbf{B}^{*}=\mathbf{B}_{\perp,i}\mathbf{B}_{\perp,i}^{*}, \tag{5.39}$$

and consequently,

$$\mathbf{A}\left(\mathbf{P}-\widehat{\mathbf{P}}_{\text{tot}}\right)\mathbf{E}^{*}+\mathbf{E}\left(\mathbf{P}-\widehat{\mathbf{P}}_{\text{tot}}\right)\mathbf{A}^{*}+\mathbf{B}_{\perp,i}\mathbf{B}_{\perp,i}^{*}=\mathbf{0}. \tag{5.40}$$

This means that this approach for approximating $\mathbf{P}$ in fact performs a restart after each iteration: the current error $\mathbf{P}-\widehat{\mathbf{P}}_{\text{tot}}$ solves a similar Lyapunov equation as the original one—only the input matrix changes to $\mathbf{B}_{\perp,i}$. This restart is the consequence of $\mathbf{F}_{\text{tot}}=\mathbf{0}$ in the residual, such that the remaining error $\mathbf{P}$–$\widehat{\mathbf{P}}_{\text{tot}}$ is completely available for the next iteration. (This basically corresponds to $\boldsymbol{G}_f(s)$ being all-pass in model order reduction.)

In order to compute an approximation $\widehat{\mathbf{P}}$ that satisfies the conditions for $\mathcal{H}_2$ pseudo-optimality, as assumed in Lemma 5.10, one can use the PORK algorithm presented in Section 4.3.4. As we are at this point not concerned with the computation of an $\mathcal{H}_2$ pseudo-optimal reduced model, it is actually not necessary to execute all steps of PORK. For the solution of Lyapunov equations in terms of $\mathcal{H}_2$ pseudo-optimality, it suffices to use a version of PORK that is specially tied for Lyapunov equations, and which will be denoted as "PORK-Lyap" hereafter. This is illustrated in Algorithm 5.1. It should be noted, that in principle $\mathbf{B}_{\perp}=\mathbf{B}-\mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r$, but due to condition *ii)* of Theorem 4.26, the different statement in Step 3 of Algorithm 5.1 follows.

Within an iterative procedure as defined by Lemma 5.10, i. e. which is based on $\mathcal{H}_2$ pseudo-optimality and which uses PORK-Lyap, all data can be recursively computed. To clarify this, the basic procedure is depicted in the following.

Algorithm 5.1 Pseudo-optimal rational Krylov for Lyapunov equations (PORK-Lyap)

```
Input: $\mathbf{E}$, $\mathbf{B}$, $\mathbf{V}$, $\mathbf{S}$, $\mathbf{L}$, such that $\mathbf{AV}-\mathbf{EVS}=\mathbf{BL}$ is satisfied (see Section 2.3)
Output: approximate solution $\widehat{\mathbf{P}}$ in terms input $\mathcal{H}_2$ pseudo-optimality and corresponding residual $\mathbf{R}=\mathbf{B}_\perp\mathbf{B}_\perp^*$
$\mathbf{P}_r^{-1}=$ lyap($\mathbf{S}^*$, $-\mathbf{L}^*\mathbf{L}$)   // direct solver for $\mathbf{S}^*\mathbf{P}_r^{-1}+\mathbf{P}_r^{-1}\mathbf{S}-\mathbf{L}^*\mathbf{L}=\mathbf{0}$, condition iii)
$\widehat{\mathbf{P}}=\mathbf{V}\left(\mathbf{P}_r^{-1}\right)^{-1}\mathbf{V}^*$
$\mathbf{B}_\perp=\mathbf{B}+\mathbf{EV}\left(\mathbf{P}_r^{-1}\right)^{-1}\mathbf{L}^*$   // condition ii)
```

- Compute $k\in\mathbb{N}^+$ new directions of a Krylov subspace, i. e. a $\mathbf{V}_i\in\mathbb{R}^{N\times k}$ whose columns form a basis of a rational input Krylov subspace with input $\mathbf{B}_{\perp,i}$ and corresponding $\mathbf{S}_i$ and $\mathbf{L}_i$.
- Plug $\mathbf{E}$, $\mathbf{B}_{\perp,i}$, $\mathbf{V}_i$, $\mathbf{S}_i$ and $\mathbf{L}_i$ into PORK-Lyap to compute the approximation $\widehat{\mathbf{P}}_i=\mathbf{V}_i\mathbf{P}_{r,i}\mathbf{V}_i^*$.
- Update the total approximation by $\mathbf{V}_{\text{tot}}\leftarrow[\mathbf{V}_{\text{tot}},\mathbf{V}_i]$ and by (5.36), i. e. $\widehat{\mathbf{P}}_{\text{tot}}\leftarrow\widehat{\mathbf{P}}_{\text{tot}}+\widehat{\mathbf{P}}_i$, and the residual $\mathbf{R}_{\text{tot}}=\mathbf{B}_{\perp,i+1}\mathbf{B}_{\perp,i+1}^*$, where $\mathbf{B}_{\perp,i+1}$ is given by Step 3 of PORK-Lyap.
- (Usually compute some norm of the residual in order to) evaluate an appropriate stopping criterion.
- If the desired accuracy is achieved, stop the algorithm; otherwise restart by calculating k additional directions.

Remark 5.11. It should be noted that if the approximate solution $\widehat{\mathbf{P}}$ is computed in the sense of $\mathcal{H}_2$ pseudo-optimality, then $\widehat{\mathbf{P}}$ is irrespective of whether we use the cumulative framework, or a single run of PORK-Lyap. Assume e. g. $m=1$ and that a sequence of interpolation points s_1, … s_k is given. Then one possibility is to directly compute the complete matrix $\mathbf{V}\in\mathbb{C}^{N\times k}$ and plug it into PORK-Lyap, in order to compute $\widehat{\mathbf{P}}$ in a single step. Another possibility is to put only one basis $\mathbf{v}_i\in\mathbb{C}^N$ (for each shift s_i) after another into PORK-Lyap, with intermediate updates of the approximation $\widehat{\mathbf{P}}_{\text{tot}}$ and the residual, as just described in the cumulative approach. The final approximations of both approaches would then be equal, $\widehat{\mathbf{P}}=\widehat{\mathbf{P}}_{\text{tot}}$. This provides an additional degree of freedom: we do not have to compute the total approximation at once, we can rather split the calculations into suitably small parts and restart the algorithm after each iteration—without affecting the outcome. It should however be stressed, that this only applies to the case that the approximation is computed in terms of $\mathcal{H}_2$ pseudo-optimality.

It would be normally necessary to present at this point promising numerical examples, which verify the suitability of combining the proposed cumulative framework with approximations in terms of $\mathcal{H}_2$ pseudo-optimality; but this can actually be omitted due to the following reason. It can be shown, that this approach in fact is equivalent to a prevalent method for solving (5.1): the *alternating directions implicit* (ADI) iteration. To be precise, the total approximation $\widehat{\mathbf{P}}_{\text{tot}}$ of the above approach equals the outcome of the ADI iteration—only the computation might differ. As various numerical examples are available in the literature, which already prove that the outcome of the ADI iteration is competitive to other methods, we can refrain from presenting further ones. Instead, we may immediately go into the theoretic details of the link between ADI and RKSM. In this respect, the above discussions can just as well be seen as the theoretical grounding for a deeper analysis and further development of the ADI iteration.

5.4 Alternating Directions Implicit (ADI) Iteration

The ADI iteration was originally introduced for the numerical solution of parabolic and elliptic differential equations by Peaceman and Rachford [151], and later identified by Ellner and Wachspress [58, 198] to be a suitable tool for solving Lyapunov equations. Assume that an initial approximation $\widehat{\mathbf{P}}_0$ is chosen, e. g. $\widehat{\mathbf{P}}_0 = \mathbf{0}$, and that $\mathbf{E} = \mathbf{I}$. Further assume that a sequence of complex shifts $s_1, s_2, \ldots, s_k$ is given. It should be noted, that unlike in the original formulation, the shifts $s_i \in \mathbb{C}$ are here chosen in the right half of the complex plane, in order to align the notation to the definition of rational Krylov subspaces as in Section 1.6.1. Then an approximation $\widehat{\mathbf{P}}$ of $\mathbf{P}$ is given by the following half-step iteration—the ADI iteration:

$$\begin{aligned} (\mathbf{A} - s_i\mathbf{I})\,\widehat{\mathbf{P}}_{i-\frac{1}{2}} &= -\mathbf{B}\mathbf{B}^T - \widehat{\mathbf{P}}_{i-1}\left(\mathbf{A}^T - s_i\mathbf{I}\right), \\ (\mathbf{A} - s_i\mathbf{I})\,\widehat{\mathbf{P}}_i^T &= -\mathbf{B}\mathbf{B}^T - \widehat{\mathbf{P}}_{i-\frac{1}{2}}\left(\mathbf{A}^T - s_i\mathbf{I}\right). \end{aligned} \tag{5.41}$$

The drawback of this approach is that the approximation $\widehat{\mathbf{P}}$ is generally dense, which renders its computation via (5.41) unsuitable for large-scale Lyapunov equations. However, it was recognized by Penzl [153] and by Li and White [128], that the iteration (5.41) can be recast into another iterative procedure, which directly generates a low-rank Cholesky factor $\mathbf{Z}$, such that $\widehat{\mathbf{P}} = \mathbf{Z}\mathbf{Z}^*$. To this end, divide $\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_k]$ into the block $\mathbf{Z}_i \in \mathbb{C}^{N \times m}$, and assume that $\widehat{\mathbf{P}}_0 = \mathbf{0}$, then (5.41) can be generalized to arbitrary but non-singular $\mathbf{E} \neq \mathbf{I}$, where the blocks $\mathbf{Z}_i$ are given by the following iteration:

$$\begin{aligned} \mathbf{Z}_1 &= \sqrt{2\operatorname{Re}(s_1)}\,(\mathbf{A} - s_1\mathbf{E})^{-1}\,\mathbf{B}, \\ \mathbf{Z}_i &= \sqrt{\frac{\operatorname{Re}(s_i)}{\operatorname{Re}(s_{i-1})}}\left(\mathbf{I} + (s_i + \overline{s}_{i-1})\,(\mathbf{A} - s_i\mathbf{E})^{-1}\,\mathbf{E}\right)\mathbf{Z}_{i-1}, \qquad i = 2, \ldots, k. \end{aligned} \tag{5.42}$$

Various contributions are available in the literature, which are based on the formulation (5.42). These include: partly heuristic approaches for selecting the shifts s_i [34, 153, 199, 210]; modifications to prevent the low-rank factor $\mathbf{Z}$ from having linearly dependent columns [95, 170]; and a slight modification which directly delivers a real low-rank factor $\mathbf{Z}$, if the sequence of shifts is closed under complex conjugation [33]. Further articles adapt (5.42) for second-order models [32], or discuss issues that occur when $\widehat{\mathbf{P}}$ is used for balanced truncation [169, 209]. For a general overview, the thesis of Saak [168] is recommended.

The formulation (5.42) is the basis for the following discussions, where a $\mathbf{Z}$, which results from the iteration (5.42), will be called the "ADI basis". Li and White [128] proved that the ADI basis actually spans a rational block-input Krylov subspace, which already indicates that there exists a link between ADI and RKSM. The actual connection, however, was discovered only very recently. Druskin et al. [55] and Flagg and Gugercin [67] independently presented proofs, that the approximations of the ADI iteration and RKSM are identical under certain constraints. The aim of the remainder of this thesis is to slightly generalize these proofs, then identify the ADI iteration as the $\mathcal{H}_2$ pseudo-optimal cumulative framework presented above, and by doing so, introducing an alternative way to compute the ADI approximation based on rational Krylov subspaces, which finally leads to a new ADI iteration with tangential interpolation. These results are based on the publications [210, 213].

5.4.1 The Sylvester Equation for the ADI Iteration

In order to translate the ADI iteration (5.42) into a framework with rational Krylov subspaces, we start with stating the $\mathbf{B}$-Sylvester equation for the ADI basis $\mathbf{Z}$, similar to Theorem 2.4.

Lemma 5.12. *Let* $\mathbf{I}$ *denote the identity matrix of dimension* $m \times m$ *and define* $\alpha_i := \sqrt{2\,\mathrm{Re}(s_i)}$, *and*

$$\mathbf{S}_{\mathrm{ADI}} = \begin{bmatrix} s_1\mathbf{I} & \alpha_1\alpha_2\mathbf{I} & \cdots & \alpha_1\alpha_k\mathbf{I} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \alpha_{k-1}\alpha_k\mathbf{I} \\ & & & s_k\mathbf{I} \end{bmatrix}, \quad \textit{and} \quad \mathbf{L}_{\mathrm{ADI}} = [\alpha_1\mathbf{I},\, \ldots,\, \alpha_k\mathbf{I}]\,. \tag{5.43}$$

Then the ADI basis $\mathbf{Z}$ *from the iteration (5.42) satisfies the Sylvester equation*

$$\mathbf{A}\mathbf{Z} - \mathbf{E}\mathbf{Z}\mathbf{S}_{\mathrm{ADI}} = \mathbf{B}\mathbf{L}_{\mathrm{ADI}}. \tag{5.44}$$

Proof. It follows from the Sylvester equation (5.44) and the definitions (5.43), that

$$(\mathbf{A} - s_i\mathbf{E})\,\mathbf{Z}_i = \alpha_i \left(\sum_{j=1}^{i-1} \alpha_j \mathbf{E}\mathbf{Z}_j + \mathbf{B}\right). \tag{5.45}$$

We prove the equivalence of the ADI iteration (5.42) and (5.45) by induction. Obviously, $\mathbf{Z}_1$ in (5.45) is equal to $\mathbf{Z}_1$ of the ADI iteration (5.42). Now assume that $\mathbf{Z}_i$ from (5.42) is given by (5.45) and substitute $-s_i = \bar{s}_i - \alpha_i^2$. Then (5.45) becomes

$$(\mathbf{A} + \bar{s}_i\mathbf{E})\,\mathbf{Z}_i = \alpha_i \left(\sum_{j=1}^{i} \alpha_j \mathbf{E}\mathbf{Z}_j + \mathbf{B}\right), \tag{5.46}$$

which is equivalent to

$$(\mathbf{A} - s_{i+1}\mathbf{E}) \left[\mathbf{I} + (s_{i+1} + \bar{s}_i)\,(\mathbf{A} - s_{i+1}\mathbf{E})^{-1}\mathbf{E}\right] \mathbf{Z}_i = \alpha_i \left(\sum_{j=1}^{i} \alpha_j \mathbf{E}\mathbf{Z}_j + \mathbf{B}\right). \tag{5.47}$$

Using $\left[\mathbf{I} + (s_{i+1} + \bar{s}_i)\,(\mathbf{A} - s_{i+1}\mathbf{E})^{-1}\mathbf{E}\right]\mathbf{Z}_i = \frac{\alpha_i}{\alpha_{i+1}}\mathbf{Z}_{i+1}$ from (5.42), shows that (5.45) is true for $\mathbf{Z}_{i+1}$, which completes the proof by induction. □

Lemma 5.12 shows that any ADI basis $\mathbf{Z}$ satisfies a $\mathbf{B}$-Sylvester equation, and hence, must span a rational Krylov subspace. This is stated in the next lemma, which, in that sense, presents a new and simpler proof than the original statement in [128].

Lemma 5.13. *Given a sequence of shifts $s_1, s_2, \ldots, s_k$, $s_i \in \mathbb{C}$, where multiplicities are allowed, define $\mathbf{A}_{s_i} = (\mathbf{A} - s_i\mathbf{E})$. Then the ADI basis $\mathbf{Z}$ from the iteration (5.42) spans the rational block-input Krylov subspace*

$$\operatorname{span}(\mathbf{Z}) = \operatorname{span}\left\{\mathbf{A}_{s_1}^{-1}\mathbf{B},\ \mathbf{A}_{s_2}^{-1}\mathbf{E}\mathbf{A}_{s_1}^{-1}\mathbf{B},\ \ldots,\ \mathbf{A}_{s_k}^{-1}\mathbf{E}\ldots\mathbf{A}_{s_2}^{-1}\mathbf{E}\mathbf{A}_{s_1}^{-1}\mathbf{B}\right\}. \tag{5.48}$$

Proof. $\mathbf{Z}$ satisfies the $\mathbf{B}$-Sylvester equation (5.44), where the pair $(\mathbf{L}_{\mathrm{ADI}}, \mathbf{S}_{\mathrm{ADI}})$ is observable. Then it follows from Theorem 2.4 that $\mathbf{Z}$ spans a rational input Krylov subspace, where the expansion points correspond to the eigenvalues of $\mathbf{S}_{\mathrm{ADI}}$. The eigenvalues of $\mathbf{S}_{\mathrm{ADI}}$ can be readily identified, due to the upper diagonal structure in (5.43). As $\mathbf{S}_{\mathrm{ADI}}$ has the $(m \times m)$-dimensional blocks $s_i\mathbf{I}$ on its diagonal, it follows that $\mathbf{Z}$ spans the block-input Krylov subspace (5.48). □

5.4.2 The ADI Iteration Implicitly Performs $\mathcal{H}_2$ Pseudo-Optimal MOR

We are now ready to state the main result of this section, which describes how the approximation $\widehat{\mathbf{P}}$ from the ADI iteration (5.42) can be generated by Krylov projection methods. This new and enlightening result unveils the link between ADI and RKSM.

Theorem 5.14. *Given a sequence of shifts $s_1, s_2, \dots, s_k$, $s_i \in \mathbb{C}$, let the columns of* $\mathbf{V}$ *form the basis of the corresponding rational block-input Krylov subspace (5.48), and let* $\mathbf{S}$ *and* $\mathbf{L}$ *be given according to Theorem 2.4. Then both approximations* $\widehat{\mathbf{P}}$ *of* $\mathbf{P}$*, that results from the ADI iteration (5.42) and from the PORK-Lyap Algorithm 5.1, are equal.*

Proof. Due to Lemma 5.13, the ADI basis $\mathbf{Z}$ spans a rational input Krylov subspace and we can plug $(\mathbf{Z}, \mathbf{S}_{\mathrm{ADI}}, \mathbf{L}_{\mathrm{ADI}})$ from Lemma 5.12 into PORK-Lyap in Algorithm 5.1. Step 1 of PORK-Lyap then requires to solve $\mathbf{S}_{\mathrm{ADI}}^{*}\mathbf{P}_{r,\mathrm{ADI}}^{-1} + \mathbf{P}_{r,\mathrm{ADI}}^{-1}\mathbf{S}_{\mathrm{ADI}} = \mathbf{L}_{\mathrm{ADI}}^{*}\mathbf{L}_{\mathrm{ADI}}$ for $\mathbf{P}_{r,\mathrm{ADI}}^{-1}$. Due to (5.43) it follows that this is solved by $\mathbf{P}_{r,\mathrm{ADI}} = \mathbf{I}$ identity. Then the outcome of PORK-Lyap is $\widehat{\mathbf{P}} = \mathbf{Z}\mathbf{P}_{r,\mathrm{ADI}}\mathbf{Z}^{*} = \mathbf{Z}\mathbf{Z}^{*}$ which equals the approximation of the ADI iteration, and which completes the proof. □

Before discussing Theorem 5.14, we will first derive a low-rank formulation of the residual from the ADI iteration, because as the approximations $\widehat{\mathbf{P}}$ of the ADI iteration and of PORK-Lyap are equal, so have to be the residuals.

Corollary 5.15. *Let* $\widehat{\mathbf{P}}$ *be given by the ADI iteration (5.42), then the residual is* $\mathbf{R} = \mathbf{B}_{\perp}\mathbf{B}_{\perp}^{*}$*, with* $\mathbf{B}_{\perp} = \mathbf{B} + \mathbf{E}\mathbf{Z}\mathbf{L}_{\mathrm{ADI}}^{*}$*, and where* $\mathbf{L}_{\mathrm{ADI}}$ *is defined in (5.43).*

Proof. Due to the proof of Theorem 5.14, $\mathbf{P}_{r,\mathrm{ADI}} = \mathbf{I}$. Then the proof readily follows from Step 3 of PORK-Lyap in Algorithm 5.1. □

Theorem 5.14 uncovers how the approximation $\widehat{\mathbf{P}}$ of the ADI iteration can be alternatively constructed within a framework that relies on projections onto rational Krylov subspaces. Corollary 5.15 additionally proposes a convenient low-rank formulation of the residual that results from the approximation $\widehat{\mathbf{P}}$ of the ADI iteration (5.42). These results, however, involve further consequences, which are discussed next.

Remark 5.16 (Link to RKSM). Theorem 5.14 identifies the ADI iteration as a particular type of RKSM: the ADI basis $\mathbf{Z}$ spans a rational Krylov subspace, and the associated $\mathbf{B}$-Sylvester equation can be stated in closed form like in (5.44). This provokes a family of reduced $\mathbf{E}_r$, $\mathbf{A}_r$ and $\mathbf{B}_r$ (as discussed in Section 2.5), that are attainable via projection onto $\mathbf{Z}$. Executing the PORK algorithm (or equivalently PORK-Lyap), just picks one member out of this family, and hence, there exists a $\mathbf{W}$, such that $\mathbf{P}_{r,\mathrm{ADI}} = \mathbf{I}$ satisfies

$$\mathbf{A}_{r,\mathrm{ADI}}\mathbf{P}_{r,\mathrm{ADI}}\mathbf{E}_{r,\mathrm{ADI}}^{*} + \mathbf{E}_{r,\mathrm{ADI}}\mathbf{P}_{r,\mathrm{ADI}}\mathbf{A}_{r,\mathrm{ADI}}^{*} + \mathbf{B}_{r,\mathrm{ADI}}\mathbf{B}_{r,\mathrm{ADI}}^{*} = \mathbf{0}, \tag{5.49}$$

with $\mathbf{E}_{r,\mathrm{ADI}} = \mathbf{W}^*\mathbf{E}\mathbf{Z}$, $\mathbf{A}_{r,\mathrm{ADI}} = \mathbf{W}^*\mathbf{A}\mathbf{Z}$ and $\mathbf{B}_{r,\mathrm{ADI}} = \mathbf{W}^*\mathbf{B}$, which yields the approximation $\widehat{\mathbf{P}} = \mathbf{Z}\mathbf{P}_{r,\mathrm{ADI}}\mathbf{Z}^* = \mathbf{Z}\mathbf{Z}^*$ of the ADI iteration. This means that there exists a reduced Lyapunov equation (5.49), which follows from a projection onto the rational Krylov subspace span($\mathbf{Z}$), and which causes the same approximation $\widehat{\mathbf{P}}$ as the ADI iteration (5.42). This renders the ADI iteration a very special type of RKSM: the degrees of freedom in RKSM are the expansion points in the Krylov subspace and the direction of projection, which is specified by $\mathbf{W}$. The ADI iteration hence can be interpreted as RKSM with a particular choice of $\mathbf{W}$. It should be noted, that although this $\mathbf{W}$ is not built up explicitly, the respective $\mathbf{E}_{r,\mathrm{ADI}}$, $\mathbf{A}_{r,\mathrm{ADI}}$ and $\mathbf{B}_{r,\mathrm{ADI}}$ from (5.49) could still be computed by the final steps of PORK in Algorithm 4.2, if desired. But as this $\mathbf{W}$, and thus also $\mathbf{E}_{r,\mathrm{ADI}}$, $\mathbf{A}_{r,\mathrm{ADI}}$ and $\mathbf{B}_{r,\mathrm{ADI}}$, are fixed in advance, it is indeed possible to compute the resulting approximation $\widehat{\mathbf{P}}$ by the ADI iteration (5.42)—without actually solving (5.49).

Remark 5.17 (Link to $\mathcal{H}_2$ pseudo-optimal MOR). Let an arbitrary $\mathbf{C} \in \mathbb{R}^{p\times N}$ be given, which induces a dynamical model with the transfer function $\boldsymbol{G}(s) = \mathbf{C}\left(s\mathbf{E} - \mathbf{A}\right)^{-1}\mathbf{B}$, and let $\mathbf{Z}$ be the ADI basis that follows from (5.42). Then using $\mathbf{S}_{\mathrm{ADI}}$ and $\mathbf{L}_{\mathrm{ADI}}$ from (5.43), the unique input $\mathcal{H}_2$ pseudo-optimal reduced model $\boldsymbol{G}_r(s)$ can be computed by PORK, and its Controllability Gramian $\mathbf{P}_{r,\mathrm{ADI}}$ provides an approximation $\widehat{\mathbf{P}} = \mathbf{Z}\mathbf{P}_{r,\mathrm{ADI}}\mathbf{Z}^*$ of $\mathbf{P}$, which equals the outcome of the ADI iteration. This means that the ADI iteration (5.42) generates the same approximation $\widehat{\mathbf{P}}$, that would follow from the input $\mathcal{H}_2$ pseudo-optimal reduced model (for arbitrary outputs). If the $\mathcal{H}_2$ pseudo-optimal reduced model $\boldsymbol{G}_r(s)$ would be constructed through PORK, then it would satisfy the conditions of Theorem 4.26, and because the columns of $\mathbf{Z}$ from (5.42) is a basis of a rational *block* Krylov subspace, $\boldsymbol{G}(s_i) = \boldsymbol{G}_r(s_i)$ would hold true and the reduced eigenvalues λ_i would be $\lambda_i = -\overline{s}_i$, $i = 1, \ldots, k$, all with geometric multiplicity m.

Remarks 5.16 and 5.17 show that a (virtual) reduced Lyapunov equation (5.49), and if a $\mathbf{C} \in \mathbb{R}^{p\times N}$ is given, additionally a (virtual) reduced model $\boldsymbol{G}_r(s)$, may be associated with the ADI iteration (5.42). This unveils the link of the ADI iteration to methods based on projections onto rational Krylov subspaces for the solution of Lyapunov equations and model order reduction. To conclude these findings: the ADI iteration qualifies as a straightforward instruction to implement RKSM in a (cumulative and) $\mathcal{H}_2$ pseudo-optimal manner; or to put it bluntly:

$$\mathrm{ADI} = \mathrm{RKSM} + \mathrm{PORK}\ (+\mathrm{CURE}), \tag{5.50}$$

where “CURE” is put in parentheses due to the following reason: it is regardless of whether or not using CURE, if reductions are solely performed with PORK, because then the final approximation does not alter, cf. Remark 5.11.

5.4.3 Tangential ADI Iteration

So far we have used the machinery "projections onto rational Krylov subspaces" only to describe the ADI iteration. In what follows, it shall be also used for the synthesis of a new ADI iteration that additionally can handle tangential interpolation.

First of all, consider again the kind of equation (5.50), where "CURE" is put in parentheses. Theorem 5.14, in fact, only proves ADI = RKSM + PORK, but due to Remark 5.11, also ADI = RKSM + PORK + CURE holds. It should therefore be possible to restart the ADI iteration (5.42) at any point, in order to integrate the cumulative idea. This is done in the next lemma, which presents a reformulation of the ADI iteration (5.42).

Lemma 5.18. *Define* $\mathbf{B}_{\perp,0} = \mathbf{B}$ *and* $\alpha_i = \sqrt{2\,\mathrm{Re}(s_i)}$*, then the ADI iteration (5.42) is equivalent to the following iteration for* $i = 1, \ldots, k$*:*

$$\begin{aligned} \mathbf{Z}_i &= \alpha_i \left(\mathbf{A} - s_i\mathbf{E}\right)^{-1} \mathbf{B}_{\perp,i-1}, \\ \mathbf{B}_{\perp,i} &= \mathbf{B}_{\perp,i-1} + \alpha_i \mathbf{E}\mathbf{Z}_i. \end{aligned} \tag{5.51}$$

Proof. It is shown in the proof of Lemma 5.12, that $\mathbf{Z}_i$ is given by

$$\left(\mathbf{A} - s_i\mathbf{E}\right) \mathbf{Z}_i = \alpha_i \left(\sum_{j=1}^{i-1} \alpha_j \mathbf{E}\mathbf{Z}_j + \mathbf{B} \right). \tag{5.52}$$

The residual is proven in Corollary 5.15 to be $\mathbf{R} = \mathbf{B}_\perp \mathbf{B}_\perp^*$, with $\mathbf{B}_\perp = \mathbf{B} + \mathbf{E}\mathbf{Z}\mathbf{L}_{\mathrm{ADI}}^*$, and by using the definition of $\mathbf{L}_{\mathrm{ADI}}$ in (5.43), it follows that $\mathbf{B}_{\perp,i} = \mathbf{B} + \sum_{j=1}^{i} \alpha_j \mathbf{E}\mathbf{Z}_j$. By substituting this in (5.52), the statement can be concluded. □

Lemma 5.18 states the cumulative version of the ADI iteration (5.42), where the restart is performed after every single shift. It would of course also be possible to execute several, say k, iterations of the original ADI iteration (5.42), not till then compute $\mathbf{B}_{\perp,k} = \mathbf{B} + \sum_{j=1}^{k} \alpha_j \mathbf{E}\mathbf{Z}_j$, and only subsequently perform the restart as described by (5.51), i. e. $\mathbf{Z}_{k+1} = \alpha_{k+1} \left(\mathbf{A} - s_{k+1}\mathbf{E}\right)^{-1} \mathbf{B}_{\perp,k}$. This means that one can freely choose when to perform a restart by using the previous $\mathbf{B}_{\perp,i-1}$, cf. (5.51), and when to apply the original ADI iteration by using the previous $\mathbf{Z}_{i-1}$, cf. (5.42).

Remark 5.19. Lemma 5.18 provides a more natural formulation of the ADI iteration than (5.42), because it incorporates the residual factor $\mathbf{B}_\perp$ as an integral part of the iteration. That means that the re-formulation updates the current residual on the way as a byproduct. Surprisingly, this works with consistent numerical effort: the main effort in both formulations (5.42) and (5.51) is to solve an LSE for each iterate $\mathbf{Z}_i$. The remaining operations in both iterations are a single matrix-vector product with $\mathbf{E}$ and a weighted sum of two $N \times m$ blocks.

The low-rank factor $\mathbf{Z}$ of both ADI iterations (5.42) and (5.51) gains a new block of m columns in each step, and hence, the larger m is, the faster the ADI basis grows with each iteration. If now a practical application requires many iterations for convergence, the final $\mathbf{Z}$ might grow too large for reasonable processing. It would instead be desirable, that for every shift s_i only one column—or as many columns as absolutely necessary—are added to the low-rank factor $\mathbf{Z}$. For projection methods with rational Krylov subspaces, this problem is solved by introducing tangential directions $\mathbf{l}_i$ for each shift s_i. Hence, our goal now is to introduce this possibility also into the ADI iteration. The finding will be denoted as *tangential (low-rank) ADI iteration* (T-LR-ADI) in the following.

To this end, the re-formulation in Lemma 5.18 is essential: in the original formulation (5.42), every iterate $\mathbf{Z}_i$ originates from its predecessor $\mathbf{Z}_{i-1}$, whereas in the new formulation (5.51), each $\mathbf{Z}_i$ originates from the recent residual factor $\mathbf{B}_{\perp,i-1}$ instead. This permits us to use individual tangential directions $\mathbf{l}_i \in \mathbb{C}^m$ for every iterate by simply replacing $\mathbf{B}_{\perp,i}$ in (5.51) by $\mathbf{B}_{\perp,i}\mathbf{l}_i$. This is indeed a valuable extension of the ADI iteration, because it prevents the ADI basis $\mathbf{Z}$ from growing too large, and thereby it may ensure that $\mathbf{Z}$ stays numerically manageable. Incorporating tangential directions therefore extends the application spectrum of the low-rank ADI iteration, if the tangential directions are suitably chosen. This completely new idea was already introduced in the preprint [210].

All that remains to show, is that replacing $\mathbf{B}_{\perp,i}$ in (5.51) by $\mathbf{B}_{\perp,i}\mathbf{l}_i$ still fulfils the conditions for input $\mathcal{H}_2$ pseudo-optimality, and furthermore, how the residual factor $\mathbf{B}_{\perp,i}$ changes by incorporating tangential directions. Different from the block iteration (5.51), real and complex conjugated shifts have to be distinguished in T-LR-ADI. Both cases are treated separately in Theorems 5.20 and 5.21.

Theorem 5.20 (Real T-LR-ADI). *Define* $\mathbf{B}_{\perp,0} := \mathbf{B}$*, and assume exclusively real shifts* $s_i \in \mathbb{R}$*, and real tangential directions* $\mathbf{l}_i \in \mathbb{R}^m$*, with unit length* $\|\mathbf{l}_i\|_2 = 1$*, for* $i = 1, \ldots, k$*. If* $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]$ *is given by the T-LR-ADI iteration*

$$\begin{aligned} \mathbf{z}_i &= \alpha_i \left(\mathbf{A} - s_i\mathbf{E}\right)^{-1} \mathbf{B}_{\perp,i-1}\mathbf{l}_i, \\ \mathbf{B}_{\perp,i} &= \mathbf{B}_{\perp,i-1} + \alpha_i \mathbf{E}\mathbf{z}_i\mathbf{l}_i^T, \end{aligned} \tag{5.53}$$

then $\widehat{\mathbf{P}}=\mathbf{Z}\mathbf{Z}^T$ *satisfies the conditions of Theorem 4.26 for input* $\mathcal{H}_2$ *pseudo-optimality.*

Proof. As the re-formulated ADI iteration comprises a restart after every single step, it is actually sufficient to show that the first iterate $\widehat{\mathbf{P}} = \mathbf{z}_1\mathbf{z}_1^T$ is equal to the outcome of PORK-Lyap, because then the conditions of Theorem 4.26 for input $\mathcal{H}_2$ pseudo-optimality will be ensured. To this end, rewrite the first equation of (5.53) as $(\mathbf{A}-s_1\mathbf{E})\,\mathbf{z}_1=\alpha_1\mathbf{B}\mathbf{l}_1$, and by comparing this with (5.44), we can identify $\mathbf{S}_{\mathrm{ADI}}=s_1$ and $\mathbf{L}_{\mathrm{ADI}}=\alpha_1\mathbf{l}_1$. As by assumption $\mathbf{l}_1^*\mathbf{l}_1=1$, the solution at Step 1 of PORK-Lyap is $\mathbf{P}_r^{-1}=1$, and the approximate solution reads as $\widehat{\mathbf{P}}=\mathbf{z}_1\left(\mathbf{P}_r^{-1}\right)^{-1}\mathbf{z}_1^T=\mathbf{z}_1\mathbf{z}_1^T$. It is left to prove that $\mathbf{B}_{\perp,1}=\mathbf{B}+\alpha_1\mathbf{E}\mathbf{z}_1\mathbf{l}_1^T$, which directly follows from Step 3 of PORK-Lyap. □

Assume a complex shift $s_1\in\mathbb{C}$ and tangential direction $\mathbf{l}_1\in\mathbb{C}^m$, and compute the first iterate $\mathbf{z}_1$ by (5.53). In order to yield a real approximation $\widehat{\mathbf{P}}=\mathbf{Z}\mathbf{Z}^*\in\mathbb{R}^{N\times N}$, the second iterate $\mathbf{z}_2$ has to be contained in $\mathrm{span}[\mathbf{z}_1,\overline{\mathbf{z}}_1]$. However, the $\mathbf{z}_2$ resulting from the direct application of (5.53) with $s_2=\overline{s}_1$ and $\mathbf{l}_2=\overline{\mathbf{l}}_1$ would not satisfy this. For this reason, T-LR-ADI requires a slight modification for complex conjugated shifts. This is done in the next theorem, which proposes an iteration similar (5.53), but which directly yields a real ADI basis for complex conjugated pairs of shifts and tangential directions.

Theorem 5.21 (Complex T-LR-ADI). *Let* $\mathbf{B}_{\perp,0}:=\mathbf{B}$, *and assume for every iterate* $i=1,\dots,k$ *a complex shift* $s_i\in\mathbb{C}$ *with nonzero imaginary part, and a tangential direction* $\mathbf{l}_i\in\mathbb{C}^m$, *with* $\|\mathbf{l}_i\|_2=1$. *Define*

$$\alpha_i=\sqrt{2\ \mathrm{Re}(s_i)},\quad \beta_i=\mathbf{l}_i^*\overline{\mathbf{l}}_i\frac{\mathrm{Re}(s_i)}{\overline{s}_i},\quad \gamma_i=\frac{1}{\sqrt{1-\beta_i\overline{\beta}_i}},\quad \delta_i=\sqrt{1+\mathrm{Re}(\beta_i)}. \tag{5.54}$$

If $\mathbf{Z}=[\mathbf{Z}_1,\dots,\mathbf{Z}_k]$ *is given by the T-LR-ADI iteration*

$$\begin{aligned}\mathbf{v}_i&=\alpha_i\left(\mathbf{A}-s_i\mathbf{E}\right)^{-1}\mathbf{B}_{\perp,i-1}\mathbf{l}_i,\\ \mathbf{Z}_i&=\frac{\sqrt{2}}{\delta_i}\left[\mathrm{Re}(\mathbf{v}_i),\ \gamma_i\left(\mathrm{Im}(\beta_i)\,\mathrm{Re}(\mathbf{v}_i)+\delta_i^2\,\mathrm{Im}(\mathbf{v}_i)\right)\right],\\ \mathbf{B}_{\perp,i}&=\mathbf{B}_{\perp,i-1}+\frac{\sqrt{2}\alpha_i}{\delta_i}\mathbf{E}\mathbf{Z}_i\left[\mathrm{Re}(\mathbf{l}_i),\ \gamma_i\left(\mathrm{Im}(\beta_i)\,\mathrm{Re}(\mathbf{l}_i)+\delta_i^2\,\mathrm{Im}(\mathbf{l}_i)\right)\right]^T,\end{aligned} \tag{5.55}$$

where each real $\mathbf{Z}_i\in\mathbb{R}^{N\times 2}$ *contains both ADI bases for the pair* $(s_i,\mathbf{l}_i)$ *and the complex conjugated pair* $(\overline{s}_i,\overline{\mathbf{l}}_i)$, *i. e.* $\mathrm{span}(\mathbf{Z}_i)=\mathrm{span}\{\mathbf{v}_i,\overline{\mathbf{v}}_i\}$, *then* $\widehat{\mathbf{P}}=\mathbf{Z}\mathbf{Z}^T$ *satisfies the conditions of Theorem 4.26 for input* $\mathcal{H}_2$ *pseudo-optimality.*

Proof. First of all, note that γ_i always exists. This is due to $\beta_i\overline{\beta}_i=\mathbf{l}_i^*\overline{\mathbf{l}}_i\mathbf{l}_i^T\mathbf{l}_i\frac{\mathrm{Re}^2(s_i)}{(s_i\overline{s}_i)}$. Direct complex analysis shows that $0\le\mathbf{l}_i^*\overline{\mathbf{l}}_i\mathbf{l}_i^T\mathbf{l}_i\le 1$, which is due to the assumption $\|\mathbf{l}_i\|_2=1$, and that $0<\frac{\mathrm{Re}^2(s_i)}{(s_i\overline{s}_i)}<1$, because s_i has non-zero imaginary part. Therefore, $0<\beta_i\overline{\beta}_i<1$, and γ_i always exists.

As the re-formulated ADI iteration comprises a restart after every single step, it is here again sufficient to show that the first iterate $\widehat{\mathbf{P}}=\mathbf{Z}_1\mathbf{Z}_1^T$ is equal to the outcome of PORK-Lyap, because then the conditions of Theorem 4.26 for input $\mathcal{H}_2$ pseudo-optimality will be ensured. To this end, choose $\mathbf{V}=[\mathbf{v}_1,\overline{\mathbf{v}}_1]$ as a basis of the desired tangential-input rational Krylov subspace, which includes the pair $(s_i,\mathbf{l}_i)$ and its complex conjugated pair $(\overline{s}_i,\bar{\mathbf{l}}_i)$. Then, the $\mathbf{B}$-Sylvester equation $\mathbf{AV}-\mathbf{EVS}_{\mathrm{ADI}}=\mathbf{BL}_{\mathrm{ADI}}$ is satisfied for $\mathbf{S}_{\mathrm{ADI}}=\mathrm{diag}(s_1,\overline{s}_1)$, and $\mathbf{L}_{\mathrm{ADI}}=\left[\mathbf{l}_1,\bar{\mathbf{l}}_1\right]\alpha_1$. As by assumption $\mathbf{l}_1^*\mathbf{l}_1=1$, the solution at Step 1 of PORK-Lyap is

$$\mathbf{P}_r^{-1}=\begin{bmatrix} 1 & \beta_1 \\ \overline{\beta}_1 & 1 \end{bmatrix}, \quad \text{and thus} \quad \mathbf{P}_r=\left(\mathbf{P}_r^{-1}\right)^{-1}=\gamma^2\begin{bmatrix} 1 & -\beta_1 \\ -\overline{\beta}_1 & 1 \end{bmatrix}. \tag{5.56}$$

In order to get a real basis, introduce the unitary basis transformation $\mathbf{T}\in\mathbb{C}^{2\times 2}$,

$$\mathbf{T}=\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -\imath \\ 1 & \imath \end{bmatrix}, \quad \mathbf{TT}^*=\mathbf{I}, \tag{5.57}$$

because then $\widehat{\mathbf{P}}=\mathbf{VP}_r\mathbf{V}^*=\mathbf{VTT}^*\mathbf{P}_r\mathbf{TT}^*\mathbf{V}^*=\widetilde{\mathbf{V}}\widetilde{\mathbf{P}}_r\widetilde{\mathbf{V}}^*$, with $\widetilde{\mathbf{P}}_r=\mathbf{T}^*\mathbf{P}_r\mathbf{T}\in\mathbb{R}^{2\times 2}$ and $\widetilde{\mathbf{V}}=\mathbf{VT}=\sqrt{2}[\mathrm{Re}(\mathbf{v}_1),\,\mathrm{Im}(\mathbf{v}_1)]\in\mathbb{R}^{N\times 2}$. It can be shown by exploiting the definitions (5.54), that $\widetilde{\mathbf{P}}_r$ and its Cholesky factorization $\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^*=\widetilde{\mathbf{P}}_r$ are given by

$$\widetilde{\mathbf{P}}_r=\mathbf{T}^*\mathbf{P}_r\mathbf{T}=\gamma_1^2\begin{bmatrix} 1-\mathrm{Re}(\beta_1) & \mathrm{Im}(\beta_1) \\ \mathrm{Im}(\beta_1) & 1+\mathrm{Re}(\beta_1) \end{bmatrix} \text{ and } \widetilde{\mathbf{R}}=\frac{1}{\delta_1}\begin{bmatrix} 1 & \gamma_1\,\mathrm{Im}(\beta_1) \\ 0 & \gamma_1\delta_1^2 \end{bmatrix}. \tag{5.58}$$

As $\widehat{\mathbf{P}}=\widetilde{\mathbf{V}}\widetilde{\mathbf{R}}\widetilde{\mathbf{R}}^*\widetilde{\mathbf{V}}^*$, the result for $\mathbf{Z}_1=\widetilde{\mathbf{V}}\widetilde{\mathbf{R}}$ can be concluded. It is left to prove the statement for $\mathbf{B}_{\perp,1}$, which follows from Step 3 of PORK-Lyap:

$$\mathbf{B}_{\perp,1}=\mathbf{B}+\mathbf{EVP}_r\mathbf{L}_{\mathrm{ADI}}^*=\mathbf{B}+\mathbf{E}\widetilde{\mathbf{V}}\widetilde{\mathbf{P}}_r\mathbf{T}^*\mathbf{L}_{\mathrm{ADI}}^*=\mathbf{B}+\mathbf{EZ}_1(\mathbf{L}_{\mathrm{ADI}}\mathbf{T}\widetilde{\mathbf{R}})^*. \tag{5.59}$$

With $\mathbf{L}_{\mathrm{ADI}}\mathbf{T}=\sqrt{2}\alpha_1[\mathrm{Re}(\mathbf{l}_1),\,\mathrm{Im}(\mathbf{l}_1)]\in\mathbb{R}^{m\times 2}$, the result follows. □

Both ADI iterations for real (5.53) and complex conjugated shifts (5.55), can be combined. An implementation of the resulting T-LR-ADI iteration is illustrated in Algorithm 5.2. There, the real low-rank factor $\mathbf{Z}=[\mathbf{Z}_1,\,\mathbf{Z}_2,\,,\ldots,\,\mathbf{Z}_k]$ is computed iteratively, where $\mathbf{Z}_i$ has one column if s_i is real, and two columns if s_i is complex. Additionally, the low-rank factor $\mathbf{B}_\perp$ of the residual is iteratively computed.

Remark 5.22. The reasoning for the convergence criterion is as follows: for the approximation $\widehat{\mathbf{P}}=\mathbf{0}$, we have the residual $\mathbf{BB}^T$, hence we might be interested in the

Algorithm 5.2 Tangential-Low-Rank-ADI (T-LR-ADI)

```
Input: $\mathbf{E}$, $\mathbf{A}$, $\mathbf{B}$, $tol$
Output: Approximation $\widehat{\mathbf{P}} = \mathbf{Z}\mathbf{Z}^T$ and residual $\mathbf{R} = \mathbf{B}_\perp \mathbf{B}_\perp^T$
 1: initial choice of $s_1 \in \mathbb{C}$ and $\mathbf{l}_1 \in \mathbb{C}^m$ with $\|\mathbf{l}_1\|_2 = 1$
 2: $\mathbf{Z} = [\,]$, $\mathbf{B}_\perp = \mathbf{B}$
 3: repeat
 4:     solve $(\mathbf{A} - s_i \mathbf{E})\,\mathbf{v} = \mathbf{B}_\perp \mathbf{l}_i$ for $\mathbf{v}$
 5:     if $s_i \in \mathbb{R}$ and $\mathbf{l}_i \in \mathbb{R}^m$ then
 6:         $\mathbf{Z}_i = \sqrt{2 s_i}\mathbf{v}, \ \mathbf{L}_i = \sqrt{2 s_i}\mathbf{l}_i$
 7:     else
 8:         $\beta = \mathbf{l}_i^* \bar{\mathbf{l}}_i \frac{\mathrm{Re}(s_i)}{\bar{s}_i}, \quad \gamma = \frac{1}{\sqrt{1 - \beta\bar{\beta}}}, \quad \delta = \sqrt{1 + \mathrm{Re}(\beta)}$
 9:         $\mathbf{Z}_i = \frac{2}{\delta}\sqrt{\mathrm{Re}(s_i)}\,[\mathrm{Re}(\mathbf{v}), \ \gamma\,(\mathrm{Im}(\beta)\,\mathrm{Re}(\mathbf{v}) + \delta^2\,\mathrm{Im}(\mathbf{v}))]$
10:         $\mathbf{L}_i = \frac{2}{\delta}\sqrt{\mathrm{Re}(s_i)}\,[\mathrm{Re}(\mathbf{l}_i), \ \gamma\,(\mathrm{Im}(\beta)\,\mathrm{Re}(\mathbf{l}_i) + \delta^2\,\mathrm{Im}(\mathbf{l}_i))]^T$
11:     end if
12:     $\mathbf{Z} = [\mathbf{Z}, \mathbf{Z}_i]$
13:     $\mathbf{B}_\perp = \mathbf{B}_\perp + \mathbf{E}\mathbf{Z}_i\mathbf{L}_i^T$
14:     determine $s_{i+1}$ and $\mathbf{l}_{i+1}$ with $\|\mathbf{l}_{i+1}\|_2 = 1$
15: until $\|\mathbf{B}_\perp\|_2 < tol\,\|\mathbf{B}\|_2$
```

relative residual error, which would yield $\|\mathbf{R}\|_2 < tol\,\left\|\mathbf{B}\mathbf{B}^T\right\|_2$. As $\mathbf{R} = \mathbf{B}_\perp \mathbf{B}_\perp^T$, this is however equivalent to Step 15 of Algorithm 5.2. Nevertheless, one might as well think of alternative criteria for convergence.

T-LR-ADI by Algorithm 5.2 represents a generalization of the block ADI iteration (5.51). To demonstrate this, assume m equal real shifts $s_1 = s_2 = \ldots = s_m$ and an orthonormal basis $\{\mathbf{l}_1, \ldots, \mathbf{l}_m\}$ as tangential directions. Then the outcome of T-LR-ADI, $[\mathbf{z}_1, \ldots, \mathbf{z}_m]$, is equal to the first iterate of the block ADI iteration (5.42) or (5.51). Therefore, the block ADI iteration is a special case of T-LR-ADI, where the latter provides an additional degree of freedom: instead of being restricted to the whole block, we may pick only certain directions of our choice.

It is left open, how to determine the shifts s_i and tangential directions $\mathbf{l}_i$. Basically, this should be done on the basis of some optimization procedure. However, this represents a research direction in its own right, which is out of the scope of this work. The interested reader is instead referred to [210], where this discussion is more detailed. There was also presented a numerical example, which justifies the idea of a tangential ADI iteration, as it indeed can outperform the ADI iteration (5.42) based on blocks.

Remark 5.23 (Parallelization). There is a final remark in order: in all formulations of the ADI iteration—that is (5.42), (5.51), (5.53) and (5.55)—one in principle has to wait until the preceding shift was processed, before the next can be used; but with the findings of this chapter, it is for the first time possible to parallelize the ADI iteration for solving Lyapunov equations. Without going into computational details, this will be demonstrated by a simple example. Assume a single input $m=1$, and that we have access to k processors, and that a set of k shifts $s_1,\ldots,s_k$ is given. The main numerical effort in the ADI iteration is to solve the k LSEs $(\mathbf{A}-s_i\mathbf{E})\,\mathbf{v}_i=\mathbf{b}$ for $\mathbf{v}_i$, so it would be desirable to distribute them and simultaneously solve them on the k processors. If we do so, a final step to merge the different solves is required. This could be done by plugging the matrices $\mathbf{V}=[\mathbf{v}_1,\ldots,\mathbf{v}_k]$, $\mathbf{S}=\operatorname{diag}(s_1,\ldots,s_k)$, and $\mathbf{L}=[1,\ldots,1]$ into PORK-Lyap and solve the Lyapunov equation of dimension k by direct methods for $\mathbf{P}_r$. The ADI basis $\mathbf{Z}$ then follows from the Cholesky factorization $\mathbf{P}_r=\mathbf{R}\mathbf{R}^*$, and by taking $\mathbf{Z}=\mathbf{V}\mathbf{R}$, whereas the residual factor is given by $\mathbf{b}_\perp=\mathbf{b}+\mathbf{E}\mathbf{V}\mathbf{P}_r\mathbf{L}^*=\mathbf{b}+\mathbf{E}\mathbf{Z}\mathbf{R}_r^*\mathbf{L}^*$. Then we could repeat the whole procedure by determining k new shifts (or possibly recycle the already given ones). In conclusion, if k processors are available, it is reasonable to distribute k LSEs and compute the joint approximation $\widehat{\mathbf{P}}$ by PORK-Lyap—instead of using one of the iterative procedures (5.42), (5.51), (5.53) or (5.55), which perform a restart after every single shift, and thereby render parallelization impossible.

5.4.4 Overview on the Link Between ADI and Krylov

The link between Krylov based methods and the ADI iteration for solving Lyapunov equations has been investigated by various authors. After introducing the low-rank formulation of the ADI iteration, Li and White [128] already proved that the ADI basis $\mathbf{Z}$ spans a rational Krylov subspace. It has been accepted since then, that both methods are connected somehow; Gallivan et al. [79] expressed this in the following way:

> *"Even though these methods [Editor: ADI iteration included] cannot directly be interpreted as interpolation techniques, they are linked to Krylov based interpolation."*

The results of this section, however, suggest that the ADI iteration may indeed be interpreted as an interpolation technique, by associating (virtual) reduced data to its constitutive functional iterations (5.42), (5.51), (5.53) or (5.55). First steps in this direction yet are due to Flagg and Gugercin [67] and Druskin et al. [55]: it was independently proven in their works, that the approximation $\widehat{\mathbf{P}}$ of the ADI iteration equals the one of RKSM with $\mathbf{W}=\mathbf{V}$, if and only if the eigenvalues of the projected matrix $\mathbf{E}_r^{-1}\mathbf{A}_r$ are the mirror images of the shifts s_i, with respect to the imaginary axis. This condition, however, can hold true only for special sets of shifts s_i (these could be computed in an IRKA-like manner). The results given here, and which has been presented in [207, 213], instead describe a constructive way how the approximation of the ADI iteration can be computed by Krylov-based projections—for arbitrary shifts. Thereby, the “oblique” nature of the ADI iteration is identified, as the Krylov-based approach generally requires oblique projections, in order to copy the approximation of the ADI iteration. It should be noted, that the “obliqueness” of the projection that the ADI iteration is associated with, can be measured with low numerical effort. This was presented in [213], where this measure is also used to estimate the quality of approximation.

The link between RKSM and ADI is proven in [55] by showing the equivalence of the ADI iteration to the so-called skeleton approximation. There are therefore three approaches for solving large-scale Lyapunov equations—ADI iteration, skeleton approximation and RKSM in terms of $\mathcal{H}_2$ pseudo-optimality—, all of which originate from completely different motivations, but still generate equal approximations.

The $\mathcal{H}_2$ pseudo-optimal nature of the ADI iteration, was already recognized by Flagg and Gugercin [67], however, only for the single-input case. It was noted, that this “proves harder to extend” to multiple inputs $m > 1$, which was also considered as an “interesting research direction to pursue”. This work presents the full generalization to multiple shifts and to both block-input and tangential-input Krylov subspaces.

A very interesting approach for solving large-scale Sylvester and Lyapunov equations was proposed by Ahmad et al. [3], which was denoted as “Krylov subspace restart scheme”. The starting point is quite similar to the approach pursued here: given a basis of a rational Krylov subspace, the family of possible reduced data is formulated like in Section 2.5, although this is carried out in [3] via an auxiliary (intermediate step of an) orthogonal projection. Then the remaining degree of freedom is determined, such that the rank of the residual is minimized—which in fact causes $\mathbf{R} = \mathbf{B}_{\perp}\mathbf{B}_{\perp}^{*}$. The benefit then is that the algorithm can be restarted with $\mathbf{B}_{\perp}$ instead of $\mathbf{B}$ and that the approximation $\widehat{\mathbf{P}}$ can be cumulated with guaranteed monotonically decreasing error. What was not recognized is, that the emerging reduced data has eigenvalues and expansion points as mirror images and hence, that this approach actually depicts an $\mathcal{H}_2$ pseudo-optimal approximation. Furthermore, the cumulative idea was proposed only for the solution of Sylvester and Lyapunov equations, whereas the possibility of also accumulating an approximate transfer function was not discovered. In addition, the equivalence to the ADI iteration was not recognized. Nevertheless, it is interesting to note that a completely different motivation—minimizing the rank of the residual—in fact yields the same approximation as the $\mathcal{H}_2$ pseudo-optimal framework discussed in this work. The “Krylov subspace restart scheme” by Ahmad et al. [3] thus can be seen as the fourth approach alongside the above mentioned ones, all of which yield equal approximate solutions $\widehat{\mathbf{P}}$.

It was already recognized in [67], that the residual $\mathbf{R}$ of the ADI iteration fulfils certain orthogonality conditions for special shifts; yet the first explicit formulation of the residual is probably stated in [55], which, however, is inappropriate for numerical computations, because an ill-conditioned Cauchy matrix is involved. The above formulation of the ADI residual, which is easy to implement, well-suited for numerical computations, and which directly includes the above statement on orthogonality, was first presented in the talk [207], then proven in [213], and in the meantime independently proven by Benner et al. [32].

The re-formulation (5.51) of the ADI iteration was independently found by Benner and Kürschner, cf. [31] and [34]. It was also presented in [210], where additionally the tangential ADI iteration was introduced.

It should be possible to directly generalize the results of this chapter to large-scale (i. e. sparse-sparse) Sylvester equations. This is omitted here for a concise presentation, but first ideas can be found e. g. in [34, 67]. Finally, it should be noted that the link between ADI and $\mathcal{H}_2$ pseudo-optimal RKSM was used in [209], where the effect of the approximations $\widehat{\mathbf{P}}$ from ADI and RKSM on the reduced order model by approximate balanced truncation was investigated.

6 Conclusions

This work treats model order reduction of linear time invariant systems using projections onto rational Krylov subspaces. Because theoretically any reduced model may be generated with projections onto rational Krylov subspaces, we may without loss of generality choose "Krylov" as the tool to construct reduced models.

It was shown in this work that it is useful to describe bases of rational Krylov subspaces through particular sparse-dense Sylvester equations, which may actually be understood as a duality. The in-depth analysis of this duality allows to define a family of reduced models that ensures moment matching at prescribed interpolation points. Moreover, a new proof of moment matching was derived in this work.

Sylvester equations are definitely the main tool in this research, as all proofs basically emanate from their detailed understanding. In this respect, the Sylvester equations themselves are mainly of theoretical interest, as they are not beneficial in their own right. They rather pave the way to new methods of model order reduction. By contrast, the matrices that form the Sylvester equations, these are $\mathbf{V}$, $\mathbf{S}$, $\mathbf{L}$, and $\mathbf{B}_\perp$, are indeed relevant for practical applications and new approaches to MOR. One example is the reshaped error model presented in Section 3.1: the sum $\boldsymbol{G}(s) - \boldsymbol{G}_r(s)$ is transformed into the product $\boldsymbol{G}_\perp(s)\boldsymbol{G}_f(s)$. This has both analytical and numerical advantages (see e. g. the rigorous upper bounds on the error proposed in [150]). Moreover, the error factorization is the basis for a paradigm shift in MOR towards the cumulative framework CURE, presented in Section 3.2. It permits the accumulation of independently reduced models, and at the same time the preservation of the aforementioned generality that Krylov-based projections offer.

The flexibility that projections onto Krylov subspaces provide is one of their main advantages compared to other methods of MOR. However, one has to make sure that these degrees of freedom do not turn into disadvantages, because if any reduced model may be generated, then also the worst one is possible. It is therefore essential to have useful guidelines how to determine these degrees of freedom. To this end, this thesis suggests a deliberate way to restrict at least half of the degrees of freedom. This is carried out by proposing the concept of "$\mathcal{H}_2$ pseudo-optimality". The label "pseudo-optimal" stems from the fact that a certain kind of optimality it automatically ensured. Moreover, forcing $\mathcal{H}_2$ pseudo-optimality requires only marginal numerical effort compared to the computation of bases of rational Krylov subspaces. One benefit then is that stability is preserved in the reduced model. Furthermore, $\mathcal{H}_2$ pseudo-optimality is a natural extension of the CURE framework, since it ensures that the approximation error decreases monotonically with each "salami slice", i. e. with each additional reduced model. To conclude, this thesis promotes to perform $\mathcal{H}_2$ pseudo-optimal reductions within the cumulative framework.

The sole remaining issue left over is the determination of interpolation points and tangential directions. However, this represents a research direction in its own right, which is why it is tackled in another thesis by Panzer [148]. There, the structure which the combination of the cumulative framework with $\mathcal{H}_2$ pseudo-optimality offers is exploited to propose an optimization procedure that has guaranteed convergence towards interpolation points that yield locally $\mathcal{H}_2$ optimal reduced models. One of the tools to derive these results are the small-scale and easy-to-evaluate matrix equations for $\mathcal{H}_2$ pseudo-optimality, which are presented in this thesis. They mark the main devices for the analysis and construction of $\mathcal{H}_2$ pseudo-optimal reduced models, and consequently, they are among the most important results of this thesis.

Although these results intended to improve MOR using projections onto rational Krylov subspaces, they may also be exploited to approximately solve large-scale Lyapunov equations. It was shown that applying the ideas of $\mathcal{H}_2$ pseudo-optimality to Lyapunov equations actually results in the same approximation as one would obtain from the prevalent ADI iteration. This thesis therefore not only provides a novel view on the ADI iteration, but it also offers tools for the analysis and improvement of the ADI iteration. One such example is the numerically efficient low-rank formulation of the residual, which was presented in this thesis, and which eases convergence analysis of the ADI iteration. Moreover, the disclosure of the link between the ADI iteration and projections onto rational Krylov subspaces enables the generalization of the ADI iteration to tangential interpolation. The benefit is that this new functionality might prevent the ADI basis from growing too large in certain cases, and it thereby ensures that the final approximation stays numerically manageable. Finally, the results of this work allow for the first time to parallelize the computations involved in the ADI iteration. If multiple processors are available, then the suggested ideas for distributing computations to these processors have the potential to massively reduce computational time of the ADI iteration.

Although $\mathcal{H}_2$ pseudo-optimality seems to be a promising contribution to MOR, there are of course still open questions that remain to be clarified. For example, the optimization procedures for the determination of interpolation points allow for improvement; especially elaborate algorithms for the optimal selection of tangential directions are still lacking. Maybe also some ideas for optimization that are contained in the mentioned literature on $\mathcal{H}_2$ pseudo-optimality based on data of transfer functions might allow to be translated into a large-scale setting based on rational Krylov subspaces.

Future work might also include some generalizations of the presented results. This might be the formulation of necessary and sufficient conditions for frequency weighted $\mathcal{H}_2$ pseudo-optimality. Moreover, the results of this research may be generalized to the solution of large-scale (sparse-sparse) Sylvester equations, which, however, should be straightforward.

Appendix A

Proof of Theorem 4.26

Equations (2.42) and (4.40) will be required in the following rewritten form:

$$\mathbf{S} = \mathbf{E}_r^{-1}\mathbf{A}_r - \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L}, \tag{A.1}$$

$$-\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} = \mathbf{E}_r^{-1}\mathbf{A}_r + \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}. \tag{A.2}$$

Proof of i) ⇔ ii): Subtract (A.2) from (A.1):

$$\mathbf{S} + \mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} = -\mathbf{E}_r^{-1}\mathbf{B}_r\left(\mathbf{L} + \mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\right). \tag{A.3}$$

If the left hand side is zero, so is the right hand side and vice versa. Setting the left hand side to zero is equivalent to condition *i)*, and setting the right hand side to zero is equivalent to condition *ii)*, if $\mathbf{B}_r$ has full column rank.
Although it was assumed in the theorem that $\mathbf{B}_r$ has full column rank, let us briefly consider the case that this is not satisfied (this is particularly the case if $n < m$): it will follow from the rest of the proof that conditions *ii)–vi)* are equivalent to each other, irrespective of whether $\mathbf{B}_r$ has full column rank or not. If now $\mathbf{B}_r$ has not full column rank, then (A.3) shows that conditions *ii)–vi)* are sufficient but not necessary for condition *i)*. □

Proof of ii) ⇔ iii): Replacing $\mathbf{E}_r^{-1}\mathbf{A}_r$ and $\mathbf{A}_r^*\mathbf{E}_r^{-*}$ in (A.2) by (A.1), and multiplying the result with $\mathbf{P}_r$ from the right leads to

$$\mathbf{S}\mathbf{P}_r + \mathbf{P}_r\mathbf{S}^* + \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L}\mathbf{P}_r + \mathbf{P}_r\mathbf{L}^*\mathbf{B}_r^*\mathbf{E}_r^{-*} + \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*} = \mathbf{0}. \tag{A.4}$$

By using

$$\begin{aligned}\left(\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^*\right)\left(\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^*\right)^* &= \\ \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*} + \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L}\mathbf{P}_r &+ \mathbf{P}_r\mathbf{L}^*\mathbf{B}_r^*\mathbf{E}_r^{-*} + \mathbf{P}_r\mathbf{L}^*\mathbf{L}\mathbf{P}_r,\end{aligned} \tag{A.5}$$

it follows that

$$\mathbf{S}\mathbf{P}_r + \mathbf{P}_r\mathbf{S}^* - \mathbf{P}_r\mathbf{L}^*\mathbf{L}\mathbf{P}_r = -\left(\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^*\right)\left(\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^*\right)^*. \tag{A.6}$$

If the left-hand side is zero, so is the right-hand side and vice versa, which proves equivalence of conditions *ii)* and *iii)*. □

Proof of ii) ⇔ iv): Noting that $\mathbf{X}$ in (4.38) is unique, we insert *iv)* in (4.38) and multiply the result with $\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}$ from the right:

$$\mathbf{A}\mathbf{V} + \mathbf{E}\mathbf{V}\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} + \mathbf{B}\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} = \mathbf{0}. \tag{A.7}$$

Using (A.2) for $\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}$ yields

$$\Leftrightarrow \quad \mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r^{-1} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} + \mathbf{B}\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} = \mathbf{0}. \tag{A.8}$$

Substituting $\mathbf{B} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{B}_r = \mathbf{B}_\perp$ we get

$$\Leftrightarrow \quad \mathbf{A}\mathbf{V} - \mathbf{E}\mathbf{V}\mathbf{E}_r^{-1}\mathbf{A}_r^{-1} = \mathbf{B}_\perp\left(-\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\right). \tag{A.9}$$

Subtracting (A.9) from (2.39) yields $\mathbf{B}_\perp(\mathbf{L} + \mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}) = \mathbf{0}$. As $\mathbf{B}_\perp$ is assumed to have full column rank, condition *ii)* follows, showing equivalence to condition *iv)*. □

Proof of ii) ⇔ v): Owing to Theorem 5.3,

$$\mathbf{A}\widehat{\mathbf{P}}\mathbf{E}^T + \mathbf{E}\widehat{\mathbf{P}}\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{B}_\perp\mathbf{B}_\perp^* + \mathbf{F}\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{F}^*. \tag{A.10}$$

Condition *v)* therefore is equivalent to $\mathbf{F}\mathbf{B}_\perp^* + \mathbf{B}_\perp\mathbf{F}^* = \mathbf{0}$. This in turn is equivalent to $\mathbf{F} = \mathbf{0}$, because on the one hand $\mathbf{B}_\perp$ is assumed to have full column rank and on the other hand $\operatorname{span}(\mathbf{B}_\perp) \neq \operatorname{span}(\mathbf{F})$. Finally, $\mathbf{F} = \mathbf{E}\mathbf{V}(\mathbf{E}_r^{-1}\mathbf{B}_r + \mathbf{P}_r\mathbf{L}^*) = \mathbf{0}$ is equivalent to condition *ii)*. □

Proof of iii) ⇔ vi): Starting from condition *iii)*

$$\mathbf{P}_r^{-1}\mathbf{S} + \mathbf{S}^*\mathbf{P}_r^{-1} - \mathbf{L}^*\mathbf{L} = \mathbf{0}, \tag{A.11}$$

we replace $\mathbf{S}$ by (A.1)

$$\mathbf{S} = \mathbf{E}_r^{-1}\mathbf{A}_r - \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L} \overset{ii)}{=} \mathbf{E}_r^{-1}\mathbf{A}_r + \mathbf{P}_r\mathbf{L}^*\mathbf{L}, \tag{A.12}$$

to show that this is equivalent to *vi)*:

$$\Leftrightarrow \quad \mathbf{P}_r^{-1}\mathbf{E}_r^{-1}\mathbf{A}_r + \mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} + \mathbf{L}^*\mathbf{L} = \mathbf{0}, \tag{A.13}$$

$$\Leftrightarrow \quad \mathbf{E}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\mathbf{E}_r^{-1}\mathbf{A}_r + \mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\mathbf{E}_r^{-1}\mathbf{E}_r + \mathbf{L}^*\mathbf{L} = \mathbf{0}. \tag{A.14}$$

By comparing the Lyapunov equations (A.14) and (4.83), we can identify $\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}\mathbf{E}_r^{-1} = \mathbf{Q}_f$, which proves equivalence of conditions *i)–vi)*. □

Proof of vii) ⇔ vi),ii): $\mathbf{P}_r$ and $\mathbf{E}_r^*\mathbf{Q}_f\mathbf{E}_r$ define the Controllability and Observability Gramian of $\boldsymbol{G}_f(s) = \mathbf{L}\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r + \mathbf{I}$, respectively. By condition *vi)*, it holds that $\mathbf{P}_r\mathbf{E}_r^*\mathbf{Q}_f\mathbf{E}_r = \mathbf{I}$. Using the result of Glover [84, Theorem 5.1], this is equivalent to the existence of a feed-through $\mathbf{D}$, such that $\mathbf{L}\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r + \mathbf{D}$ is all-pass. Taking $\mathbf{D} = \mathbf{I}$ and employing condition *ii)*, $\mathbf{B}_r^*\mathbf{E}_r^{-*} + \mathbf{L}\mathbf{P}_r = 0$, we can check with [84, Theorem 5.1] that $\boldsymbol{G}_f(s)\boldsymbol{G}_f^*(-\overline{s}) = \mathbf{I}$. (Please note that there is a little typo in [84, Theorem 5.1], as it is printed as: $\boldsymbol{G}_f(s)\boldsymbol{G}_f^*(-s) = \mathbf{I}$.) □

Proof of i) ⇒ viii): It is obvious that condition *viii)* is necessary for condition *i)*. □

Proof of i) ⇐ viii) if $\mathbf{V}$ *spans single- or block-input Krylov subspace:* We only prove the case that $\mathbf{V}$ spans a rational block-input Krylov subspace, as the single-input case then is directly included. Consider equation (A.1):

$$\mathbf{S} = \mathbf{E}_r^{-1}\mathbf{A}_r - \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{L}. \tag{A.15}$$

Condition *viii)* leads to the interpretation of (A.15) as a pole placement problem in control theory: we are searching for the "feedback" $\mathbf{L}$ such that the eigenvalues of $\mathbf{E}_r^{-1}\mathbf{A}_r$ are mirrored along the imaginary axis. Because condition *viii)* requires that all eigenvalues are assigned, it follows that the pair $(\mathbf{E}_r^{-1}\mathbf{A}_r, \mathbf{E}_r^{-1}\mathbf{B}_r)$ must be controllable. Generally, the multivariable pole placement problem has many solutions. However, due to block Krylov subspaces, all to-be-assigned eigenvalues of $\mathbf{S}$ have geometric multiplicity m, and the Jordan blocks to each eigenvalue have equal dimensions, cf. Theorem 2.4 and Corollary 2.6. In this particular case, O'Reilly and Fahmy [147, Corollary 8] proved that the feedback $\mathbf{L}$ becomes unique. Consider again equation (A.2) in order to identify the desired feedback:

$$-\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1} = \mathbf{E}_r^{-1}\mathbf{A}_r + \mathbf{E}_r^{-1}\mathbf{B}_r\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}. \tag{A.16}$$

As the left-hand side shares the desired Jordan normal form, (A.16) constitutes the same pole placement problem as (A.15). Due to uniqueness of the "feedback" $\mathbf{L}$, we can identify $\mathbf{L} = -\mathbf{B}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}$ and $\mathbf{S} = -\mathbf{P}_r\mathbf{A}_r^*\mathbf{E}_r^{-*}\mathbf{P}_r^{-1}$ which proves that *viii)* $\Rightarrow$ *i)* and *viii)* $\Rightarrow$ *ii)*, if $\mathbf{V}$ spans a rational *block*-input Krylov subspace □

Appendix B

Proof of Theorem 4.27

Proof of i): Define $\mathbf{A}_{\mathrm{H}} = \mathbf{P}_r^{-1}\mathbf{E}_r^{-1}\mathbf{A}_r\mathbf{P}_r$, $\mathbf{B}_{\mathrm{H}} = -\mathbf{P}_r^{-1}\mathbf{E}_r^{-1}\mathbf{B}_r$, and $\mathbf{C}_{\mathrm{H}} = -\mathbf{C}_r\mathbf{P}_r$, then

$$\mathbf{C}_{\mathrm{H}}\left(s\mathbf{I} - \mathbf{A}_{\mathrm{H}}\right)^{-1}\mathbf{B}_{\mathrm{H}} = \mathbf{C}_r\mathbf{P}_r\mathbf{P}_r^{-1}\left(s\mathbf{I} - \mathbf{E}_r^{-1}\mathbf{A}_r\right)^{-1}\mathbf{P}_r\mathbf{P}_r^{-1}\mathbf{E}_r^{-1}\mathbf{B}_r = \boldsymbol{G}_r(s) \tag{B.1}$$

is an admissible state-space realization of the reduced model. If the conditions of Theorem 4.26 are satisfied, then $\mathbf{A}_{\mathrm{H}} \overset{i)}{=} -\mathbf{S}^*$ and $\mathbf{B}_{\mathrm{H}} \overset{ii)}{=} \mathbf{L}^*$ and hence, $\boldsymbol{G}_r(s) = \mathbf{C}_{\mathrm{H}}\left(s\mathbf{I} - (-\mathbf{S}^*)\right)^{-1}\mathbf{L}^*$.

We first prove the case that $\mathbf{V}$ spans a rational tangential-input Krylov subspace with k expansion points s_i, $i = 1, \dots, k$ and respective tangential directions $\mathbf{L}_i \in \mathbb{C}^{m \times m_i}$. We may assume without loss of generality that in this case

$$\mathbf{S} = \begin{bmatrix} s_1\mathbf{I} & & \\ & \ddots & \\ & & s_k\mathbf{I} \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & \dots & \mathbf{L}_k \end{bmatrix}, \quad \mathbf{C}_{\mathrm{H}} = \begin{bmatrix} \mathbf{C}_1 & \dots & \mathbf{C}_k \end{bmatrix}, \tag{B.2}$$

which yields $\boldsymbol{G}_r(s) = \mathbf{C}_{\mathrm{H}}\left(s\mathbf{I} - (-\mathbf{S}^*)\right)^{-1}\mathbf{L}^* = \sum_{i=1}^{k} \frac{\mathbf{C}_i\mathbf{L}_i^*}{s+\overline{s}_i}$, from which we can identify the reduced eigenvalues $\lambda_i = -\overline{s}_i$ and the respective input residues $\mathbf{B}_i = \mathbf{L}_i^*$. Then the tangential interpolation $\boldsymbol{G}(s_i)\mathbf{L}_i = \boldsymbol{G}_r(s_i)\mathbf{L}_i$, $i = 1, \dots, k$, due to Theorem 2.4 becomes $\boldsymbol{G}(-\overline{\lambda}_i)\mathbf{B}_i^* = \boldsymbol{G}_r(-\overline{\lambda}_i)\mathbf{B}_i^*$, $i = 1, \dots, k$, which proves $\mathcal{H}_2$ pseudo-optimality due to Theorem 4.19.

As was noted before Corollary 4.24, the above result already includes single-, block- and tangential-input Krylov subspaces, if each expansion point is used only once. For the generalization to higher multiplicities, consider the case that $\mathbf{V}$ spans a rational tangential-input Krylov subspace with a single expansion point s_0. Without loss of generality, we may assume in this case

$$\mathbf{S} = \begin{bmatrix} s_0\mathbf{I} & \mathbf{I} & & \\ & \ddots & \ddots & \\ & & \ddots & \mathbf{I} \\ & & & s_0\mathbf{I} \end{bmatrix}, \quad \mathbf{L} = [\, \mathbf{L}_1 \ \ \ldots \ \ \mathbf{L}_k \,], \quad \mathbf{C}_{\mathrm{H}} = [\, \mathbf{C}_1 \ \ \ldots \ \ \mathbf{C}_k \,], \tag{B.3}$$

from which it follows that $\mathbf{A}_{\mathrm{H}} = -\mathbf{S}^*$ and $\mathbf{B}_{\mathrm{H}} = \mathbf{L}^*$ are in the form assumed in Theorem 4.22, and that the reduced eigenvalue is $\lambda = -\overline{s}_0$ and that the input residues are $\mathbf{B}_i = \mathbf{L}_i^*$. Therefore, the tangential interpolation

$$\left(\mathbf{M}_0^{s_0} - \widehat{\mathbf{M}}_0^{s_0}\right)\mathbf{L}_1 = \mathbf{0} \tag{B.4}$$

$$\left(\mathbf{M}_0^{s_0} - \widehat{\mathbf{M}}_0^{s_0}\right)\mathbf{L}_2 + \left(\mathbf{M}_1^{s_0} - \widehat{\mathbf{M}}_1^{s_0}\right)\mathbf{L}_1^* = \mathbf{0} \tag{B.5}$$

$$\vdots$$

$$\sum_{i=0}^{k-1}\left(\mathbf{M}_i^{s_0} - \widehat{\mathbf{M}}_i^{s_0}\right)\mathbf{L}_{k-i} = \mathbf{0} \tag{B.6}$$

become the necessary and sufficient conditions for $\mathcal{H}_2$ pseudo-optimality stated in Theorem 4.22.

As noted before Theorem 4.22, the combination of different expansion points in $\mathbf{V}$ with higher multiplicities requires cumbersome notation and is omitted for the sake of a concise presentation. □

Proof of ii): The gradient of J with respect to $\mathbf{C}_r$ is given by [187, 203]: $\nabla_{\mathbf{C}_r} J = 2\left(\mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X}\right) = 2\mathbf{C}\left(\mathbf{V}\mathbf{P}_r - \mathbf{X}\right)$. By condition *iv)* of Theorem 4.26, namely $\mathbf{X} = \mathbf{V}\mathbf{P}_r$, it directly follows that $\nabla_{\mathbf{C}_r} J = 0$. □

Proof of existence of $\mathbf{V}$*:* If $\boldsymbol{G}_r(s)$ is input $\mathcal{H}_2$ pseudo-optimal, then it satisfies the interpolatory conditions of Theorems 4.19 or 4.22. Then define the matrix $\mathbf{S}$ in Jordan canonical form, with the mirror images $-\overline{\lambda}_i$ of the reduced poles as eigenvalues, and the matrix $\mathbf{L}$, with the residues $\mathbf{B}_i^*$ as columns. Let $\mathbf{V}$ solve the $\mathbf{B}$-Sylvester equation (2.15), and define the family $\boldsymbol{G}_{\mathbf{F}}(s)$ by (2.44), i.e. the reduced model $\boldsymbol{G}_r(s) = \mathbf{C}_r\left(s\mathbf{E}_r - \mathbf{A}_r\right)^{-1}\mathbf{B}_r$ with $\mathbf{A}_r = \mathbf{S} + \mathbf{F}\mathbf{L}$, $\mathbf{E}_r = \mathbf{I}$, $\mathbf{B}_r = \mathbf{F}$ and $\mathbf{C}_r = \mathbf{C}\mathbf{V}$ with the free parameter $\mathbf{F}$. Then it holds $\boldsymbol{G}(-\overline{\lambda}_i)\mathbf{B}_i^* = \boldsymbol{G}_r(-\overline{\lambda}_i)\mathbf{B}_i^*$ due to Theorem 2.15 and we have to show that there exists an $\mathbf{F}$, such that the conditions of Theorem 4.26 are satisfied. To this end, let $\mathbf{P}_r$ be the unique solution of the Lyapunov equation given by condition *iii)* of Theorem 4.26 and choose $\mathbf{F} = -\mathbf{P}_r\mathbf{L}^*$. Then it is left to prove that $\mathbf{P}_r$ indeed is the Controllability

Gramian of the constructed reduced model, which is defined as

$$\begin{align} \mathbf{A}_r\mathbf{P}_r + \mathbf{P}_r\mathbf{A}_r^* + \mathbf{B}_r\mathbf{B}_r^* &\overset{\mathbf{A}_r=\mathbf{S}+\mathbf{F}\mathbf{L}}{=} \mathbf{S}\mathbf{P}_r + \mathbf{P}_r\mathbf{S}^* + \mathbf{F}\mathbf{L}\mathbf{P}_r + \mathbf{P}_r\mathbf{L}^*\mathbf{F}^* + \mathbf{B}_r\mathbf{B}_r^* \tag{B.7}\\ &\overset{\mathbf{B}_r=\mathbf{F}=-\mathbf{P}_r\mathbf{L}^*}{=} \mathbf{S}\mathbf{P}_r + \mathbf{P}_r\mathbf{S}^* - \mathbf{P}_r\mathbf{L}^*\mathbf{L}\mathbf{P}_r. \tag{B.8} \end{align}$$

Equation (B.8) is equal to zero by construction of $\mathbf{P}_r$, and hence, the reduced model $\boldsymbol{G}_r(s)$ satisfies the conditions of Theorem 4.26, which completes the proof. □

References

[1] M. I. Ahmad, M. Frangos, and I. M. Jaimoukha. “Second order $\mathcal{H}_2$ optimal approximation of linear dynamical systems”. In: *18th IFAC World Congress.* Milano, Italy, 2011 (cf. p. 70).

[2] M. I. Ahmad, I. M. Jaimoukha, and M. Frangos. “$\mathcal{H}_2$ optimal model reduction of linear dynamical systems”. In: *49th IEEE Conference on Decision and Control.* Atlanta, USA, 2010 (cf. p. 70).

[3] M. I. Ahmad, I. M. Jaimoukha, and M. Frangos. “Krylov Subspace Restart Scheme for Solving Large-Scale Sylvester Equations”. In: *American Control Conference.* Baltimore, USA, 2010 (cf. pp. 40, 45, 59, 105, 110, 134).

[4] M. I. Ahmad, D. B. Szyld, and M. B. van Gijzen. *Preconditioned multishift BiCG for $\mathcal{H}_2$-optimal model reduction.* Tech. rep. 12-06-15, Department of Mathematics, Temple University, 2012 (cf. p. 18).

[5] P. R. Aigrain and E. M. Williams. “Synthesis of n-reactance networks for desired transient response”. *Journal of Applied Physics,* 20.6 (1949), pp. 597–600 (cf. p. 76).

[6] B. Anderson and Y. Liu. “Controller reduction: concepts and approaches”. *IEEE Transactions on Automatic Control,* 34.8 (1989), pp. 802–812 (cf. p. 5).

[7] B. Anic, C. A. Beattie, S. Gugercin, and A. C. Antoulas. “Interpolatory weighted-$\mathcal{H}_2$ model reduction”. *Automatica,* 49 (2013), pp. 1275–1280 (cf. pp. 6, 77).

[8] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems.* SIAM, 2005 (cf. pp. 7, 8, 10, 11, 13, 27, 35, 40, 61, 69, 113).

[9] A. C. Antoulas. “On pole placement in model reduction”. *at-Automatisierungstechnik,* 9 (2007), pp. 443–448 (cf. p. 91).

[10] A. C. Antoulas and D. C. Sorensen. “Lyapunov, Lanczos, and inertia”. *Linear Algebra and Its Applications,* 326.1 (2001), pp. 137–150 (cf. p. 109).

[11] A. C. Antoulas, D. C. Sorensen, and Y. Zhou. “On the decay rate of Hankel singular values and related issues”. *Systems & Control Letters,* 46.5 (2002), pp. 323–342 (cf. pp. 9, 13).

[12] W. E. Arnoldi. “The principle of minimized iterations in the solution of the matrix eigenvalue problem”. *Quart. Appl. Math.* 9 (1951), pp. 17–29 (cf. p. 40).

[13] A. Astolfi. “Model reduction by moment matching, steady-state response and projections”. In: *IEEE Conference on Decision and Control.* 2010, pp. 5344–5349 (cf. pp. 41, 45, 46).

[14] A. Astolfi. “A new look at model reduction by moment matching for linear systems”. In: *46th IEEE Conference on Decision and Control.* 2007, pp. 4361–4366 (cf. pp. 38, 41, 45).

[15] A. Astolfi. “Model reduction by moment matching for linear and nonlinear systems”. *IEEE Transactions on Automatic Control,* 55.10 (2010), pp. 2321–2336 (cf. pp. 38, 41, 44, 45).

[16] Z. Bai. “Krylov subspace techniques for reduced-order modeling of large scale dynamical systems”. *Applied Numerical Mathematics,* 43 (2002), pp. 9–44 (cf. p. 7).

[17] V. Balakrishnan, Q. Su, and C.-K. Koh. “Efficient balance-and-truncate model reduction for large scale systems”. In: *American Control Conference.* Vol. 6. 2001, pp. 4746–4751 (cf. p. 14).

[18] R. Bartels and G. Stewart. “Solution of the matrix equation AX+XB= C”. *Comm. ACM,* 15.9 (1972), pp. 820–826 (cf. p. 12).

[19] U. Baur, P. Benner, and L. Feng. “Model order reduction for linear and nonlinear systems: a system-theoretic perspective”. *Archives of Computational Methods in Engineering,* (2014) (cf. pp. 7, 11, 20).

[20] C. A. Beattie, S. Gugercin, A. C. Antoulas, and E. Gildin. “Controller reduction by Krylov projection methods”. In: *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems.* 2004 (cf. p. 5).

[21] C. A. Beattie and S. Gugercin. “A trust region method for optimal $\mathcal{H}_2$ model reduction”. In: *IEEE Conference on Decision and Control.* 2009 (cf. pp. 62, 70, 78).

[22] C. A. Beattie and S. Gugercin. “Krylov-based minimization for optimal $\mathcal{H}_2$ model reduction”. In: *46th IEEE Conference on Decision and Control.* 2007, pp. 4385–4390 (cf. p. 78).

[23] C. A. Beattie and S. Gugercin. “Model reduction by rational interpolation”. 2014 (cf. pp. 7, 20, 75).

[24] C. A. Beattie and S. Gugercin. “Realization-independent $\mathcal{H}_2$-approximation”. In: *51st IEEE Conference on Decision and Control.* 2012, pp. 4953–4958 (cf. pp. 62, 77, 80, 83, 95, 98, 104).

[25] B. Beckermann. “An error analysis for rational Galerkin projection applied to the Sylvester equation”. *SIAM Journal on Numerical Analysis,* 49.6 (2011), pp. 2430–2450 (cf. pp. 14, 114).

[26] D. Bender. “Lyapunov-like equations and Reachability/Observability Gramians for descriptor systems”. *IEEE Transactions on Automatic Control,* 32.4 (1987), pp. 343–348 (cf. p. 11).

[27] P. Benner, J.-R. Li, and T. Penzl. “Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems”. *Numerical Linear Algebra with Applications,* 15.9 (2008), pp. 755–777 (cf. pp. 14, 105).

[28] 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,* 36.1 (2013), pp. 32–52 (cf. pp. 14, 109, 114).

[29] P. Benner and T. Damm. “Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems”. *SIAM Journal on Control and Optimization,* 49.2 (2011), pp. 686–711 (cf. p. 11).

[30] P. Benner, S. Gugercin, and K. Willcox. *A survey of model reduction methods for parametric systems.* Preprint MPIMD/13-14. Max Planck Institute Magdeburg, 2013 (cf. p. 6).

[31] P. Benner and P. Kürschner. *Computing real low-rank solutions of Sylvester equations by the factored ADI method.* Preprint MPIMD/13-05. Max Planck Institute Magdeburg, 2013 (cf. p. 135).

[32] P. Benner, P. Kürschner, and J. Saak. “An improved numerical method for balanced truncation for symmetric second-order systems”. *Mathematical and Computer Modelling of Dynamical Systems,* 19.6 (2013), pp. 593–615 (cf. pp. 14, 124, 135).

[33] P. Benner, P. Kürschner, and J. Saak. “Efficient handling of complex shift parameters in the low-rank Cholesky factor ADI method”. *Numerical Algorithms,* 62.2 (2013), pp. 225–251 (cf. pp. 14, 105, 124).

[34] P. Benner, P. Kürschner, and J. Saak. *Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations.* Preprint MPIMD/13-18. Max Planck Institute Magdeburg, 2013, pp. 123–143 (cf. pp. 14, 124, 135).

[35] P. Benner, E. S. Quintana-Ortí, and G. Quintana-Ortí. “Balanced truncation model reduction of large-scale dense systems on parallel computers”. *Mathematical and Computer Modelling of Dynamical Systems,* 6.4 (2000), pp. 383–405 (cf. p. 10).

[36] B. Besselink et al. “A comparison of model reduction techniques from structural dynamics, numerical mathematics and systems and control”. *Journal of Sound and Vibration,* 332.19 (2013), pp. 4403–4422 (cf. p. 7).

[37] S. Bhattacharyya and E. De Souza. “Pole assignment via Sylvester’s equation”. *Systems & Control Letters,* 1.4 (1982), pp. 261–263 (cf. p. 29).

[38] A. Bouhamidi, M. Hached, M. Heyouni, and K. Jbilou. “A preconditioned block Arnoldi method for large Sylvester matrix equations”. *Numerical Linear Algebra with Applications,* 20.2 (2011), pp. 208–219 (cf. p. 14).

[39] J. Brandts. “Computing tall skinny solutions of AX-XB=C”. *Mathematics and Computers in Simulation,* 61.3–6 (2003), pp. 385–397 (cf. p. 27).

[40] A. Bryson and A. Carrier. “Second-order algorithm for optimal model order reduction”. *Journal of Guidance Control and Dynamics,* 13.5 (1990), pp. 887–892 (cf. p. 77).

[41] A. Bunse-Gerstner, D. Kubalińska, and G. Vossen. *Equivalences between necessary optimality conditions for H2-norm optimal model reduction.* Berichte aus der Technomathematik 07-06. Universität Bremen, 2007 (cf. pp. 74, 76).

[42] A. Bunse-Gerstner, D. Kubalińska, G. Vossen, and D. Wilczek. "H2-norm optimal model reduction for large scale discrete dynamical MIMO systems". *Journal of computational and applied mathematics,* 233.5 (2010), pp. 1202–1216 (cf. p. 77).

[43] R. Cavin and S. Bhattacharyya. "Robust and well-conditioned eigenstructure assignment via Sylvester's equation". In: *American Control Conference.* 1982, pp. 1053–1057 (cf. p. 29).

[44] Y. Chahlaoui, K. A. Gallivan, A. Vandendorpe, and P. Van Dooren. "Dimension reduction of large-scale systems". In: *Dimension Reduction of Large-Scale Systems.* Ed. by P. Benner, D. C. Sorensen, and V. Mehrmann. Springer, 2005. Chap. Model reduction of second-order systems, pp. 149–172 (cf. p. 11).

[45] Y. Chahlaoui, D. Lemonnier, A. Vandendorpe, and P. Van Dooren. "Second-order balanced truncation". *Linear Algebra and Its Applications,* 415.2 (2006), pp. 373–384 (cf. pp. 6, 11).

[46] V. Chellaboina, W. M. Haddad, D. S. Bernstein, and D. A. Wilson. "Induced convolution operator norms of linear dynamical systems". *Mathematics of Control, Signals and Systems,* 13.3 (2000), pp. 216–239 (cf. p. 61).

[47] E. Chiprout and M. S. Nakhla. "Analysis of interconnect networks using complex frequency hopping (CFH)". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 14.2 (1995), pp. 186–200 (cf. p. 20).

[48] B. N. Datta. "Krylov subspace methods for large-scale matrix problems in control". *Future Generation Computer Systems,* 19.7 (2003), pp. 1253–1263 (cf. pp. 13, 109, 113).

[49] S. Datta, D. Chakraborty, and B. Chaudhuri. "Partial pole placement with controller optimization". *IEEE Transactions on Automatic Control,* 57.4 (2012), pp. 1051–1056 (cf. p. 30).

[50] C. De Villemagne and R. E. Skelton. "Model reductions using a projection formulation". *International Journal of Control,* 46.6 (1987), pp. 2141–2169 (cf. p. 20).

[51] J. Deutscher and C. Harkort. "Structure-preserving approximation of distributed-parameter second-order systems using Krylov subspaces". *Mathematical and Computer Modelling of Dynamical Systems,* 20.4 (2014), pp. 395–413 (cf. p. 6).

[52] J. Doyle, B. Francis, and A. Tannenbaum. *Feedback Control Theory.* Macmillan Publishing Co., 1990 (cf. p. 4).

[53] V. Druskin and V. Simoncini. "Adaptive rational Krylov subspaces for large-scale dynamical systems". *Systems & Control Letters,* 60 (2011), pp. 546–560 (cf. pp. 14, 110, 114).

[54] V. Druskin, V. Simoncini, and M. Zaslavsky. “Adaptive tangential interpolation in rational Krylov subspaces for MIMO dynamical systems”. *SIAM J. Matrix Anal. & Appl.* 35.2 (2014), pp. 476–498 (cf. pp. 14, 114).

[55] V. Druskin, L. Knizhnerman, and V. Simoncini. “Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation”. *SIAM Journal on Numerical Analysis,* 49.5 (2011), pp. 1875–1898 (cf. pp. 14, 124, 133–135).

[56] C. Eckart and G. Young. “The approximation of one matrix by another of lower rank”. *Psychometrika,* 1.3 (1936), pp. 211–218 (cf. p. 115).

[57] E. Eitelberg and G. Roppenecker. “Comments on ’A necessary condition for optimization in the frequency domain’ and on ’Optimization and pole placement for a single input controllable system’”. *International Journal of Control,* 38.2 (1983), pp. 493–494 (cf. p. 30).

[58] N. S. Ellner and E. L. Wachspress. “New ADI model problem applications”. In: *ACM Fall Joint Computer Conference.* Los Alamitos, CA, USA, 1986, pp. 528–534 (cf. p. 123).

[59] D. Enns. “Model reduction with balanced realizations: an error bound and a frequency weighted generalization”. In: *23rd IEEE Conference on Decision and Control.* Vol. 23. 1984, pp. 127–132 (cf. p. 6).

[60] P. Feldmann and R. Freund. “Efficient linear circuit analysis by Padé approximation via the Lanczos process”. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 14.5 (1995), pp. 639–649 (cf. p. 20).

[61] K. Fernando and H. Nicholson. “On a fundamental property of the cross-Gramian matrix”. *IEEE Transactions on Circuits and Systems,* 31.5 (1984), pp. 504–505 (cf. p. 11).

[62] K. Fernando and H. Nicholson. “On the cross-Gramian for symmetric MIMO systems”. *IEEE Transactions on Circuits and Systems,* 32.5 (1985), pp. 487–489 (cf. p. 11).

[63] K. Fernando and H. Nicholson. “On the structure of balanced and other principal representations of SISO systems”. *IEEE Transactions on Automatic Control,* 28.2 (1983), pp. 228–231 (cf. p. 11).

[64] A. Ferrante, W. Krajewski, A. Lepschy, and U. Viaro. “Convergent algorithm for L2 model reduction”. *Automatica,* 35.1 (1999), pp. 75–79 (cf. p. 77).

[65] G. M. Flagg, C. A. Beattie, and S. Gugercin. “Convergence of the iterative rational Krylov algorithm”. *Systems & Control Letters,* 61 (2012), pp. 688–691 (cf. pp. 75, 77).

[66] G. M. Flagg, C. A. Beattie, and S. Gugercin. “Interpolatory $\mathcal{H}_\infty$ model reduction”. *Systems & Control Letters,* 62.7 (2013), pp. 567–574 (cf. p. 5).

[67] G. M. Flagg and S. Gugercin. “On the ADI method for the Sylvester equation and the optimal-$\mathcal{H}_2$ points”. *Applied Numerical Mathematics,* 64 (2013), pp. 50–58 (cf. pp. 14, 124, 133–135).

[68] M. Frangos and I. M. Jaimoukha. “Adaptive rational interpolation: Arnoldi and Lanczos-like equations”. *European Journal of Control,* 14 (2008), pp. 342–354 (cf. pp. 40, 48, 50).

[69] M. Frangos and I. M. Jaimoukha. “Adaptive rational interpolation: restarting methods for a modified rational Arnoldi algorithm”. In: *European Control Conference.* Budapest, Hungary, 2009 (cf. p. 40).

[70] M. Frangos and I. M. Jaimoukha. “Adaptive rational Krylov algorithms for model reduction”. In: *European Control Conference.* Kos, Greece, 2007 (cf. p. 40).

[71] M. Frangos and I. M. Jaimoukha. “Rational interpolation: modified rational Arnoldi algorithm and Arnoldi-like equations”. In: *46th IEEE Conference on Decision and Control.* New Orleans, USA, 2007 (cf. p. 40).

[72] R. W. Freund. “Krylov-subspace methods for reduced-order modeling in circuit simulation”. *Journal of Computational and Applied Mathematics,* 123 (2000), pp. 395–421 (cf. pp. 6, 19).

[73] R. W. Freund. “Model reduction methods based on Krylov subspaces”. *Acta Numerica,* 12 (2003), pp. 267–319 (cf. pp. 7, 20).

[74] D. Gaier. *Lectures on Complex Approximation.* Birkhäueser, 1987 (cf. pp. 80, 103).

[75] K. Gallivan, A. Vandendorpe, and P. Van Dooren. “Model reduction via truncation: an interpolation point of view”. *Linear Algebra and Its Applications,* 375 (2003), pp. 115–134 (cf. pp. 20, 44).

[76] K. A. Gallivan, A. Vandendorpe, and P. Van Dooren. “Model reduction of MIMO systems via tangential interpolation”. *SIAM Journal on Matrix Analysis and Applications,* 26.2 (2004), pp. 328–349 (cf. pp. 30, 31, 35, 38).

[77] K. A. Gallivan, A. Vandendorpe, and P. Van Dooren. “Model reduction and the solution of Sylvester equations”. In: *17th International Symposium on Mathematical Theory of Networks and Systems.* Kyoto, Japan, 2006 (cf. pp. 19, 36, 38).

[78] K. A. Gallivan, A. Vandendorpe, and P. Van Dooren. “On the generality of multipoint Padé approximations”. In: *15th IFAC World Congress on Automatic Control.* 2002, p. 6 (cf. pp. 20, 44).

[79] K. A. Gallivan, A. Vandendorpe, and P. Van Dooren. “Sylvester equations and projection-based model reduction”. *Journal of Computational and Applied Mathematics,* 162.1 (2004), pp. 213–229 (cf. pp. 7, 30, 31, 35, 38, 133).

[80] W. Gawronski and J. Juang. “Model reduction in limited time and frequency intervals”. *International Journal of Systems Science,* 21.2 (1990), pp. 349–376 (cf. p. 6).

[81] Y. Genin and A. Vandendorpe. “On the embedding of state space realizations”. *Mathematics of Control, Signals, and Systems,* 19.2 (2007), pp. 123–149 (cf. p. 20).

[82] A. Ghafoor and V. Sreeram. “A survey/RReview of frequency-weighted balanced model reduction techniques”. *Journal of Dynamic Systems, Measurement, and Control,* 130.6 (2008) (cf. p. 6).

[83] E. G. Gilbert. “Linear System Approximation by Mean Square Error Minimization in the Time Domain”. PhD thesis. University of Michigan, 1957 (cf. pp. 102, 103).

[84] K. Glover. “All optimal Hankel-norm approximations of linear multivariable systems and their L-error bounds”. *International Journal of Control,* 39.6 (1984), pp. 1115–1193 (cf. p. 143).

[85] E. J. Grimme. “Krylov Projection Methods for Model Reduction”. PhD thesis. Dep. of Electrical Eng., Uni. Illinois at Urbana Champaign, 1997 (cf. pp. 17, 20, 25, 47).

[86] E. J. Grimme, D. C. Sorensen, and P. Van Dooren. “Model reduction of state space systems via an implicitly restarted Lanczos method”. *Numerical Algorithms,* 12 (1 1996), pp. 1–31 (cf. p. 6).

[87] S. Gugercin and A. C. Antoulas. “A survey of balancing methods for model reduction”. In: *European Control Conference.* 2003 (cf. p. 11).

[88] S. Gugercin, A. C. Antoulas, C. A. Beattie, and E. Gildin. “Krylov-based controller reduction for large-scale systems”. In: *43rd IEEE Conference on Decision and Control.* Vol. 3. 2004, pp. 3074–3077 (cf. p. 5).

[89] S. Gugercin. “An iterative SVD-Krylov based method for model reduction of large-scale dynamical systems”. *Linear Algebra and Its Applications,* 428.8–9 (2008), pp. 1964–1986 (cf. pp. 77, 104).

[90] S. Gugercin and A. C. Antoulas. “A survey of model reduction by balanced truncation and some new results”. *International Journal of Control,* 77.8 (2004), pp. 748–766 (cf. pp. 6, 7).

[91] S. Gugercin and A. C. Antoulas. “Model reduction of large-scale systems by least squares”. *Linear Algebra and Its Applications,* 415 (2006), pp. 290–321 (cf. p. 80).

[92] S. Gugercin, A. C. Antoulas, and C. A. Beattie. “$\mathcal{H}_2$ model reduction for large-scale linear dynamical systems”. *SIAM Journal on Matrix Analysis and Applications,* 30.2 (2008), pp. 609–638 (cf. pp. 61, 70, 71, 74, 76, 77, 80).

[93] S. Gugercin and J.-R. Li. “Smith-type methods for balanced truncation of large sparse systems”. In: *Dimension Reduction of Large-Scale Systems.* Springer, 2005, pp. 49–82 (cf. pp. 13, 14).

[94] S. Gugercin, R. V. Polyuga, C. A. Beattie, and A. Van Der Schaft. “Structure-preserving tangential interpolation for model reduction of port-Hamiltonian systems”. *Automatica,* 48.9 (2012), pp. 1963–1974 (cf. pp. 6, 77, 104).

[95] S. Gugercin, D. C. Sorensen, and A. C. Antoulas. “A modified low-rank Smith method for large-scale Lyapunov equations”. *Numerical Algorithms,* 32.1 (2003), pp. 27–55 (cf. pp. 14, 124).

[96] S. Gugercin, T. Stykel, and S. Wyatt. “Model reduction of descriptor systems by interpolatory projection methods”. *SIAM J. Sci. Comput.* 35.5 (2013), B1010–B1033 (cf. p. 20).

[97] Y. Halevi. “Can any reduced order model be obtained via projection?” In: *American Control Conference.* Boston, Massachusetts, 2004 (cf. p. 20).

[98] S. Hammarling. “Numerical solution of the stable, non-negative definite Lyapunov equation”. *IMA Journal of Numerical Analysis,* 2.3 (1982), pp. 303–323 (cf. p. 12).

[99] M. Heinkenschloss, T. Reis, and A. C. Antoulas. “Balanced truncation model reduction for systems with inhomogeneous initial conditions”. *Automatica,* 47.3 (2011), pp. 559–564 (cf. p. 11).

[100] M. Heyouni. “Extended Arnoldi methods for large low-rank Sylvester matrix equations”. *Applied Numerical Mathematics,* 60.11 (2010), pp. 1171–1182 (cf. pp. 14, 114).

[101] G. Hirzinger and G. Kreisselmeier. “On optimal approximation of high-order linear systems by low-order models”. *International Journal of Control,* 22.3 (1975), pp. 399–408 (cf. p. 76).

[102] A. S. Hodel and B. Tenison. “Numerical solution of the Lyapunov equation by approximate power iteration”. *Linear Algebra and Its Applications,* 236 (1996), pp. 205–230 (cf. pp. 13, 113, 114).

[103] C. Hsu, U. Desai, and R. Darden. “Reduction of large-scale systems via generalized Gramians”. In: *22nd IEEE Conference on Decision and Control.* Vol. 22. 1983, pp. 1409–1410 (cf. p. 11).

[104] D. C. 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,* 30.12 (1985), pp. 1201–1211 (cf. pp. 10, 73, 76).

[105] T. C. Ionescu and A. Astolfi. “Families of moment matching based, structure preserving approximations for linear port Hamiltonian systems”. *Automatica,* 49.8 (2013), pp. 2424–2434 (cf. p. 45).

[106] T. C. Ionescu and A. Astolfi. “Families of reduced order models that achieve nonlinear moment matching”. In: *American Control Conference.* 2013, pp. 5518–5523 (cf. p. 45).

[107] T. C. Ionescu and A. Astolfi. “On moment matching with preservation of passivity and stability”. In: *IEEE Conference on Decision and Control.* 2010, pp. 6189–6194 (cf. p. 45).

[108] T. C. Ionescu, A. Astolfi, and P. Colaneri. “Families of moment matching based, low order approximations for linear systems”. *Systems & Control Letters,* 64 (2014), pp. 47–56 (cf. p. 45).

[109] I. M. Jaimoukha and E. M. Kasenally. “Implicitly restarted Krylov subspace methods for stable partial realizations”. *SIAM Journal on Matrix Analysis and Applications,* 18.3 (1997), pp. 633–652 (cf. p. 6).

[110] I. M. Jaimoukha and E. M. Kasenally. “Krylov subspace methods for solving large Lyapunov equations”. *SIAM Journal on Numerical Analysis,* 31.1 (1994), pp. 227–251 (cf. pp. 13, 113).

[111] I. M. Jaimoukha and E. M. Kasenally. “Oblique projection methods for large scale model reduction”. *SIAM Journal on Matrix Analysis and Applications,* 16.2 (1995), pp. 602–627 (cf. pp. 13, 113).

[112] K. Jbilou. “ADI preconditioned Krylov methods for large Lyapunov matrix equations”. *Linear Algebra and Its Applications,* 432.10 (2010), pp. 2473–2485 (cf. p. 14).

[113] K. Jbilou and A. Riquet. “Projection methods for large Lyapunov matrix equations”. *Linear Algebra and Its Applications,* 415.2 (2006), pp. 344–358 (cf. pp. 13, 113).

[114] T. Kailath. *Linear Systems.* Prentice-Hall, Inc., New Jersey, 1980 (cf. p. 4).

[115] D. W. Kammler. “Least squares approximation of completely monotonic functions by sums of exponentials”. *SIAM Journal on Numerical Analysis,* 16.5 (1979), pp. 801–818 (cf. p. 74).

[116] J. Kautsky, N. K. Nichols, and P. Van Dooren. “Robust pole assignment in linear state feedback”. *International Journal of Control,* 41:5 (1985), pp. 1129–1155 (cf. p. 29).

[117] H. Kimura. “Optimal L2-approximation with fixed poles”. *Systems & Control Letters,* 2.5 (1983), pp. 257–261 (cf. p. 104).

[118] L. Knizhnerman and V. Simoncini. “Convergence analysis of the extended Krylov subspace method for the Lyapunov equation”. *Numerische Mathematik,* 118.3 (2011), pp. 567–586 (cf. pp. 14, 114).

[119] W. Krajewski, A. Lepschy, G. Mian, and U. Viaro. “Optimality conditions in multivariable L2 model reduction”. *Journal of the Franklin Institute,* 330.3 (1993), pp. 431–439 (cf. p. 76).

[120] W. Krajewski, A. Lepschy, M. Redivo-Zaglia, and U. Viaro. “A program for solving the L2 reduced-order model problem with fixed denominator degree”. *Numerical Algorithms,* 9.2 (1995), pp. 355–377 (cf. p. 77).

[121] W. Krajewski and U. Viaro. “Iterative-interpolation algorithms for L2 model reduction”. *Control and Cybernetics,* 38.2 (2009), pp. 543–554 (cf. pp. 77, 103).

[122] J. Lam and H. Tam. “Robust partial pole-placement via gradient flow”. *Optimal Control Applications and Methods,* 18 (1979), pp. 371–379 (cf. p. 30).

[123] A. Laub, M. Heath, C. Paige, and R. Ward. “Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms”. *IEEE Transactions on Automatic Control,* 32.2 (1987), pp. 115–122 (cf. p. 11).

[124] A. Laub, L. Silverman, and M. Verma. "A note on cross-Grammians for symmetric realizations". *Proceedings of the IEEE,* 71.7 (1983), pp. 904–905 (cf. p. 11).

[125] S. Lefteriu and A. C. Antoulas. "A new approach to modeling multiport systems from frequency-domain data". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 29.1 (2010), pp. 14–27 (cf. pp. 36, 60).

[126] A. Lepschy, G. Mian, G. Pinato, and U. Viaro. "Rational L2 approximation: a non-gradient algorithm". In: *32nd IEEE Conference on Decision and Control.* Vol. 3. 1991, pp. 2321–2323 (cf. p. 77).

[127] J.-R. Li, F. Wang, and J. K. White. "An efficient Lyapunov equation-based approach for generating reduced-order models of interconnect". In: *36th annual ACM/IEEE Design Automation Conference.* 1999, pp. 1–6 (cf. p. 14).

[128] J.-R. Li and J. White. "Low rank solution of Lyapunov equations". *SIAM Journal on Matrix Analysis and Applications,* 24.1 (2002), pp. 260–280 (cf. pp. 14, 15, 105, 123–125, 133).

[129] Y. Lin and V. Simoncini. "Minimal residual methods for large scale Lyapunov equations". *Applied Numerical Mathematics,* 72 (2013), pp. 52–71 (cf. p. 14).

[130] T. N. Lucas. "Optimal discrete model reduction by multipoint Padé approximation". *Journal of the Franklin Institute,* 330.5 (1993), pp. 855–867 (cf. p. 77).

[131] T. N. Lucas. "Optimal model reduction by multipoint Padé approximation". *Journal of the Franklin Institute,* 330.1 (1993), pp. 79–93 (cf. p. 77).

[132] T. N. Lucas. "Sub-optimal discrete model reduction by multipoint Padé approximation". *Journal of the Franklin Institute,* 333.1 (1996), pp. 57–69 (cf. p. 104).

[133] T. N. Lucas. "Suboptimal model reduction by multi-point Padé approximation". *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering,* 208.2 (1994), pp. 131–134 (cf. p. 104).

[134] A. Mayo and A. C. Antoulas. "A framework for the solution of the generalized realization problem". *Linear Algebra and Its Applications,* 425.2–3 (2007), pp. 634–662 (cf. p. 36).

[135] R. McDonough and W. Huggins. "Best least-squares representation of signals by exponentials". *IEEE Transactions on Automatic Control,* 13.4 (1968), pp. 408–412 (cf. p. 104).

[136] V. Mehrmann and T. Stykel. "Balanced truncation model reduction for large-scale systems in descriptor form". *Dimension Reduction of Large-Scale Systems,* (2005), pp. 83–115 (cf. p. 11).

[137] L. Meier and D. G. Luenberger. "Approximation of Linear Constant Systems". *IEEE Transactions on Automatic Control,* 12.5 (1967), pp. 585–588 (cf. pp. 76, 79, 104).

[138] S. A. Melchior, P. Van Dooren, and K. A. Gallivan. “Model reduction of linear time-varying systems over finite horizons”. *Applied Numerical Mathematics,* 77 (2014), pp. 72–81 (cf. p. 77).

[139] D. G. Meyer and S. Srinivasan. “Balancing and model reduction for second-order form linear systems”. *IEEE Transactions on Automatic Control,* 41.11 (1996), pp. 1632–1644 (cf. p. 11).

[140] G. Miller. “Least-squares rational Z-transform approximation”. *Journal of the Franklin Institute,* 295.1 (1973), pp. 1–7 (cf. p. 76).

[141] H. B. Minh, C. Batlle, and E. Fossas. “A new estimation of the lower error bound in balanced truncation method”. *Automatica,* 50.8 (2014), pp. 2196–2198 (cf. p. 10).

[142] B. C. Moore. “Principal component analysis in linear systems: controllability, observability and model reduction”. *IEEE Transactions on Automatic Control,* AC-26 (1981), pp. 17–32 (cf. p. 11).

[143] C. Mullis and R. Roberts. “Synthesis of minimum roundoff noise fixed point digital filters”. *IEEE Transactions on Circuits and Systems,* 23.9 (1976), pp. 551–562 (cf. p. 11).

[144] N. K. Nichols and P. Van Dooren. “Robust pole assignment and optimal stability margins”. *Electronics Letters,* 20.16 (1984), pp. 660–661 (cf. p. 29).

[145] A. Odabasioglu, M. Celik, and L. T. Pileggi. “PRIMA: passive reduced-order interconnect macromodeling algorithm”. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 17.8 (1998), pp. 645–654 (cf. p. 6).

[146] M. R. Opmeer. “Model order reduction by balanced proper orthogonal decomposition and by rational interpolation”. *IEEE Transactions on Automatic Control,* 57.2 (2012), pp. 472–477 (cf. pp. 13, 14).

[147] J. O’Reilly and M. M. Fahmy. “The minimum number of degrees of freedom in state feedback control”. *International Journal of Control,* 41.3 (1985), pp. 749–768 (cf. p. 143).

[148] H. K. F. Panzer. “Model Order Reduction by Krylov Subspace Methods with Global Error Bounds and Automatic Choice of Parameters”. PhD thesis. Technische Universität München, 2014 (cf. pp. xii, 22, 36, 46, 55, 59, 76, 78, 92, 93, 99, 101, 103, 138).

[149] H. K. F. Panzer, S. Jaensch, T. Wolf, and B. Lohmann. “A greedy rational Krylov method for $\mathcal{H}_2$-pseudooptimal model order reduction with preservation of stability”. In: *American Control Conference.* 2013, pp. 5532–5537 (cf. pp. 47, 52, 59, 78, 93, 103).

[150] H. K. F. Panzer, T. Wolf, and B. Lohmann. “$\mathcal{H}_2$ and $\mathcal{H}_\infty$ error bounds for model order reduction of second order systems by Krylov subspace methods”. In: *European Control Conference.* 2013, pp. 4484–4489 (cf. pp. 103, 114, 137).

[151] D. W. Peaceman and H. H. Rachford. “The numerical solution of parabolic and elliptic differential equations”. *Journal of the Society for Industrial and Applied Mathematics,* 3.1 (1955), pp. 28–41 (cf. p. 123).

[152] T. Penzl. “Numerical solution of generalized Lyapunov equations”. *Advances in Computational Mathematics,* 8.1–2 (1998), pp. 33–48 (cf. p. 12).

[153] T. Penzl. “A cyclic low-rank Smith method for large sparse Lyapunov equations”. *SIAM Journal on Scientific Computing,* 21.4 (2000), pp. 1401–1418 (cf. pp. 14, 123, 124).

[154] T. Penzl. “Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case”. *Systems & Control Letters,* 40 (2000), pp. 139–144 (cf. p. 13).

[155] J. R. Phillips and L. M. Silveira. “Poor man’s TBR: a simple model reduction scheme”. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 24.1 (2005), pp. 43–55 (cf. pp. 13, 14).

[156] L. T. Pillage and R. A. Rohrer. “Asymptotic waveform evaluation for timing analysis”. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 9.4 (1990), pp. 352–366 (cf. p. 20).

[157] C. Poussot-Vassal and P. Vuillemin. “An iterative eigenvector tangential interpolation algorithm for large-scale LTI and a class of LPV model approximation”. In: *European Control Conference.* Zurich, 2013, pp. 4490–4495 (cf. pp. 7, 77).

[158] T. Reis and T. Selig. “Balancing transformations for infinite-dimensional systems with nuclear Hankel operator”. *Integral Equations and Operator Theory,* 79.1 (2014), pp. 67–105 (cf. p. 11).

[159] T. Reis and T. Stykel. “Balanced truncation model reduction of second-order systems”. *Mathematical and Computer Modelling of Dynamical Systems,* 14.5 (2008), pp. 391–406 (cf. pp. 6, 11).

[160] T. Reis and T. Stykel. “Lyapunov balancing for passivity-preserving model reduction of RC circuits”. *SIAM Journal on Applied Dynamical Systems,* 10.1 (2011), pp. 1–34 (cf. pp. 6, 11).

[161] J. B. Riggs and T. F. Edgar. “Least squares reduction of linear systems using impulse response”. *International Journal of Control,* 20.2 (1974), pp. 213–223 (cf. pp. 61, 76, 104).

[162] G. Roppenecker. “Reglerentwurf durch sukzessive Polvorgabe”. *Regelungstechnik,* (1983) (cf. p. 30).

[163] G. Roppenecker. “Vollständige modale Synthese und ihre Anwendung zum Entwurf strukturbeschränkter Zustandsrückführungen”. In: *Fortschrittberichte der VDI Zeitschriften.* Vol. 8. 59. Verein Deutscher Ingenieure (VDI), 1983 (cf. pp. 29, 30).

[164] G. Roppenecker and P. Kocher. “Vollständige Modale Synthese optimaler Zustandsregelungen”. *at - Automatisierungstechnik,* 36.8 (1988), pp. 295–300 (cf. p. 30).

[165] A. Ruhe. “Rational Krylov algorithms for nonsymmetric eigenvalue problems. II. Matrix pairs”. *Linear Algebra and Its Applications,* 197 (1994), pp. 283–295 (cf. p. 20).

[166] Y. Saad. “Numerical solution of large Lyapunov equations”. In: *Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proc. MTNS-89.* Birkhauser, 1990, pp. 503–511 (cf. pp. 13, 113).

[167] Y. Saad. “Projection and deflation method for partial pole assignment in linear state feedback”. *IEEE Transactions on Automatic Control,* 33.3 (1988), pp. 290–297 (cf. p. 30).

[168] J. Saak. “Efficient Numerical Solution of Large Scale Algebraic Matrix Equations in PDE Control and Model Order Reduction”. PhD thesis. Chemnitz University of Technology, 2009 (cf. p. 124).

[169] J. Saak, P. Benner, and P. Kürschner. “A goal-oriented dual LRCF-ADI for balanced truncation”. In: *Vienna Conference on Mathematical Modelling (MATHMOD).* 2012 (cf. pp. 114, 124).

[170] J. Sabino. “Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method”. PhD thesis. Rice Univ. Houston, 2007 (cf. pp. 14, 114, 124).

[171] M. Safonov and R. Chiang. “A Schur method for balanced-truncation model reduction”. *IEEE Transactions on Automatic Control,* 34.7 (1989), pp. 729–733 (cf. p. 11).

[172] B. Salimbahrami and B. Lohmann. “Order reduction of large scale second order systems using Krylov subspace methods”. *Linear Algebra and Its Applications,* 415.23 (2006), pp. 385–405 (cf. p. 6).

[173] H. Sandberg and A. Rantzer. “Balanced truncation of linear time-varying systems”. *IEEE Transactions on Automatic Control,* 49.2 (2004), pp. 217–229 (cf. p. 11).

[174] V. Simoncini. “A new iterative method for solving large-scale Lyapunov matrix equations”. *SIAM Journal on Scientific Computing,* 29.3 (2007), pp. 1268–1288 (cf. pp. 14, 114).

[175] V. Simoncini. “Computational methods for linear matrix equations”. *Preprint,* (2014), pp. 1–58 (cf. pp. 14, 109, 113, 114).

[176] V. Simoncini and V. Druskin. “Convergence analysis of projection methods for the numerical solution of large Lyapunov equations”. *SIAM Journal on Numerical Analysis,* 47.2 (2009), pp. 828–843 (cf. pp. 13, 113).

[177] D. C. Sorensen and A. C. Antoulas. “The Sylvester equation and approximate balanced reduction”. *Linear Algebra and Its Applications,* 351–352 (2002), pp. 671–700 (cf. pp. 11, 27).

[178] D. C. Sorensen and Y. Zhou. “Direct methods for matrix Sylvester and Lyapunov equations”. *Journal of Applied Mathematics,* 2003.6 (2003), pp. 277–303 (cf. p. 12).

[179] J. T. Spanos, M. H. Milman, and D. L. Mingori. “A new algorithm for L2 optimal model reduction”. *Automatica,* 28.5 (1992), pp. 897–909 (cf. p. 104).

[180] T. Stykel. “Balanced truncation model reduction for semidiscretized Stokes equation”. *Linear Algebra and Its Applications,* 415.2 (2006), pp. 262–289 (cf. p. 11).

[181] T. Stykel. “Gramian-based model reduction for descriptor systems”. *Mathematics of Control, Signals, and Systems (MCSS),* 16.4 (2004), pp. 297–319 (cf. p. 11).

[182] T. Stykel and V. Simoncini. “Krylov subspace methods for projected Lyapunov equations”. *Applied Numerical Mathematics,* 62.1 (2012), pp. 35–50 (cf. p. 14).

[183] C. Therapos. “Balancing transformations for unstable nonminimal linear systems”. *IEEE Transactions on Automatic Control,* 34.4 (1989), pp. 455–457 (cf. p. 11).

[184] C. Therapos. “Low-order modelling via constrained least squares minimisation”. *Electronics Letters,* 24.9 (1988), pp. 549–550 (cf. p. 104).

[185] M. Tombs and I. Postlethwaite. “Truncated balanced realization of a stable nonminimal state-space system”. *International Journal of Control,* 46.4 (1987), pp. 1319–1330 (cf. p. 11).

[186] P. Trnka et al. “Structured model order reduction of parallel models in feedback”. *IEEE Transactions on Control Systems Technology,* 21.3 (2013), pp. 739–752 (cf. p. 5).

[187] P. Van Dooren, K. A. Gallivan, and P.-A. Absil. “$\mathcal{H}_2$-optimal model reduction of MIMO systems”. *Applied Mathematics Letters,* 21.12 (2008), pp. 1267–1273 (cf. pp. 75, 76, 146).

[188] 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,* 31.5 (2010), pp. 2738–2753 (cf. pp. 74–77, 87).

[189] A. Vandendorpe. “Model Reduction of Linear Systems, an Interpolation Point of View”. PhD thesis. Université Catholique De Louvain, 2004 (cf. pp. 19, 25, 30–32, 35, 38).

[190] A. Varga. “Efficient minimal realization procedure based on balancing”. *IMACS Symp. on Modelling and Control of Technological Systems,* 2 (1991), pp. 42–47 (cf. p. 11).

[191] A. Varga. “Periodic Lyapunov equations: some applications and new algorithms”. *International Journal of Control,* 67.1 (1997), pp. 69–88 (cf. p. 11).

[192] A. Varga and B. Anderson. “Accuracy-enhancing methods for balancing-related frequency-weighted model and controller reduction”. *Automatica,* 39.5 (2003), pp. 919–927 (cf. p. 5).

[193] P. Vilbé and L. Calvez. “Constrained $l2$ suboptimal model reduction”. *Electronics Letters,* 23.25 (1987), pp. 1340–1342 (cf. p. 104).

[194] P. Vilbé, L. Calvez, and M. Sévellec. "Suboptimal model reduction via least-square approximation of time-response by its derivatives and integrals". *Electronics Letters,* 28.2 (1992), pp. 174–175 (cf. p. 104).

[195] P. Vilbé, L. Calvez, M. Sévellec, and C. Nouet. "$L2$-optimal numerator via Routh table". *Electronics Letters,* 28.14 (1992), pp. 1306–1308 (cf. p. 104).

[196] G. Vossen. "$H_{2,\alpha}$-norm optimal model reduction for optimal control problems subject to parabolic and hyperbolic evolution equations". *Optimal Control Applications and Methods,* Published online (2013) (cf. p. 76).

[197] G. Vossen, A. Bunse-Gerstner, D. Kubalińska, and D. Wilczek. *Necessary optimality conditions for H2-norm optimal model reduction.* Berichte aus der Technomathematik 07-05. Universität Bremen, 2007 (cf. pp. 69, 71, 76).

[198] E. L. Wachspress. "Iterative solution of the Lyapunov matrix equation". *Applied Mathematics Letters,* 1.1 (1988), pp. 87–90 (cf. p. 123).

[199] E. L. Wachspress. "The ADI minimax problem for complex spectra". *Applied Mathematics Letters,* 1.3 (1988), pp. 311–314 (cf. pp. 14, 124).

[200] E. L. Wachspress. *The ADI model problem.* Springer, New York, 2013 (cf. p. 14).

[201] J. Walsh. *Interpolation and Approximation by Rational Functions in the Complex Plane.* 3rd. American Mathematical Society, 1960 (cf. pp. 80, 103).

[202] G. Wang, V. Sreeram, and W. Liu. "A new frequency-weighted balanced truncation method and an error bound". *IEEE Transactions on Automatic Control,* 44.9 (1999), pp. 1734–1737 (cf. p. 6).

[203] D. A. Wilson. "Optimum solution of model-reduction problem". In: *Proceedings of the Institution of Electrical Engineers.* Vol. 117. 6. 1970, pp. 1161–1165 (cf. pp. 70, 72, 76, 103, 104, 146).

[204] D. Wilson. "Model reduction for multivariable systems". *International Journal of Control,* 20.1 (1974), pp. 57–64 (cf. pp. 98, 104).

[205] D. Wilson and R. Mishra. "Optimal reduction of multivariable systems". *International Journal of Control,* 29.2 (1979), pp. 267–278 (cf. p. 76).

[206] T. Wolf, B. Lohmann, R. Eid, and P. Kotyczka. "Passivity and structure preserving order reduction of linear port-Hamiltonian systems using Krylov subspaces". *European Journal of Control,* 16.4 (2010), pp. 401–406 (cf. p. 6).

[207] T. Wolf, H. K. F. Panzer, and B. Lohmann. "ADI-Lösung großer Ljapunow-Gleichungen mittels Krylov-Methoden und neue Formulierung des Residuums". In: *Talk given at the GMA Fachausschuss 1.30.* 2012 (cf. pp. 110, 134, 135).

[208] T. Wolf, H. K. F. Panzer, and B. Lohmann. "H2 pseudo-optimality in model order reduction by Krylov subspace methods". In: *European Control Conference.* 2013 (cf. pp. 62, 77, 80, 85, 89, 91, 103, 105).

[209] T. Wolf, H. K. F. Panzer, and B. Lohmann. "Model reduction by approximate balanced truncation: a unifying framework". *at-Automatisierungstechnik,* 61.8 (2013), pp. 545–556 (cf. pp. 11, 124, 135).

[210] T. Wolf, H. Panzer, and B. Lohmann. *ADI iteration for Lyapunov equations: a tangential approach and adaptive shift selection.* 2013. URL: http://arxiv.org/abs/1312.1142 (cf. pp. 14, 110, 124, 129, 132, 135).

[211] T. Wolf, H. K. F. Panzer, and B. Lohmann. "Gramian-based error bound in model reduction by Krylov-subspace methods". In: *18th IFAC World Congress.* Milano, Italy, 2011, pp. 3587–3592 (cf. pp. 25, 39, 47, 48, 50).

[212] T. Wolf, H. K. F. Panzer, and B. Lohmann. "Sylvester equations and a factorization of the error system in Krylov-based model reduction". In: *Vienna Conference on Mathematical Modelling (MATHMOD).* 2012 (cf. pp. 25, 35, 38, 39, 41, 45–48).

[213] T. Wolf and H. K. F. Panzer. *The ADI iteration for Lyapunov equations implicitly performs $\mathcal{H}_2$ pseudo-optimal model order reduction.* 2013. URL: http://arxiv.org/abs/1309.3985 (cf. pp. 14, 80, 83, 104, 110, 124, 134, 135).

[214] T. Wolf, H. K. F. Panzer, and B. Lohmann. "On the residual of large-scale Lyapunov equations for Krylov-based approximate solutions". In: *American Control Conference.* 2013 (cf. pp. 14, 110, 114).

[215] S. Wyatt. "Issues in Interpolatory Model Reduction: Inexact Solves, Second-order Systems and DAEs". PhD thesis. Virginia Polytechnic Institute and State University, 2012 (cf. pp. 18, 46).

[216] Y. Xu and T. Zeng. "Fast optimal H2 model reduction algorithms based on Grassmann manifold optimization". *International Journal of Numerical Analysis & Modeling,* 10.4 (2013), pp. 972–991 (cf. p. 78).

[217] Y. Xu and T. Zeng. "Optimal $\mathcal{H}_2$ model reduction for large scale MIMO systems via tangential interpolation". *International Journal of Numerical Analysis & Modeling,* 8.1 (2011) (cf. p. 75).

[218] B. Yan, S. X.-D. Tan, P. Liu, and B. McGaughy. "SBPOR: second-order balanced truncation for passive order reduction of RLC circuits". In: *44th Annual Design Automation Conference.* 2007, pp. 158–161 (cf. p. 11).

[219] W.-Y. Yan and J. Lam. "An approximate approach to $\mathcal{H}_2$ optimal model reduction". *IEEE Transactions on Automatic Control,* 44.7 (1999), pp. 1341–1358 (cf. pp. 61, 78).

[220] T. Zeng. "Alternating direction algorithm for optimal H2 model reduction". In: *Fourth International Conference on Intelligent Control and Information Processing.* 2013, pp. 722–725 (cf. p. 78).

[221] K. Zhou, G. Salomon, and E. Wu. "Balanced realization and model reduction for unstable systems". *International Journal of Robust and Nonlinear Control,* 9.3 (1999), pp. 183–198 (cf. p. 11).

[222] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control.* Vol. 40. Prentice Hall New Jersey, 1996 (cf. pp. 4, 29, 49, 70).