Diploma Thesis

Model Reduction of Positive Systems

Author:
Christian Grussler

Supervisor:
Prof. Dr. Tobias Damm

Co-Supervisor:
Prof. Anders Rantzer

Technical University Kaiserslautern
Department of Mathematics
Technomathematics Group

November 29, 2012

Abstract - The present work is concerned with the model order reduction of positive linear systems. Generally, well-established model reduction methods applied to positive systems do not guarantee the positivity of the approximation. To this end we examine some of the reasons for this behaviour and present a new method based on Balanced Truncation. Further, we compare the results of this approach with those of reduction methods especially developed for positive linear systems.

Contents

Abstract i

Notation and Symbols iv

Introduction vii

1. Positive Linear Systems 1
- 1.1. Continuous Time Systems . . . 1
- 1.2. Discrete Time Systems . . . 7

2. Positive Realization 11
- 2.1. Reachability, Observability and Realizability . . . 12
- 2.2. Second-Order Realization and Special Cases . . . 20

3. Balanced Truncation 24
- 3.1. Standard Balanced Truncation . . . 29
- 3.2. Balanced Truncation Algorithm . . . 35
- 3.3. Singular Perturbation Balanced Truncation . . . 36
- 3.4. Balanced Truncation of Positive Systems . . . 38
  - 3.4.1. Balanced Truncation with respect to Lyapunov Inequalities . . . 40

4. Model Reduction of Positive Systems based on the Bounded Real Lemma 45
- 4.1. Iterative Linear Matrix Approach I . . . 46
- 4.2. Algorithm: Iterative Linear Matrix Approach I . . . 49
- 4.3. Iterative Linear Matrix Approach II . . . 54
- 4.4. Algorithm: Iterative Linear Matrix Approach II . . . 56

5. Krylov Subspace Methods 59
- 5.1. Arnoldi Iteration . . . 59
- 5.2. Lancozos Iteration & Biorthogonalization Algorithm . . . 61
- 5.3. Model Reduction via Coefficient Matching . . . 64
- 5.4. Coefficient Matching for Positive Systems . . . 67
- 5.5. Iterative Rational Krylov Algorithm . . . 70

6. Symmetric Balanced Truncation 82
- 6.1. Symmetric Balanced Truncation Algorithm . . . 86
- 6.2. Symmetric Balanced Truncation for MIMO-systems . . . 87

7. Numerical Examples **89**
7.1. Water reservoirs 90
7.2. Compartmental Networks 92
7.3. Heat Equation 94

Conclusions and Open Problems **98**

A. Appendix **99**
A.1. Cones 99

Bibliography **103**

Notation and Symbols

Below is a list of frequently used symbols and notations.

$\mathbb{R}^n$	n-dimensional real space
$\mathbb{C}^n$	n-dimensional complex space
$i\mathbb{R}$	imaginary axis
$\mathbb{R}^n_+$	nonnegative orthant
$\mathbb{R}^{n\times n}$	all $n \times n$ real matrices
$\mathbb{C}^{n\times n}$	all $n \times n$ complex matrices
$A \gg 0$	each element is positive
$A \geqq 0$	each element is nonnegative
$A > 0$	positive definite
$A \geq 0$	semi-positive definite
A^{-1}	inverse
$A^\sharp$	Moore-Penrose inverse
$diag(a_1, \ldots, a_n)$	diagonal $n \times n$ matrix with a_i as its diagonal entries
$\sigma(A)$	spectrum
$tr(A)$	trace
$rg(A)$	range
$\ker(A)$	kernel (nullspace)
$rk(A)$	rank
$\det(A)$	determinant
A^T	transpose
$\lambda_i(A)$	i-th eigenvalue of A
I	identity matrix of dimension n
a_{ij}	(i,j)-th entry of A
$\overline{\alpha}$	complex conjugate of $\alpha \in \mathbb{C}$
$\Re(\lambda)$	real part of $\lambda \in \mathbb{C}$
$\lvert\alpha\rvert$	absolute value of $\alpha \in \mathbb{C}$
$\lvert A\rvert$	component wise absolute value of a matrix $A \in \mathbb{C}^{n\times n}$
$\lVert A\rVert_2$	2-norm
$\langle \alpha, \beta \rangle$	scalar product
$\lVert G\rVert_\infty$	$\mathcal{H}_\infty$-norm of $G(s)$
$\mathcal{H}_2$	$\mathcal{H}_2$ space
$\mathscr{F}[x(t)]$	Fourier transform of $x(t)$
$\mathscr{L}[x(t)]$	Laplace transform of $x(t)$
$:=$	defined as

Introduction

This thesis is about model order reduction of positive linear systems and aims to give a comparison between well-established approaches and those that were especially developed for the treatment of positive systems.

Mathematical modelling of biological, chemical and physical systems often yields complex high-dimensional models resulting e.g. from system identification [13] or discretized partial differential equations [4] [26]. A serious problem of these models is that they are hard to analyse and simulate, which is why lower-dimensional systems are preferred over complex ones. Approximating high-order models by reduced ones is the essential idea of model order reduction in control and has received considerable attention during the past decades e.g. in [11] [12] [23] [24] [30]. In case of a linear time-invariant system, the model reduction problem can be described as the approximation of a system

$$G: \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{0.1}$$

with state variables $x \in \mathbb{R}^n$, input $u \in \mathbb{R}^m$ and output $y \in \mathbb{R}^p$, for small m, p and large dimension n.[4] [12] Amongst the many optimality criteria for linear approximations, most common is to consider the error with respect to the $\mathcal{H}_\infty$-norm, concerning the the frequency domain, or the $\mathcal{H}_2$-norm regarding the input-output behaviour.[12] [30] For this purpose different reduction methods have been developed, but most famous became those based on projection approaches, such as Balanced Truncation and Krylov subspace methods. Both will be discussed in Chapter 3 and 5, respectively.

A class of linear systems which is of particular interest, is given by the so-called positive linear systems and will be introduced in Chapter 1. These systems are characterized by the nonnegativity of the output and state variables for every nonnegative input and initial state. Systems with such positivity constraints are often found in the context of measured quantities e.g. temperature or mass flow (see Chapter 7).
In order to perform a good simulation it is natural to preserve these properties after performing model order reduction. Unfortunately, we will see in Chapter 2, that positive systems are defined on cones instead of linear subspaces and therefore conventional reduction methods are not able to guarantee the preservation of the positivity. In fact, it will turn out, that this is in a large part also a positive realization problem.

As a consequence, new methods have been developed in [7][14][22], which will be presented and discussed in Chapter 3 and 4. Furthermore, we will investigate in Chapter 3 and 5 for which positive systems Balanced Truncation and Krylov subspace methods

can still be applied and use these results to develop a new approach in Chapter 6. This approach is be based on a lesser-known symmetry characterization of balanced single-input-single-output systems. Concluding, we compare the quality among all the presented methods in Chapter 7.

1. Positive Linear Systems

In this chapter, we look at the two basic concepts of positive linear systems, external (input-output) and internal (input-state-output).
For this purpose we first introduce the notion of *stricly positive (nonnegative) vectors* and correspondingly *strictly positive (nonnegative) matrices*. The entries of a real vector $v \in \mathbb{R}^n$ and a real valued matrix $A \in \mathbb{R}^{n \times m}$ are denoted by v_i and a_{ij}, respectively. We say

$$\begin{aligned} v \text{ is a } \textit{strictly positive vector}, \quad v \gg 0 \quad &:\Leftrightarrow \quad v_i > 0 \text{ for all } i \\ v \text{ is a } \textit{nonnegative vector}, \quad v \geqq 0 \quad &:\Leftrightarrow \quad v_i \geq 0 \text{ for all } i \end{aligned}$$

$$\begin{aligned} A \text{ is a } \textit{strictly positive matrix}, \quad A \gg 0 \quad &:\Leftrightarrow \quad a_{ij} > 0 \text{ for all } (i,j) \\ A \text{ is a } \textit{nonnegative matrix}, \quad A \geqq 0 \quad &:\Leftrightarrow \quad a_{ij} \geq 0 \text{ for all } (i,j) \end{aligned}$$

Observe the difference between the notation of a strictly positive (nonnegative) matrix and a *(semi-)positive definite matrix*, which we denote as $A > 0$ and $A \geq 0$, respectively. Naturally, we use all these notations to describe the relation between two arbitrary elements, e.g. $A \geq B$ is defined by $A - B \geq 0$.
A real vector valued function $u(t) \in \mathbb{R}^n$ is called *nonnegative* if and only if $u(t) \geqq 0 \ \forall t$.

1.1. Continuous Time Systems

The notations above allow us to give the definition of *external positivity* of a linear system. In this section we will focus on continuous time systems and discuss the discrete case in the subsequent section.

Definition 1.1 (*Externally positive linear system*)

A linear system (A, B, C, D) is called ***externally positive*** *if and only if its forced output (i.e. the output corresponding to a zero initial state) is nonnegative for every nonnegative input.[6]*

Recalling the well-known representation of the impulse response matrix of a continuous-time linear system:

$$g(t) := Ce^{At}B + D\delta(t), \tag{1.1}$$

where $\delta(t)$ denotes a delta-dirac impulse, we are able to give a better mathematical description of an externally positive system.

Theorem 1.1 (*External positivity*)

A linear system (A, B, C, D) is externally positive if and only if its impulse response is nonnegative, i.e. $g(t) \geqq 0,\ \forall t \geq 0$.[6]

Proof: ► Necessity: Since $\delta(t)$ is a positive input, $g(t) \geqq 0, \forall t \geq 0$ by definition of external positivity.

► Sufficiency: Assuming $g(t), u(t) \geqq 0$, the output to the system with zero initial state is given by:

$$y(t) = \int_0^t g(\tau)u(t-\tau)d\tau \ \Rightarrow\ y(t) \geqq 0.$$

■

Since the transfer function $G(s)$ of a state-space system (A, B, C, D) is nothing else than the Laplace transformation of $g(t)$, i.e.

$$G(s) = C(sI - A)^{-1}B + D = \mathscr{L}[g(t)] := \int_0^\infty g(t)e^{-st}dt,$$

it follows by the nonnegativity of $g(t)$ that

$$G(0) \geqq 0 \ \Rightarrow\ |G(i\omega)| \leqq \int_0^\infty |g(t)||e^{-i\omega t}|dt \leqq \int_0^\infty |g(t)|dt = \int_0^\infty g(t)dt = G(0),$$

where $|\ \cdot\ |$ denotes the component wise absolute value.

According to [30], the $\mathcal{H}_\infty$-*norm* of a system $G(s)$ is defined as

$$\|G\|_\infty := \sup_\omega \overline{\sigma}\{G(i\omega)\} = \sup_\omega \|G(i\omega)\|_2,$$

where $\overline{\sigma}$ denotes the largest singular value of $G(i\omega)$. Consequently by $\|A\|_2 \leq \||A|\|_2$ we get the following lemma.

Lemma 1.1 (*Gain of a positive system*)

For the $\mathcal{H}_\infty$-norm of the transfer function of a positive system it holds:

$$\|G\|_\infty = \|G(0)\|_2.$$

A closer look at $g(t)$ tells us immediately, that if $e^{At}, B, C, D \geqq 0 \Rightarrow g(t) \geqq 0$. In the next example we will see, this is not a necessary assumption for external positivity, but it will be for our second positivity definition, the so called *internal positivity*.

Example 1.1 (*External, but not internally positive system*)
Let us consider the following single-input-single-output (SISO) system.

$$A := \begin{pmatrix} -1 & 0 \\ -1 & -2 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c := \begin{pmatrix} 0 & -1 \end{pmatrix}.$$

This system has eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -2$. Thus the impulse response can be expressed by

$$g(t) = \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t}.$$

Moreover, $g(0) = cb = 0$, which yields $\alpha = -\beta$ and then

$$\dot{g}(t) = \alpha(\lambda_1 e^{\lambda_1 t} - \lambda_2 e^{\lambda_2 t}).$$

Since $\dot{g}(0) = cAb = 1$, we have $\alpha = \dfrac{1}{\lambda_1 - \lambda_2} > 0$ and thus, $g(t) \geqq 0$, $\forall t \geq 0$.

An obvious disadvantage of Theorem 1.1 is, that in many cases it is not possible to check easily whether the impulse response of a system is nonnegative or not. Still there is a useful exclusion criteria for external positivity based on the poles of the system.

Lemma 1.2 (*Dominant Pole*)
The transfer function of an external positive system has at least one real dominant pole.[5]

Proof: A dominant complex pole leads to a long-term behaviour of $g(t)$ which is oscillating and thus the impulse responses becomes negative. ■

As a consequence of Lemma 1.2 we can also give a condition on the zeros in case of a SISO system.

Lemma 1.3 (*Real Zeros*)
The real zeros of an externally positive SISO system are smaller than the dominant pole.[6]

Proof: Assume there is a zero $z_0 \in \mathbb{R}$ that is greater than the real dominant pole. By this assumption z_0 lies within the radius of convergence of $G(s)$ and therefore by the nonnegativity of $g(t)$

$$G(z_0) = \int_0^\infty g(t) e^{-z_0 t} dt > 0,$$

which is a contradiction. Thus, there cannot exist any zero, that is greater or equal to the real dominant pole. ■

For *internal positivity* it is much easier to check whether a system is externally positive or not (see Theorem 1.3). Therefore we will draw most of our attention to this class of systems and introduce them with the next definition.

Definition 1.2 (*Internally Positive Linear System*)

*A linear system (A, B, C, D) is called **(internally) positive** if and only if its state and output are nonnegative for every nonnegative input and every nonnegative initial state.[6]*

The difference between internal and external positivity is obviously the additional condition of the nonnegativity of the state vector and hence every positive linear system is externally positive. The definition requires explicitly a nonnegative output for *every* nonnegative inital state and nonnegative input. Thus it suffices to consider the particular pair $[x(0) = e_i,\ u(0) = 0]$ [6], where e_i denotes the i-th unit vector in $\mathbb{R}^n$.

Looking at the state-space-representation of a (cont.) linear system (0.1), it becomes clear, that $C \geqq 0$ has to hold in order to provide internal positivity: if $x(0) = e_i$, $u(0) = 0$ and $c_{ij} < 0$ for at least one j, it will lead immediately to a negative output and therefore $C \geqq 0$ is a necessary condition for (internal) positivity.

A similar consideration can be done for D by switching the roles of $x(0)$ and $u(0)$. Together $C, D \geqq 0$ are sufficient and necessary conditions for a nonnegative output under the assumption of a nonnegative state and input.

All that remains is the analysis of the state-equation:

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{1.2}$$

For this purpose we start with some important definitions and results for a certain class of matrices, the so called Metzler matrices (or sometimes $-Z$-matrices) [2].

Definition 1.3 (*Metzler- and Z-matrix*)

*If $A + \alpha I \geqq 0$ with $\alpha \in \mathbb{R}^n$, then A is called a **Metzler matrix** or **essentially nonnegative**, short $A \overset{e}{\geqq} 0$.*

*A is called a **Z-matrix** if $-A \overset{e}{\geqq} 0$ and hence a **Metzler matrix** is also called a **-Z-matrix**.*

A *Metzler matrix* is in a manner of speaking a matrix with nonnegative off-diagonal entries.

Remark: If $A + \alpha I \geqq 0$, then it follows by the series representation of the exponential function.

$$0 \leqq e^{A+\alpha I} = e^{A} e^{\alpha I} \;\Leftrightarrow\; e^{A} \geqq 0 \tag{1.3}$$

Thus the matrix exponential of a $-Z$-matrix is always nonnegative and therefore $e^{At} \geqq 0$ for all $t \geq 0$.

An important subclass of Metzler matrices are the so called $-M$-matrices [2].

Definition 1.4 (*M-matrix*)

A matrix $A \in \mathbb{R}^n$ is called an ***M-matrix****, if $-A \overset{e}{\geqq} 0$ and $\Re(\lambda) \geq 0$, $\forall \lambda \in \sigma(A)$, where $\Re(\lambda)$ denotes the real part of λ.*

Analogous: A is a ***-M-matrix*** *if $A \overset{e}{\geqq} 0$ and $\Re(\lambda) \leq 0$, $\forall \lambda \in \sigma(A)$.*

This is, all the eigenvalues of a $-M$-matrix are suited in the left complex plane and the matrix is stable. The most important results about asymptotically stable $-M$-matrices are summarized in the following theorem.

Theorem 1.2 (*Asymptotically stable $-M$-matrix*)

Let A be a $-Z$-matrix, then the following statements are equivalent [2]:

(i) A is a nonsingular $-M$-matrix

(ii) $\Re(\lambda) < 0, \forall \lambda \in \sigma(A)$

(iii) If $A = B - \alpha I$ with $B \geqq 0$, then $\rho(B) < \alpha$.

(iv) $(-A)^{-1} \geqq 0$

(v) $\exists D > 0$ diagonal : $AD + DA^T < 0$.

Proof: (i) $\Rightarrow$ (ii): By the nonsingularity of A it follows, that its determinant

$$\det(A) = \prod_{i=1}^{n} \lambda_i \neq 0 \quad \text{with} \quad \lambda_i \in \sigma(A).$$

Hence, by definition of a $-M$-matrix we can conclude, that all real eigenvalues of A have to be strictly negative and only strictly imaginary eigenvalues can fulfil $\Re(\lambda) = 0$.
Let $\alpha > 0$ be sufficiently large, such that $A + \alpha I \geqq 0$, then $\lambda + \alpha \in \sigma(A + \alpha I)$, $\forall \lambda \in \sigma(A)$. Consequently, if A possesses a strictly imaginary eigenvalue $\tilde{\lambda}$, then $\Re(\tilde{\lambda} + \alpha) > \lambda_0$, $\forall \lambda_0 \in \sigma(A) \cap \mathbb{R}$ and therefore $|\tilde{\lambda} + \alpha| > |\lambda_0|$. But this contradicts the Frobenius-Perron-Theorem (Theorem 1.5).

(ii) $\Rightarrow$ (i): Clear by '(i) $\Rightarrow$ (ii)'.

(ii) $\Rightarrow$ (iii): Again by Theorem 1.5 there exists a $\lambda_0 \in \sigma(B)$, such that $\lambda_0 = \rho(B) \geq 0$. Since $\lambda_0 = \lambda + \alpha$ for a $\lambda \in \sigma(A)$ and $\Re(\lambda) < 0$ by assumption, we conclude $\alpha > 0$ and consequently by Pythagoras $|\lambda + \alpha|^2 < \alpha^2$.

(iii) $\Rightarrow$ (ii): In the same way as '(ii) $\Rightarrow$ (iii)'.

(iii) $\Rightarrow$ (iv): Since (ii) and (iii) are equivalent, it follows from '(ii) $\Rightarrow$ (i)', that $A = B - \alpha I$ is invertible. By assumption $A^{-1} = (B - \alpha I)^{-1} = \frac{1}{\alpha}(\frac{1}{\alpha}B - I)^{-1}$ and $\rho(\frac{1}{\alpha}B) < 1$.
By the well-known Neumann series theorem [17] we conclude

$$(-A)^{-1} = \sum_{i=0}^{\infty} \frac{1}{\alpha^{i+1}} B^i \geqq 0.$$

(iv) $\Rightarrow$ (v): Let $x := (-A)^{-1}e$, with $e = \big(1, \ldots, 1\big)^T$.
Then by assumption

$$D_x := diag\big(x_1, \ldots, x_n\big) := \begin{pmatrix} x_1 & & \\ & \ddots & \\ & & x_n \end{pmatrix} > 0$$

and thus

$$AD_x e = Ax = -e \ll 0 \tag{1.4}$$

Observe, since $(-A)^{-1}A = -I$ and A is $-Z$-matrix, it follows by assumption that $a_{ii} < 0$. By (1.4) we conclude then, that AD is strictly diagonally dominant. The same can be done for A^T: let $y := (-A)^{-T}e$, then $D_y := diag\big(y_1, \ldots, y_n\big) > 0$ and $A^T D_y e = Ay = e \ll 0$. Consequently, $e^T D_y A D_x = D_x e \ll 0$ and $D_y A D_x e = D_y e \ll 0$ and therefore $D_y A D_x$ is row and column diagonally dominant.
With the help of Gershgorin's circle theorem [3] we receive

$$P := D_y A D_x + D_x A^T D_y < 0,$$

because P is strictly diagonally dominant with $p_{ii} < 0$. Multiplying from both sides with $D_y^{-1} = D_y^{-T}$ leads to

$$AD_x D_y^{-1} + D_y^{-1} D_x A^T = D_y^{-T} P D_y^{-1} < 0.$$

Hence, $D := D_x D_y^{-1}$ concludes the proof.

(v) $\Rightarrow$ (ii): Follows directly by Lemma 3.1. ■

Now we are ready to turn back to equation (1.2) and describe the properties of (A, B) in order to satisfy the condition of a nonnegative state vector. We make a consideration similar to what has been done in [16] or [6]. Let us start with the case $u(t) = 0$:

$$\dot{x}(t) = Ax(t).$$

If $x(t)$ lies on the boundary of the positive orthant $\mathbb{R}^n_+$, i.e if $x(t)$ contains an element x_i equal to zero, then the corresponding derivative $\dot{x}_i(t)$ has to be nonnegative in order to keep x_i nonnegative. This is obviously equivalent to the fact that the off-diagonal entries have to be nonnegative and hence A has to be a Metzler matrix. Since we only consider asymptotically stable systems, we can conclude with the help of Theorem 1.2, that A has to be a nonsingular $-M$-matrix and hence its diagonal entries have to be strictly negative as seen in the proof to Theorem 1.2.
As for D, a negative entry in B and a corresponding $u(t) \geqq 0$ would lead to a violation of $\dot{x}_i(t) \geqq 0$, which is why we can conclude that B has to be nonnegative.
In conclusion we have shown that A being a Metzler matrix and $B \geqq 0$ are sufficient and necessary conditions to assure a nonnegative state vector. Altogether we summarize the following characterization of positive linear systems.

Theorem 1.3 (*Continuous Positive Linear System*)

A (cont.) linear system (A, B, C, D) is positive if and only if A is a $-M$-matrix and $B, C, D \geqq 0$.

1.2. Discrete Time Systems

For discrete time systems the definitions of external and internal positivity remain the same. The difference compared to continuous time systems is, that a discrete system represents its own recursive solution algorithm [10]. The solution to the state can be given explicitly by the recursion

$$x(t) = A^{t-t_0}x(t_0) + \sum_{k=t_0}^{t-1} A^{t-k-1}Bu(k). \tag{1.5}$$

with an initial state $x(t_0)$ at time t_0. Consequently, we can consider immediately $x(t+1)$, instead of $\dot{x}(t)$. Analogous to the delta-dirac impulse, a pulse in discrete time is defined as

$$\delta_d(t) := \begin{cases} 1 \text{ for } t = 0 \\ 0 \text{ for } t > 0 \end{cases} \tag{1.6}$$

and thus for discrete time the impulse response ($x(0) = 0$), is given by

$$g_d(t) = C\sum_{k=0}^{t-1} A^{t-k-1}B\delta_d(k) + D\delta_d(t) = \begin{cases} D & \text{for } t = 0 \\ CA^{t-1}B & \text{for } t > 0 \end{cases} \tag{1.7}$$

Obviously, for an *discrete externally positive* system it is necessary that

$$CA^{t-1}B \geqq 0 \text{ for } t > 0 \quad \text{and} \quad D \geqq 0.$$

By recursion (1.5) this is also a sufficient condition, i.e Theorem 1.1 is valid in discrete time, too. In the same way, as in the continuous case, the transfer function of a discrete system is given by the $\mathscr{Z}$-transformation of its impulse response, i.e.

$$G(z) = C(zI - A)^{-1}B + D = \mathscr{Z}[g_d(t)] = \sum_{t=0}^{\infty} \frac{g_d(t)}{z^t}.$$

Notice, by setting $g_t := g_d(t)$, the series expansion of the continuous time impulse response can be written as

$$Ce^{At}B + D\delta(t) = C\sum_{i=0}^{\infty} \frac{A^i t^i}{i!}B + D\delta(t) = \sum_{i=1}^{\infty} g_i \frac{t^{(i-1)}}{(i-1)!} + D\delta(t), \tag{1.8}$$

and g_t is also known as *Markov coefficients*.

For a *discrete internally positive* system it is clear, that $A, B, C, D \geqq 0$ is sufficient. The necessity of this condition can be readily seen by considering $x(t+1)$ instead of $\dot{x}(t)$ in the proof to Theorem 1.3. Then in case that $x(t)$ is on the boundary of $\mathbb{R}^n_+$, e.g. $x(t) = e_i$, $x(t+1)_i = a_{ii}$ has to remain positive. Thus we can state the discrete analogous of Theorem 1.3 as follows.

Theorem 1.4 (*Discrete Positive Linear System*)

A discrete linear system (A, B, C, D) is positive if and only if $A, B, C, D \geqq 0$.

In Lemma 1.2 we have discovered that for a (cont.) positive system, A must have a real dominant pole. With an extension of the so-called *Frobenius-Perron-Theorem* [16] [18] one can make the same conclusion for the discrete case.

Theorem 1.5 (*Frobenius-Perron Extension*)

Let $A \gg 0$, then there exists a real $\lambda_0 > 0$ and a $x_0 \gg 0$ such that

(i) $Ax_0 = \lambda_0 x_0$

(ii) $\lambda_0 > |\lambda|, \quad \forall \lambda \in \sigma(A) \setminus \{\lambda_0\}.$

In case of $A \geqq 0$, the same statements can be made by replacing the strict relations with $\geq$ and $\geqq$, respectively.

Proof: ▶ We start with the case $A \gg 0$:
Let λ_0 denote the maximal value for which $Ax - \lambda x \geqq 0$, for some $x \in \mathbb{R}^n_+ \setminus \{0\}$. It is obvious, that a lower bound for λ_0 is provided by $\lambda = 0$, but it is also possible to find an upper bound. Let $\|A\|_\infty := \max_i \sum_{j=1}^{n} |a_{ij}|$, then

$$\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty \;\Rightarrow\; (Ax)_i \leq \|A\|_\infty \max_i\{x_i\}, \forall x \geq 0$$

and $\|A\|_\infty$ is an upper bound for λ.
Let $x_0 \in \mathbb{R}^n_+ \setminus \{0\}$ be a vector fulfilling $Ax_0 - \lambda_0 x_0 \geqq 0$, then by $A \gg 0$ it follows that $\lambda_0 > 0$ and $Ax > 0$ for all $x \in \mathbb{R}^n_+ \setminus \{0\}$. Consequently, $A(Ax_0 - \lambda_0 x_0) > 0$ and equal to zero if and only if $Ax_0 = \lambda_0 x_0$. By looking at

$$0 < A(Ax_0 - \lambda_0 x_0) = A(Ax_0) - \lambda_0(Ax_0),$$

we observe, that λ_0 is not maximal regarding the vector $Ax_0 \gg 0$ and this a contradiction to the maximality of λ_0. Thus $Ax_0 = \lambda_0 x_0 \gg 0$, which concludes the proof of (i) for $A \gg 0$.
Considering any other eigenvalue $\lambda \in \sigma(A) \setminus \{\lambda_0\}$ with eigenvector y, it is easy to see that

$$A|y| - |Ay| \geq 0$$

and equivalently

$$A|y| - |\lambda||y| \geq 0.$$

Then by definition of λ_0 it must hold $\lambda_0 \geq |\lambda|$. The strict inequality is readily seen by an sufficiently small eigenvalue shift $\alpha > 0$, such that $A - \alpha I \gg 0$. Using the first part we conclude

$$|\lambda - \alpha| \leq \lambda_0 - \alpha.$$

If there existes a complex $\lambda \in \sigma(A) \setminus \{\lambda_0\}$ such that $|\lambda| = \lambda_0$, then by Pythagoras $|\lambda - \alpha|^2 > (\lambda_0 - \alpha)^2$ and we have a contradiction to the maximality of $\lambda_0 - \alpha$. This concludes the whole proof for $A \gg 0$.

► Now we treat the case $A \geqq 0$:

Let Δ be a strictly positive matrix. Then $A_k := A + \frac{1}{k}\Delta$ with $k \geq 1$ defines a sequence of strictly positive matrices converging towards A.
Thus, by the first part of the proof, we know there exists a strictly dominant eigenvalue $\lambda_k > 0$ of A_k. By Gershgorin's circle theorem [3] it follows from the definition of A_k, that $\lambda_1 \geq \lambda_2 \geq \cdots \geq r$, where r denotes the spectral radius of A. Therefore $\{\lambda_k\}_{k\geq 1}$ defines a monotonically decreasing convergent sequence.
Since $\{v_k\}_{k\geq 1}$ defines a bounded sequence within the compact unit-ball, we can extract a convergent subsequence $\{v_{k_i}\}_{k_i \geq 1}$, according to the well-known theorem of Bolzano-Weierstraß.
Consequently, by the positivity of v_k and λ_k we conclude

$$\lim_{i\to\infty} v_{k_i} = v^* \geqq 0 \text{ with } \|v^*\| = 1 \quad \text{and} \quad \lim_{i\to\infty} \lambda_{k_i} = \lambda^* \geq r.$$

In the end we get

$$Av^* = \lim_{i\to\infty} A_{k_i} v_{k_i} = \lim_{i\to\infty} \lambda_{k_i} v_{k_i} = \lambda^* v^*$$

and therefore $\lambda_0 := \lambda^* = r$ and $v_0 := v^*$. ■

Notice, that if $A^m x_0 = \lambda_0 x_0$ for some $m > 0$, then $Ax_0 = \sqrt[m]{\lambda_0} x_0$. Hence, if $A^m \gg 0$, we can apply Theorem 1.5 and get the same statements as for $A \gg 0$.

Corollary 1.1

Let $A \geqq 0$ and assume $A^m \gg 0$ for some $m > 0$. Then we can make for A the same conclusions as for a positive matrix in Theorem 1.5.

In order to find an answer, when such an $m > 0$ exists, we need to look at a certain class of matrices, called *irreducible matrices* [18].

Definition 1.5 (*Reducible matrix*)

Let $A \in \mathbb{R}^{n\times n}$, then A is called ***reducible*** *if there exists a permutation matrix π, such that*

$$\pi^T A \pi = \begin{pmatrix} B_1 & * \\ & B_2 \end{pmatrix}$$

with square matrices B_1 and B_2. If A is not reducible, then it is called ***irreducible****.*

Interesting property of the largest eigenvalue for irreducible nonnegative matrices is given by the following lemma and theorem.

Lemma 1.4

If A is irreducible nonnegative matrix with a multiple dominant eigenvalue, then $tr(A) = 0$. [18]

Theorem 1.6

A is a irreducible nonnegative matrix with unique largest eigenvalue if and only if $A^m \gg 0$ for some $m > 0$. [18]

For reducible nonnegative matrices this statements hold generally not true, what we can see for example if we assume A to be the identity. However, reducible nonnegative matrices have the property of having multiple nonnegative eigenvalues, which we can conclude from the following lemma.

Lemma 1.5

Let A be a reducible nonnegative matrix, then there exists a permutation matrix π such that

$$\pi^T A \pi = \begin{pmatrix} B_1 & * & * & * \\ & B_2 & * & * \\ & & \ddots & * \\ & & & B_k \end{pmatrix},$$

where each B_i is irreducible or equal to zero. [18]

This means, the eigenvalues of $A \geqq 0$ are given by the eigenvalues of $B_i \geqq 0$ and it is possible to diagonalize $\pi^T A \pi$ by a blockdiagonal matrix. Hence, according to Theorem 1.5, there must exists at least one nonnegative eigenvector to each B_i corresponding to the largest eigenvalue of B_i.

2. Positive Realization

A clear drawback of Theorem 1.3 is the fact, that a simple state-space transformation can already destroy the nonnegativity of (B, C, D) and the Metzler matrix property of A. In this case all that is left is the nonnegativity of the impulse response. On the other side, as we demonstrate in Example 2.1, the nonnegativity of the impulse response does not guarantee a minimal positive realization. This chapter will treat the problem of positive realizability.

For a first order system with transfer function

$$G(s) = \frac{1}{s+\alpha_1} M \text{ with } M \in \mathbb{R}^{k\times m} \tag{2.1}$$

it can be seen, that the nonnegativity of the impulse response $g(t)$ implies the positivity of the minimal realization of $G(s)$. By the nonngetivity of $g(t)$ it follows $M \geqq 0$ and its rank $rk(M)$ of M must be equal to one, due to the fact that $G(s)$ is a system of first order. For instance by applying Singular Value decomposition, we can decompose M into two positive vectors $C \in \mathbb{R}^k$ and $B^T \in \mathbb{R}^m$: let $M = U\Sigma V^T$ with

$$\Sigma = \begin{pmatrix} \sigma_1 & \\ & 0 \end{pmatrix} \in \mathbb{R}^{k\times m},\ U \in \mathbb{R}^{k\times k} \text{ and } V \in \mathbb{R}^{m\times m}.$$

then $MM^T = U\Sigma^2 U^T \geqq 0$ and $M^T M = V\Sigma^2 V^T \geqq 0$. We conclude that in each case the first column u_1 of U and v_1 of V is an eigenvector to the largest eigenvalue σ_1. Consequently, by Theorem 1.5, $u_1, v_1 \geqq 0$ up to a negligible sign-change and therefore we can define

$$A := -\alpha_1, \quad B := \sqrt{\sigma_1}|v_1|^T, \quad C := \sqrt{\sigma_1}|u_1|, \tag{2.2}$$

which is a positive minimal realization of $G(s)$.

This implication does not hold for systems of higher orders in general and hence, the minimal realization of a positive system does <u>not</u> need to be just a transformation of a positive realization. In order to get the positive realization of an (internally) positive system, we may need to increase the state-space dimension.

Example 2.1 (Nonpositive minimal realization)
Let us consider the system (A, b, c) with

$$A := \begin{pmatrix} -2 & 0 & 0 & 1 \\ 1 & -2 & 0 & 0 \\ 0 & 1 & -2 & 0 \\ 0 & 0 & 1 & -2 \end{pmatrix}, \quad b := c^T := \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

The transfer function of this system is given by

$$G(s) = \frac{2s^2 + 7s + 7}{(s+1)(s^2 + 4s + 5)} = \frac{2s^2 + 7s + 7}{s^3 + 5s^2 + 9s + 5},$$

which has poles at -1 and $-2 \pm i$. However, by a straight forward calculation of the characteristic polynomial of a 3×3-Metzler matrix $\tilde{A}$ and a comparison of its coefficients with $s^3 + 5s^2 + 9s + 5$, it follows

$$-\tilde{a}_{11} - \tilde{a}_{22} - \tilde{a}_{33} = 5 \quad \text{and} \quad \tilde{a}_{11}\tilde{a}_{22} + \tilde{a}_{11}\tilde{a}_{33} + \tilde{a}_{22}\tilde{a}_{33} \geq 9.$$

This gives

$$(-4 - \tilde{a}_{22} - \tilde{a}_{33})(\tilde{a}_{22} + \tilde{a}_{33}) \geq 9$$

or equivalently

$$(\tilde{a}_{22} + \tilde{a}_{33} + 2)^2 \leq -5.$$

Hence, the system does not have a minimal realization, which is positive.

2.1. Reachability, Observability and Realizability

In the following we want to investigate where the reasons lie, that not every externally positive SISO-systems system has a positive minimal realization. Furthermore, we will show that for second-order systems, external positivity is equivalent to internal positivity. For this purpose we start with the definitions of reachable and observable sets with respect to a nonnegative input.

Let us consider the linear time-invariant SISO-system (A, b, c^T)

$$\begin{aligned} \dot{x} &= Ax + bu, \\ y &= c^T x, \end{aligned} \tag{2.3}$$

then, as in [1] [6] [19], the reachable and observable sets, with respect to nonnegative inupts, are defined as follows.

Definition 2.1 (*Reachable set*)

Let $X_\infty(A, b)$ be the set of all points that can be reached within finite time from the origin by nonnegative inputs, i.e.

$$X_\infty(A, b) := \{x \mid x = \int_0^t e^{A(t-\tau)} bu(\tau) d\tau, t \geq 0, u \geq 0 \text{ integrable}\}.$$

Then we define the reachable set X_r as

$$X_r := X_r(A, b) := \overline{X}_\infty(A, b).$$

Definition 2.2 (*Observable set*)

Let $X_o(c^T, A)$ be the set of all states, that cause a nonnegative output for all $t \geq 0$ if $u(t) = 0$, i.e.

$$X_o := X_o(c^T, A) := \{x \mid \langle c, e^{At}x\rangle \geq 0, \forall t \geq 0\},$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product. Then X_o is called the observable set.

Note, do not confuse X_r and X_o with the reachable and observable subspaces of a system, which are given by the ranges of P and Q in (3.2) and (3.4), presented in the next chapter.

By looking at the definition of $X_\infty(A, b)$ it is easy to see that $X_\infty(A, b)$ is a convex cone because of the linearity of the system (see Appendix): assume x_1, $x_2 \in X_\infty(A, b)$ are two states steered by u_1 and u_2 in time $t_1 \geq t_2$. Then by setting

$$\tilde{u}_2(\tau) := \begin{cases} 0, 0 \geq \tau < t_1 - t_2 \\ u_2(\tau), t_1 - t_2 \leq \tau \leq t_1 \end{cases}$$

we can steer the system from the origin to

$$x_0 := \alpha x_1 + \beta x_2, \ \alpha, \beta \geq 0$$

by the positive input

$$u_0 := \alpha u_1 + \beta \tilde{u_2}.$$

Since X_o is the dual set to $\{e^{A^T t}c, \ \forall t \geq 0\}$, it is closed and convex (Lemma A.1). We can conclude the same by its property of being the dual cone to X_r. In order to do this we need to show, that every $x \in X_r$ can be approximated similar as the points in X_o.[1] [19]

Lemma 2.1

If C_r denotes the smallest convex cone containing the set

$$\{x \mid x = e^{At}b, \ t \geq 0\}$$

then

$$X_r = \overline{C}_r.$$

Proof: For a real-valued interval $[0, t] \subset \mathbb{R}$, it holds

$$\int_0^t e^{A(t-\tau)} b\delta(\tau - \tau_0) d\tau = \begin{cases} e^{A(t-\tau_0)}b, \ \tau_0 \in]0, t[\\ \frac{1}{2}e^{A(t-\tau_0)}b, \ \tau_0 = 0 \text{ or } \tau = b \text{ for } t > 0 \\ 0, \ \text{otherwise.} \end{cases} \tag{2.4}$$

Let $v \in \{x \mid x = e^{At}b, \forall\ t \geq 0\}$, then by equation (2.4) we find with $u(\tau) := 2\delta(\tau)$ a nonnegative input such that $v \in X_\infty(A,b)$. Since C_r and X_r are both convex, we conclude $\overline{C}_r \subset X_r$.

Suppose $x \in X_\infty(A,b)$. Since $e^{At}b \in C_r\ \forall t \geq 0$ we get because of its cone property $e^{A(t-\tau)}bu(\tau)\Delta \in X_r$ with $0 \leq \tau \leq t$ and $\Delta \geq 0$. Consequently if we approximate

$$x = \int_0^t e^{A(t-\tau)}bu(\tau)d\tau$$

by its Riemann-sum S_N, we get by convexity of C_r

$$x \approx S_N := \sum_{i=1}^{N} e^{A(t-\tau_i)}bu(\tau_i)(\tau_i - \tau_{i-1}) \in C_r \text{ with } \tau_0 = 0 < \tau_1 < \cdots < \tau_N = t.$$

Since S_N defines a convergent series of C_r, its limit lies in $\overline{C}_r$ and therefore

$$x = \lim_{N\to\infty} S_N \in \overline{C}_r.$$

■

Observe the similarity between X_o and C_r. By definition of the dual cone and the fact that every dual cone is closed, it holds

$$\begin{aligned} X_r(A,b)^* = C_r^* &= \{y \mid \langle e^{At}b, y\rangle \geq 0,\ \forall t \geq 0\} \\ &= \{y \mid \langle b, e^{A^T t}y\rangle \geq 0,\ \forall t \geq 0\} \\ &= X_o(b^T, A^T). \end{aligned}$$

Since X_r is closed and convex, we get by Theorem A.1, $X_r^{**} = X_r$.

Lemma 2.2 (*Dual Cone*)

$$X_r(A,b)^* = X_o(b^T, A^T) \quad \textit{and} \quad X_o(c^T, A)^* = X_r(A^T, c).\ \textit{[19]}$$

In the following we want to find out more about the geometric structure of X_r and X_o in case of external positivity and relate them to (externally positive) minimal realizations.

Lemma 2.3

Let (A, b, c^T) be an externally positive system, then $X_r \subseteq X_o$.[19]

Proof: By Lemma 2.1 and its proof we know, that a vector $x \in X_r$ can be arbitrarily close approximated by a nonnegative finite linear combination of C_r

$$x = \sum_{i=1}^{N} e^{At_i}b\Delta_i.$$

According to Theorem 1.1

$$c^T e^{At} b \geq 0 \ \forall t \geq 0$$

and therefore

$$c^T e^{At} x = \sum_{i=1}^{N} c^T e^{A(t+t_i)} b \Delta_i \geq 0, \ \forall t \geq 0 \ \Rightarrow \ x \in X_o.$$

■

Theorem 2.1 (*Minimal Realization*)

(A, b, c^T) is a minimal realization if and only if $X_r(A, b)$ is solid and $X_o(c^T, A)$ is pointed.[19]

Proof: By Lemma 2.2 and Theorem A.2 we only have to show the statement for either X_r or X_o and the other follows by duality. Notice, in fact we will basically use the same arguments as to prove Theorem A.2.

Suppose X_o is not pointed, then as in the proof of Theorem A.2 there must be a line $\alpha v \in X_o \ \forall \alpha \in \mathbb{R}$ and $v \in \mathbb{R}^n_+ \setminus \{0\}$ for which $c^T e^{At} \alpha v \geq 0$ holds.
Consequently,

$$c^T A v = 0 \ \forall t \geq 0,$$

and thus $Qv = 0$, which means that v has no influence on the output and cannot be observed. This contradicts the minimality of the realization.

In the other case, if the realization is not observable, there must exist a nonzero vector $w \in \mathbb{R}^n$ such that

$$c^T e^{At} w = 0 \ \forall t \geq 0,$$

and as before there would exist a line in X_o, such that $\alpha w \in X_o \ \forall \alpha \in \mathbb{R}$, which contradicts the pointedness. ■

A direct consequence of Lemma 2.3 and Theorem 2.1 is the following result about minimal realizations of externally positive systems. The same follows from the results in [19], but by assuming external positivity, we can give a much shorter proof here.

Theorem 2.2

Let (A, b, c^T) be a minimal realization of an externally positive system. Then X_r and X_o are proper cones.

Proof: By Theorem 2.1 we only need to show the pointedness of X_r and solidness of X_o.

Let us assume X_r is not pointed. Then

$$\exists x \in \mathbb{R}^n : x \in X_r \cap -X_r.$$

According to Lemma 2.3, $X_r \subseteq X_o$ and hence

$$x \in X_o \cap -X_o,$$

which means X_o would be pointed. Since we have assumed to have a minimal realization this is a contradiction to Theorem 2.1. Again, the same statement for X_o follows by the duality of X_r and X_o. ■

These results show which strict requirements have to be fulfilled for a minimal realization of a positive system. The questions is now, by looking at X_r and X_o, what is the distinction between internal and external positivity of a minimal realization.

Theorem 2.3 (*Minimal Positive Realization*)

Let (A, b, c^T) be a minimal realization of a strictly proper transfer function $G(s)$. Then $G(s)$ possesses a positive realization if and only if there exists a polyhedral cone X_p, such that

(i) $(A + \lambda I)X_p \subset X_p$ *for some* $\lambda \geq 0$,

(ii) $X_r \subset X_p \subset X_o$.*[19]*

In the proof of the theorem as well as in the subsequent conclusions we will need the following lemma.

Lemma 2.4

(i) $X_r(A, b) = X_r(A + \lambda I, b), \quad X_o(c, A) = X_o(c, A + \lambda I), \ \forall \lambda \in \mathbb{R}$

(ii) $e^{At} X_r \subset X_r, \quad e^{At} X_o \subset X_o, \ \forall t \geq 0$.*[19]*

Proof: The first statement follows by $c^T e^{(A+\lambda I)t} x = e^{\lambda t} c^T e^{At} x$, the definition and cone property of X_o and Lemma 2.2.
For the second statement we just need to notice, that $e^{At} e^{A\tilde{t}} = e^{A(t+\tilde{t})}$ and apply it to the definition of X_r and X_o. ■

Now we are ready to prove the main theorem of this chapter.

Proof to Theorem 2.3: ► Sufficiency: By definition of a polyhedral cone we can write X_p as $X_p = P\mathbb{R}^k_+$ with $P \in \mathbb{R}^{n \times k}$. Consequently from assumption (i) we conclude

$$(A + \lambda I)P = PK_A, \text{ with } K_A \in \mathbb{R}^{k \times k}_+,$$

and define

$$A_p := K_A - \lambda I \overset{e}{\geqq} 0.$$

By the definition of C_r and assumption (ii), $b \in X_r \subset X_p$. Hence, there exists a vector $b_p \in \mathbb{R}_+^k$ such that

$$Pb_p = b.$$

Again, with the same arguments it holds $c \in X_r(A^T, c) = X_o(c^T, A)$. Since $X_p \subset X_o$ it follows by the definition of a dual set, that $X_o^* \subset X_p^*$ and consequently $c \in X_p^*$. Then by the definition of the dual set, there must exist a vector $c_p \in \mathbb{R}_+^k$ such that

$$c_p = P^T c.$$

Noticing that $AP = PA_p$ and thus $A^k P = A^{k-1}(AP) = A^{k-1}PA_p = \cdots = PA_p^k$ we can compare the impulse responses of the system (A, b, c^T) and (A_p, b_p, c_p^T) as follows

$$c^T e^{At} b = c^T \sum_{k=0}^{\infty} \frac{A^k t^k}{k!} P b_p = c^T P \sum_{k=0}^{\infty} \frac{A_p^k t^k}{k!} b = c_p^T e^{A_p t} b_p.$$

Hence, it holds for the transfer functions

$$c_p^T (sI - A_p)^{-1} b_p = c^T (sI - A)^{-1} b,$$

and (A_p, b_p, c_p^T) is a positive realization of $G(s)$.

► Necessity: Assume (A_p, b_p, c_p^T) is a positive realization of $G(s)$ of dimension N.
By setting $\lambda := max_{i=1,\ldots,n}\{-a_{p_{ii}}\}$ we define a nonnegative matrix $\tilde{A}_p$ by

$$\tilde{A}_p := A_p + \lambda I \quad \text{and} \quad \tilde{A} := A + \lambda I.$$

It is readily noted, that $(\tilde{A}_p, b_p, c_p^T)$ and $(\tilde{A}, b, c^T)$ are both realization of $G(s - \lambda)$, where $(\tilde{A}, b, c^T)$ is a minimal realization. With the help of Lemma 2.4 (i) we get

$$X_r(A, b) = X_r(\tilde{A}, b) \quad \text{and} \quad X_o(c^T, A) = X_o(c^T, \tilde{A}),$$

and thus in order to conclude the prove it is sufficient to show that there exists a polyhedral cone $X_p \subset \mathbb{R}^n$ such that

(i) $\tilde{A} X_p \subset X_p$,

(ii) $X_r(\tilde{A}, b) \subset X_p \subset X_o(c^T, \tilde{A})$.

We will find such a polyhedral cone by doing so for the observable part of $(\tilde{A}_p, b_p, c_p^T)$.
Let us transform the system $(\tilde{A}_p, b_p, c_p^T)$ into the observable canonical form:

$$T_o^{-1} \tilde{A}_p T_o = \begin{pmatrix} A_{11} & 0 \\ * & * \end{pmatrix}, \quad T_o^{-1} b_p = \begin{pmatrix} b_1 \\ * \end{pmatrix}, \quad c_p^T T_o = \begin{pmatrix} c_1^T & 0 \end{pmatrix}.$$

Since (A_{11}, b_1, c_1) is observable, we can transform it into the controllable canonical form

$$A_{aa} := T_c^{-1} A_{11} T_c = \begin{pmatrix} \tilde{A}^T & * \\ 0 & * \end{pmatrix}, \quad b_a := T_c^{-1} b_1 = \begin{pmatrix} c \\ 0 \end{pmatrix}, \quad c_a := c_1^T T_c = \begin{pmatrix} b^T & * \end{pmatrix}.$$

and retrieve the minimal realization $(\tilde{A}^T, c, b^T)$. The reason why we use this form instead of $(\tilde{A}, b, c^T)$, is that by transposing we get the same matrix structure as for the observable form. Notice, this is only possible for SISO-systems. Then, by defining

$$T := \begin{pmatrix} T_c & \\ & I \end{pmatrix} T_o,$$

we get a system

$$T^{-1}\tilde{A}_pT = \begin{pmatrix} A_{aa} & 0 \\ * & * \end{pmatrix}, \quad T^{-1}b_p = \begin{pmatrix} b_a \\ * \end{pmatrix}, \quad c_p^TT = \begin{pmatrix} c_a^T & 0 \end{pmatrix} \text{ and } T^{-1} = \begin{pmatrix} (T^{-1})_a \\ * \end{pmatrix}.$$

We have partitioned the system in exactly this way because of the zero-matrices that we will need in the following and the fact, that we can express $(\tilde{A}_p, b_p, c_p)$ in terms of $(\tilde{A}, b, c)$. Given a matrix Q, we define by $cone\{Q\}$ the polyhedral cone generated by the columns of Q. If we set $X_k := (cone\{(T^{-1})_a\})^*$ then $X_k^* = cone\{(T^{-1})_a\}$ and $A_{aa}X_k^* = cone\{A_{aa}(T^{-1})_a\}$. By looking at

$$\begin{pmatrix} (T^{-1})_a \\ * \end{pmatrix}\tilde{A}_p = \begin{pmatrix} (T^{-1})_a\tilde{A}_p \\ * \end{pmatrix} = \begin{pmatrix} A_{aa} & 0 \\ * & * \end{pmatrix}\begin{pmatrix} (T^{-1})_a \\ * \end{pmatrix}$$

we observe

$$(T^{-1})_a\tilde{A}_p = A_{aa}(T^{-1})_a.$$

Hence,

$$A_{aa}X_k^* = cone\{(T^{-1})_a\tilde{A}_p\} \subset cone\{(T^{-1})_a\} = X_k^*,$$

where the inclusion follows by $\tilde{A}_p \geqq 0$. Consequently, by the definition of a dual cone $A_{aa}^TX_k \subset X_k$. Because b and c are nonnegative we get

$$b_a = (T^{-1})_ab \in X_k^* \text{ and because } c_a^T(T^{-1})_a = c,\ c_a \in X_k.$$

Let us write the polyhedral cone X_k as

$$X_k = cone\left\{\begin{pmatrix} P \\ * \end{pmatrix}\right\} \text{ and define } X_p := cone\{P\},$$

where P denotes the matrix of the first n rows, corresponding to $\tilde{A}$ in A_{aa}^T. Then by considering the corresponding parts of P in b_a and c_a, it is straightforward to see that

$$\tilde{A}X_p \subset X_p,\ b \in X_p \text{ and } c^T \in X_p^*. \tag{2.5}$$

Thus condition (i) is verified and it is left so show condition (ii).
By (2.5) we get $\tilde{A}^ib \in X_p, \forall i \geq 0$ and therefore by Lemma A.2 $e^{\tilde{A}t}b = \sum_{i=0}^{\infty} \tilde{A}^ib \in X_p, \forall t \geq 0$.
Together with Lemma 2.1 we get conclude $X_r(\tilde{A}, b) \subset X_p$.
In the same way we can show the second inclusion. By the $\tilde{A}$-invariance of X_p it follows again $\tilde{A}^TX_p^* \subset X_p^*$ and hence as before $X_r(\tilde{A}^T, c) \subset X_p^*$. From the definition of a dual cone, Theorem A.1 and Lemma 2.2 we get $X_p^{**} = X_p \subset X_r(\tilde{A}^T, c)^* = X_o(c^T, \tilde{A})$. ■

In Lemma 2.3 we got, that X_r is a subset of X_o and hence it would be a perfect candidate for X_p. Unfortunately we know, not every externally positive system has a positive realization of dimension equal to the order. In this case X_r is either not polyhedral or not $(A + \lambda I)$-invariant. The next lemma will show the relation between these two properties.

Lemma 2.5

Let $X_p \subseteq \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Then $e^{At}X_p \subset X_p \ \forall t \geq 0$ if and only if there exists a $\lambda \geq 0$, such that $(A+\lambda I)X_p \subset X_p$.[19]

Proof: ▶ Sufficiency: Let us assume there exists an $x \in e^{At}X_p \setminus X_p$, then by the cone property of X_p

$$e^{\eta t}e^{At} = e^{(A+\eta I)t} \notin X_p, \forall \eta.$$

Consequently, by series expansion it holds

$$e^{(A+\eta I)t} = \sum_{i=0}^{\infty} \frac{(A+\eta I)^i x}{i!} t^i \notin X_p, \forall \eta, \forall t \geq 0$$

Together with Lemma A.2 this is a contradiction to $(A+\lambda I)X_p \in X_p$. Observe for this direction a closed convex cone X_p would have been sufficient.

▶ Necessity: Let X_p be generated by the set $\{p_1, \ldots, p_N\}$ and X_p^* by $\{q_1, \ldots, q_{N'}\}$. Hence by definition of the dual cone and by assumption

$$\langle q_j, p_i\rangle \geq 0 \text{ and } \langle q_j, e^{At}p_i\rangle \geq 0, \forall t \geq 0,\ 1 \leq i \leq N,\ 1 \leq j \leq N' \tag{2.6}$$

Again by definition of X_p^*, the existence of a $\lambda \geq 0$ such that $(A+\lambda)X_p \subset X_p$, is equivalent to

$$\exists \lambda \geq 0 : \langle (A+\lambda I)p_i, q_j\rangle = \langle q_j, (A+\lambda I)p_i\rangle \geq 0,\ 1 \leq i \leq N,\ 1 \leq j \leq N'.$$

If $\langle q_j, Ap_i\rangle \geq 0$ then it is obvious by (2.6), that

$$\langle q_j, (A+\lambda_{ij}I)p_i\rangle \geq 0, \forall \lambda_{ij} \geq 0,$$

and we set $\lambda_{ij} = 0$. If $\langle q_j, Ap_i\rangle < 0$ and $\langle q_j, p_i\rangle > 0$ we can define

$$\lambda_{ij} := -\frac{\langle q_j, Ap_i\rangle}{\langle q_j, p_i\rangle} > 0,$$

and it holds $\langle q_j, (A+\lambda_{ij}p_i\rangle \geq 0$. For the case $\langle q_j, p_i\rangle = 0$, we have to take a look at the series expansion of $\langle q_j, e^{At}p_i\rangle$

$$0 \leq \langle q_j, e^{At}p_i\rangle = \langle q_j, p_i\rangle + \langle q_j, Ap_i\rangle t + R(t) = \langle q_j, Ap_i\rangle t + R(t), \forall t \geq 0 \text{ and } R(t) \in \mathcal{O}(t^2).$$

Hence, by dividing by $t \geq 0$ yields

$$\langle q_j, Ap_i\rangle + \tilde{R}(t), \forall t \geq 0 \text{ with } \tilde{R}(t) \in \mathcal{O}(t),$$

and we can conclude $\langle q_j, Ap_i\rangle \geq 0$, because for sufficiently small $t \geq 0$ it holds

$$|\tilde{R}(t)| \leq |\langle q_j, Ap_i\rangle|.$$

Setting $\lambda := \max_{i,j} \lambda_{ij}$ concludes the proof. ■

As a consequence of Lemma 2.4 (ii) and Lemma 2.5 it follows, that if X_r is polyhedral then it is also $(A - \lambda I)$-invariant. Unfortunately, neither the location of a polyhedral cone X_p nor the determination if X_r is polyhedral, is an easy solvable problem and to the authors knowledge there exists no systematic way.

An exception is the case of a second-order system. In this case we know from Lemma A.3 and Theorem 2.2 that X_r is always polyhedral.

Corollary 2.1 (*First- and Second Order Realizability*)
Every first- and second-order externally positive system has a positive realization.

Still, in the view of positivity preserving model order reduction, the same problems will remain and thus it is more advisable to stick to the preservation of the matrix properties, as we will do now to get an explicit expression of a second-order system.

2.2. Second-Order Realization and Special Cases

In the following we will show an easy way to get a second-order positive realization and discuss some special cases of higher dimensional positive realizations.
By equation (1.8) and the rules for Laplace transformation we know that we can write any (discrete- and continuous-time) transfer function,

$$G(p) = \frac{\beta_1 p^{n-1} + \beta_2 p^{n-2} + \cdots + \beta_n}{p^n + \alpha_1 p^{n-1} + \cdots + \alpha_n},$$

as a series of Markov coefficients

$$G(p) = \sum_{i=1}^{\infty} \frac{g_i}{p^i}. \tag{2.7}$$

Applying polynomial long division on $G(p)$ and comparing the coefficients with equation (2.7) yields that for the first n Markov coefficients[6]

$$\beta_i = g_i + \sum_{k=1}^{i-1} \alpha_k g_{i-k}, \ i = 1, \ldots, n. \tag{2.8}$$

Among many canonical realizations of a SISO-tranfer function $G(s)$, the most well-known are the *observable canonical form*

$$A_o = \begin{pmatrix} 0 & 0 & \cdots & 0 & -\alpha_n \\ 1 & 0 & \cdots & 0 & -\alpha_{n-1} \\ 0 & 1 & \cdots & 0 & -\alpha_{n-2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -\alpha_1 \end{pmatrix} \quad b_o = \begin{pmatrix} \beta_n \\ \beta_{n-1} \\ \beta_{n-2} \\ \vdots \\ \beta_1 \end{pmatrix}, \quad c_o = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

and the *controllable canonical form*

$$A_c = A_o^T, \quad b_c = c_o^T, \quad c_c = b_o^T.$$

Beside those two canonical forms, there exist two further interesting realizations called *Markov form* and *Dual Markov form* [6], which follow directly from the first two by consideration of equation (2.8).
The Markov form is given by

$$A_M = A_o, \quad b_M = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad c_M^T = \begin{pmatrix} g_1 & g_2 & \cdots & g_{n-1} & g_n \end{pmatrix}$$

and its dual by

$$A_{M^*} = A_M^T, \quad b_{M^*} = c_M, \quad c_{M^*}^T = b_M^T.$$

Observe, by the proof to Theorem 1.2 none of these realizations is suitable for a continuous-time positive system, since in this case all the diagonal entries have to be smaller than 0. Fortunately for a discrete system we only need $A \geqq 0$.

Theorem 2.4

Let $G(z)$ be the transfer function of a discrete externally positive system with $\alpha_i \leq 0$, $i = 1, ..., n$, then the system is positively realizable with dimension n by the Markov form and its dual.[6]

For a positive continuous-time system (A, b, c^T) it follows by the definition of a $-M$-matrix, that there must exist an $\gamma > 0$, such that $A + \gamma I \geqq 0$. Hence, the system $(A + \gamma I, b, c^T)$ possesses nonnegative Markov coefficients and its transfer function is given by

$$G_\alpha(s) = c^T(sI - A + \gamma I)^{-1}b = c^T((s - \gamma)I - A)^{-1}b = G(s - \gamma).$$

Applying Lemma 2.4 leads to the following theorem.

Theorem 2.5

Let $G(s)$ be the transfer function of a continuous-time externally positive system. If there exists an $\gamma > 0$, such that the Markov coefficients g_{γ_i} of $G(s-\gamma)$ are nonnegative and the corresponding α_{γ_i} are nonpositive, then there exists a positive realization of dimension n.[6]

Observe, γ cannot be chosen arbitrarily: assume we have have a Markov form positive realization (A_M, b_M, c_M^T) of $G(z-\gamma)$, then the positive continuous-time realization is given by $(A_M - \gamma I, b_M, c_M^T)$. By looking at the Markov form we notice, that the trace of $A_M - \gamma$ is given by

$$tr(A_M - \gamma) = -(n-1)\gamma + (-\gamma - \alpha_1) \text{ with } \alpha_1 \leq 0.$$

Consequently,

$$\gamma = -\frac{tr(A_M + \gamma) + \alpha_1}{n} \geq -\frac{tr(A_M + \gamma)}{n}$$

and we have found a lower bound. Since the trace of a matrix is invariant under similarity transformations and equal to the sum of all eigenvalues, we can conclude the following result.

Theorem 2.6

Assume $G(s)$ is the transfer function of a continuous-time externally positive system and there exists a $\gamma > 0$, such that the Markov coefficients g_{γ_i} of $G(s-\gamma)$ are nonnegative and the corresponding α_{γ_i} are nonpositive. Then it has to hold

$$\gamma \geq -\frac{1}{n}\sum_{i=1}^{n} p_i,$$

where p_i denote the poles of $G(s)$.

Now let us consider the second-order case. Corollary 2.1 tells us, that each externally positive transfer function

$$G(s) = \frac{\beta_1 s + \beta_2}{s^2 + \alpha_1 s + \alpha_2} \tag{2.9}$$

can be positively realized with a state-space dimension equal to 2. By Lemma 1.2 we know that the dominant pole of $G(s)$ has to be real and therefore $G(s)$ consists of two real poles $p_1, p_2 < 0$. Since $G(s-\gamma)$ can then be written as

$$G(s-\gamma) = \frac{\beta_1(s-\gamma) + \beta_2}{(s-(\gamma+p_1))(s-(\gamma+p_2))}$$

it follows for α_{γ_1} and α_{γ_2} that

$$\alpha_{\gamma_1} = -(\gamma+p_1) - (\gamma+p_2) = -2\gamma - (p_1+p_2) \quad \text{and} \quad \alpha_{\gamma_2} = (\gamma+p_1)(\gamma+p_2).$$

Observe, since $\alpha_{\gamma_1}, \alpha_{\gamma_2} \leq 0$, α_{γ_1} gives the same condition as Theorem 2.6.
Let $\gamma := -\frac{p_1+p_2}{2}$ then

$$\alpha_{\gamma_2} = (-\frac{p_1+p_2}{2} + p_1)(-\frac{p_1+p_2}{2} + p_2) = (\frac{p_1-p_2}{2})(\frac{p_2-p_1}{2}) = -\frac{(p_1-p_2)^2}{4} \leq 0,$$

and thus we found a discrete realization given by

$$A_M = \begin{pmatrix} 0 & \frac{(p_1-p_2)^2}{4} \\ 1 & 0 \end{pmatrix}, \quad b_M = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c_M^T = \begin{pmatrix} g_{\gamma_1} & g_{\gamma_2} \end{pmatrix}.$$

Together with equation (2.8) we get the following continuous-time positive reali-zation

$$A_p = \begin{pmatrix} \frac{p_1+p_2}{2} & \frac{(p_1-p_2)^2}{4} \\ 1 & \frac{p_1+p_2}{2} \end{pmatrix}, \quad b_p = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c_p^T = \begin{pmatrix} \beta_1 & \beta_2 - \beta_1 \frac{p_1+p_2}{2} \end{pmatrix}. \tag{2.10}$$

It easy to show that another realization of $G(s)$ is given by

$$A_p = \begin{pmatrix} p_2 & 0 \\ \beta_2 + \beta_1 p_1 & p_1 \end{pmatrix}, \quad b_p = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad c_p^T = \begin{pmatrix} \beta_1 & 1 \end{pmatrix}.[6] \tag{2.11}$$

Since $\beta_1 > 0$ by equation (2.8) and $p_2 \leq p_1 < 0$, all we need to verify for internal positivity is that

$$\beta_2 + \beta_1 p_1 > 0.$$

This is equivalent to show that $p_1 > -\frac{\beta_2}{\beta_1}$. Since $-\frac{\beta_2}{\beta_1}$ represents a real zero of the system, the internal positivity follows by Lemma 1.3.

Notice, for both realizations we avoid to show the nonnegativity of the impulse response. In comparison to Theorem 2.3, we also do not need to find the boundaries of X_r. Thus, we found an shorter and more applicable proof for the equivalence of external and internal positivity for second-order continuous-time systems.

Theorem 2.7 (*Second Order Positive Realization*)

Let $G(s)$ be a second-order transfer function given by equation (2.9). Then $G(s)$ is externally positive if and only if $\beta_1 > 0$ and the system possesses a real dominant pole p_1 such that $\beta_2 + \beta_1 p_1 > 0$.
If $G(s)$ is externally positive, then it possesses an internally positive realization given in (2.10) or (2.11).

All these problems of positive realizability, that we have encountered in this section and the section before, are basically the main difficulties we have to deal with, when we want to preserve the (internal) positivity of a system after performing model order reduction. Especially problematic is, that even if we have an externally positive reduced system, we cannot estimate how large its positive realization gets, as seen in Example 2.1.

3. Balanced Truncation

Amongst the different reduction methods one has turned out to be simple and efficient, the so-called Balanced Truncation. The main advantage of Balanced Truncation is its interpretation with the help of energy functions and the providence of a good error bound estimation in the $\mathcal{H}_\infty$-norm.

The easiest way to perform model order reduction is to remove successively uncontrollable and unobservable states in order to gain a minimal realization. In fact, this means nothing else than getting a realization with identical reachable and observable space. This idea can be interpreted and generalized with the help of energy functions and Lypunov equations.

It is a well-known result, if we define P by

$$AP + PA^T = -BB^T, \sigma(A) \subset \mathbb{C}^- \tag{3.1}$$

$$P = \int_0^\infty e^{At} BB^T e^{A^T t} dt, \tag{3.2}$$

the range $rg(P)$ of P is equal to the reachable subspace. P is called the *Controllability Gramian*. The same consideration can be done for the observable subspace.

Let Q be defined by

$$A^T P + PA = -C^T C, \sigma(A) \subset \mathbb{C}^- \tag{3.3}$$

$$Q = \int_0^\infty e^{A^T t} C^T C e^{At} dt, \tag{3.4}$$

then $rg(Q)$ is again equal to the observable subspace and Q is called the *Observability Gramian*.

The equations (3.1) and (3.3) give obviously an indirect way of testing, whether a system is controllable/observable and are named after its discoverer *Lypunov equations*. Lyapunov equations are especially important in the context of stability, which we want to explain with the next Lemma.

Lemma 3.1

Let P be the solution to

$$AP + PA^T = -H, \tag{3.5}$$

then

1. $\Re(\lambda_i(A)) \leq 0$ *if* $P > 0$ *and* $H \geq 0$
2. $\Re(\lambda_i(A)) < 0$ *if* $P > 0$ *and* $H > 0$,

where $\lambda_i(A)$ denotes the i-th eigenvalue of A.[30]

Proof: Let v be an eigenvector to the eigenvalue λ of A^T, i.e. $A^T v = \lambda v$. Then by assumption

$$\bar{v}^T(AP + PA^T)v = 2(\bar{\lambda} + \lambda)\bar{v}^T P v = -\bar{v}^T H v \leq 0$$

and since $P > 0$ it follows that $\bar{\lambda} + \lambda = \Re(\lambda) \leq 0$. ■

Remark: The solution to a Lyapunov equation as in (3.5) is unique if and only if $\lambda_i(A) + \bar{\lambda}_j(A) \neq 0, \forall i, j$. Hence, the solution can be attained by solving a system of equations and it is not required to solve the integral explicitly.[30]

Let us assume the situation of an uncontrollable system (A, B, C, D) with Controllability Gramian $P = \begin{pmatrix} P_1 & \\ & 0 \end{pmatrix}$, $P_1 > 0$. Inserting P into equation (3.1) and partitioning the system matrices A, B and C, leads to

$$0 = AP + PA^T + BB^T = \begin{pmatrix} A_{11}P_1 + P_1A_{11}^T + B_1B_1^T & P_1A_{21}^T + B_1B_2^T \\ A_{21}P_1 + B_2B_1^T & B_2B_2^T \end{pmatrix} \Rightarrow B_2 = 0, A_{21} = 0.$$

Consequently, the transfer function $G(s) = C(sI - A)^{-1}B + D = C_1(sI - A_{11})^{-1}B_1$.
The result is not surprising, but we can see how simple it is to reduce uncontrollable states of a system with the help of Lyapunov equations. Since this will be the essential idea of this chapter, we summarize it in the following Lemma.

Lemma 3.2

Let (A, B, C, D) be the state-space realization of a stable system with transfer function $G(s)$ and Controllability Gramian $P = \begin{pmatrix} P_1 & \\ & 0 \end{pmatrix}$, $P_1 > 0$. Partitioning the system according to P into

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \; B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \; C = \begin{pmatrix} C_1 & C_2 \end{pmatrix},$$

such that

$$A_{11}P_1 + P_1A_{11}^T = -B_1B_1^T,$$
$$A_{11}^TP_1 + P_1A_{11} = -C_1^TC_1,$$

results in a controllable state-space system (A_{11}, B_1, C_1, D), which is also a realization of $G(s)$.[30]

Remark: By switching the roles of P and $Q = \begin{pmatrix} Q_1 & \\ & 0 \end{pmatrix}$ the same can be done, which leads to an observable state-space realization (A_{11}, B_1, C_1, D) of $G(s)$.

Beside the range of P, the interesting property of the Controllability Gramians is the interpretation, that it measures how difficult it is to reach a certain state in a stable system.

Lemma 3.3 (*Control of Minimal Energy*)

Let x_0 be a reachable state, i.e. $x_0 \in rg(P)$. Among all controls u, steering the system from 0 to $x(0) = x_0$ over the interval $]-\infty, 0]$, $u(t) = B^Te^{-A^Tt}P^\sharp x_0$ minimizes the energy $E_c(u) = \int_{-\infty}^{0} \parallel u(\tau) \parallel^2 d\tau = x_0P^\sharp x_0$.[28]

The minimization is done over the interval $]-\infty, 0]$ and thus all possible controls steering the system to x_0 over an interval $[t_1, 0]$, with $t_1 < 0$, are considered.
Further, $P^\sharp$ denotes the Moore-Penrose pseudoinverse of P, which results from its Singular Value Decomposition by

$$P = U^T \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} U, \quad \Sigma = diag\left(s_1, \ldots, s_n\right), \quad s_1 \geq s_2 \geq \cdots s_n > 0, \quad U^{-1} = U^T$$

and

$$P^\sharp := U^T \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U.$$

Hence, depending on the eigenvalues of P, there are states requiring more energy compared to others to reach them. Now, the idea could be to transform the system with $x = U\xi$, resulting in the system $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) := (U^T AU, U^T B, CU, D)$

$$\begin{cases} \dot{\xi}(t) = \tilde{A}\xi(t) + \tilde{B}u(t), \\ y(t) = \tilde{C}\xi(t) + \tilde{D}u(t), \end{cases}$$

with diagonal Controllability Gramian $\tilde{P} = UPU^T = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$.

Then $\tilde{P}^\sharp = \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix}$ and $E_c(u) = \xi_0^T \tilde{P}^\sharp \xi_0 = \dfrac{1}{s_i}$ for $\xi_0 = e_i$. Therefore its maximum is attained in $\dfrac{1}{s_n}$. We may conclude to proceed as for uncontrollable states and suppress those states ξ_i, that correspond to small values in $\tilde{P}$. This would lead to a stable system (Lemma 3.1), but in many cases also to a very big $\mathcal{H}_\infty$-error between the original and the truncated system.[30]

The same consideration is valid for the observability of a state. The more influence a state has on the system output, the easier it is to observe. If $x(0) = x_0$ denotes the state to observe and we set the input $u \equiv 0$, then the output of the system is given by $y(t) = Ce^{At}x_0$. As before for u we consider the energy $E_o(y)$ of y,

$$E_o(y) = \int_0^\infty y(\tau)^T y(\tau) d\tau = \int_0^\infty x_0^T e^{A^T\tau} C^T C e^{A\tau} x_0 d\tau = x_0^T Q x_0$$

and note, the smaller $x_0^T Q x_0$, the harder it is to observe x_0.

Analogous to P, we could again diagonalize $Q = U^T \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} U$ and transform the system by $x = U^T\xi$ with new Observability Gramian $\tilde{Q} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$. Unfortunately, the neglection of states, which are hard to observe (small E_o), can lead to a big $\mathcal{H}_\infty$-error as well.[30]

Example 3.1 (High Truncation Error)
Let

$$G(s) := \begin{pmatrix} \dfrac{4}{s+1} & 0 \\ 0 & \dfrac{4}{s+1} \end{pmatrix}$$

be a system with a state-space representation

$$A := \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix}, \quad B := \begin{pmatrix} 2 & 0 \\ 0 & \frac{2}{\alpha} \end{pmatrix}, \quad C := \begin{pmatrix} 2 & 0 \\ 0 & 2\alpha \end{pmatrix}.$$

Then it is easy to see, that the Gramians are given by

$$P = \begin{pmatrix} 1 & \\ & \frac{1}{\alpha^2} \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & \\ & \alpha^2 \end{pmatrix}.$$

Consequently, the weaker the second state to reach/observe the easier it is to observe/reach. Truncating this state leads to a system

$$A := -2, \quad B := \begin{pmatrix} 2 & 0 \end{pmatrix}, \quad C := \begin{pmatrix} 2 \\ 0 \end{pmatrix},$$

with transfer function

$$G_1(s) := \begin{pmatrix} \frac{4}{s+1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Then by Theorem 1.1 it follows, that $\|G\|_\infty = \|G - G_1\|_\infty = 4$, which gives a relative error of 100 %.

Notice, $E_o(y) = \int_0^\infty y(\tau)^T y(\tau) d\tau$ is nothing else than the scalar product in the well-known Hilbert Space L^2 [30]. By Parsevals formula [21] we can describe this in the frequency domain as

$$\|y\|_2^2 = \int_0^\infty \bar{y}^T(\tau) y(\tau) d\tau = \frac{1}{2\pi} \int_{-\infty}^\infty \overline{Y}^T(i\omega) Y(i\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^\infty \|Y(i\omega)\|_2^2 d\omega,$$

where $Y(i\omega) = \mathscr{F}[y(t)]$ denotes the Fourier-Transformation of $y(t)$. In words the equations says, that the total energy of a signal in time-domain is equal to its total energy in frequency domain.
Using that $Y(i\omega) = G(i\omega)U(i\omega)$, we can conclude the following inequality

$$\|y\|_2^2 = \frac{1}{2\pi} \int_{-\infty}^\infty \|Y(i\omega)\|_2^2 d\omega \leq \frac{1}{2\pi} \int_{-\infty}^\infty \|G(i\omega)\|_2^2 \|U(i\omega)\|_2^2 d\omega \leq \|G\|_\infty^2 \|u\|_2^2.$$

Observe, by looking at the Fourier transformation of a sinusoid

$$\mathscr{F}(sin(\omega_0 t)) = \sqrt{2\pi} \frac{\delta(\omega - \omega_0) - \delta(\omega + \omega_0)}{2i}$$

and the well-known properties of $\delta(t)$ as in equation 2.4, we find an input to give equality at the maximizing frequency ω_0 of a SISO-system. In the MIMO-case such an input can also be constructed by considering the Singular Value Decomposition of $G(i\omega)$ as done in [30]. Thus

$$\|G\|_\infty = \sup_{\|u\|_2 \neq 0} \frac{\|y\|_2}{\|u\|_2}. \tag{3.6}$$

Soon, this fact will be very important to prove the error-bound of a truncated system.

3.1. Standard Balanced Truncation

The problem with the considered truncation approaches is obviously, that a state which is hard to observe does not need to be hard to reach and vice versa. Consequently the neglection of states, which are hard to observe and to reach at the same time, is the only feasible way. This leads us to the concept of *balancing* a system, which is based on the following theorem.

Theorem 3.1 (*Balancing Transformation Matrix*)

Let P and Q be two real positive semi-definite matrices. Then there exists a non-singular matrix T such that

$$P_b := T^{-1}PT^{-T} = diag\left(\Sigma, \Sigma_p, 0, 0\right), \quad Q_b := T^TQT = diag\left(\Sigma, 0, \Sigma_q, 0\right),$$

with diagonal Σ, Σ_p, $\Sigma_q > 0$ [30]

Proof: We only show the proof for the case $P, Q > 0$ and refer otherwise to [30].
Let P be decomposed by Singular Value Decomposition into

$$P = U\Sigma_P U^T,$$

and define

$$L := U\Sigma_P^{\frac{1}{2}}.$$

By another Singular Value Decomposition of L^TQL into

$$L^TQL = V\Sigma^2V^T,$$

and we define

$$T := LV\Sigma^{-\frac{1}{2}}.$$

Then it it straightforward to verify that

$$T^{-1}PT^{-1} = \Sigma^{\frac{1}{2}}V^TL^{-1}LL^TL^{-T}V\Sigma^{\frac{1}{2}} = \Sigma$$

and

$$T^TQT = \Sigma^{-\frac{1}{2}}V^TL^TQLV\Sigma^{\frac{1}{2}} = \Sigma$$

■

Let P and Q be the Gramians of a linear system (A, B, C, D) and T the corresponding matrix given in Theorem 3.1. Transforming the system by the equation $x = T\xi$ results in the new state-space representation

$$(A_b, B_b, C_b, D_b) := (T^{-1}AT, T^{-1}B, CT, D), \tag{3.7}$$

with the Gramians P_b and Q_b as defined in Theorem 3.1. The zero matrices in P_b and Q_b correspond to uncontrollable and unobservable states, which can be neglected without causing an error (Proposition 3.2). Thus the important information is collected in

$$\Sigma = diag\left(\sigma_1 I_{k_1}, \ldots, \sigma_N I_{k_N}\right), \ \sigma_1 > \sigma_2 > \cdots > \sigma_N > 0, \ k_i > 0, \ i = 1 \ldots N \tag{3.8}$$

and $\{\sigma_1, \ldots, \sigma_N\}$ are called the Hankel Singular Values of (A_b, B_b, C_b, D_b).

Observe,

$$P_b Q_b = T^{-1} P Q T = \begin{pmatrix} \Sigma_1^2 & \\ & 0 \end{pmatrix} \Rightarrow \{\sigma_1^2, \ldots, \sigma_N^2\} = \sigma(PQ) \setminus \{0\} \tag{3.9}$$

and the columns of T have to be eigenvectors of PQ.

Finally we are able to identify states, which are both hard to reach and to observe. A state-space realization (A_b, B_b, C_b, D_b), possessing the identical Controllability and Observability Gramians is then called a *Balanced Realization*.

The final step is to decide which states to truncate and to partition the system according to those.

If our balanced state-space system is given by

$$\begin{cases} \dot{\xi}(t) = A_b \xi(t) + B_b u(t), \\ y(t) = C_b \xi(t) + D_b u(t), \end{cases}$$

we know, states that correspond to small Hankel Singular Values have the least influence and cause the smallest error when truncated. Hence, the question that is left is, how big the error might become. An answer to this has been given e.g. in [24] and [30].

Theorem 3.2 (*Balanced Truncation and Error Bound*)

Suppose (A_b, B_b, C_b, D_b) is the balanced realization of an asymptotically stable system with transfer function $G(s)$, Gramians $\Sigma = diag\left(\Sigma_1, \Sigma_2\right)$,

$$\Sigma_1 = diag\left(\sigma_1 I_{k_1}, \ldots, \sigma_r I_{k_r}\right), \Sigma_2 = diag\left(\sigma_{r+1} I_{k_{r+1}}, \ldots, \sigma_N I_{k_N}\right)$$

and Hankel Singular Values $\sigma_1 > \cdots > \sigma_r > \sigma_{r+1} > \cdots > \sigma_N > 0$.

Partitioning the system matrices A_b, B_b and C_b accordingly to Σ_1 results in a truncated system $(A_r, B_r, C_r, D_r) := (A_{11}, B_1, C_1, D)$ with transfer function $G_r(s)$ which is balanced, controllable, observable and asymptotically stable.

Moreover, it holds for the $\mathcal{H}_\infty$-error

$$\|G(s) - G_r(s)\|_\infty \leq 2 \sum_{i=r+1}^{N} \sigma_i. \tag{3.10}$$

Proof: Note, since A is asymptotically stable and $\Sigma > 0$, we assume implicitly, that (A, B) is controllable and (A, C) observable. According to Lemma 3.2, this is not a restriction. Partitioning the system into

$$\begin{pmatrix}\dot{\xi}_1\\ \dot{\xi}_2\end{pmatrix} = \begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + \begin{pmatrix}B_1\\ B_2\end{pmatrix}u,$$
$$y = \begin{pmatrix}C_1 & C_2\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + Du,$$

gives the reduced system

$$\dot{\xi}_r = A_{11}\xi_r + B_1 u,$$
$$y_r = C_1\xi_r + Du,$$

and the following Lyapunov equations

$$A_{11}\Sigma_1 + \Sigma_1 A_{11}^T = -B_1 B_1^T, \tag{3.11}$$
$$A_{11}^T\Sigma_1 + \Sigma_1 A_{11} = -C_1^T C_1, \tag{3.12}$$
$$A_{21}\Sigma_1 + \Sigma_2 A_{12}^T = -B_2 B_1^T, \tag{3.13}$$
$$A_{12}^T\Sigma_1 + \Sigma_2 A_{21} = -C_2^T C_1. \tag{3.14}$$

Since $B_1 B_1^T \geq 0$ it follows by Lemma 3.1 that A_{11} is a stable matrix. Now let us assume there exists a purely imaginary eigenvalue $i\omega$ of A_{11} with an eigenbasis collected in the matrix V, i.e.

$$A_{11} = i\omega V \quad \text{and} \quad \overline{V}^T A_{11}^T = -i\omega\overline{V}^T.$$

By multiplying equation (3.12) from the left with $\overline{V}^T$ and from the right with V we get

$$i\omega\overline{V}^T\Sigma_1 V - i\omega\overline{V}^T\Sigma_1 V = -\overline{V}^T C_1^T C_1 V \;\Leftrightarrow\; \overline{V}^T C_1^T C_1 V = 0$$

and hence

$$C_1 V = 0. \tag{3.15}$$

Thus, multiplying (3.12) with V only from the right side yields

$$A_{11}^T\Sigma_1 V = -i\omega\Sigma_1 V \quad \text{and} \quad \overline{V}^T\Sigma_1 A_{11} = i\omega\overline{V}^T\Sigma_1. \tag{3.16}$$

Using these equations after multiplying equation (3.11) from the left with $\overline{V}^T\Sigma_1$ and from the right with $\Sigma_1 V$ leads to

$$i\omega\overline{V}^T\Sigma_1^2 V - i\omega\overline{V}^T\Sigma_1^2 V = -\overline{V}^T\Sigma_1 B_1 B_1^T\Sigma_1 V \;\Leftrightarrow\; \overline{V}^T\Sigma_1 B_1 B_1^T\Sigma_1 V = 0$$

and therefore

$$B_1^T\Sigma_1 V = 0. \tag{3.17}$$

As before this yields

$$A_{11}\Sigma_1^2 V = i\omega\Sigma_1^2 V \tag{3.18}$$

by multiplying equation (3.11) only from the right side with $\Sigma_1 V$. Obviously, this means $\Sigma_1^2 V$ consists of eigenvectors of A_{11} to the eigenvalue $i\omega$, which is why there must exist a matrix M such that $\Sigma_1^2 V$ can be expressed as

$$\Sigma_1^2 V = VM.$$

Furthermore, $rg(V)$ is an Σ_1^2-invariant subspace and thus for every eigenvalue $\lambda \in \sigma(M)$ with eigenvector w_λ it holds

$$\Sigma_1^2 V w_\lambda = \lambda V w_\lambda,$$

which shows, that $\lambda \in \sigma(\Sigma_1^2)$ with an eigenvector in $rg(V)$. Hence, we can choose V such

$$M = \hat{\Sigma}_1^2$$

with a diagonal $\hat{\Sigma}_1$, whose diagonal entries are subset of those belonging to Σ_1. Further, by multiplication of equation (3.13) with $\Sigma_1 V$ from the right we get from equation (3.17), that

$$A_{21}\Sigma_1^2 V + \Sigma_2 A_{12}^T \Sigma_1 V = 0.$$

Similar, if we multiply equation (3.14) with Σ_2 from the left and V from the right, we can conclude by equation (3.15), that

$$\Sigma_2^2 A_{21} V + \Sigma_2 A_{12}^T \Sigma_1 V = 0. \tag{3.19}$$

Together we receive

$$A_{21}\Sigma_1^2 V = \Sigma_2^2 A_{21} V \;\Leftrightarrow\; A_{21} V \hat{\Sigma}_1^2 = \Sigma_2^2 A_{21} V \;\Leftrightarrow\; (\Sigma_2^2)^{-1} A_{21} V \hat{\Sigma}_1^2$$

After assumption Σ_1 and Σ_2 do not have common diagonal entries and hence the same holds for $\hat{\Sigma}_1^2$ and Σ_2^2. This leaves us with the conclusion

$$A_{21} V = 0,$$

which allows us to write

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} V \\ 0 \end{pmatrix} = i\omega \begin{pmatrix} V \\ 0 \end{pmatrix}.$$

But this is a contradiction, since we have assumed A to be asymptotically stable and therefore A_{11} cannot possesses a purely imaginary eigenvalue. Consequently, the reduced system is asymptotically stable with the Gramians Σ_1, which also means, that (A_{11}, B_1) is controllable and (A_{11}, C_1) observable.

Now, we take care of the error bound estimation. For this purpose we rewrite (3.1) as

$$A^T\Sigma^{-1} + \Sigma^{-1}A = -\Sigma^{-1}BB^T\Sigma^{-1}.$$

Since

$$\begin{pmatrix} \Sigma^{-1}BB^T\Sigma^{-1} & -\Sigma^{-1}B \\ -B^T\Sigma^{-1} & I \end{pmatrix} = \begin{pmatrix} \Sigma^{-1}B \\ -I \end{pmatrix} \begin{pmatrix} B^T\Sigma^{-1} & -I \end{pmatrix} \geq 0$$

we get

$$\begin{pmatrix} A^T\Sigma^{-1} + \Sigma^{-1}A & \Sigma^{-1}B \\ B^T\Sigma^{-1} & -I \end{pmatrix} \leq 0$$

or equivalently

$$\begin{pmatrix} A^T\Sigma^{-1} + \Sigma^{-1}A & \Sigma^{-1}B \\ B^T\Sigma^{-1} & 0 \end{pmatrix} \leq \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}.$$

Rewriting the left side of this inequality results in

$$\begin{pmatrix} A & B \\ I & 0 \end{pmatrix}^T \begin{pmatrix} 0 & \Sigma^{-1} \\ \Sigma^{-1} & 0 \end{pmatrix} \begin{pmatrix} A & B \\ I & 0 \end{pmatrix} \leq \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}. \tag{3.20}$$

Let us define

$$z(t) := A_{21}\xi_r(t) + B_2u(t),$$

and assume that $\xi(0) = 0$ and $\xi_r(0) = 0$. Multiplying (3.20) from the right with

$$\begin{pmatrix} \xi_1 + \xi_r \\ \xi_2 \\ 2u \end{pmatrix}$$

and from the left with its transpose, gives then

$$\begin{pmatrix} \dot{\xi}_1 + \dot{\xi}_r \\ \dot{\xi}_2 + z \\ \xi_1 + \xi_r \\ \xi_2 \end{pmatrix}^T \begin{pmatrix} 0 & 0 & \Sigma_1^{-1} & 0 \\ 0 & 0 & 0 & \Sigma_2^{-2} \\ \Sigma_1^{-1} & 0 & 0 & 0 \\ 0 & \Sigma_2^{-1} & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{\xi}_1 + \dot{\xi}_r \\ \dot{\xi}_2 + z \\ \xi_1 + \xi_r \\ \xi_2 \end{pmatrix} \leq 4u^Tu$$

or equivalently

$$2(\dot{\xi}_1 + \dot{\xi}_r)^T\Sigma_1^{-1}(\xi_1 + \xi_r) + 2(\dot{\xi}_2 + z)^T\Sigma_2^{-1}\xi_2 \leq 4u^Tu \tag{3.21}$$

By applying partial integration we get

$$2\int_0^T (\dot{\xi}_1(t) + \dot{\xi}_r(t))^T\Sigma_1^{-1}(\xi_1(t) + \xi_r(t))dt = (\xi_1(T) + \xi_r(T))^T\Sigma_1^{-1}(\xi_1(T) + \xi_r(T))$$

and

$$2\int_0^T \dot{\xi}_2^T(t)\Sigma_1^{-1}\xi_2(t)dt = \xi_2^T(T)\Sigma_1^{-1}\xi_2(T).$$

Consequently, integrating over the inequality (3.21) gives

$$\begin{pmatrix} \xi_1(T) + \xi_r(T) \\ \xi_2(T) \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} \xi_1(T) + \xi_r(T) \\ \xi_2(T) \end{pmatrix} + 2\int_0^T z^T(t)\Sigma_2^{-1}\xi_2(t)dt \leq 4\int_0^T u^T(t)u(t)dt$$

and by the positive definiteness of Σ^{-1} we get

$$2\int_0^\infty z^T(t)\Sigma_2^{-1}\xi_2(t)dt \leq 4\|u\|_2^2. \tag{3.22}$$

A similar consideration can be done for (3.3). Rewriting the equation as

$$\begin{pmatrix} A \\ I \end{pmatrix}^T \begin{pmatrix} 0 & \Sigma \\ \Sigma & 0 \end{pmatrix} \begin{pmatrix} A \\ I \end{pmatrix} = -C^TC$$

and multiplying it from the right with

$$\begin{pmatrix} \xi_1 - \xi_r \\ \xi_2 \end{pmatrix}$$

and from the left with its transposed, leads to

$$\begin{pmatrix} \dot{\xi}_1 - \dot{\xi}_r \\ \dot{\xi}_2 - z \\ \xi_1 - \xi_r \\ \xi_2 \end{pmatrix}^T \begin{pmatrix} 0 & 0 & \Sigma_1 & 0 \\ 0 & 0 & 0 & \Sigma_2 \\ \Sigma_1 & 0 & 0 & 0 \\ 0 & \Sigma_2 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{\xi}_1 - \dot{\xi}_r \\ \dot{\xi}_2 - z \\ \xi_1 - \xi_r \\ \xi_2 \end{pmatrix} = (y - y_r)^T(y - y_r).$$

As before partial integration yields

$$\begin{pmatrix}\xi_1(T)-\xi_r(T)\\ \xi_2(T)\end{pmatrix}^T \Sigma \begin{pmatrix}\xi_1(T)-\xi_r(T)\\ \xi_2(T)\end{pmatrix} - 2\int_0^T z^T(t)\Sigma_2\xi_2(t)dt \leq -\int_0^T (y-y_r)^T(t)(y-y_r)(t)dt$$

and consequently

$$-2\int_0^\infty z^T(t)\Sigma_2\xi_2(t)dt + \|y-y_r\|_2^2 \leq 0. \tag{3.23}$$

Suppose now, we perform the truncation successively for each Hankel Singular Value, starting with the states belonging to σ_N and calling the truncated system G_{N-1}. Then we can assume $\Sigma_2 = \sigma_N I$ and we get by multiplying (3.22) with σ_N^2 and adding it to (3.23), that

$$\|y-y_{N-1}\|_2 \leq 2\sigma_N\|u\|_2,$$

which is according to (3.6) equivalent to

$$\|G-G_{N-1}\|_\infty \leq 2\sigma_N.$$

By expanding $G-G_r$ to a telescope sum and using the triangle inequality, we get

$$\|G-G_r\|_\infty = \|G-G_{N-1}+G_{N-1}+\cdots+G_{r+1}-G_r\|_\infty \leq 2\sum_{i=r+1}^{N}\sigma_i.$$

■

Note, the error bound does only depend on the Hankel Singular Values given by (3.9). By contrast, the multiplicity k_i of σ_i is not important. But if the system possesses Singular Values $\sigma_i \approx \sigma_j$, we cannot assume, that the influence on the upper error bound will be close to $2\sigma_i$. Instead we have to expect $2\sigma_i + 2\sigma_j \approx 4\sigma_i$.

Example 3.2 (Close Hankel Singular Values [30])
Let us consider the transfer function

$$G(s) = \sum_{i=1}^{n}\frac{b_i}{s+a_i}$$

with $a_i, b_i > 0$. This is obviously a positive system with $\|G(s)\|_\infty = \sum_{i=1}^n \frac{b_i}{a_i}$ and a realization

$$A := diag\left(-a_1,\ldots,-a_n\right) \quad \text{and} \quad B^T = C = \begin{pmatrix}\sqrt{b_1} & \ldots & \sqrt{b_n}\end{pmatrix}.$$

Then by (3.1) and (3.3), it follows

$$AP + PA^T = A^TQ + QA = BB^T = C^TC,$$

which leads to P and Q with entries

$$p_{ij} = q_{ij} = \left(\frac{\sqrt{b_ib_j}}{a_i+a_j}\right).$$

Thus $\sigma_i = \lambda_i(P) = \lambda_i(Q)$ and

$$\sum_{i=1}^{n} \sigma_i = tr(P) = \sum_{i=1}^{n} \frac{b_i}{2a_i} = \frac{1}{2}\|G\|_\infty.$$

By choosing $a_i = b_i = \alpha^{2i}$ we attain $P = Q \to \frac{1}{2}I$ as $\alpha \to \infty$ and therefore $\sigma_i \to \frac{1}{2}$. This shows the tightness of the error bound.

Observe, scaling a system $G(s)$ by k, i.e. $\tilde{G}(s) := kG(s)$, also scales the reduced-order system and the Hankel Singular Values by the factor k. Consequently for a for very small/large k, we attain a very small/large error. In order to perform a fair comparison we need to consider the relative error

$$\frac{\|G - G_r\|_\infty}{\|G\|_\infty} \leq \frac{2}{\|G\|_\infty} \sum_{i=r+1}^{N} \sigma_i = 2 \sum_{i=r+1}^{N} \tilde{\sigma}_i, \text{ with } \tilde{\sigma}_i = \frac{\sigma_1}{\|G\|_\infty}.$$

For asymptotic stability it might be important to truncate all states, that correspond to the same Hankel Singular Value.

Example 3.3 (Unstable Balanced Truncation)
The system (A, B, C) given by

$$A := \begin{pmatrix} -2 & 0 & 0 \\ 0 & 0 & 0.5 \\ 0 & -0.5 & -2 \end{pmatrix}, \quad C = B = diag\left(2, 0, 2\right),$$

is clearly a balanced asymptotically stable system with $P = Q = diag\left(2, 1, 1\right)$. Thus, if we truncated only the third state, we would obtain an unstable system.

3.2. Balanced Truncation Algorithm

Let $T = \begin{pmatrix} T_1 & T_2 \end{pmatrix}$ be the balancing matrix of Theorem 3.1, partitioned according to Σ_1 of Theorem 3.2, and $T^{-1} = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$. Then the balanced system is given by

$$\begin{aligned} A_b &= T^{-1}AT = \begin{pmatrix} S_1AT_1 & S_1AT_2 \\ S_2AT_1 & S_2AT_2 \end{pmatrix}, \\ B_b &= T^{-1}B = \begin{pmatrix} S_1B \\ S_2B \end{pmatrix}, \\ C_b &= CT = \begin{pmatrix} CT_1 & CT_2 \end{pmatrix}, \\ D_b &= D. \end{aligned}$$

Thus the reduced system can be written as

$$A_r = S_1AT_1, \quad B_r = S_1B, \quad C_r = CT_1, \quad D_r = D \tag{3.24}$$

and we observe, instead of balancing the whole system we only need to find T_1 and S_1 with $S_1T_1 = I$. For $P, Q > 0$ we saw in the proof to Theorem 3.1 how to get them in a systematic way. In case of $P, Q \geq 0$, we can proceed in almost the same manner to achieve a balanced truncated system, without balancing the original one.

As before, P can be decomposed into

$$P = U \begin{pmatrix} \Sigma_P & 0 \\ 0 & 0 \end{pmatrix} U^T,$$

and we define

$$L := U \begin{pmatrix} \Sigma_P^{\frac{1}{2}} & 0 \\ 0 & 0 \end{pmatrix}.$$

The difference compared to $P > 0$ is obviously, that L does not have full rank. By considering the Singular Value Decomposition of Q

$$Q = U_Q \begin{pmatrix} \Sigma_Q & 0 \\ 0 & 0 \end{pmatrix} U_Q^T$$

it is clear, that $\sigma(L^TQL) = \sigma(PQ)$. Thus, Singular Value Decomposition yields

$$L^TQL = V \begin{pmatrix} \Sigma_1^2 & 0 \\ 0 & 0 \end{pmatrix} V^T.$$

Let $V = \begin{pmatrix} v_1, \ldots, v_{N_\sigma}, \ldots, v_n \end{pmatrix}$, with $N_\sigma := k_1 + \cdots + k_N$ corresponding to the notations of Theorem 3.2. Then we can define the matrices T_1 and S_1 as follows

$$T_1 = L \begin{pmatrix} v_1, \cdots, v_{r_\sigma} \end{pmatrix} diag \begin{pmatrix} \sigma_1 I_{k_1}, \ldots, \sigma_r I_{k_r} \end{pmatrix}^{-\frac{1}{2}},$$

$$S_1 = diag \begin{pmatrix} \sigma_1 I_{k_1}, \ldots, \sigma_r I_{k_r} \end{pmatrix}^{\frac{1}{2}} \begin{pmatrix} u_1^T \\ \vdots \\ u_{r_\sigma}^T \end{pmatrix} L^\sharp = T_1^\sharp$$

with $r_\sigma := k_1 + \cdots + k_r$. Notice, if we choose $r_\sigma = N_\sigma$, we get

$$S_1 P S_1^T = \Sigma_1^{\frac{1}{2}} V^T L^\sharp L L^T L^{\sharp^T} V \Sigma_1^{\frac{1}{2}} = \Sigma_1,$$

$$T_1^T Q T_1 = \Sigma_1^{-\frac{1}{2}} V^T L^T Q L V \Sigma_1^{-\frac{1}{2}} = \Sigma_1,$$

and end up with a balanced realization, that has truncated all uncontrollable and unobservable states, i.e. a minimal realization.

3.3. Singular Perturbation Balanced Truncation

A property of Balanced Truncation, as we have introduced it now, is that

$$\lim_{s \to \infty} G_r(s) = \lim_{s \to \infty} G(s).$$

This is easy to see since $D_r = D$. With a bit more care it is possible to get $G_r(0) = G(0)$ instead, meaning that the stationary property is preserved.

Corollary 3.1

Suppose we are in the position of Theorem 3.2. The system of the truncated states (A_{22}, B_2, C_2, D) is also balanced and asymptotically stable.

Proof: In the same way as for A_{11}. ■

Let us partition the balanced system, which has already truncated all the uncontrollable and observable states, as follows

$$\begin{pmatrix}\dot{\xi}_1\\ \dot{\xi}_2\end{pmatrix} = \begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + \begin{pmatrix}B_1\\ B_2\end{pmatrix}u, \tag{3.25}$$
$$y = \begin{pmatrix}C_1 & C_2\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix} + Du.$$

We know the system is in stationary state if and only if there is no change of ξ_1 and ξ_2 after some time or equivalently, when $\dot{\xi}_1 = 0$ and $\dot{\xi}_2 = 0$. Thus replacing ξ_2 by its static relationship will preserve the stationary property of $G(s)$. Setting $\dot{\xi}_2 = 0$ gives

$$0 = A_{21}\xi_1 + A_{22}\xi_2 + B_2u \;\Leftrightarrow\; \xi_2 = -A_{22}^{-1}(A_{21}\xi_1 + B_2u)$$

and is well defined by Corollary 3.1. Inserting this expression of ξ_2 in (3.25) results in a reduced system

$$\begin{aligned}\dot{\xi}_1 &= A_r\xi_1 + B_ru\\ y &= C_r\xi_1 + D_ru\end{aligned}$$

with

$$\begin{aligned}A_r &= A_{11} - A_{12}A_{22}^{-1}A_{21}, \qquad B_r = B_1 - A_{12}A_{22}^{-1}B_2,\\ C_r &= C_1 - C_2A_{22}^{-1}A_{21}, \qquad D_r = D - C_2A_{22}^{-1}B_2.\end{aligned} \tag{3.26}$$

This variation of Balanced Truncation, preserving the stationary property, is called *Singular Perturbation Balanced Truncation* [10] and leaves Theorem 3.2 unchanged. For a better distinction we refer in the following to our first variant as *Standard Balanced Truncation*.

In contrary to *Standard Balanced Truncation*, where it is sufficient to calculate the terms in (3.24), *Singular Perturbation Balanced Truncation* requires to calculate all the terms of the balanced realization. Since Balanced Truncation is independent of the state-space representation of the original system, we just need to choose $r_\sigma = N_\sigma$ in the forgone section.

Remark: Generally, neither of both Balanced Truncation methods preserve the physically meaning of the state.

3.4. Balanced Truncation of Positive Systems

Let us turn our focus back to positive systems. In general Balanced Truncation applied to positive systems does not result in positive truncated systems, as we can see in the following example.

Example 3.4 (Nonpositive Reduced System)
Consider the following positive system

$$A := \begin{pmatrix} -3 & 1 & 0 \\ 0 & -4 & 1 \\ 0 & 0 & -3 \end{pmatrix}, \quad B := \begin{pmatrix} 3 \\ 0 \\ 2 \end{pmatrix}, \quad C := \begin{pmatrix} 5 & 4 & 1 \end{pmatrix}.$$

By reduction to second order we obtain the system

$$A_2 := \begin{pmatrix} -2.57 & 0.34 \\ -0.34 & -2.82 \end{pmatrix}, \quad B_2 := \begin{pmatrix} 4.13 \\ 0.27 \end{pmatrix}, \quad C_2 := \begin{pmatrix} 4.13 & -0.27 \end{pmatrix},$$

which has poles in $-2.70 \pm 0.31i$. According to Lemma 1.2, this cannot be a positive system.

Remark: We already know a minimal representation of a positive system does not have to have the same amount of states as its positive realization. In turn, neglecting the states, that correspond to the zero matrices in P_b and Q_b can have the effect of destroying the positive realization, though the procedure does not cause an error in the transfer function.

The reason why we could not consider the easier case of n = 2 in Example 3.4 is a consequence of the next theorem.

Theorem 3.3 (***Positive First Order Balanced Truncation***)
Let (A_1, B_1, C_1, D_1) be the reduced first order system attained by Standard Balanced Truncation of a positive system (A, B, C, D). Then (A_1, B_1, C_1, D_1) is always positive and asymptotically stable with first order positive realization $(A_1, |B_1|, |C_1|, D_1)$.

Proof: Let P and Q be the Gramians to a positive system (A, B, C, D), explicitly given by the equations (3.2) and (3.4).
By implication (1.3) we know

$$e^{At} \geqq 0 \ \forall t \geq 0 \ \Rightarrow \ e^{At}B, Ce^{At} \geqq 0 \ \forall t \geq 0.$$

Hence P and Q are nonnegative matrices and we conclude $PQ \geqq 0$. In (3.9) we noticed the Hankel Singular Values of a system are eigenvalues of PQ and the columns of T given by Theorem 3.1 correspond to its eigenvectors.
We first consider the case if σ_1 is a unique Hankel Singular Value. By Theorem 1.5 there exists a nonnegative right-eigenvector v_1 to the largest eigenvalue σ_1, i.e.

$$PQv_1 = \sigma_1 v_1 \quad \text{with} \quad T = \begin{pmatrix} v_1, & \dots & , v_n \end{pmatrix}.$$

If we denote the rows of T^{-1} by w_i^T, i.e. $T^{-1} = \begin{pmatrix} w_1^T \\ \vdots \\ w_n^T \end{pmatrix}$ and recall that

$$Q_b = T^T Q T = \begin{pmatrix} \Sigma & & \\ & 0 & \\ & & \Sigma_q \\ & & & 0 \end{pmatrix} \quad \Leftrightarrow \quad QT = T^{-T} \begin{pmatrix} \Sigma & & \\ & 0 & \\ & & \Sigma_q \\ & & & 0 \end{pmatrix},$$

we conclude

$$0 \leqq Qv_1 = w_1\sigma_1 \quad \Rightarrow \quad w_1 \geqq 0,$$

and w_1 is a nonnegative left-eigenvector of PQ to the eigenvalue σ_1.

Hence by Theorem 3.2,

$$A_1 = w_1^T A v_1 < 0, \quad B_1 = w_1^T B \geqq 0, \quad C_1 = Cv_1 \geqq 0, \quad D_1 = D \geqq 0.$$

In case of a σ_1 with multiplicity $k_1 > 1$, we have seen in Example 3.3, that $A_1 = 0$ is possible. On the other hand, since the k_1-th order reduced system, which belongs to all σ_1, is according to Theorem 3.2 asymptotically stable, there must exist at least one asymptotically stable first approximation. We want to show now, that in this case positivity is still preserved.

We start with the case, that PQ is irreducible. Since $\sigma(PQ)$ contains a multiple σ_1^2, it follows by Theorem 1.6, that $tr(PQ) = 0$. This is obviously a contradiction, which is why PQ can only be reducible.

If PQ is reducible, then it follows by Lemma 1.5, that there exist k_1 linear independent nonnegative eigenvectors to the eigenvalue σ_1^2. Hence, as in the case of a unique σ_1 we can obtain an asymptotically stable first order approximation with $B_1, C_1, D_1 \geqq 0$.

In all the cases Theorem 1.3 concludes the proof. ■

Remark: Theorem 3.3 is in general not transferable to Singular Perturbation Balanced Truncation. For example, if we truncate the system in Example 3.4 to first order, then we result in a system $(A_1, B_1, C_1, D_1) = (-2.61, 4.16, 4.16, 0.03)$.

Observe, a reduced order system resulting of Balanced Truncation is independent of the state-space representation of the original system. Hence, Theorem 3.3 gives a new way of testing, whether it is possible that a system possess a positive realization or not.

Corollary 3.2

If for the reduced first order system $G_1(s) = \dfrac{1}{(s-\alpha_1)} M$ of a transfer function $G(s)$ does not hold $M \geqq 0$, then $G(s)$ is not a positive system.

Remark: By reducing the truncated system of Example 3.4 to first order, it is clear that this can only be a necessary condition.

3.4.1. Balanced Truncation with respect to Lyapunov Inequalities

A first order approximation is not always sufficient, of course. In Chapter 6 we will give an extension of Theorem 3.3 to higher orders in case of a SISO-system.
Before doing so we want to investigate some methods that have already dealt with higher order approximations. Until today, there are to the author's knowledge three methods [7], [14] and [22] concerning model order reduction of positive systems. The method in [22] is based on Balanced Truncation with respect to *Lyapunov Inequalities* and will be discussed in this section.

The idea is, instead of considering the Lyapunov equations (3.1) and (3.3), to regard

$$\begin{aligned} AP + PA^T + BB^T &\leq 0, \\ A^TQ + QA + C^TC &\leq 0, \end{aligned} \tag{3.27}$$

with $P, Q \geq 0$.

In the same way as for the Gramians we can apply Theorem 3.1 to any solution pair (P, Q) satisfying the Lyapunov Inequalities (3.27) and obtain a balanced system

$$(A_b, B_b, C_b, D_b) := (T^{-1}AT, T^{-1}B, CT, D)$$

with

$$P_b = T^{-1}PT^{-T} = diag\left(\Sigma, \Sigma_p, 0, 0\right) \quad \text{and} \quad Q_b = T^TQT = diag\left(\Sigma, 0, \Sigma_q, 0\right), \tag{3.28}$$

fulfilling

$$\begin{aligned} A_bP_b + P_bA_b^T + B_bB_b^T &\leq 0, \\ A_b^TQ_b + Q_bA_b + C_b^TC_b &\leq 0, \end{aligned} \tag{3.29}$$

and

$$\Sigma = diag\left(\sigma_1 I_{k_1}, \ldots, \sigma_N I_{k_N}\right) \text{ for some } \sigma_1 > \sigma_2 > \cdots > \sigma_N > 0.$$

In this case we call $\{\sigma_1, \ldots, \sigma_N\}$ the *Generalized Hankel Singular Values*, because for the truncation they will play the same role as the Hankel Singular Values.

In fact, it is readily seen, that the proof in Theorem 3.2 does not change for Lyapunov Inequalities and thus performing truncation on (A_b, B_b, C_b, D_b) with respect to the *Generalized Hankel Singular Values* leaves the statements of Theorem 3.2 and Corollary 3.1 almost completely invariant: the difference is, that P and Q do not necessarily represent the controllable and observable subspaces any more and hence we cannot assume the minimality of the truncated system. Since this is obviously a generalization of Balanced Truncation, we refer to it as *Generalized Balanced Truncation*.

By the consideration of Lyapunov Inequalities we gain a degree of freedom that allows us to use P and Q to shape a certain balancing transformation matrix T. In order to guarantee a positive truncated system, the first idea would be to get a matrix T such that

$$B_b = T^{-1}B \geqq 0 \quad \text{and} \quad C_b = CT \geqq 0.$$

The easiest way to fulfil these requirements is to attain a matrix $T \geqq 0$ with $T^{-1} \geqq 0$. This brings us to the definition of a certain class of matrices, called *Monomial Matrices* [2].

Definition 3.1 (*Monomial Matrix*)

Let A be a matrix that can be expressed by the matrix product $A = \pi D$, where D is diagonal and invertible and π a permutation matrix. The matrix A is then called monomial or generalized permutation matrix.

In the following Lemma we will see why this class of matrices is so important for our idea.

Lemma 3.4

If A is a nonnegative matrix, then its inverse A^{-1} is nonnegative if and only if A is monomial.[2]

Proof: ►Sufficiency: Clear by $A^{-1} = D^{-1}\pi$.
►Necessity: If $A = (a_1, \ldots, a_n)$ and $A^{-1} = (s_1, \ldots, s_n)^T$, then it has to hold

$$s_i^T a_j = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Since $a_i, s_i \geqq 0 \; \forall i$, it must hold by linear independence of $\{s_i, i = 1, \ldots, n\}$, that a_i contains at least $(n-1)$ zeros. Consequently A is a monomial matrix. ■

Hence our problem reduces to find P and Q fulfilling (3.27), such that the eigenvectors of PQ can be represented by a permutation matrix. This is the case if and only if PQ is diagonal. Fortunately we know by Theorem 1.2 (v), if A is an asymptotically stable -M-matrix, there exist diagonal matrices $P, Q > 0$ such that

$$\begin{aligned} AP + PA^T < 0 \quad &\Leftrightarrow \quad \exists \lambda_p > 0 : AP + PA^T \leq -\lambda_p I, \\ A^T Q + QA < 0 \quad &\Leftrightarrow \quad \exists \lambda_q > 0 : A^T Q + QA \leq -\lambda_q I. \end{aligned}$$

Scaling P and Q to $\tilde{P} := \dfrac{P \parallel B \parallel_2^2}{\lambda_p}$ and $\tilde{Q} := \dfrac{Q \parallel C \parallel_2^2}{\lambda_q}$ provides us with feasible diagonal solutions to the Lyapunov Inequalities in (3.27).

Obviously $\tilde{P}\tilde{Q}$ is diagonal and applying Theorem 3.1 leads then to a monomial transformation matrix $T = \pi\bar{T}$ with diagonal $\bar{T} > 0$ and permutation matrix π. As for normal Balanced Truncation we get a balanced system (A_b, B_b, C_b, D_b) as defined in (3.7) with

$$T^{-1}P_bQ_bT = \begin{pmatrix} \Sigma_1 & \\ & 0 \end{pmatrix} \quad \Rightarrow \quad \bar{T} = \begin{pmatrix} \bar{T}_1 & \\ & I \end{pmatrix}$$

and because of the permutation we can assume w.l.o.g. that the Generalized Hankel Singular Values are in descending order.

Multiplying a $-Z$-matrix with a positive diagonal matrix preserves the sign of each matrix element, thus $\bar{A} := \bar{T}^{-1}A\bar{T}$ is $-Z$-matrix. It is straightforward to see, that this also holds for $A_b = \pi^T A\pi$ and consequently (A_b, B_b, C_b, D_b) is a positive system by Theorem 1.3. *Standard Truncation* of such a system yields an approximation (A_r, B_r, C_r, D_r), which is again positive, because $B_r, C_r, D_r \geqq 0$ and A_r is a $-M$-matrix as the principle minor of a $-M$-matrix.

The same conclusions can be done for the *Singular Perturbation Truncation*. Since A_{21},$A_{21} \geqq 0$ and $-A_{22}^{-1} \geqq 0$ by Theorem 1.2 (iv), we see immediately B_r,C_r,$D_r \geqq 0$ as defined in (3.26). The $-M$-property of A_r can be seen by noticing, that

$$A_r = A_{11} - A_{12}A_{22}^{-1}A_{21}$$

is the Schur complement of a $-M$-matrix: it is well-known [30] that

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_r^{-1} & -A_r^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}A_r^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}A_r^{-1}A_{12}A_{22}^{-1} \end{pmatrix} \leqq 0 \tag{3.30}$$

and thus $-A_r^{-1} \geqq 0$. Since $-A_{12}A_{22}^{-1}A_{21} \geqq 0$ it follows that A_r is a $-Z$-matrix and therefore by Theorem 1.2 (iv), A_r must be a nonsingular $-M$-matrix.
For the reduced system we can summarize the following result.

Theorem 3.4

Let (A_b, B_b, C_b, D_b) be the balanced realization of an asymptotically stable positive system $G(s)$ with respect to the Lyapunov Inequalities given in (3.29) and diagonal solutions P_b, $Q_b \geq 0$ as in (3.28). Then regardless of whether applying Standard Balanced Truncation or Singular Perturbation Balanced Truncation the reduced-order system (A_r, B_r, C_r, D_r) is again asymptotically stable and positive.
For the error-bound it holds the same as in Theorem 3.2.

Note, for the computation of the reduced system, it is not necessary to compute the balanced realization itself. Let us assume without loss of generality $\bar{T} = \bar{T}_1$ and define

$$\bar{A} := \pi^T A\pi, \quad \bar{B} := \pi^T B, \quad \bar{C} := C\pi, \quad \bar{D} := D.$$

By splitting $\bar{T}_1$, $\bar{A}$, $\bar{B}$ and $\bar{C}$ according to the truncation candidates into

$$\bar{T}_1 = \begin{pmatrix} \bar{T}_{11} & \\ & \bar{T}_{12} \end{pmatrix}, \bar{A} = \begin{pmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} \bar{B}_1 \\ \bar{B}_2 \end{pmatrix}, \quad \bar{C} = \begin{pmatrix} \bar{C}_1 & \bar{C}_2 \end{pmatrix},$$

we can rewrite the resulting balanced system as

$$\begin{pmatrix} \dot{\xi}_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} \bar{T}_{11}^{-1}\bar{A}_{11}\bar{T}_{11} & \bar{T}_{11}^{-1}\bar{A}_{12}\bar{T}_{12} \\ \bar{T}_{12}^{-1}\bar{A}_{21}\bar{T}_{11} & \bar{T}_{12}^{-1}\bar{A}_{22}\bar{T}_{12} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \bar{T}_{11}^{-1}\bar{B}_1 \\ \bar{T}_{12}^{-1}\bar{B}_2 \end{pmatrix} u,$$
$$y = \begin{pmatrix} \bar{C}_1\bar{T}_{11} & \bar{C}_2\bar{T}_{12} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \bar{D}u.$$

We observe, the truncated system $(\bar{T}_{11}^{-1}\bar{A}_{11}\bar{T}_{11}, \bar{T}_{11}^{-1}\bar{B}_1, \bar{C}_1\bar{T}_{11}, \bar{D})$ results by the transformation $x = \bar{T}_{11}^{-1}\xi$ of $(\bar{A}_{11}, \bar{B}_1, \bar{C}_1, \bar{D})$, which is a positive system itself. Thus, for the computation of the reduced model, it is sufficient to determine the permutation matrix π, which can e.g. be done by calculating the Singular Value Decomposition of PQ with

$$PQ = \pi \begin{pmatrix} \Sigma_1 & \\ & 0 \end{pmatrix} \pi^T.$$

It can easily been seen, that the same holds in case of appyling Singular Perturbation Balanced Truncation.

Observe, since the reduced system $(\bar{A}_{11}, \bar{B}_1, \bar{C}_1, \bar{D})$ is just a permutation of the original system, the Gramians are preserved up to a permutation. The advantage is, that we are keeping the physical meaning of each state. On the other hand we are basically left with the same problems as for an unbalanced system in the sense of Lyapunov Equalities. Thus we have to expect large Generalized Hankel Singular Values and a big truncation error. In fact, during numerical experiments, especially for SISO-systems, it turned out, that this method has poor approximation properties compared to the first order reduction via Balanced Truncation with respect to Lyapunov equalities.

Example 3.5 (Large Truncation Error)
Let us consider for instance the system

$$A := -diag\begin{pmatrix} 1, 1, 1 \end{pmatrix}, \quad B^T = C = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$$

with transfer function

$$G(s) = \frac{3}{s+1}.$$

By application of Balanced Truncation we obtain, according to Theorem 3.3, a minimal first order positive realization. In contrast, Generalized Balanced Truncation leads to

$$G_1(s) = \frac{1}{s+1},$$

which gives an absolute error $\|G - G_1\|_\infty = 2$.

In order to minimize the error of Generalized Balanced Truncation, it is essential to minimize the Generalized Hankel Singular Values, i.e. to attain many small values in PQ. A diagonal solution, as given in the proof to Theorem 1.2, is therefore not advisable. Instead Semidefinite Programming, a subfield of convex optimization, can be used to solve the Lyapunov inequalities. For the reason, that minimization of all eigenvalues is clearly not a convex problem, i.e. cannot be expressed as a convex function, an alternative is to minimize the trace and use a two step procedure as proposed in [22]. This procedure is based on the following algorithm.

Algorithm 3.1 (*Minimization of the Generalized Hankel Singular Values*)

(i) *For $j = 0$ let P_0 be the solution to (3.27) s.t. $\min tr(P)$.*

(ii) *For any j and fixed P_{j-1}, solve (3.27) for Q_j s.t. $\min tr(P_{j-1}Q)$.*

(iii) *For fixed Q_j find P_j s.t. $\alpha_j = \min tr(P_{j-1}Q)$ and (3.27).*
If $\frac{\alpha_{j-1} - \alpha_j}{\alpha_j} < TOL_\alpha$, for a prescribed tolerance TOL_α, then we have obtained optimal $P = P_j$ and $Q = Q_j$.
Otherwise set $j := j + 1$ and continue with (ii).

In the first step we apply Algorithm 3.1 to the whole system. Subsequently, we make a decision about the truncation candidates. The second step serves the purpose of getting a sharper error bound and does the same as in the first step, but restricted to those values in P and Q, that correspond to the truncation candidates.

Remark: The minimization of the trace does not guarantee, that we choose the best truncation candidates. Still, empirically it suffices in most cases.

4. Model Reduction of Positive Systems based on the Bounded Real Lemma

Beside the Hankel Singular Values in case of Balanced Truncation, there exists another well-known condition for a bound of the $\mathcal{H}_\infty$ truncation error, given by the so-called Bounded Real Lemma [30]. In this chapter we discuss two iterative methods, developed in [7] and [14], which are based on this lemma.

Theorem 4.1 (*Bounded Real Lemma*)

Let (A, B, C, D) be a state-space representation of $G(s)$. Then $G(s)$ is asymptotically stable and satisfies $\|G\|_\infty < \gamma$ if and only if there exists a matrix $P > 0$, such that

$$\begin{pmatrix} A^T P + PA & PB & C^T \\ B^T P & -\gamma I & D^T \\ C & D & -\gamma I \end{pmatrix} < 0. \tag{4.1}$$

Let

$$G_r : \begin{cases} \dot{x_r}(t) = A_r x_r(t) + B_r u(t), \\ y_r(t) = C_r x_r(t) + Du(t) \end{cases}$$

be a reduced-order approximation of (0.1). By denoting, $\hat{x} = \begin{pmatrix} x^T & x_r^T \end{pmatrix}^T$ and $e = y - y_r$, we can represent the error system $(G(s) - G_r(s))$ as

$$G_e : \begin{cases} \dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \hat{B}u(t), \\ e(t) = \hat{C}\hat{x}(t) + \hat{D}u(t) \end{cases} \tag{4.2}$$

with

$$\hat{A} = \begin{pmatrix} A & 0 \\ 0 & A_r \end{pmatrix}, \quad \hat{B} = \begin{pmatrix} B \\ B_r \end{pmatrix}, \quad \hat{C} = \begin{pmatrix} C & -C_r \end{pmatrix} \quad \text{and} \quad \hat{D} = D - D_r.$$

Then the $\mathcal{H}_\infty$-error of the approximation can be described as

$$\|G - G_r\|_\infty = \|G_e\|_\infty,$$

and by Theorem 4.1 it follows, that $\|G_e\|_\infty < \gamma$ if and only if there exists a matrix $\hat{P}$, such that

$$\Pi := \begin{pmatrix} \hat{A}^T\hat{P} + \hat{P}\hat{A} & \hat{P}\hat{B} & \hat{C}^T \\ \hat{B}^T\hat{P} & -\gamma I & \hat{D}^T \\ \hat{C} & \hat{D} & -\gamma I \end{pmatrix} < 0. \tag{4.3}$$

In contrary to just finding $\hat{P}$, which could be solved by semidefinite programming, in this situation $\hat{P}$ is coupled by its product with $\hat{A}$ and $\hat{B}$ and consequently with A_r and B_r, which are variables themselves. Our aim is now to resolve this coupling by introducing a new matrix variable with flexible structure.

4.1. Iterative Linear Matrix Approach I

Let us collect the system matrices of the reduced system in a matrix

$$\mathscr{G}_r := \begin{pmatrix} A_r & B_r \\ C_r & D_r \end{pmatrix}$$

and define

$$\begin{aligned} \bar{A} &:= \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad \bar{B} := \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad \bar{C} := \begin{pmatrix} C & 0 \end{pmatrix}, \quad \bar{D} := D, \\ \bar{F} &:= \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix}, \quad \bar{M} := \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix}, \quad \bar{N} := \begin{pmatrix} 0 \\ I \end{pmatrix}, \quad \bar{H} := \begin{pmatrix} 0 & -I \end{pmatrix}., \end{aligned} \tag{4.4}$$

Then we can express the error system G_e in terms of $\mathscr{G}_r$ as follows

$$\hat{A} = \bar{A} + \bar{F}\mathscr{G}_r\bar{M}, \quad \hat{B} = \bar{B} + \bar{F}\mathscr{G}_r\bar{N}, \quad \hat{C} = \bar{C} + \bar{H}\mathscr{G}_r\bar{M}, \quad \hat{D} = \bar{D} + \bar{H}\mathscr{G}_r\bar{N}. \tag{4.5}$$

Observe, since $\Pi < 0$ it holds that $\Pi^{-1} < 0$ and hence

$$\begin{pmatrix} \hat{P}\bar{F} \\ 0 \\ \bar{H} \end{pmatrix}^T \Pi^{-1} \begin{pmatrix} \hat{P}\bar{F} \\ 0 \\ \bar{H} \end{pmatrix} < 0.$$

Thus for any matrix $S > 0$, there must exist an $\alpha > 0$ such that

$$-\alpha S - \begin{pmatrix} \hat{P}\bar{F} \\ 0 \\ \bar{H} \end{pmatrix}^T \Pi^{-1} \begin{pmatrix} \hat{P}\bar{F} \\ 0 \\ \bar{H} \end{pmatrix} < 0. \tag{4.6}$$

By Schur complement and its application to the inverse of a negative definite matrix, as given in (3.30), we get

$$\Pi_e := \begin{pmatrix} \hat{A}^T\hat{P} + \hat{P}\hat{A} & \hat{P}\hat{B} & \hat{C}^T & \hat{P}\bar{F} \\ \hat{B}^T\hat{P} & -\gamma I & \hat{D}^T & 0 \\ \hat{C} & \hat{D} & -\gamma I & \bar{H} \\ \bar{F}^T\hat{P} & 0 & \bar{H}^T & -X \end{pmatrix} < 0 \quad \text{with} \quad X := \alpha S.$$

Consequently the existence of the matrices $\hat{P}, X > 0$, such that $\Pi_e < 0$, is equivalent to the Bounded Real Lemma. We want to use this fact to construct an equivalent expression in terms of the matrices given in (4.4). For this purpose we define

$$T := \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ -U & I & -V & 0 \end{pmatrix},$$

where $U := \mathscr{G}_r \bar{M}$ and $V := \mathscr{G}_r \bar{N}$. T is obviously invertible because of its full row rank, which is why we can define Π_p by

$$\Pi_p := T^T \Pi_e T < 0.$$

Expanding Π_p yields

$$\Pi_p = \begin{pmatrix} \bar{A}^T\hat{P} + \hat{P}\bar{A} - U^T X U & \hat{P}\bar{F} + U^T X & \hat{P}\bar{B} - U^T X V & \bar{C}^T \\ \bar{F}^T\hat{P} + XU & -X & XV & \bar{H}^T \\ \bar{B}^T\hat{P} - V^T X U & V^T X & -V^T X V - \gamma I & \bar{D}^T \\ \bar{C} & \bar{H} & \bar{D} & -\gamma I \end{pmatrix} < 0$$

and we observe, that $\hat{P}$ is completely decoupled from $\mathscr{G}_r$. Instead we have constructed a new coupling with X, which is however much more flexible, since X is arbitrary up to a scaling factor. Let us summarize this result in the following Lemma.

Lemma 4.1

Let G_e be the error system given in (4.2) and expressed in terms of (4.4). Then G_e is asymptotically stable and satisfies $\|G_e\|_\infty < \gamma$ if and only if there exist matrices $\hat{P}, X > 0$ such that $\Pi_p < 0$.

Observe, so far all the results hold for any reduced-order system. If we could fixate U and V, our problem could be easily solved by convex optimization. In the following we want to decouple U and V from $\mathscr{G}_r$ and treat them as variables. At the same time we want to incorporate the required positivity constraints.

Let L be a matrix of the same size and partitioning as $\mathscr{G}_r$, such that

$$L := \begin{pmatrix} L_1 & L_2 \\ L_3 & L_4 \end{pmatrix} \in \mathbb{P} \tag{4.7}$$

i.e. L_1 is a nonsingular $-M$-matrix and $L_2, L_3, L_4 \geqq 0$. By assuming

$$\mathscr{G}_r = X^{-1} L \quad \Leftrightarrow \quad L = X \mathscr{G}_r \tag{4.8}$$

with diagonal $X > 0$, we can rewrite Π_p as

$$\begin{pmatrix} \bar{A}^T\hat{P} + \hat{P}\bar{A} & \hat{P}\bar{F} + \bar{M}^T L^T & \hat{P}\bar{B} & \bar{C}^T \\ \bar{F}^T\hat{P} + L\bar{M} & -X & L\bar{N} & \bar{H}^T \\ \bar{B}^T\hat{P} & \bar{N}^T L^T & -\gamma I & \bar{D}^T \\ \bar{C} & \bar{H} & \bar{D} & -\gamma I \end{pmatrix} - \begin{pmatrix} U^T XU & 0 & U^T XV & 0 \\ 0 & 0 & 0 & 0 \\ V^T XU & 0 & V^T XV & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} < 0 \quad (4.9)$$

Expressing the second term of (4.9) with the help of

$$\Phi := \begin{pmatrix} \bar{M} & 0 & \bar{N} \end{pmatrix}, \quad \Psi := \begin{pmatrix} U & 0 & V \end{pmatrix},$$

leads to

$$\begin{aligned} -\begin{pmatrix} U^T XU & 0 & U^T XV \\ 0 & 0 & 0 \\ V^T XU & 0 & V^T XV \end{pmatrix} &= -\Phi^T \mathscr{G}_r^T X \mathscr{G}_r \Phi \\ &= -\Phi^T \mathscr{G}_r^T X \mathscr{G}_r \Phi + (\Psi - \mathscr{G}_r\Phi)^T X (\Psi - \mathscr{G}_r\Phi) \\ &= -\Psi^T L\Phi - \Phi^T L^T \Psi + \Psi^T X \Psi. \end{aligned} \quad (4.10)$$

Consequently (4.9) is equivalent to

$$\Pi_p = \begin{pmatrix} \Pi_{p_{11}} & \hat{P}\bar{F} + \bar{M}^T L^T & \Pi_{p_{13}} & \bar{C}^T \\ \bar{F}^T\hat{P} + L\bar{M} & -X & L\bar{N} & \bar{H}^T \\ \Pi_{p_{13}}^T & \bar{N}^T L^T & \Pi_{p_{33}} & \bar{D}^T \\ \bar{C} & \bar{H} & \bar{D} & -\gamma I \end{pmatrix} < 0 \quad (4.11)$$

with

$$\begin{aligned} \Pi_{p_{11}} &:= \bar{A}^T\hat{P} + \hat{P}\bar{A} - U^T L\bar{M} + \bar{M}^T L^T U + U^T XU \\ \Pi_{p_{13}} &:= \hat{P}\bar{B} - \bar{M}L^T V - U^T L\bar{N} + U^T XV \\ \Pi_{p_{33}} &:= -V^T L\bar{N} - \bar{N}^T L^T V + V^T XV - \gamma I \end{aligned}$$

in case that (A_r, B_r, C_r, D_r) is a positive system. It is readily seen, that we can generalize this result to arbitrary matrices $\tilde{U}$ and $\tilde{V}$ as follows. If we define $\tilde{\Psi} := \begin{pmatrix} \tilde{U} & 0 & \tilde{V} \end{pmatrix}$, we can write (4.10) as

$$\begin{aligned} -\Phi^T \mathscr{G}_r^T X \mathscr{G}_r \Phi &\leq -\Phi^T \mathscr{G}_r^T X \mathscr{G}_r \Phi + (\tilde{\Psi} - \mathscr{G}_r\Phi)^T X (\tilde{\Psi} - \mathscr{G}_r\Phi) \\ &= -\tilde{\Psi}^T L\Phi - \Phi^T L^T \tilde{\Psi} + \tilde{\Psi}^T X \tilde{\Psi}. \end{aligned} \quad (4.12)$$

Thus, if

$$\tilde{\Pi}_p(\tilde{U}, \tilde{V}) := \begin{pmatrix} \tilde{\Pi}_{p_{11}} & \hat{P}\bar{F} + \bar{M}^T L^T & \tilde{\Pi}_{p_{13}} & \bar{C}^T \\ \bar{F}^T\hat{P} + L\bar{M} & -X & L\bar{N} & \bar{H}^T \\ \tilde{\Pi}_{p_{13}}^T & \bar{N}^T L^T & \tilde{\Pi}_{p_{33}} & \bar{D}^T \\ \bar{C} & \bar{H} & \bar{D} & -\gamma I \end{pmatrix} < 0$$

where

$$\begin{aligned}\tilde{\Pi}_{p_{11}} &:= \bar{A}^T\hat{P} + \hat{P}\bar{A} - \tilde{U}^T L\bar{M} + \bar{M}^T L^T\tilde{U} + \tilde{U}^T X\tilde{U}\\ \tilde{\Pi}_{p_{13}} &:= \hat{P}\bar{B} - \bar{M}L^T\tilde{V} - \tilde{U}^T L\bar{N} + \tilde{U}^T X\tilde{V}\\ \tilde{\Pi}_{p_{33}} &:= -\tilde{V}^T L\bar{N} - \bar{N}^T L^T\tilde{V} + \tilde{V}^T X\tilde{V} - \gamma I\end{aligned}$$

and

$$\mathscr{G}_r = X^{-1}L \tag{4.13}$$

it follows by Lemma 4.1, that $\|G_e\|_\infty < \gamma$ with a positive solution (A_r, B_r, C_r, D_r). Let us summarize this result in the following theorem.

Theorem 4.2

G_e is asymptotically stable and satisfies $\|G_e\|_\infty < \gamma$ with a positive system (A_r, B_r, C_r, D_r) if and only if there exists a $\hat{P} > 0$, matrices $\tilde{U}$ and $\tilde{V}$, a diagonal $X > 0$ and $L \in \mathbb{P}$ such that $\tilde{\Pi}_p(\tilde{U}, \tilde{V}) < 0$. Then we can write $\mathscr{G}_r = X^{-1}L$.

As mentioned before, if we already knew $\tilde{U}$ and $\tilde{V}$, solving $\tilde{\Pi}_p(\tilde{U}, \tilde{V}) < 0$ for P, X and L is a convex problem. Thus our problem reduces to how to choose them properly.

4.2. Algorithm: Iterative Linear Matrix Approach I

Now, we want to propose a method how to find $\tilde{U}$ and $\tilde{V}$ iteratively. For this purpose we notice first, if X, L and $\hat{P}$ are fix, there must exist an $\alpha \in \mathbb{R}$ for every $\tilde{U}$ and $\tilde{V}$, such that

$$\tilde{\Pi}_p(\tilde{U}, \tilde{V}) < \alpha \begin{pmatrix} I & & & \\ & 0 & & \\ & & I & \\ & & & 0 \end{pmatrix}. \tag{4.14}$$

By (4.10) and (4.12) it follows, that α attains its minimum for

$$\tilde{U} = X^{-1}L\bar{M} \quad \text{and} \quad \tilde{V} = X^{-1}L\bar{N}. \tag{4.15}$$

On the other hand it is clear, that for fixed $\tilde{U}$ and $\tilde{V}$, we can always attain solutions $\hat{P}$,X, L and α, satisfying (4.14), by convex optimization.

Consequently, if we start with $\tilde{U} = \tilde{U}_0$ and $\tilde{V} = \tilde{V}_0$ and find solutions $\hat{P}$, X and L fulfilling (4.14) for the smallest possible α, we can decrease α monotonically by updating $\tilde{U}$ and $\tilde{V}$ as in (4.15). When α reaches a nonnegative level we have found a positive approximation (A_r, B_r, C_r, D_r) accordingly to Theorem 4.2. The case that α converges to a positive value will be treated later.

A good way of choosing $\tilde{U}_0$ and $\tilde{V}_0$ can be found by considering the expression of the

error system matrices in (4.5) and using (4.3). If we cannot find solutions $\hat{Q}$, V_0, U_0 fulfilling

$$\Pi_P(U_0,V_0) := \begin{pmatrix} (\bar{A}+\bar{F}U_0)^T\hat{Q}+\hat{Q}(\bar{A}+\bar{F}U_0) & \hat{Q}(\bar{B}+\bar{F}V_0) & (\bar{C}+\bar{H}U_0)^T \\ (\bar{B}+\bar{F}V_0)^T\hat{Q} & -\gamma I & (\bar{D}+\bar{H}V_0)^T \\ \bar{C}+\bar{H}U_0 & \bar{D}+\bar{H}V_0 & -\gamma I \end{pmatrix} < 0,$$

then, according to Theorem 4.1, there does not exist any solution satisfying this error bound. As before, finding these solutions is not a convex problem because of the products $\hat{Q}\bar{F}U_0$ and $\hat{Q}\bar{F}V_0$. We can overcome this obstacle by defining $W_0 := U_0\hat{Q}$ and considering its dual problem, i.e. we apply the Bounded Real Lemma to G_e^T, which leaves the error bound unchanged. Thus we get

$$\begin{pmatrix} \bar{A}\hat{Q}+\bar{F}W_0+\hat{Q}\bar{A}^T+\bar{F}^TW_0^T & \hat{Q}\bar{C}^T+W_0^T\bar{H}^T & \bar{B}+\bar{F}V_0 \\ \bar{C}\hat{Q}+\bar{H}W_0 & -\gamma I & \bar{D}+\bar{H}V_0 \\ (\bar{B}+\bar{F}V_0)^T & (\bar{D}+\bar{H}V_0)^T & -\gamma I \end{pmatrix} < 0 \tag{4.16}$$

and finding the solutions $\hat{Q}$, V_0 and W_0 is a convex problem.
In this case the connection between U_0, V_0 and $\mathscr{G}_r$ is not important to us, why solving for W_0 is feasible. Especially because of our positivity constraints on $\mathscr{G}_r$, we cannot do the same for achieving a positive reduced-order system.

We have already seen, if there exist a diagonal $X > 0$, $\hat{P} > 0$ and $\mathscr{G}^*$, such that $\Pi_p < 0$, then we conclude by (4.12), for sufficiently small

$$\|(\Psi_0 - \mathscr{G}^*\Phi)^T X(\Psi_0 - \mathscr{G}^*\Phi)\|_2 \quad \text{with} \quad \Psi_0 := \begin{pmatrix} U_0 & 0 & V_0 \end{pmatrix}$$

it holds $\tilde{\Pi}_p(U_0,V_0) < 0$, with $L = X\mathscr{G}^*$. Since X can be very large, as seen in (4.6), we need a way of minimizing $\|\Psi_0 - \mathscr{G}^*\Phi\|_2$. Such a method can be concluded from the next theorem.

Theorem 4.3 (*Initial Optimization*)

Let $\epsilon > 0$ be sufficiently small and Φ and Ψ_0 as defined before, then the following statements are equivalent:

(i) There exists a solution $\mathscr{G}^*$, such that $\|G_e\|_\infty < \gamma$ and $\|\Psi_0 - \mathscr{G}^*\Phi\|_2 \leq \epsilon$.

(ii) $\|\Psi_0\Phi_\perp\|_2 \leq \epsilon$ and $\Pi_P(U_0,V_0) < 0$.

where $\Phi_\perp$ denotes a matrix consisting of a basis of the kernel of Φ, i.e. $\Phi\Phi_\perp = 0$.

Proof: $(i) \Rightarrow (ii)$: Φ as defined before is explicitly given as

$$\Phi = \begin{pmatrix} 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix}$$

and thus $\Phi_\perp$ has the form

$$\Phi_\perp = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix}.$$

Since, $\Phi_\perp^T \Phi_\perp = I$, it follows that $\|\Phi_\perp\|_2 = 1$ and by assumption that

$$\|\Psi_0 \Phi_\perp\|_2 = \|\Psi_0 \Phi_\perp - \mathscr{G}^* \Phi \Phi_\perp\|_2 \leq \|\Psi_0 - \mathscr{G}^* \Phi\|_2 \|\Phi_\perp\|_2 \leq \epsilon$$

Let $\hat{Q}$ be a solution that satisfies $\Pi_P(\mathscr{G}^* \bar{M}, \mathscr{G}^* \bar{N}) < 0$. Then $\Pi_P(U_0, V_0)$, with the same $\hat{Q}$, can be expressed as

$$\Pi_P(U_0, V_0) = \Pi_P(\mathscr{G}^* \bar{M}, \mathscr{G}^* \bar{N}) + \Xi,$$

where

$$\Xi := \begin{pmatrix} (U_0 - \mathscr{G}^* \bar{M})^T \bar{F}^T \hat{Q} + \hat{Q} \bar{F} (U_0 - \mathscr{G}^* \bar{M}) & \bar{F}(V_0 - \mathscr{G}^* \bar{N}) & (U_0 - \mathscr{G}^* \bar{M})^T \bar{H}^T \\ (V_0 - \mathscr{G}^* \bar{N})^T \bar{F}^T & 0 & (V_0 - \mathscr{G}^* \bar{N})^T \bar{H}^T \\ \bar{H}(U_0 - \mathscr{G}^* \bar{M}) & \bar{H}(V_0 - \mathscr{G}^* \bar{N}) & 0 \end{pmatrix}$$

Consequently, if $\Psi_0 - \mathscr{G}^* \Phi = \begin{pmatrix} U_0 - \mathscr{G}^* \bar{M} & 0 & V_0 - \mathscr{G}^* \bar{N} \end{pmatrix}$ is sufficiently small, it follows by the negative definiteness of $\Pi_P(\mathscr{G}^* \bar{M}, \mathscr{G}^* \bar{N})$, that $\Pi_P(U_0, V_0) < 0$ with the solution $\hat{Q}$.

$(ii) \Rightarrow (i)$: By choosing $\mathscr{G}^* = \Psi_0 \Phi^T$ and noticing, that $\Phi \Phi^T = I$, we can write

$$(\Psi_0 - \mathscr{G}^* \Phi) \begin{pmatrix} \Phi^T & \Phi_\perp \end{pmatrix} = \begin{pmatrix} 0 & \Psi_0 \Phi_\perp \end{pmatrix}.$$

It is obvious, that $\begin{pmatrix} \Phi^T & \Phi_\perp \end{pmatrix}$ is an invertible matrix with $\begin{pmatrix} \Phi \\ \Phi_\perp^T \end{pmatrix} \begin{pmatrix} \Phi^T & \Phi_\perp \end{pmatrix} = I$.
Consequently,

$$\Psi_0 - \mathscr{G}^* \Phi = \begin{pmatrix} 0 & \Psi_0 \Phi_\perp \end{pmatrix} \begin{pmatrix} \Phi^T & \Phi_\perp \end{pmatrix}^{-1}$$

and by assumption

$$\|\Psi_0 - \mathscr{G}^* \Phi\|_2 = \| \begin{pmatrix} 0 & \Psi_0 \Phi_\perp \end{pmatrix} \|_2 \| \begin{pmatrix} \Phi^T & \Phi_\perp \end{pmatrix}^{-1} \|_2 \leq \epsilon.$$

Showing that $\|G_e\|_\infty < \gamma$ follows as before by $\Pi_P(U_0, V_0) < 0$ and

$$\Pi_P(\mathscr{G}^* \bar{M}, \mathscr{G}^* \bar{N}) = \Pi_P(U_0, V_0) - \Xi < 0$$

for a sufficiently small ϵ. ■

Observe, if we can find U_0 and V_0, such that $\|\Psi \Phi_\perp\|_2 = 0$ with $\Pi_P(U_0, V_0) < 0$, then $\tilde{\Pi}_p(U_0, V_0) < 0$. Further, for fixed $\hat{Q}$, it is a convex problem to solve $\Pi_P(U_0, V_0) < 0$, such that $\mathscr{G}^* = \Psi \Phi_\perp$ represents a positive system. Thus, we can already include our positivity constraints into the optimization of the initial values.

Let us summarize the algorithm as we have developed it so far. The condition $\|\Psi_0 \Phi_\perp\|_2 = \bar{\sigma}(\Psi_0 \Phi_\perp) \leq \epsilon$ will be expressed with help of its Schur complement equivalence.

Algorithm 4.1 (*Initial Value Optimization*)

(i) *Solve the initial problem (4.16) to get* U_0 *and* V_0 *and set* $j = 0$.

(ii) *For any* j *and fixed* $\Psi_j = \begin{pmatrix} U_j & 0 & V_j \end{pmatrix}$ *find a solution* $\hat{Q}_j$ *to* $\Pi_P(U_j, V_j) < 0$.

(iii) *For fixed* $\hat{Q}_j$ *find an optimized* $\Psi_j^* = \begin{pmatrix} U_j^* & 0 & V_j^* \end{pmatrix}$ *and* ϵ_j^* *such that*

$$\epsilon_j^* = \min_{\Psi_j = [U_j \ 0 \ V_j]} \epsilon_j \ s.t. \begin{cases} \Pi_P(U_j, V_j) < 0 \\ \begin{pmatrix} -\epsilon_j I & (\Psi_j \Phi_\perp)^T \\ \Psi_j \Phi_\perp & -I \end{pmatrix} < 0 \\ \Psi_j \Phi_\perp \in \mathbb{P} \end{cases}$$

(iv) *If* $\dfrac{|\epsilon_j^* - \epsilon_{j-1}^*|}{|\epsilon_j^*|} < TOL_{init}$ *or* $j = MAX_{iter}$, *where* TOL_{init} *is a prescribed tolerance and* MAX_{iter} *a maximal number of iterations, then optimal initial matrices* $U_0^* = U_j^*$ *and* $V_0^* = V_j^*$ *are obtained.*
Otherwise set $\Psi_{j+1} := \Psi_j^*$, $j := j + 1$ *and go to step (ii).*

Algorithm 4.2 (*Iterative LMI Approach*)

(i) *Set* $j = 0$ *and* $U_j := U_0^*$ *and* $V_j := V_0^*$ *obtained by Algorithm 4.1*

(ii) *For any* j *and fixed* U_j *and* V_j *find optimal* $\hat{P}^* > 0$, *diagonal* $X_j^* > 0$, $L_j^* \in \mathbb{P}$ *and* α_j^*, *such that*

$$\alpha_j^* = \min_{\hat{P}_j, X_j, L_j} \alpha \ s.t. \ \tilde{\Pi}_p(U_j, V_j) < \alpha_j \begin{pmatrix} I & & & \\ & 0 & & \\ & & I & \\ & & & 0 \end{pmatrix}.$$

(iii) *If* $\alpha_j^* \leq 0$, *then an optimal* $\mathscr{G}_r^* = (X_j^*)^{-1} L_j^*$ *is found.*
If $\dfrac{|\alpha_j^* - \alpha_{j-1}^*|}{|\alpha_j^*|} < TOL_\alpha$ *or* $j = MAX_{iter}$, *for a prescribed tolerance* TOL_α *and an maximal number of iterations* MAX_{iter}, *then* α_j^* *has probably converged to a positive value and we stop without a solution.*
Otherwise, set $j := j + 1$,

$$U_{j+1} := (X_j^*)^{-1} L_j^* \bar{M}, \quad V_{j+1} := (X_j^*)^{-1} L_j^* \bar{N}$$

and continue with step (ii).

Now we want to treat the case, when we cannot find an $\alpha \leq 0$. For the initial value determination we already considered the dual problem G_e^T. By defining $U_d := \bar{F}\mathscr{G}_r$ and $V_d := \bar{H}\mathscr{G}_r$ we can write

$$\hat{A}^T := \bar{A}^T + \bar{M}^T U_d^T, \quad \hat{B}^T := \bar{B}^T + \bar{N}^T U_d^T, \quad \hat{C} := \bar{C}^T + \bar{M}^T V_d^T, \quad \hat{D}^T := \bar{D}^T + \bar{N}^T V_d.$$

Then analogously to G_e, we can define

$$\tilde{\Pi}_d(\tilde{U}, \tilde{V}) = \begin{pmatrix} \tilde{\Pi}_{d_{11}} & \tilde{\Pi}_{d_{12}} & \tilde{\Pi}_{d_{13}} & \bar{B} \\ \tilde{\Pi}_{d_{12}}^T & -Z & L^T\bar{H}^T & \bar{N} \\ \tilde{\Pi}_{d_{13}}^T & \bar{H}L & \tilde{\Pi}_{d_{33}} & \bar{D} \\ \bar{B}^T & \bar{N}^T & \bar{D}^T & -\gamma I \end{pmatrix} < 0$$

where

$$\begin{aligned} \tilde{\Pi}_{d_{11}} &:= \bar{A}\hat{Q} + \hat{Q}\bar{A}^T - \tilde{U}L^T\bar{F}^T + \bar{F}L\tilde{U}^T + \tilde{U}Z\tilde{U}^T \\ \tilde{\Pi}_{d_{12}} &:= \hat{Q}\bar{M}^T + \bar{F}L \\ \tilde{\Pi}_{d_{13}} &:= \hat{Q}\bar{C}^T - \bar{F}L\tilde{V}^T - \tilde{U}L^T\bar{H}^T + \tilde{U}Z\tilde{V}^T \\ \tilde{\Pi}_{d_{33}} &:= -\tilde{V}L^T\bar{H}^T - \bar{H}L\tilde{V}^T + \tilde{V}Z\tilde{V}^T - \gamma I \end{aligned}$$

Theorem 4.4

G_e^T is asymptotically stable and satisfies $\|G_e^T\|_\infty < \gamma$ with a positive system (A_r, B_r, C_r, D_r) if and only if there exists a $\hat{Q} > 0$, matrices $\tilde{U}$ and $\tilde{V}$, a diagonal $Z > 0$ and $L \in \mathbb{P}$ such that $\tilde{\Pi}_d(\tilde{U}, \tilde{V}) < 0$. Then we can write $\mathscr{G}_r = LZ^{-1}$.

In the same way as for the primal algorithm, we can obtain initial matrices U_0 and V_0, by applying the Bounded Real Lemma to G_e. The Dual Iterative LMI Approach can then be given as follows.

Algorithm 4.3 (*Dual Iterative LMI Approach*)

(i) *Set $j = 0$ and $U_j := U_0$ and $V_j := V_0$.*

(ii) *For any j and fixed U_j and V_j find optimal $\hat{Q}^* > 0$, diagonal $Z_j^* > 0$, $L_j^* \in \mathbb{P}$ and β_j^*, such that*

$$\beta_j^* = \min_{\hat{Q}_j, Z_j, L_j} \beta \text{ s.t. } \tilde{\Pi}_d(U_j, V_j) < \beta_j \begin{pmatrix} I & & & \\ & 0 & & \\ & & I & \\ & & & 0 \end{pmatrix}.$$

(iii) *If $\beta_j^* \leq 0$, then an optimal $\mathscr{G}_r^* = L_j^*(Z_j^*)^{-1}$.*

(iv) *If $\dfrac{|\beta_j^* - \beta_{j-1}^*|}{|\beta_j^*|} < TOL_\beta$ or $j = MAX_{iter}$, for a prescribed tolerance TOL_β and a maximal number of iterations MAX_{iter}, then β_j^* is probably converged to a positive value or converges so slow, that we stop without a solution.*

(v) *Otherwise, set $j := j + 1$,*

$$U_{j+1} := \bar{F} L_j^*(Z_j^*)^{-1}, \quad V_{j+1} := \bar{H} L_j^*(Z_j^*)^{-1}$$

and continue with step (ii).

The motivation behind the additional consideration of the dual approach is, that an optimal solution in primal direction does not imply the optimality in dual direction and vice versa. Consequently if α converges to a positive value with $\mathscr{G}_r^\alpha := \mathscr{G}_r$, we can define $U_0 = \bar{F}\mathscr{G}_r^\alpha$ and $V_0 = \bar{H}\mathscr{G}_r^\alpha$ to use them as the initial matrices in the dual approach. Conversely, if β converges to a positive value with $\mathscr{G}_r^\beta := \mathscr{G}_r$, we can to the same by defining $U_0 = \mathscr{G}_r^\beta \bar{M}$ and $V_0 = \mathscr{G}_r^\beta \bar{N}$ and applying the primal approach. This procedure can be repeated until either α and β converge or one of them becomes nonnegative.

4.3. Iterative Linear Matrix Approach II

In this section we want to have a look at another approach to decouple A_r and B_r from $\hat{P}$, as presented in [7]. It is readily seen, since the Schur complement of negative definite matrix is also negative definite, that the existence of $\hat{P} > 0$ such that $\Pi < 0$ holds if and only if there exists $\tilde{P} > 0$, such that

$$\Theta := \begin{pmatrix} \hat{A}^T\tilde{P} + \tilde{P}\hat{A} + \hat{C}^T\hat{C} & \tilde{P}\hat{B} + \hat{C}^T\hat{D} \\ \hat{B}^T\hat{P} + \hat{D}^T\hat{C} & -\gamma^2 I + \hat{D}^T\hat{D} \end{pmatrix} < 0. \tag{4.17}$$

$\tilde{P}$ and $\hat{P}$ fulfil the relation $\tilde{P} = \gamma\hat{P}$. Further, if (4.17) is valid, then there must exist a sufficiently small $\epsilon > 0$, such that

$$\begin{pmatrix} \hat{A}^T\tilde{P} + \tilde{P}\hat{A} + \hat{C}^T\hat{C} + \epsilon\hat{A}^T\tilde{P}\hat{A} & \tilde{P}\hat{B} + \hat{C}^T\hat{D} \\ \hat{B}^T\hat{P} + \hat{D}^T\hat{C} & -\gamma^2 I + \hat{D}^T\hat{D} + \epsilon\hat{B}^T\tilde{P}\hat{B} \end{pmatrix} < 0. \tag{4.18}$$

In order to apply Schur complement equivalence to (4.18), we notice

$$\Theta - \begin{pmatrix} I & \epsilon\hat{B} \\ I + \epsilon\hat{A} & 0 \end{pmatrix}^T \begin{pmatrix} -\epsilon^{-1}\tilde{P} & 0 \\ 0 & -\epsilon^{-1}\tilde{P} \end{pmatrix} \begin{pmatrix} I & \epsilon\hat{B} \\ I + \epsilon\hat{A} & 0 \end{pmatrix} = \\ = \begin{pmatrix} -2\epsilon^{-1}\tilde{P} + \hat{C}^T\hat{C} & \hat{C}^T\hat{D} \\ \hat{D}^T\hat{C} & -\gamma^2 I + \hat{D}^T\hat{D} \end{pmatrix}$$

which leads to a decoupling of $\hat{A}$ and $\hat{B}$ with $\tilde{P}$ as follows

$$\begin{pmatrix} -2\epsilon^{-1}\tilde{P} + \hat{C}^T\hat{C} & \hat{C}^T\hat{D} & I & I + \epsilon\hat{A}^T \\ \hat{D}^T\hat{C} & -\gamma^2 I + \hat{D}^T\hat{D} & \epsilon\hat{B}^T & 0 \\ I & \epsilon\hat{B} & -\epsilon P^{-1} & 0 \\ I + \epsilon\hat{A} & 0 & 0 & -\epsilon P^{-1} \end{pmatrix} < 0.$$

If we define then

$$\tilde{A}_r := \epsilon A_r, \quad \tilde{B}_r := \epsilon B_r, \quad X := \epsilon^{-1}P \quad \tilde{X} := \epsilon P^{-1}$$

with

$$X = \begin{pmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{pmatrix}, \quad \tilde{X} = \begin{pmatrix} \tilde{X}_{11} & \tilde{X}_{12} \\ \tilde{X}_{12}^T & \tilde{X}_{22} \end{pmatrix},$$

it follows by another Schur complement equivalence for $\hat{C}$ and $\hat{D}$, that

$$\Theta_e := \begin{pmatrix} -2X_{11} & -2X_{12} & 0 & I & 0 & I + \epsilon A^T & 0 & C^T \\ -2X_{12}^T & -2X_{22} & 0 & 0 & I & 0 & I + \tilde{A}_r^T & -C_r^T \\ 0 & 0 & -\gamma^2 I & \epsilon B^T & \tilde{B}_r^T & 0 & 0 & D^T - D_r^T \\ I & 0 & \epsilon B & -\tilde{X}_{11} & -\tilde{X}_{12} & 0 & 0 & 0 \\ 0 & I & \tilde{B}_r & -\tilde{X}_{12}^T & -\tilde{X}_{22} & 0 & 0 & 0 \\ I + \epsilon A & 0 & 0 & 0 & 0 & -\tilde{X}_{11} & -\tilde{X}_{12} & 0 \\ 0 & I + \tilde{A}_r & 0 & 0 & 0 & -\tilde{X}_{12}^T & -\tilde{X}_{22} & 0 \\ C & -C_r & D - D_r & 0 & 0 & 0 & 0 & I \end{pmatrix} < 0.$$

Theorem 4.5

Let (A, B, C, D) be an asymptotically stable positive system with transfer function $G(s)$. Then a reduced-order asymptotically stable positive system G_r with $\|G_e\|_\infty < \gamma$ exists if and only if we can find $\epsilon > 0$, $X, \bar{X} > 0$, a $-M$-matrix $\tilde{A}_r$ and $\tilde{B}_r, C_r, D_r \geqq 0$ such that $\Theta_e < 0$ and $X\bar{X} = I$.

Observe, in Θ_e all the variables are decoupled. Hence without the requirement $X\bar{X} = I$, we would be left with a convex problem.

4.4. Algorithm: Iterative Linear Matrix Approach II

In the following we will see how to treat this coupling of X and $\bar{X}$ with the help of the so-call *Convex Cone Linearization Algorithm (CCL)* [9]. The basic idea of this algorithm is to minimize $tr(X\bar{X})$ with respect to positive definite matrices $X, \bar{X} \in \mathbb{R}^{n\times n}$ fulfilling

$$\begin{pmatrix} X & I \\ I & \bar{X} \end{pmatrix} \geq 0. \tag{4.19}$$

If (4.19) holds, then by considering its Schur complement, we conclude

$$X - \bar{X} \geq 0 \;\Leftrightarrow\; \bar{X}^{\frac{1}{2}} X \bar{X}^{\frac{1}{2}} - I \geq 0, \tag{4.20}$$

and thus

$$tr(X\bar{X}) = tr(\bar{X}^{\frac{1}{2}} X \bar{X}^{\frac{1}{2}}) \geq n. \tag{4.21}$$

It is obvious, that equality can be achieved if $X\bar{X} = I$, but by considering the diagonalization $\bar{X}^{\frac{1}{2}} X \bar{X}^{\frac{1}{2}} = T^T D T$ with $D \geq 0$, it follows by (4.20) and (4.21) that

$$D - I \geq 0 \text{ and } tr(X\bar{X} - I) = tr(D - I) \geq 0.$$

Hence, $tr(X\bar{X}) = n$ if and only if $X\bar{X} = I$ and we can reduce our problem of finding $X\bar{X} = I$ to the minimization of $tr(X\bar{X})$ with respect to (4.19).
Since minimizing $tr(X\bar{X})$ is not a convex problem either, it will be solved by considering its linearisation. At a given, feasible point $(X_0, \bar{X}_0)$ a linear approximation can be given as

$$tr(X\bar{X}) \approx c + tr(X\bar{X}_0 + \bar{X}X_0),\ c \in \mathbb{R}$$

From Theorem 3.1 we know, that the product of two matrices $P, Q \geq 0$ has exclusively nonnegative eigenvalues and hence $tr(PQ) \geq 0$. Applying this to $X\bar{X}_0$ and $\bar{X}X_0$ leads to

$$tr(X\bar{X}_0 + \bar{X}X_0) = tr(X\bar{X}_0) + tr(\bar{X}X_0) \geq 0. \tag{4.22}$$

Thus, the smaller $tr(X\bar{X}_0 + \bar{X}X_0)$ the smaller $tr(X\bar{X})$. The idea of the CCL-algorithm is to minimize $tr(X\bar{X}_0 + \bar{X}X_0)$, which a convex problem, because X and $\bar{X}$ are decoupled. The whole CCL-algorithm can be described as follows.

Algorithm 4.4 (*Convex Cone Linear Approximation Algorithm*)

(i) *Let $(X_0, \bar{X}_0)$ be a feasible solution of (4.19) and set $j = 0$.*

(ii) *For any j and fixed $(X_j, \bar{X}_j)$ find an optimal solution $(X^*, \bar{X}^*)$ to*

$$\mathscr{P}_k : \min_{(X,\bar{X})} tr(X\bar{X}_j + \bar{X}X_j) \text{ s.t. } (4.19).$$

(iii) *If a stopping criterion is fulfilled, then an optimal $(X^*, \bar{X}^*)$ is found. Otherwise set $j := j + 1$,*

$$X_{j+1} := X^*, \quad \bar{X}_{j+1} := \bar{X}^*$$

and go to step (ii).

By defining $t_j := tr(X_{j+1}\bar{X}_j + \bar{X}_{j+1}X_j)$ and by the optimality of of t_j with respect to $\mathscr{P}_k$, it follows immediately that

$$t_j \leq tr(X_j\bar{X}_{j-1} + \bar{X}_jX_{j-1}) = t_{j-1}$$

Thus $\{t_j\}$ is a monotonically decreasing sequence, which converges according to (4.22).

Now we are ready to give the whole algorithm in order to fulfil the requirements of Theorem 4.5.

Algorithm 4.5 (*CCL-based LMI Approach*)

(i) *For given reduced order r and error bound γ, let $(X_0, \bar{X}_0)$ be a feasible solution s.t. (4.19) and $\Theta_e < 0$. If the solutions exists set $j = 0$, otherwise stop without a solution.*

(ii) *For any j and fixed $(X_j, \bar{X}_j)$ find an optimal solution $(X^*, \bar{X}^*, \bar{A}_r, \bar{B}_r, C_r, D_r, \epsilon)$*

$$t_j := \min_{X,\bar{X}} tr(X\bar{X}_j + \bar{X}X_j) \text{ s.t. } \begin{cases} (4.19) \\ \Theta_e < 0 \end{cases}$$

(iii) *Set $A_r := \epsilon^{-1}\bar{A}_r, B_r := \epsilon^{-1}\bar{B}_r, \tilde{P} := \epsilon X^*$ and plug them into Θ. If $\Theta < 0$, then a reduced order system, satisfying the prescribed error bound, is found. If $\frac{t_j - t_{j-1}}{t_j} > TOL_\delta$ or $j < MAX_{iter}$, for a prescribed tolerance TOL_δ and the maximal number of iterations MAX_{iter}, set $j := j + 1$,*

$$X_{j+1} := X^*, \quad \bar{X}_{j+1} := \bar{X}^*$$

and go to step (ii). Otherwise stop without a solution.

Remark: The CCL-algorithm also works without the constraint in (4.19), which is why it could be used as an alternative to minimize $tr(PQ)$ in order to minimize the Generalized Hankel Singular Values in Subsection 3.4.1. From the experience of numerical examples this does not add any advantage. In contrary, by using Algorithm 3.1 instead of the CCL-algorithm, the convergence of the algorithm discussed in this section is much slower.

In Chapter 7 we will see, even though both methods in this chapter are based on the Bounded Real Lemma, the method of Section 4.1 gives significantly better results than the just presented one. Still, we should notice, that both methods are based on LMIs, which restricts its application to low dimensional systems due to the high numerical effort of performing the required optimizations. A conventional solver e.g. *SeDuMi* possesses a complexity of $\mathcal{O}(n^2m^{2.5} + m^{3.5})$, where n stands for the number of decision variables and m for the number of rows in the LMI [20].

5. Krylov Subspace Methods

In the previous chapter we have encountered the problem, that the performance of model order reduction methods can depend strongly on the dimension of the system we would like to reduce. However, the occurrence of systems consisting of several thousand states, called *large-scale systems*, is not unusual. Applying LMI approaches as well as Balanced Truncation to such systems requires far more computational power than we have at our disposal today. A way of getting around this problem is given by the so-called *Krylov subspace methods* [4][12], which will be covered in this chapter.

The problem of large scale systems originates from the context of solving a system of linear equations

$$Ax = b, A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$$

with "very large" dimension n. Since direct solving methods such as LU- and QR-decomposition et al. possess a complexity of $\mathcal{O}(n^3)$, those methods can easily reach the limit of computational power. Overcoming this problem in case of a system with random A is probably impossible. However, if A is sparse, i.e. A contains "sufficiently many" zeros, it is feasible to approximate the solution.
For this purpose iterative methods with complexity $\mathcal{O}(n^2)$ have been developed .[27] A particular class within those iterative methods are the Krylov subspace methods [27].
In the following we will not discuss the solvers itself, but we are interested in their fundamental concepts, which will be used for developing a model order reduction method.

5.1. Arnoldi Iteration

The basic idea of all Krylov subspace methods is the (orthogonal) projection of A onto the Krylov subspace $\mathcal{K}_m(A, b)$, which we define now.

Definition 5.1 (*Krylov subspace*)

For $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ the Krylov subspace of dimension m is defined as

$$\mathcal{K}_m(A, b) := span\{b, Ab, \ldots, A^{m-1}b\}.$$

The orthogonal projection of A onto $\mathcal{K}_m(A, b)$ should be interpreted by the following linear operator $\mathcal{K}_m(A, b) \to \mathcal{K}_m(A, b)$: for given $x \in \mathcal{K}_m(A, b)$ apply A to it and perform an orthogonal projection of Ax back into $\mathcal{K}_m(A, b)$.

The orthogonal projector of $\mathbb{R}^n$ onto $\mathcal{K}_m(A,b)$ can be described with the help of a modified Gram-Schmidt iteration applied to $\mathcal{K}_m(A,b)$, known as the *Arnoldi iteration algorithm*.

Given $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n \setminus \{0\}$ we can set similar to Gram-Schmidt

$$\begin{aligned} w_0 &:= b, & v_1 &:= \frac{w_0}{\|w_0\|_2} \\ w_1 &:= Av_1 - \langle v_1, Av_1\rangle v_1, & v_2 &:= \frac{w_1}{\|w_1\|_2} \\ &\vdots & &\vdots \\ w_m &:= Av_m - \textstyle\sum_{j=1}^{m}\langle v_j, Av_m\rangle v_j, & v_{m+1} &:= \frac{w_m}{\|w_m\|_2} \end{aligned}$$

and attain an orthonormal basis $\{v_1,\dots,v_m\}$ of $\mathcal{K}_m(A,b)$ if $w_i \neq 0$ for $1 \leq i \leq m-1$. Otherwise the Krylov subspace dimension is smaller then m, i.e. $\mathcal{K}_m(A,b) \subseteq \mathcal{K}_k(A,b)$ for some $k < m$ and we consider the projection onto $\mathcal{K}_k(A,b)$ instead.

By defining

$$H_m := \begin{pmatrix} \langle v_1, Av_1\rangle & \langle v_1, Av_2\rangle & \cdots & \langle v_1, Av_m\rangle \\ \|w_1\|_2 & \langle v_2, Av_2\rangle & \cdots & \langle v_2, Av_m\rangle \\ & \ddots & \ddots & \vdots \\ & & \|w_{m-1}\|_2 & \langle v_m, Av_m\rangle \end{pmatrix} \quad \text{and} \quad V_m := \begin{pmatrix} v_1 & \dots & v_m \end{pmatrix},$$

we can write

$$AV_m = V_m H_m + w_m e_m^T.$$

Observe, since $V_m^T V_m = I$ and $V_m^T w_m = 0$, we obtain

$$V_m^T A V_m = H_m = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1m} \\ h_{21} & h_{22} & \cdots & h_{2m} \\ & \ddots & \ddots & \vdots \\ & & h_{m,m-1} & h_{mm} \end{pmatrix}$$

and A is an upper Hessenberg matrix with respect to the orthonormal basis of $\mathcal{K}_m(A,b)$. Further, if $x \in \mathbb{R}^n$, then $V_m V_m^T x \in \mathcal{K}_m(A,b)$ and

$$V_m^T(x - V_m V_m^T x) = 0.$$

Obviously, $V_m V_m^T$ is an orthogonal projection and the projection of Ax onto $\mathcal{K}_m(A,b)$ can be written as $V_m V_m^T Ax$. Therefore, the linear operator $V_m V_m^T A$ describes the orthogonal projection of A onto $\mathcal{K}_m(A,b)$. Moreover, if we write $x \in \mathcal{K}_m(A,b)$ in the basis $\{v_1,\dots,v_m\}$ as $V_m\xi$, then

$$V_m V_m^T Ax = V_m V_m^T A V_m \xi = V_m \eta,$$

with

$$\eta := H_m \xi.$$

Again, we can say that $V_m \eta$ is the representation of η in the standard basis and consequently H_m can be interpreted as the matrix representation $V_m V_m^T Ax$ with respect to the basis $\{v_1, \ldots, v_m\}$.

The Algorithm to the *Arnoldi Iteration* can be described efficiently in the following form.

Algorithm 5.1 (*Arnoldi Iteration Algorithm*)

Let $b \neq 0$ *be arbitrary and set* $v_1 := \frac{b}{\|b\|_2}$.
FOR $j = 1, \ldots, m$
$z := Av_j$
$h_{ij} := \langle v_i, z \rangle,\ i = 1, \ldots, j$
$w_j := z - \sum_{i=1}^{j} h_{ij} v_i$
$h_{j+1,j} := \|w_j\|_2$

IF $h_{j+1,j} = 0$: ***STOP***
ELSE $v_{j+1} := \frac{w_j}{h_{j+1,j}}$
END
END

5.2. Lancozos Iteration & Biorthogonalization Algorithm

In case that A is symmetric, it is readily seen that

$$H_m = H_m^T = \begin{pmatrix} \alpha_1 & \beta_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{pmatrix} \text{ with } \beta_i > 0 \ \forall i. \tag{5.1}$$

The Arnoldi Algorithm simplifies to the so-called *Lanczos Iteration Algorithm*, which requires the determination of maximal three entries per iteration.

Algorithm 5.2 (*Lanczos Iteration Algorithm*)

Let $b \neq 0$ *be arbitrary and set* $v_1 := \dfrac{b}{\|b\|_2}$, $v_0 = 0$ *and* $\beta_0 = 0$.
FOR $j = 1, \ldots, m$
$z := Av_j$
$\alpha_j := \langle v_j, z \rangle$
$z := z - \alpha_j v_j - \beta_{j-1} v_{j-1}$
$\beta_j := \|z\|_2$

IF $\beta_j = 0$: ***STOP***
ELSE $v_{j+1} := \dfrac{z}{\beta_j}$
END
END

This idea can be generalized to the so-called *Biorthogonalization Algorithm*. If we insist on getting a tridiagonal H_m, even if A is not symmetric, we need to give up the use of unitary transformation matrices.

Let $A = VTV^{-1}$ for a nonsingular, but generally not unitary V and a tridiagonal T. If we define $W := V^{-T}$ and take the transpose of A, we receive the equivalent equation $A^T = WT^TW^{-1}$ and it is obvious that $W^TV = V^{-1}V = I$. Though the columns of V do not form an orthogonal basis, they are orthogonal to the columns of W. The central idea of the Biorthogonalization Algorithm is to find such matrices V and W with *biorthogonal* columns.

In the view of the Arnoldi and Lanczos Iteration, our aim is to determine matrices

$$V_m = (v_1, \ldots, v_m), \quad W_m = (w_1, \ldots, w_m)$$

such that

$$W_m^T V_m = I \quad \text{and} \quad W_m^T A V_m = H_m = \begin{pmatrix} \alpha_1 & \gamma_1 & & \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_{m-1} \\ & & \beta_{m-1} & \alpha_m \end{pmatrix}.$$

We will find such matrices by computing two biorthogonal bases

$$v_1, \ldots, v_m \text{ of } \mathcal{K}_m(A, v_1) \quad \text{and} \quad w_1, \ldots, w_m \text{ of } \mathcal{K}_m(A^T, w_1).$$

Again, Biorthogonalization can be performed by a modification of Gram-Schmidt.

Given $A \in \mathbb{R}^{n \times n}$ and $v_1, w_1 \in R^n$ with $\langle v_1, w_1 \rangle = 1$, we set

$$\tilde{v}_{k+1} := Av_k - \sum_{j=1}^{k} \langle Av_k, w_j \rangle v_j, \quad \tilde{w}_{k+1} := A^T w_k - \sum_{j=1}^{k} \langle A^T w_k, v_j \rangle w_j$$

and

$$v_{k+1} := \frac{\tilde{v}_{k+1}}{\beta_k}, \quad w_{k+1} := \frac{\tilde{w}_{k+1}}{\gamma_k}$$

such that

$$\frac{\langle \tilde{v}_{k+1}, \tilde{w}_{k+1} \rangle}{\beta_k \gamma_k} = \langle v_{k+1}, w_{k+1} \rangle = 1.$$

It is easy to see, that $v_k \perp span\{w_1, \ldots, w_{k-1}\} = \mathcal{K}_{k-1}(A^T, w_1)$, why

$$\sum_{j=1}^{k-2} \langle Av_k, w_j \rangle v_j = \sum_{j=1}^{k-2} \langle A^T w_k, v_j \rangle w_j = 0.$$

Then by defining

$$\alpha_k := \langle Av_k, w_k \rangle$$

H_m attains the desired tridiagonal form.

As before, since

$$V_m W_m^T x = x,$$

for $x \in \mathcal{K}_m(A^T, v_1)$, we can interpret H_m as the matrix representation of the projection of A onto $\mathcal{K}_m(A^T, v_1)$.

The Biorthogonalization Algorithm can be described efficiently as follows.

Algorithm 5.3 (*Biorthogonalization Algorithm*)

Let $v_1, w_1 \neq 0$ *be arbitrary with* $\langle v_1, w_1 \rangle = 1$. *Set* $\beta_0 = \gamma_0 = 0$ *and* $v_0 = w_0 = 0$.
FOR $j = 1, \ldots, m$

$$\begin{aligned}
v_{j+1} &:= Av_j \\
w_{j+1} &:= A^T w_j \\
\alpha_j &:= \langle v_{j+1}, w_j \rangle \\
v_{j+1} &:= v_{j+1} - \alpha_j v_j - \beta_{j-1} v_{j-1} \\
w_{j+1} &:= w_{j+1} - \alpha_j w_j - \gamma_{j-1} w_{j-1} \\
\beta_j &:= |\langle v_{k+1}, w_{k+1} \rangle|^{\frac{1}{2}} \\
\gamma_j &:= sign(\langle v_{k+1}, w_{k+1} \rangle) \beta_j
\end{aligned}$$

IF $\beta_j = 0$: ***STOP***

ELSE $v_{j+1} := \frac{v_{j+1}}{\beta_j}$, $w_{j+1} := \frac{w_{j+1}}{\gamma_j}$

END

END

5.3. Model Reduction via Coefficient Matching

Now we are interested in how the forgoing theory applies to model order reduction of linear control systems. For this purpose let us consider SISO-systems represented as

$$G: \begin{cases} \dot{x}(t) = Ax(t) + bu(t), \\ y(t) = cx(t) \end{cases} \tag{5.2}$$

where $A \in \mathbb{R}^{n\times n}$ and $b, c^T \in \mathbb{R}^n$.

By applying Arnoldi iteration to $\mathcal{K}_m(A,b)$ for some $m < n$ with

$$rg(V_m) = \mathcal{K}_m(A,b) \quad \text{and} \quad V_m^T V_m = I,$$

we can define a reduced order system as

$$\hat{A} := V_m^T A V_m \in \mathbb{R}^{m\times m}, \quad \hat{b} := V_m^T b \in \mathbb{R}^m \quad \text{and} \quad \hat{c} = cV_m \in \mathbb{R}^{1\times m}.$$

For the reduced order system $(\hat{A}, \hat{b}, \hat{c})$ we will not be able to find an error bound or guarantee its stability, instead the motivation behind this procedure is given by (2.7) and the following theorem.

Theorem 5.1 (*Markov Coefficient Matching by Arnoldi*)

If V_m is obtained from the application of Arnoldi's Iteration Algorithm to A and b, then

$$\hat{c}\hat{A}^{k-1}\hat{b} = cA^{k-1}b, \ k = 1, \dots, m$$

i.e. the first m Markov coefficients of $(\hat{A}, \hat{b}, \hat{c})$ and (A, b, c) match.[4]

Proof: In (5.1) we have seen, that $V_m V_m^T$ is the orthogonal projection of $\mathbb{R}^n$ onto $\mathcal{K}_m(A,b)$. Thus

$$b \in rg(V_m) \ \Rightarrow \ V_m\hat{b} = V_m V_m^T b = b.$$

We can continue in the same and get

$$Ab \in rg(V_m) \ \Rightarrow \ V_m\hat{A}\hat{b} = V_m V_m^T A V_m V_m^T b = V_m V_m^T Ab = Ab$$

$$\vdots$$

$$A^{k-1}b \in rg(V_m) \ \Rightarrow \ V_m\hat{A}^{k-1}\hat{b} = V_m V_m^T A^{k-1} b = A^{k-1}b, \ 1 \le k \le m$$

In conclusion, since $\hat{c} = cV_m$ it follows

$$\hat{c}\hat{A}^{k-1}\hat{b} = cV_m\hat{A}^{k-1}\hat{b} = cA^{k-1}b, \ k = 1, \dots, m$$

which concludes the proof. ■

If $\hat{G}(s)$ denotes the transfer function of $(\hat{A}, \hat{b}, \hat{c})$, then with the help of (2.7), the H_∞ error between $G(s)$ and $\hat{G}(s)$ can be given as

$$\|G - \hat{G}\|_\infty = \left\| \sum_{k=m+1}^{\infty} \frac{\hat{c}\hat{A}^{k-1}\hat{b} - cA^{k-1}b}{s^k} \right\|_\infty .$$

Thus, if the Markov coefficients of $G(s)$ and $\hat{G}(s)$ reach their limits of zero sufficiently fast, a small error can be attained.

Notice, the range of the Kalman controllability matrix is equal to $\mathcal{K}_n(A, b)$. Hence, if the state-space realization of a system is not completely controllable, there must exist a maximal $m < n$, such that $rk(\mathcal{K}_m(A, b)) = m$. In this case $\mathcal{K}_m(A, b)$ is A-invariant and we get

$$AV_m = V_m H_m.$$

This is exactly the same problem as the construction of a *Kalman Controllability Decompositon*. If $\{w_1, \ldots, w_{n-m}\}$ is chosen such that $\{v_1, \ldots, v_m, w_1, \ldots, w_{n-m}\}$ is an orthonormal basis of $\mathbb{R}^n$, we can define

$$T := \begin{pmatrix} V_m & W \end{pmatrix} := \begin{pmatrix} V_m & w_1 & \cdots & w_{n-m} \end{pmatrix}$$

and express AW as

$$AW = VA_2 + WA_3.$$

Thus

$$AT = \begin{pmatrix} AV_m & AW \end{pmatrix} = \begin{pmatrix} AV_m & VA_2 + WA_3 \end{pmatrix} = \begin{pmatrix} V_m & W \end{pmatrix} \begin{pmatrix} H_m & A_2 \\ 0 & A_3 \end{pmatrix}.$$

Since $b \in \mathcal{K}_m(A, b)$ we get

$$b = V_m b_1 = \begin{pmatrix} V_m & W \end{pmatrix} \begin{pmatrix} b_1 \\ 0 \end{pmatrix}$$

and consequently

$$T^T AT = \begin{pmatrix} \hat{A} & A_2 \\ 0 & A_3 \end{pmatrix} \quad \text{and} \quad T^T B = \begin{pmatrix} \hat{b} \\ 0 \end{pmatrix}.$$

In conclusion, Arnoldi's Algorithm can be used in order to remove uncontrollable states, which does not cause any error. By doing the same to the transposed system, we can go on and remove unobservable states.
Further, from the proof to Theorem 5.1 we know, that

$$V_m \begin{pmatrix} \hat{b} & \hat{A}\hat{b} & \cdots & \hat{A}^{m-1}\hat{b} \end{pmatrix} = \begin{pmatrix} b & Ab & \cdots & A^{m-1}b \end{pmatrix},$$

and hence

$$rk(\begin{pmatrix} \hat{b} & \hat{A}\hat{b} & \cdots & \hat{A}^{m-1}\hat{b} \end{pmatrix}) = m.$$

This means together with the controllability of $(\hat{A}, \hat{b}, \hat{c})$, that a reduced system, resulting from Arnoldi's Algorithm, is always controllable.

From the construction of the Biorthogonalization Algorithm we know, that Biorthogonalization can take care of both, the controllable and the observable subspace. Thus, instead of applying Arnoldi's Algorithm successively to remove uncontrollable and unobservable states, the idea could be to use Biorthogonalization with

$$v_1 := \frac{b}{\sqrt{|bc|}} \quad \text{and} \quad w_1 := \frac{c^T}{\sqrt{|bc|}},$$

and define then a reduced system as

$$\hat{A} := W_m^T A V_m, \quad \hat{b} := W_m^T b \quad \text{and} \quad \hat{c} := c V_m.$$

In this case we would expect a lower error compared to Arnoldi's Algorithm, which can indeed be motivated by the statement of the next theorem.

Theorem 5.2 (*Markov Coefficient Matching by Biorthogonalization*)

If V_m and W_m are obtained from the application of the Biorthogonalization Algorithm to A and (5.3) , then

$$\hat{c}\hat{A}^{k-1}\hat{b} = cA^{k-1}b, \ k = 1, \ldots, 2m$$

for $\hat{A}$, $\hat{b}$, $\hat{c}$ as defined in (5.3).[4]

Proof: Since $b \in rg(V_m)$ and $W_m^T V_m = I$, it follows immediately by the choice of v_1 and w_1, that

$$V_m \hat{b} = V_m W_m^T b = V_m \begin{pmatrix} w_1^T b \neq 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = v_1 w_1^T b = \frac{b}{\sqrt{|bc|}} \frac{c}{\sqrt{|bc|}} b = b$$

Consequently, we can conclude

$$\begin{aligned} Ab \in rg(V_m) \ &\Rightarrow \ V_m \hat{A}\hat{b} = V_m W_m^T A V_m W_m^T b = V_m W_m^T A b = Ab \\ &\vdots \\ A^{k-1}b \in rg(V_m) \ &\Rightarrow \ V_m \hat{A}^{k-1}\hat{b} = V_m W_m^T A^{k-1} b = A^{k-1} b, \ 1 \leq k \leq m \end{aligned}$$

The same can be done with c, i.e. since $c^T \in rg(W_m)$ and $V_m^T W_m = I$, it follows, that

$$\hat{c} W_m^T = c V_m W_m^T = (V_m^T c^T)^T W_m = c.$$

Thus

$$\begin{aligned} (cA)^T \in rg(W_m) \ &\Rightarrow \ \hat{c}\hat{A} W_m^T = c V_m W_m^T A V_m W_m^T = cA V_m W_m^T = (V_m^T (cA)^T) W_m^T = cA \\ &\vdots \\ (cA^{l-1})^T \in rg(W_m) \ &\Rightarrow \ \hat{c}\hat{A}^{l-1} W_m^T = cA^{l-1} V_m W_m^T = cA^{l-1}, \ 1 \leq l \leq m \end{aligned}$$

Summarizing both results leads to

$$\hat{c}\hat{A}^{l+k-1}\hat{b} = \hat{c}\hat{A}^{l-1}W_m^T\hat{A}V_m\hat{A}^{k-1}\hat{b} = cA^{l-1}AA^{k-1}b = cA^{k+l-1}b,\ 1 \leq k,l \leq m$$

which concludes the proof. ■

As before, the proof shows, that the reduced system $(\hat{A},\hat{b},\hat{c})$ is always controllable. Additionally we have

$$\begin{pmatrix}\hat{c}^T & \hat{A}^T\hat{c}^T & \cdots & (\hat{A}^T)^{m-1}\hat{c}\end{pmatrix}^T W_m^T = \begin{pmatrix}c^T & A^Tc^T & \cdots & (A^T)^{m-1}c\end{pmatrix}^T$$

which implies that $(\hat{A},\hat{b},\hat{c})$ is also observable.

Comparing the results of numerical experiments indicate, that Balanced Truncation generally performs a great deal better, regardless of whether Arnoldi or Biorthogonalization has been used.

Nevertheless, those approaches are iterative methods, which means that an approximation can be attained very efficiently. Unfortunately, neither Arnoldi nor Biorthogonalziation guarantees the stability of the reduced model.

Example 5.1 (Unstable Biorthogonalization)
Let us consider the following system

$$A := \begin{pmatrix}-11 & 3 & 4\\ 3 & -9 & 5\\ 4 & 5 & -19\end{pmatrix},\quad b := \begin{pmatrix}5\\5\\0\end{pmatrix},\quad c := \begin{pmatrix}1 & 4 & 4\end{pmatrix}.$$

Then $cAb = 20$ and Biorthogonalization cannot even attain a stable first order approximation.

For the sake of completeness, we want to mention, that there is a simple way of dealing with MIMO-systems, which is called *Block Arnoldi Algorithm* and can be outlined as follows.

If $B \in \mathbb{R}^{n\times q}$ with $B = \begin{pmatrix}b_1 & \cdots & b_q\end{pmatrix}$, then Arnoldi iteration will be applied separately to A and b_i in order to get the matrices $V_m^{(i)}$. Subsequently, those matrices will be collected in one matrix $V := \begin{pmatrix}V_m^{(1)} & \cdots & V_m^{(q)}\end{pmatrix}$. Choosing an orthonormal basis of V can be done by the computation of a reduced QR-factorization of V. Finally, we choose the first m columns of Q as V_m to define the reduced system.

This method can work for positive MIMO systems, but generally it does not preserve all the useful properties, that we have in case of a SISO system.

5.4. Coefficient Matching for Positive Systems

As for Balanced Truncation, Coefficient Matching generally does not preserve internal positivity.

Example 5.2 (External positivity not preserved)
If consider the system

$$A := \begin{pmatrix} -26 & 3 & 9 \\ 3 & -23 & 7 \\ 7 & 7 & -31 \end{pmatrix}, \quad b := \begin{pmatrix} 9.6 \\ 2.1 \\ 4.6 \end{pmatrix}, \quad c := \begin{pmatrix} 4.9 & 9.9 & 1.7 \end{pmatrix},$$

then Biorthogonalization leads to a second-order system

$$A := \begin{pmatrix} -12.78 & -4.10 \\ 4.10 & -14.52 \end{pmatrix}, \quad b = c^T := \begin{pmatrix} 8.70 \\ 0 \end{pmatrix},$$

which has poles in $-13.65 \pm 4.00i$. Thus, by Lemma 1.2, this is neither an external nor an internally positive system.

Still, in many examples of linear positive SISO systems, Coefficient Matching performs fairly well. In the following we will find an explanation for this behaviour, which is given by some interesting properties of the just presented Coefficient Matching methods.
As mentioned in the introduction to this chapter, large-scale system usually possess a sparse A matrix, as they result from the discretization of partial differential equations. In Chapter 7 we will see, that those discretization matrices often have the additional property of being symmetric.
In this case we know from (5.1), that Coefficient Matching by Arnoldi results in a symmetric Metzler matrix $\hat{A}$. If A is asymptotically stable and $\hat{b}, \hat{c} \geqq 0$, we have found a positive approximation.

Observe, if A is symmetric and asymptotically stable, then A must be negative definite. By *Lanczos Iteration Algorithm* we can conclude, that $\alpha_j = v_j^T A v_j < 0$. Since the first column of V_m is given by $v_1 = \dfrac{b}{\|b\|_2} \geqq 0$, we can give the following analogue to Theorem 3.3.

Theorem 5.3 (*First Order Positive Coefficient Matching*)

Let $(\hat{A}_1, \hat{b}_1, \hat{c}_1)$ be the reduced first order system attained from Coefficient Matching by Arnoldi for a positive system (A, b, c). Then $(\hat{A}_1, \hat{b}_1, \hat{c}_1)$ is always positive and asymptotic stability can be guaranteed in case that $A = A^T$ or $b^T A b < 0$.

Through out all the discussed reduction methods in the chapters before, the Metzler Matrix property was basically the main difficulty. Here we haven seen, that the symmetry of A makes it very likely, that *Coefficient Matching by Arnoldi* results in an asymptotically approximation, where $\hat{A}$ is guaranteed to be a Metzler Matrix.
The nonnegativity of $\hat{b}$ is easily assured by the fact, that

$$\hat{b} = V_m^T b = \begin{pmatrix} \|b\|_2 & 0 & \cdots & 0 \end{pmatrix}^T \geqq 0.$$

Hence, we are left with the problem of a nonnegative $\hat{c}$, which generally cannot be guaranteed. However, as we have just seen for $\hat{b}$, if $c^T = kb$ with $k \geq 0$, then $\hat{c} \geqq 0$.

Theorem 5.4

Let (A, b, c) be a positive system with $A = A^T$ and $b = kc^T$ for $k > 0$. Then Coefficient Matching by Arnoldi always results in an asymptotically stable positive system $(\hat{A}, \hat{b}, \hat{c})$.

Theorem 5.4 naturally transfers to the use of Biorthogonalization. However, by using Biorthogonalization, the symmetry of A does not assure the asymptotic stability of the first order reduced system, as seen in Example 5.1. Instead we require $cAb < 0$. Further, Biorthogonalization always assures the nonnegativity of $\hat{b}$ and $\hat{c}$ by $W_m^T V_m = I$. Nevertheless, if $A = A^T$ and $b \neq kc^T$ we need to pay the price, that we might lose the Metzler Matrix property of $\hat{A}$ for a very low order.

Example 5.3 (Low order Biorthogonalization)
Consider the system

$$A := \begin{pmatrix} -9 & 4 & 0 & 0 \\ 4 & -12 & 3 & 0 \\ 0 & 3 & -6 & 1 \\ 0 & 0 & 1 & -4 \end{pmatrix}, \quad b := \begin{pmatrix} 3 \\ 3 \\ 0 \\ 1 \end{pmatrix}, \quad c := \begin{pmatrix} 1 & 2 & 4 & 1 \end{pmatrix},$$

then Biorthogonalization yields a second-order system

$$A_b := \begin{pmatrix} -2.70 & -1.90 \\ 1.90 & -5.90 \end{pmatrix}, \quad b_b = c^T := \begin{pmatrix} 3.16 \\ 0 \end{pmatrix},$$

which is clearly not symmetric. In contrast Arnoldi gives

$$A_a := \begin{pmatrix} -6.37 & 2.78 & 0 \\ 2.78 & -10.03 & 4.97 \\ 0 & 4.97 & -9.89 \end{pmatrix}, \quad b_a := \begin{pmatrix} 4.36 \\ 0 \\ 0 \end{pmatrix}, \quad c_a := \begin{pmatrix} 2.29 & 3.03 & 2.59 \end{pmatrix}.$$

The first-order approximation by Biorthogonalization leads to a relative error of 0.17, whereas the third-order system obtained by Arnoldi gives $4.39 \cdot 10^{-3}$. Thus, there are examples for positive systems, where Arnoldi performs a great deal better, than Biorthogonalization, though the approximation was stable.

In Chapter 7 we will discuss examples, where Biorthogonalization preserves the positivity for any reduced order and Arnoldi has to stop at a much earlier stage, due to the violation of $c \geqq 0$.

In case of a non-symmetric A, Arnoldi will usually not return a symmetric $\hat{A}$, because the consideration of the additional entries in H_m makes it very unlikely to preserve the positivity for higher orders. Here, Biorthogonalization is clearly preferable because of its guaranteed band matrix structure with $\beta_j = \pm\gamma_j$. A direct consequence of this is, if a

system is transformed by a non-unitary transformation matrix, then Coefficient Matching does mostly not results in a positive system. In Chapter 6 we will see how to treat cases where the symmetry properties are destroyed or not given a priori.

Notice, Theorem 5.4 holds still true, if b and c contain negative elements. Hence, any system with $A = A^T$, $b = kc^T$ and $k > 0$ must be a positive system and can be realized by Arnoldi/Lanczos. In this sense, Arnoldi/Biorthogonalization/Lanczos cannot only be used for approximation, but also interpreted as positive realization algorithms. In the view of Chapter 2, where we have seen, that a positive realization for systems of high orders is hard to obtain, this is an observation of great importance.

A further interesting property arises, when we focus on the external positivity. Since our methods match the first m, respectively $2m$ Markov coefficients, we know from (1.8), that the error between the impulse responses of G and $\hat{G}$ can be given as

$$|g(t) - \hat{g}(t)| = \left| \sum_{k=m+1}^{\infty} \hat{g}_i - g_i \frac{t^{(k-1)}}{(k-1)!} \right| \leq \sum_{k=m+1}^{\infty} |\hat{g}_i - g_i| \frac{t^{(k-1)}}{(k-1)!}$$

As for the H_∞-error we can conclude, if the Markov coefficients reach their limits of zero sufficiently fast, a small error can be obtained and thus external positivity preserved. Unfortunately, we have seen in Example 5.2, that also for these methods we can find counter examples for the preservation of external positivity.

5.5. Iterative Rational Krylov Algorithm

An alternative way of preserving the external positivity of a SISO-system can be found by minimizing the error between the outputs of the original and the reduced system. This is the central idea of the so-called $\mathcal{H}_2$ Model Order Reduction [12].

Let

$$G_r : \begin{cases} \dot{x_r}(t) = A_r x_r(t) + b_r u(t), \\ y_r(t) = c_r x_r(t) \end{cases}$$

be a reduced-order approximation of 5.2, with $A_r \in \mathbb{R}^{r \times r}$ and $b_r, c_r^T \in \mathbb{R}^n$. If $u(t)$ is an input, such that $\|u\|_2 \leq 1$, then we can describe the error between the output of $G(s)$

and $G_r(s)$ by the inverse Laplace-transformation as follows

$$\begin{aligned}\max_{t>0}|y(t)-y_r(t)| &= \max_{t>0}\left|\frac{1}{2\pi}\int_{-\infty}^{\infty}(Y(i\omega)-Y_r(i\omega))e^{i\omega t}d\omega\right| \leq \frac{1}{2\pi}\int_{-\infty}^{\infty}|Y(i\omega)-Y_r(i\omega)|d\omega \\ &\leq \left(\frac{1}{2\pi}\int_{-\infty}^{\infty}|G(i\omega)-G_r(i\omega)|^2d\omega\right)^{\frac{1}{2}}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}|U(i\omega)|^2d\omega\right)^{\frac{1}{2}} \\ &\leq \left(\frac{1}{2\pi}\int_{-\infty}^{\infty}|G(i\omega)-G_r(i\omega)|^2d\omega\right)^{\frac{1}{2}}\|u\|_2 \\ &\leq \left(\frac{1}{2\pi}\int_{-\infty}^{\infty}|G(i\omega)-G_r(i\omega)|^2d\omega\right)^{\frac{1}{2}} =: \|G-G_r\|_{\mathcal{H}_2}.\end{aligned}$$

The $\mathcal{H}_2$-norm results from the scalar product of the well-known $\mathcal{L}_2(i\mathbb{R})$ space [30], which is a Hilbert space consisting of all matrix-valued functions G on $i\mathbb{R}$ fulfilling

$$\int_{-\infty}^{\infty} tr(\overline{G(i\omega)}G(i\omega))d\omega < \infty.$$

The scalar product of this space is given by

$$\langle G,H\rangle := \frac{1}{2\pi}\int_{-\infty}^{\infty} tr(\overline{G(i\omega)}H(i\omega))d\omega$$

for $G,H \in \mathcal{L}_2(i\mathbb{R})$. It can be shown [30], that the set of all matrix functions $G(s) \in \mathcal{L}_2(i\mathbb{R})$, which are analytic in the open right half plane, $\Re(s) > 0$, builds a closed subspace of $\mathcal{L}_2(i\mathbb{R})$, which is called $\mathcal{H}_2$. Moreover, the set of all strictly proper and real rational transfer functions represents a subspace of $\mathcal{H}_2$ [30].
Due to our restriction to real stable SISO-systems we define

$$\langle G,H\rangle_{\mathcal{H}_2} := \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{G(i\omega)}H(i\omega)d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty}G(-i\omega)H(i\omega)d\omega.$$

and

$$\|G\|_{\mathcal{H}_2} = \sqrt{\langle G,H\rangle_{\mathcal{H}_2}}.$$

Notice, since $\langle G,H\rangle_{\mathcal{H}_2} = \langle H,G\rangle_{\mathcal{H}_2}$ it follows, that $\langle G,H\rangle_{\mathcal{H}_2}$ must be real.

In the following we will present a method, which is called *Iterative Rational Krylov Algorithm* (IRKA) [12]. This method locally minimizes $\|G-G_r\|_{\mathcal{H}_2}$, aiming to preserve stability and due to its relation to Krylov subspaces, it can be applied to large-scale systems.

The main idea behind the *Iterative Rational Krylov Algorithm* is a *Moment Matching* approach called *Rational Interpolation*, which we want to discuss now. Moment Matching consists of finding a reduced system G_r that interpolates the values of $G(s)$, and maybe additionally some derivative values, at given points $\{\sigma_1,\ldots,\sigma_r\} \subset \mathbb{C}\setminus\sigma(A)$, called *shifts*.

Since the Iterative Rational Krylov Algorithm uses simple Hermite interpolation, our problem reduces to the location of G_r, so that

$$G_r(\sigma_k) = G(\sigma_k) \quad \text{and} \quad G_r^{'}(\sigma_k) = G_r^{'}(\sigma_k),\ 1 \leq k \leq r$$

or equivalently

$$c_r(\sigma_k I - A_r)^{-1} b_r = c(\sigma_k I - A)^{-1} b \quad \text{and} \quad c_r(\sigma_k I - A_r)^{-2} b_r = c(\sigma_k I - A)^{-2} b,\ 1 \leq k \leq r.$$

The expression $c(\sigma_k I - A)^{-(j+1)} b$ is called the j-th moment of $G(s)$ in σ_k and we want to match the first two moments. We will see, that this problem can be solved iteratively and is strongly related to Arnoldi/Biorthogonalization.

In order to show this we consider a reduced order model G_r constructed by the so-called *Galerkin approximation*. The Galerkin approximation has its origin in the solution approximation of partial differential equations [26] and represents a generalization of the Krylov subspace methods discussed in Section 5.3. In case of a linear system this method works as follows. Let $\mathcal{V}_r$ and $\mathcal{W}_r$ be given r-dimensional subspaces of $\mathbb{R}^n$, such that $\mathcal{V}_r \cap \mathcal{W}_r^{\perp} = \{0\}$, where $\mathcal{W}_r^{\perp}$ denotes the orthogonal complement of $\mathcal{W}_r$. The idea is to find an $v(t) \in \mathcal{V}_r$, such that

$$\dot{v}(t) - Av(t) - bu(t) \perp W_r \text{ for all } u(t). \tag{5.3}$$

In this case the output of the reduced order system is defined as $y_r := cv(t)$. Let $V_r, W_r \in \mathbb{R}^{n \times r}$ denote matrices whose columns consist of a basis of $\mathcal{V}_r$, respectively $\mathcal{W}_r$. Then we can write $v(t) = V_r x_r(t)$ with $x_r(t) \in \mathbb{R}^r$ and we get from (5.3), that

$$W_r^T(V_r \dot{x}_r - AV_r x_r(t) - bu(t)) = 0.$$

This leads to a reduced order system given by

$$A_r := (W_r^T V_r)^{-1} W_r^T A V_r, \quad b_r := (W_r^T V_r)^{-1} W_r^T b \quad \text{and} \quad c_r := cV_r. \tag{5.4}$$

Observe, the nonsingularity of $W_r^T V_r$ is a direct consequence of $\mathcal{V}_r \cap \mathcal{W}_r^{\perp} = \{0\}$, because this is equivalent to the kernel, $ker(W_r^T V_r)$, of $W_r^T V_r$ being trivial. Further, if we require the biorthogonality of W_r and V_r, then by choosing $\mathcal{V}_r = \mathcal{K}_r(A, b)$ and $\mathcal{W}_r = \mathcal{K}_r(A^T, c^T)$, we get the same results as for Biorthogonalization/Arnoldi. With the following Lemma we will relate this procedure to our interpolation problem.

Lemma 5.1

If $\sigma \in \mathbb{C} \setminus \{\sigma(A) \cup \sigma(A_r)\}$, then the following statements hold.

(i) If $(\sigma I - A)^{-1} b \in \mathcal{V}_r$, then $G_r(\sigma) = G(\sigma)$.

(ii) If $(\overline{\sigma} I - A^T)^{-1} c^T \in \mathcal{W}_r$, then $G_r(\sigma) = G(\sigma)$.

(iii) If both assumptions of (i) and (ii) hold, then $G_r(\sigma) = G(\sigma)$ and $G_r'(\sigma) = G'(\sigma)$.

Proof: Our aim is to express $G(\sigma)-G_r(\sigma)$ in terms of $(\sigma I-A)^{-1}b$ and $(\sigma I-A^T)^{-1}c^T$. This can be done efficiently by identifying $\mathcal{V}_r$ and $\mathcal{W}_r$ by means of the following linear projections

$$P_r(\sigma) := V_r(\sigma I - A_r)^{-1}(W_r^T V_r)^{-1} W_r^T(\sigma I - A)$$

and

$$\tilde{P}_r(\sigma) := (\sigma I - A)P_r(\sigma)(\sigma I - A)^{-1}.$$

To verify that these linear mappings are indeed projections, we need to notice, that

$$\begin{aligned}(\sigma I - A_r)^{-1}(W_r^T V_r)^{-1} W_r^T(\sigma I - A)V_r &= (\sigma I - A_r)^{-1}(\sigma I - W_r^T A V_r)\\ &= (\sigma I - A_r)^{-1}(\sigma I - A_r) = I.\end{aligned}$$

Thus

$$\begin{aligned}P_r(\sigma)^2 &= V_r[(\sigma I - A_r)^{-1}(W_r^T V_r)^{-1} W_r^T(\sigma I - A)V_r](\sigma I - A_r)^{-1}(W_r^T V_r)^{-1} W_r^T(\sigma I - A)\\ &= V_r(\sigma I - A_r)^{-1}(W_r^T V_r)^{-1} W_r^T(\sigma I - A) = P_r(\sigma)\end{aligned}$$

and $\tilde{P}_r(\sigma)^2 = \tilde{P}_r(\sigma)$. Further, if $v \in \mathcal{V}_r$, we can express it as $v = V_r x$ with $x \in \mathbb{R}^r$ and

$$P_r(\sigma)v = V_r(\sigma I - A_r)^{-1}(W_r^T V_r)^{-1} W_r^T(\sigma I - A)V_r x = V_r x = v$$

Hence, $P_r(\sigma)$ is a linear projection on $\mathcal{V}_r$ and $\mathcal{V}_r = rg(P_r(\sigma))$. In the same way we can show, that $\overline{\tilde{P}_r(\sigma)}^T$ is a linear projection on $\mathcal{W}_r$ and $\mathcal{W}_r = rg\left(\overline{\tilde{P}_r(\sigma)}^T\right)$. Consequently, since $\mathcal{W}_r^\perp$ is the orthogonal complement of $\mathcal{W}_r$ we can conclude, that $\mathcal{W}_r^\perp = ker(\tilde{P}_r(\sigma))$. By rewriting $G_r(\sigma)$ as

$$G_r(\sigma) = cP_r(\sigma)(\sigma I - A)^{-1}b = c(\sigma I - A)^{-1}\tilde{P}_r(\sigma)b$$

we attain

$$\begin{aligned}G(\sigma) - G_r(\sigma) &= c(I - P_r(\sigma))(\sigma I - A)^{-1}b\\ &= c(\sigma I - A)^{-1}(I - \tilde{P}_r(\sigma))b\\ &= c(\sigma I - A)^{-1}(I - \tilde{P}_r(\sigma))^2 b\\ &= c(\sigma I - A)^{-1}(I - \tilde{P}_r(\sigma))(\sigma I - A)(I - P_r(\sigma))(\sigma I - A)^{-1}b \qquad (5.5)\end{aligned}$$

From the theory of linear projections [25], we know $rg(P_r(\sigma)) = ker(I - P_r(\sigma))$ and correspondingly $ker(\tilde{P}_r(\sigma)) = rg(I - \tilde{P}_r(\sigma))$, which shows together with (5.5), that (i) and (ii) hold.
In order to show the same for the derivatives, we need to notice, as long as σ is not an eigenvalue of A or A_r, we can consider $P_r(\sigma)$ and $\tilde{P}_r(\sigma)$ as matrix-valued functions in σ, which are analytic in a sufficiently small neighbourhood of σ. Then for sufficiently small ϵ it follows in the same way as before, that

$$\mathcal{V}_r = rg(P_r(\sigma+\epsilon)) \quad \text{and} \quad \mathcal{W}_r^\perp = rg(I - \tilde{P}_r(\sigma+\epsilon)).$$

Thus, together with the series expansion of $[(\sigma+\epsilon)I - A]^{-1}$

$$(\sigma I + \epsilon I - A)^{-1} = (\sigma I - A)^{-1} - \epsilon(\sigma I - A)^{-2} + \mathcal{O}(\epsilon^2)$$

we get

$$\begin{aligned} c[(\sigma+\epsilon)I - A]^{-1}(I - \tilde{P}_r(\sigma+\epsilon)) &= c[(\sigma I - A)^{-1} - \epsilon(\sigma I - A)^{-2} + \mathcal{O}(\epsilon^2)](I - \tilde{P}_r(\sigma+\epsilon)) \\ &= -\epsilon c(\sigma I - A)^{-2}(I - \tilde{P}_r(\sigma+\epsilon)) + \mathcal{O}(\epsilon^2) \end{aligned}$$

and analogously

$$(I - P_r(\sigma+\epsilon))[(\sigma+\epsilon)I - A]^{-1}b = -\epsilon(I - P_r(\sigma+\epsilon))(\sigma I - A)^{-2}b + \mathcal{O}(\epsilon^2).$$

Consequently by (5.5)

$$G(s) - G_r(s) = \mathcal{O}(\epsilon^2),$$

which concludes the proof. ■

Remark: For complex σ the $rg(V_r)$ and $rg(W_r)$ is considered over the complex space and therefore the restriction to $V_r, W_r \in \mathbb{R}^{n\times r}$ is feasible. Equivalently, the same method can be performed with complex-valued matrices V_r and W_r, where W_r^T is substituted by $\overline{W}_r^T$. However, in this case we would generally not end up with a real-valued state-space representation.

Theorem 5.5 (*Moment Matching*)

Let $G(s)$ be a linear system with state-space representation (A, b, c), $\{\sigma_1, \dots, \sigma_r\}$ a set of distinct shifts, which is closed under conjugation (i.e. all shifts occur in conjugate pairs) and

$$\begin{aligned} \mathcal{V}_r &:= span\{(\sigma_1 I - A)^{-1}b, \dots, (\sigma_r I - A)^{-1}b\} \\ \mathcal{W}_r &:= span\{(\sigma_1 I - A^T)^{-1}c^T, \dots, (\sigma_r I - A^T)^{-1}c^T\} \end{aligned}$$

linear subspaces in $\mathbb{C}^n$. Then by choosing real matrices V_r, W_r with $\mathcal{V}_r = rg(V_r)$ and $\mathcal{W}_r = rg(W_r)$, the defined reduced system $G_r(s)$ given in (5.4) matches the first two moments of $G(s)$ in σ_k for $k = 1, \dots, r$.

Remark: The vectors $(\sigma_i I - A)^{-1}b$ and $(\sigma_1 I - A^T)^{-1}c^T$ do not require the numerically expensive computation of the inverse. Instead a matrix factorization approach, such as LU-decomposition should be used. Further, the rank of W_r and V_r are not necessarily equal to the chosen reduced order. Thus, Moment Matching often achieves a reduced system of even smaller dimension.

The question is now, how to choose the shifts properly in order to minimize the $\mathcal{H}_2$-error. For this purpose we want to look at a residue description of the scalar product in $\mathcal{H}_2$.

Let $f(s)$ be a meromorphic function on an open set $D \in \mathbb{C}$, i.e. $f(s)$ is a complex function, which is holomorphic on D except for its poles. Then we denote by $Res_\lambda[f(s)]$ the residue of $f(s)$ at a pole λ and thus

$$Res_\lambda[f(s)] = \frac{1}{(k-1)!}\frac{d^{(k-1)}}{ds^{(k-1)}}[(s-\lambda)^k f(s)], \tag{5.6}$$

where k is the order of λ.

Lemma 5.2 (*Residue description of the $\mathcal{H}_2$-norm*)

Let $G(s)$ and $H(s)$ be two strictly proper, asymptotically stable transfer functions with poles $\{\lambda_1, \ldots, \lambda_n\}$, respectively $\{\mu_1, \ldots, \mu_m\}$, then

$$\langle G, H \rangle_{\mathcal{H}_2} = \sum_{k=1}^{m} Res_{\mu_k}[G(-s)H(s)] = \sum_{k=1}^{n} Res_{\lambda_k}[H(-s)G(s)].$$

Proof: The proof is just the application of the well-known Residue Theorem [8] to $G(-s)H(s)$. By assumption it is clear, that the only singularities of $G(-s)H(s)$ in the left half plane are the poles of $H(s)$. For sufficiently large $R > 0$, we can enclose them by the left half semicircular contour

$$\Gamma_R := \{z \in \mathbb{C} \mid z = i\omega \text{ with } \omega \in [-R, R]\} \cup \left\{ z \in \mathbb{C} \;\middle|\; z = Re^{i\theta} \text{ with } \theta \in \left[\frac{\pi}{2}, \frac{3\pi}{2}\right] \right\}.$$

If we define then $\gamma_R(\omega) := i\omega$ with $\omega \in [-R, R]$ we can write by the definition of the curve integral

$$\langle G, H \rangle_{\mathcal{H}_2} = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(-i\omega)H(i\omega)d\omega = \lim_{R\to\infty} \frac{1}{2i\pi} \int_{\gamma_R} G(-s)H(s)ds.$$

Further, since $G(s)$ and $H(s)$ are strictly proper, we can estimate

$$\lim_{R\to\infty} \left| \frac{1}{2\pi i} \int_{Re^{i\theta}} G(-s)H(s)ds \right| \leq \lim_{R\to\infty} \frac{1}{2\pi} \sup_{s=Re^{i\theta}} |G(-s)H(s)| R\pi = 0$$

and thus by the Residue Theorem

$$\langle G, H \rangle_{\mathcal{H}_2} = \lim_{R\to\infty} \frac{1}{2i\pi} \int_{\Gamma_R} G(-s)H(s)ds = \sum_{k=1}^{m} Res_{\mu_k}[G(-s)H(s)].$$

■

A direct consequence of Lemma 5.2 and (5.6) is the following corollary about the $\mathcal{H}_2$-norm, which we will use in order to explain the $\mathcal{H}_2$-error.

Corollary 5.1

If $G(s)$ is a strictly proper, asymptotically stable transfer function with simple poles $\{\lambda_1, \ldots, \lambda_n\}$, then

$$\|G\|_{\mathcal{H}_2} = \left(\sum_{k=1}^{n} Res_{\lambda_k}[G(s)]G(-\lambda_k) \right)^{\frac{1}{2}}.$$

Let $\{\lambda_1, \dots, \lambda_n\}$ and $\{\hat{\lambda}_1, \dots, \hat{\lambda}_r\}$ be the poles of $G(s)$ and a reduced order model $G_r(s)$, respectively, and assume that the poles of $G_r(s)$ are distinct. Further, let ϕ_i and $\hat{\phi}_j$ be defined as follows

$$\phi_i := Res_{\lambda_i}[G(s)] \quad \text{for } i = 1, \dots, n \quad \text{and} \quad \hat{\phi}_j := Res_{\hat{\lambda}_i}[G_r(s)] \quad \text{for } j = 1, \dots, r.$$

Then the $\mathcal{H}_2$-error of the approximation can be expressed as

$$\begin{aligned} \|G - G_r\|_{\mathcal{H}_2}^2 &= \sum_{i=1}^{n} Res_{\lambda_i}[(G(s) - G_r(s))(G(-s) - G_r(-s))] \\ &\quad + \sum_{j=1}^{r} Res_{\hat{\lambda}_j}[(G(s) - G_r(s))(G(-s) - G_r(-s))] \\ &= \sum_{i=1}^{n} \phi_i(G(-\lambda_i) - G_r(-\lambda_i)) + \sum_{j=1}^{r} \hat{\phi}_j(G(-\hat{\lambda}_i) - G_r(-\hat{\lambda}_i)). \end{aligned} \tag{5.7}$$

From (5.7) we observe, that the $\mathcal{H}_2$-error arises due to the mismatches of $G(s)$ and $G_r(s)$ at $-\lambda_i$ and $-\hat{\lambda}_i$. Since $-\hat{\lambda}_i$ is a priori unknown, the idea could be to set $\sigma_i := -\lambda_i$, where $\lambda_1, \dots, \lambda_r$ denote the poles with the largest residues. It has been shown in [11], that this selection of shifts performs quite well. However, in the following we will show, that the interpolation in $-\hat{\lambda}_i$ is of greater importance, because as we will see, this represents a necessary condition for the optimal $\mathcal{H}_2$ model reduction.

As mentioned initially, our aim is to find an r-dimensional reduced order system G_r, which is stable and optimal in the sense, that it minimizes the $\mathcal{H}_2$-error. Thus, if we define $\Omega_r := \{\hat{G}_r \mid \hat{G}_r \text{ stable with dimension r}\}$ we can write our optimization problem as

$$\|G - G_r\|_{\mathcal{H}_2} = \min_{\hat{G}_r \in \Omega_r} \|G - \hat{G}_r\|_{\mathcal{H}_2}. \tag{5.8}$$

Obviously, Ω_r is not a convex set, why (5.8) may possesses multiple *local minimizers* and as a practical matter the global minimizer is hard to obtain.

Definition 5.2 (*Local minimizer*)

A reduced system G_r is called a ***local minimizer****, if*

$$\|G - G_r\|_{\mathcal{H}_2} \leq \|G - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2},$$

for all sufficiently small $\epsilon > 0$ and for all r-dimensional stable dynamical systems $\hat{G}_r^{(\epsilon)}$ with $\|G_r - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2} = \mathcal{O}(\epsilon)$.

We already know, that $\mathcal{H}_2$ is a Hilbert space and thus one could think about using the well-known characterization of the element of best approximation in a Hilbert space [17]. Unfortunately, this is not possible, since Ω_r is not a closed subspace of $\mathcal{H}_2$. In order to overcome this problem, we have to restrict our solutions to a certain class, which is given in the next theorem.

Theorem 5.6

Let $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_r\}$ be a set of distinct points in the open left half plane and define $\mathcal{P}_r(\hat{\lambda})$ to be the set of all r-th order strictly proper rational functions with poles at $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_r\}$. Then the following holds

(i) $\mathcal{P}_r(\hat{\lambda})$ is a (closed) (r-1)-dimensional subspace of $\mathcal{H}_2$.

(ii) $G_r \in \mathcal{P}_r(\hat{\lambda})$ is an element of best approximation, i.e.

$$\|G - G_r\|_{\mathcal{H}_2} = \min_{\hat{G}_r \in \mathcal{P}_r(\hat{\lambda})} \|G - \hat{G}_r\|_{\mathcal{H}_2}$$

if and only if

$$\langle G - G_r, H \rangle_{\mathcal{H}_2} = 0 \quad \forall H \in \mathcal{P}_r(\hat{\lambda}).$$

Moreover, G_r exists and is unique.

Proof: The first statement is obvious due to the strict properness and a fixed denominator of the elements of $\mathcal{P}_r(\hat{\lambda})$. The second statement follows directly from the first one by the characterization of the element of best approximation in a Hilbert space [17]. ■

This result can be used to give a necessary condition for the optimality of a reduced order system G_r with simple poles.

Theorem 5.7 (*Local Minimizer*)

Let G_r be a local minimizer to G, possessing only simple poles. Then

$$\langle G - G_r, G_r H_1 + H_2 \rangle_{\mathcal{H}_2} = 0$$

for all real systems H_1 and H_2 having the same simple poles as G_r.

Proof: By writing

$$\langle G - G_r, G_r H_1 + H_2 \rangle_{\mathcal{H}_2} = \langle G - G_r, G_r H_1 \rangle_{\mathcal{H}_2} + \langle G - G_r, H_2 \rangle_{\mathcal{H}_2} = 0$$

we can observe by Theorem 5.6, that this is equivalent to $\langle G - G_r, G_r H_1 \rangle_{\mathcal{H}_2} = 0$. Further, let $\{\mu_1, \ldots, \mu_{m_r}\} \subset \mathbb{R}$ denote the real poles of $G_r(s)$ and $\{\mu_{m_r+1}, \ldots, \mu_{m_r+m_c}\} \subset \mathbb{C} \setminus \mathbb{R}$ the complex poles in the upper half plane. Then by partial fraction decomposition we can write

$$H_1(s) = \sum_{i=1}^{m_r} \frac{a_i}{s - \mu_i} + \sum_{i=m_r+1}^{m_r} \frac{b_i s + c_i}{(s - \mu_i)(s - \overline{\mu}_i)} = \sum_{i=1}^{m_r} \frac{a_i}{s - \mu_i} + \sum_{i=m_r+1}^{m_r} \frac{b_i(s - \alpha_i) + c_i}{(s - \alpha_i)^2 + \beta_i^2},$$

for some $a_i, b_i, c_i \in \mathbb{R}$ and $\mu_i = \alpha_i + i\beta_i$.

This allows us to express $\langle G - G_r, G_r H_1 \rangle_{\mathcal{H}_2}$ as follows

$$\begin{aligned}\langle G - G_r, G_r H_1 \rangle_{\mathcal{H}_2} = &\sum_{i=1}^{m_r} a_i \left\langle G - G_r, \frac{G_r(s)}{s-\mu_i} \right\rangle_{\mathcal{H}_2} \\ &+ \sum_{i=m_r+1}^{m_r+m_c} b_i \left\langle G - G_r, \frac{(s-\alpha_i)G_r(s)}{(s-\alpha_i)^2+\beta_i^2} \right\rangle_{\mathcal{H}_2} \\ &+ \sum_{i=m_r+1}^{m_r+m_c} c_i \left\langle G - G_r, \frac{G_r(s)}{(s-\alpha_i)^2+\beta_i^2} \right\rangle_{\mathcal{H}_2}.\end{aligned}$$

In the following we want to show, that each of these terms it equal to zero.

Let $\{\hat{G}_r^{(\epsilon)}\}$ be the set of real stable transfer functions as defined in Definition 5.2 with $\|G_r - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2} \leq C\epsilon$ for some constant $C > 0$. Then for all sufficiently small $\epsilon > 0$

$$\begin{aligned}\|G - G_r\|_{\mathcal{H}_2}^2 &\leq \|G - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2}^2 \leq \|(G - G_r) + (G_r - \hat{G}_r^{(\epsilon)})\|_{\mathcal{H}_2}^2 \\ &= \|G - G_r\|_{\mathcal{H}_2}^2 + 2\left\langle G - G_r, G_r - \hat{G}_r^{(\epsilon)} \right\rangle_{\mathcal{H}_2} + \|G_r - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2}^2.\end{aligned}$$

Hence, for all all sufficiently small $\epsilon > 0$

$$0 \leq 2\left\langle G - G_r, G_r - \hat{G}_r^{(\epsilon)} \right\rangle_{\mathcal{H}_2} + \|G_r - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2}^2. \tag{5.9}$$

Assume now, that

$$\left\langle G - G_r, \frac{G_r(s)}{s-\mu_i} \right\rangle_{\mathcal{H}_2} \neq 0.$$

By writing $G_r(s) = \frac{P_{r-1}(s)}{(s-\mu_i)Q_{r-1}(s)}$, for real polynomials $P_{r-1}(s)$ and $Q_{r-1}(s)$ of degree $r-1$, we can define

$$\hat{G}_r^{(\epsilon)}(s) := \frac{P_{r-1}(s)}{(s-\mu_i-(\pm\epsilon))Q_{r-1}(s)},$$

where the sign of $\pm\epsilon$ matches that of $\left\langle G - G_r, \frac{G_r(s)}{s-\mu_i} \right\rangle_{\mathcal{H}_2}$. Then series expansion of $\hat{G}_r^{(\epsilon)}$ yields

$$\hat{G}_r^{(\epsilon)}(s) = G_r(s) \pm \epsilon \frac{G_r(s)}{s-\mu_i} + \mathcal{O}(\epsilon^2),$$

which leads to

$$G_r(s) - \hat{G}_r^{(\epsilon)}(s) = \mp\epsilon \frac{G_r(s)}{s-\mu_i} + \mathcal{O}(\epsilon^2).$$

Consequently, we get

$$\left\langle G - G_r, G_r - \hat{G}_r^{(\epsilon)} \right\rangle_{\mathcal{H}_2} = -\epsilon \left| \left\langle G - G_r, \frac{G_r(s)}{s-\mu_i} \right\rangle_{\mathcal{H}_2} \right| + \mathcal{O}(\epsilon^2)$$

and

$$\|G_r - \hat{G}_r^{(\epsilon)}\|_{\mathcal{H}_2}^2 = \epsilon^2 \left\| \frac{G_r(s)}{s-\mu_i} \right\|_{\mathcal{H}_2}^2 + \mathcal{O}(\epsilon^3),$$

which gives together with (5.9)

$$0 < \left| \left\langle G - G_r, \frac{G_r(s)}{s - \mu_i} \right\rangle_{\mathcal{H}_2} \right| \leq \tilde{C}\epsilon$$

for some constant $\tilde{C} > 0$. Thus for $\epsilon \to 0$ we have a contradiction to (5.5).

Again by writing $G_r(s) = \frac{P_{r-1}(s)}{((s-\alpha_i)^2+\beta_i^2)Q_{r-2}(s)}$, for a real polynomial $Q_{r-2}(s)$ of degree $r-2$, we can show in the same way, that

$$\left\langle G - G_r, \frac{(s-\alpha_i)G_r(s)}{(s-\alpha_i)^2+\beta_i^2} \right\rangle_{\mathcal{H}_2} = 0 \quad \text{with} \quad \hat{G}_r^{(\epsilon)} := \frac{P_{r-1}(s)}{(s-\alpha_i-(\pm\epsilon))^2+\beta_i^2)Q_{r-2}}$$

and

$$\left\langle G - G_r, \frac{G_r(s)}{(s-\alpha_i)^2+\beta_i^2} \right\rangle_{\mathcal{H}_2} = 0 \quad \text{with} \quad \hat{G}_r^{(\epsilon)} := \frac{P_{r-1}(s)}{(s-\alpha_i)^2-(\pm\epsilon)+\beta_i^2)Q_{r-2}},$$

which concludes the proof. ■

In (5.7) we have already noticed the the importance of choosing the shifts as $\sigma_i := -\hat{\lambda}_i$ for a reduced system G_r with simple poles $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_r\}$. Together with Theorem 5.7 it can be shown, that this is a necessary condition for the optimal $\mathcal{H}_2$ model reduction.

Theorem 5.8

Let $G_r(s)$ be an r-dimensional minimizer of the optimal $\mathcal{H}_2$ model reduction problem given in (5.8) and assume the poles $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_r\}$ of $G_r(s)$ to be simple. Then $G_r(s)$ interpolates $G(s)$ and its derivatives at $\{-\hat{\lambda}_1, \ldots, -\hat{\lambda}_r\}$, i.e.

$$G_r(-\hat{\lambda}_i) = G(-\hat{\lambda}_i) \quad \text{and} \quad G_r'(-\hat{\lambda}_i) = G'(-\hat{\lambda}_i) \quad \text{for } i = 1, \ldots, r.$$

Proof: Applying Theorem 5.7 with $H_1 = 0$ and arbitrary H_2 yields

$$\begin{aligned} \langle G - G_r, H_2 \rangle_{\mathcal{H}_2} &= \sum_{i=1}^{r} Res_{\hat{\lambda}_i}[(G(-s) - G_r(-s))H_2(s)] \\ &= \sum_{i=1}^{r} Res_{\lambda_i}[H_2(s)](G(-\hat{\lambda}_i) - G_r(-\hat{\lambda}_i)) = 0 \end{aligned}$$

Since $Res_{\hat{\lambda}_i}[H_2(s)]$ is chosen arbitrarily we can conclude $G(-\hat{\lambda}_i) = G_r(-\hat{\lambda}_i)$. Now let us consider the case $H_2 = 0$ and H_1 arbitrary. Then by assumption of Theorem 5.7, $G_r(s)H_1(s)$ possesses a double pole in $\hat{\lambda}_i$ and together with $G(-\hat{\lambda}_i) = G_r(-\hat{\lambda}_i)$, we can write the residue

of $(G(-s) - G_r(-s))G_r(s)H_1(s)$ at $\hat{\lambda}_i$ as

$$\begin{aligned} Res_{\hat{\lambda}_i}[(G(-s) - G_r(-s))G_r(s)H_1(s)] &= \lim_{s\to\hat{\lambda}_i} \frac{d}{ds}[(s-\hat{\lambda}_i)^2(G(-s) - G_r(-s))G_r(s)H_1(s)] \\ &= \lim_{s\to\hat{\lambda}_i} (G(-s) - G_r(-s))\frac{d}{ds}[(s-\hat{\lambda}_i)^2 G_r(s)H_1(s)] \\ &\quad - \lim_{s\to\hat{\lambda}_i} (G(-s)^{'} - G_r(-s)^{'})[(s-\hat{\lambda}_i)^2 G_r(s)H_1(s)] \\ &= -(G^{'}(-\hat{\lambda}_i) - G_r^{'}(-\hat{\lambda}_i)) \lim_{s\to\hat{\lambda}_i} [(s-\hat{\lambda}_i)^2 G_r(s)H_1(s)] \\ &= -(G^{'}(-\hat{\lambda}_i) - G_r^{'}(-\hat{\lambda}_i)) Res_{\hat{\lambda}_i}[G_r(s)] Res_{\hat{\lambda}_i}[H_1(s)]. \end{aligned}$$

Thus, we have

$$\begin{aligned} \langle G - G_r, G_r H_1\rangle_{\mathcal{H}_2} &= \sum_{i=1}^{r} Res_{\hat{\lambda}_i}[(G(-s) - G_r(-s))G_r(s)H_1(s)] \\ &= \sum_{i=1}^{r} -(G^{'}(-\hat{\lambda}_i - G_r^{'}(-\hat{\lambda}_i)) Res_{\hat{\lambda}_i}[G_r(s)] Res_{\hat{\lambda}_i}[H_1(s)] = 0. \end{aligned}$$

As before, by the arbitrariness of $Res_{\hat{\lambda}_i}[H_1(s)]$ we conclude $G^{'}(-\hat{\lambda}_i) = G_r^{'}(-\hat{\lambda}_i)$. ■

Remark: Observe, by (5.6) it is readily seen, that we can obtain analogous results for the case of higher order poles, which correspond to the interpolation of higher derivatives.

Although we know now, that the interpolation of $G(s)$ at $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_r\}$ is necessary for the optimality of the $\mathcal{H}_2$-error, there is still the problem remaining, that $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_r\}$ are a priori unknown. In the following we will solve this problem with the help of the well-known *Newton's method*. For this purpose we have to rewrite our problem as a function of $\{\sigma_1, \ldots, \sigma_r\}$. Let us define

$$\boldsymbol{\sigma} := \begin{pmatrix} \sigma_1 & \cdots & \sigma_r \end{pmatrix}^T \quad \text{and} \quad \boldsymbol{\lambda}(\boldsymbol{\sigma}) := \begin{pmatrix} \tilde{\lambda}_1 & \cdots & \tilde{\lambda}_r \end{pmatrix}^T,$$

where $\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r\}$ denote the poles of G_r and G_r interpolates $G(s)$ and $G^{'}(s)$ at $\{\sigma_1, \ldots, \sigma_r\}$. Observe, $\boldsymbol{\lambda}(\boldsymbol{\sigma})$ defines a complex function from $\mathbb{C}^r \to \mathbb{C}^r$. By defining a complex function

$$\boldsymbol{g}(\boldsymbol{\sigma}) := \boldsymbol{\lambda}(\boldsymbol{\sigma}) + \boldsymbol{\sigma}.$$

from $\mathbb{C}^r \to \mathbb{C}^r$, we get for $\boldsymbol{g}(\boldsymbol{\sigma}) = 0$, that $\boldsymbol{\lambda}(\boldsymbol{\sigma}) = -\boldsymbol{\sigma}$, which is equivalent to Theorem 5.8. Thus $\boldsymbol{g}(\boldsymbol{\sigma})$ is the required candidate for Newton's method, which appears as

$$\boldsymbol{\sigma}_{k+1} = \boldsymbol{\sigma}_k - (I + J)^{-1}(\boldsymbol{\sigma}_k + \boldsymbol{\lambda}(\boldsymbol{\sigma}_k)),$$

where J is the Jacobian matrix of $\boldsymbol{\lambda}(\boldsymbol{\sigma})$.

In [12] it has been shown, that instead of computing J explicitly, $J = 0$ is a feasible choice, due to small entries of the Jacobian matrix in the neighbourhood of an $\mathcal{H}_2$ optimal $\boldsymbol{\sigma}$. This suggests the shift update strategy $\sigma_{k+1} = -\lambda_i(A_r)$ and leads to an *Iterative Rational Krylov Algorithm* (IRKA).

Algorithm 5.4 (***Iterative Rational Krylov Algorithm***)

(i) *Set $j = 0$ and choose initial shifts $\{\sigma_1^{(0)}, \ldots, \sigma_r^{(0)}\}$.*

(ii) *For any j choose real matrices V_r and W_r such that*

$$rg(V_r) := span\{(\sigma_1^{(j)} I - A)^{-1} b, \ldots, (\sigma_r^{(j)} I - A)^{-1} b\}$$
$$rg(W_r) := span\{(\sigma_1^{(j)} I - A^T)^{-1} c^T, \ldots, (\sigma_r^{(j)} I - A^T)^{-1} c^T\}$$

.

(iii) *Define $A_r := (W_r^T V_r)^{-1} W_r^T A V_r$ and set $\sigma_i^{(j)} := -\lambda_i(A_r)$ for $i = 1, \ldots, r$.*

(iv) *If $\dfrac{|\sigma_i^{(j)} - \sigma_i^{(j-1)}|}{|\sigma_i^{(j)}|} < TOL_\sigma$, for a prescribed tolerance TOL_σ, then an optimal G_r is given by*

$$A_r := (W_r^T V_r)^{-1} W_r^T A V_r, \quad b_r := (W_r^T V_r)^{-1} W_r^T b, \quad c_r := c V_r.$$

Otherwise, set $j := j + 1$ and continue with step (ii).

As explained earlier, a reasonable choice of the initial shifts could be $\sigma_i := -\lambda_i$, where $\lambda_1, \ldots, \lambda_r$ denote those poles with the largest residue. Unfortunately, the determination of the residues can be very expensive numerically in case of large-scale systems.

Another approach of choosing the initial shifts is to generate them randomly. During numerical experiments this turned out to work very efficient. IRKA always converged to a stable solution after a small number of iterations.

Instead of using $J = 0$, J can also be computed explicitly as shown in the end of [12], which we will not discuss here.

Compared to the Coefficient Matching approaches of Section 5.3, IRKA provides similar good results as balanced truncation and results in stable approximations. In many examples IRKA preserves the external positivity of a system, especially when it comes to sparse matrices. Still, as for the other Krylov subspaces method this does generally not hold true, as shown in the following example.

Example 5.4 (IRKA not externally positive)
By considering the same example as in Example 3.4, IRKA results in the second order system

$$A_2 := \begin{pmatrix} -2.83 & -0.32 \\ 0.35 & -2.56 \end{pmatrix}, \quad b_2 := \begin{pmatrix} -3.55 \\ 0.23 \end{pmatrix}, \quad c_2 := \begin{pmatrix} -5.04 & -3.92 \end{pmatrix},$$

with poles at $-2.70 \pm 0.31i$. Thus the reduced system cannot be externally positive.

6. Symmetric Balanced Truncation

In practice a way to combine Balanced Truncation with Krylov subspace methods is to reduce a system with order much greater than 1000 to $m \approx 1000$ and then apply balanced truncation to attain a system of order $r < 100$.[4] Unfortunately, neither balanced truncation nor Krylov subspace methods have to result in a positive system. Apart from this combination of Krylov subspace methods and Balanced Truncation, we will present a further relation in this chapter with focus on the positive realization property of the Krylov subspaces methods. To this end we describe a symmetry characterization of balanced SISO-systems, which can be used to obtain an extension of Theorem 3.3 to higher order approximations. Moreover, an algorithm will be proposed in order to use this result in the context of large-scale positive systems.

In Section 5.4 we have observed the advantage of dealing with systems that consist of a symmetric A-matrix. Further, we noticed, by showing how to use Arnoldi/Lanczos for the purpose of obtaining a minimal realization, that the determination of a minimal realization can always preserve the symmetry property of A. Hence, it naturally arises the question, if Balanced Truncation, which can be used to attain a minimal realization, does also preserve the symmetry.

In general we can answer to this question with no, as seen in Example 3.4. Instead let us start our investigation with a situation similar to Theorem 5.4. If we have a system (A, B, C), not necessarily SISO, with $A = A^T$ and $B = kC^T$ for some $k > 0$, we call the system *symmetric*. It follows immediately from (3.2) and (3.4), that

$$P = \int_0^\infty e^{At} B B^T e^{A^T t} dt = k^2 \int_0^\infty e^{A^T t} C^T C e^{At} dt = k^2 Q.$$

By diagonalization of kP as $kP = T^T \Sigma T$ we can write $PQ = k^2 P^2 = \tilde{T}^{-1} \Sigma^2 \tilde{T}$ with $\tilde{T} = \frac{1}{\sqrt{|k|}} T$. Obviously, $\tilde{T}$ is a balancing transformation matrix and consequently the balanced system is given by

$$(A_b, B_b, C_b) := (\tilde{T}^{-1} A T, \tilde{T}^{-1} B, CT) = (T^T A T, \sqrt{|k|}(CT)^T, \sqrt{|k|} CT).$$

Observe, balancing the system did not only preserve the symmetry property of A, it also added $B_b = C_b^T$, i.e. (A_b, B_b, C_b) is a symmetric system with $k = 1$. Such a system is sometimes called *state-space symmetric* [15]. Thus, every symmetric system can be identified with a state-space symmetric one. The application of the truncation step to (A_b, B_b, C_b) leads to the following lemma.

Lemma 6.1 (*Balance Truncation of symmetric systems*)

Balanced Truncation preserves the symmetry of any symmetric system.

In case of a SISO-system Lemma 6.1 this equivalent, to what we have we have seen in Section 5.4. For what follows, we summarize this result in the following theorem.

Theorem 6.1 (*Balanced Truncation of symmetric systems*)

Let (A, b, c) be a symmetric SISO-system, then Balanced Truncation followed by Lanczos will always result in a positive reduced-order system.

Theorem 6.1 motivates the question, how much a small perturbation of the symmetry effects the symmetry of the balanced system. To that end let us consider the following example.

Example 6.1 (Symmetry Perturbation)
Let us consider for instance

$$A := \begin{pmatrix} -2 & 1+0.1 \\ 1 & -2 \end{pmatrix}, \quad b = c^T := \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Then balancing yield the symmetric system

$$A := \begin{pmatrix} -1.39 & -0.85 \\ -0.85 & -2.62 \end{pmatrix}, \quad b = c^T := \begin{pmatrix} -0.95 & -0.31 \end{pmatrix}$$

Obviously, for some perturbations of the symmetry the balanced system remains symmetric. Since every system can be considered a symmetric system with some "large" perturbation, we know from Example 3.4 that this is not always true. The reason for both cases is a consequence of the next theorem, which presents a symmetry characterization regarding balanced SISO-systems and is the central idea of this chapter.

Theorem 6.2 (*Absolute symmetry of a balanced system*)

Let $G(s)$ be the transfer function of an arbitrary SISO-system. Then there exists a balanced realization (A, b, c) of $G(s)$, such that (A, b, c) is ***absolutely symmetric****, i.e.*

$$|A| = |A^T| \quad \textit{and} \quad |b| = |c^T|.$$

Proof: Let (A, b, c) have simple Hankel singular values $\{\sigma_1, \ldots, \sigma_n\}$. Then by definition of a balanced system, its Lyapunov equations can be written as

$$\begin{aligned} A\Sigma + \Sigma A^T = -bb^T \quad &\Leftrightarrow \quad a_{ij}\sigma_j + \sigma_i a_{ji} = -b_i b_j, \quad 1 \le i, j \le n \\ A^T\Sigma + \Sigma A = -c^T c \quad &\Leftrightarrow \quad a_{ij}\sigma_i + \sigma_j a_{ji} = -c_i c_j, \quad 1 \le i, j \le n \end{aligned} \tag{6.1}$$

with $\Sigma := diag(\sigma_1, \ldots, \sigma_n)$. In particular, we get for $i = j$

$$2a_{ii}\sigma_i = -b_i^2 = -c_i^2 \quad \Rightarrow \quad b_i = \pm c_i \quad 1 \leq i \leq n. \tag{6.2}$$

Further, if $i \neq j$ we can deduce from (6.1)

$$\begin{pmatrix} \sigma_j & \sigma_i \\ \sigma_i & \sigma_j \end{pmatrix} \begin{pmatrix} a_{ij} \\ a_{ji} \end{pmatrix} = \begin{pmatrix} b_i b_j \\ c_i c_j \end{pmatrix}.$$

Solving this linear system for $\begin{pmatrix} a_{ij} & a_{ji} \end{pmatrix}^T$, yields together with (6.2)

$$\begin{pmatrix} a_{ij} \\ a_{ji} \end{pmatrix} = \frac{1}{\sigma_j^2 - \sigma_i^2} \begin{pmatrix} \sigma_j & -\sigma_i \\ -\sigma_i & \sigma_j \end{pmatrix} \begin{pmatrix} b_i b_j \\ \pm b_i b_j \end{pmatrix} = \frac{b_i b_j}{\sigma_j^2 - \sigma_i^2} \begin{pmatrix} \sigma_j \mp \sigma_i \\ -\sigma_i \pm \sigma_j \end{pmatrix} = \frac{b_i b_j}{\sigma_j^2 - \sigma_i^2} \begin{pmatrix} \sigma_j \mp \sigma_i \\ \pm(\sigma_j \mp \sigma_i) \end{pmatrix}$$

and hence $a_{ij} = \pm a_{ji}$.

In case of multiple Hankel singular values we can assume w.l.o.g., that $\Sigma := diag(\sigma_1 I_k, \sigma_2, \ldots, \sigma_n)$ for some $k > 1$. Then by partitioning $A = \begin{pmatrix} A_1 & * \\ * & * \end{pmatrix}$ and $b = \begin{pmatrix} B_1 \\ * \end{pmatrix}$ accordingly to $\sigma_1 I_k$, we can write

$$\sigma_1(A_1 + A_1^T) = B_1 B_1^T.$$

Thus diagonalizing $B_1 B_1^T = U^T \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix} U$, with $\lambda > 0$, yields

$$\sigma_1(UA_1U^T + UA_1^T U^T) = UB_1B_1^T U^T = \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix}$$

and it follows for $\tilde{A} := UA_1U^T$, that $\tilde{a}_{ij} = -\tilde{a}_{ji}$, $1 \leq i, j \leq k$. Defining $T := diag(U, I)$ gives a balanced absolute symmetric realization

$$(\tilde{A}, \tilde{b}, \tilde{c}) := (TAT^T, TB, CT^T),$$

which concludes the proof. ■

The important consequence of Theorem 6.2 is, if Balanced Truncation of a positive SISO-system is performed up to an order where the reduced system is still symmetric, then Lanczos will return a positive approximation. In the following we will refer to this method as *Symmetric Balanced Truncation*. In worst case this procedure just ends up with a first order positive approximation, which is why Symmetric Balanced Truncation can be considered an extension of Theorem 3.3.

For the identification of the symmetric part of the balanced realization it is not necessary to look at A itself. Instead it can be concluded from the proof to Theorem 6.2 that $b_i b_j = -c_i c_j$ if and only if $a_{ij} = -a_{ji}$. Thus in case of a positive system, the i-th leading principle minor of A becomes non-symmetric, when $b_i = -c_i$ for the first time.

Compared to Arnoldi/Biorthogonlization and Generalized Balanced Truncation, this procedure works for any transfer function, independent of its state-space representation. Thus we lose the importance of requirements such as the internal positivity or the

symmetry of a system.
Still, we know from Example 3.4, that the symmetry can only be a sufficient condition. Beyond the possibility of applying Symmetric Balanced Truncation to positive systems, it can also be used to attain positive approximations of systems, that are not even externally positive. This could be of interest e.g. if the model of an (externally) positive system was attained by system identification and contains small errors, which violate the positivity.

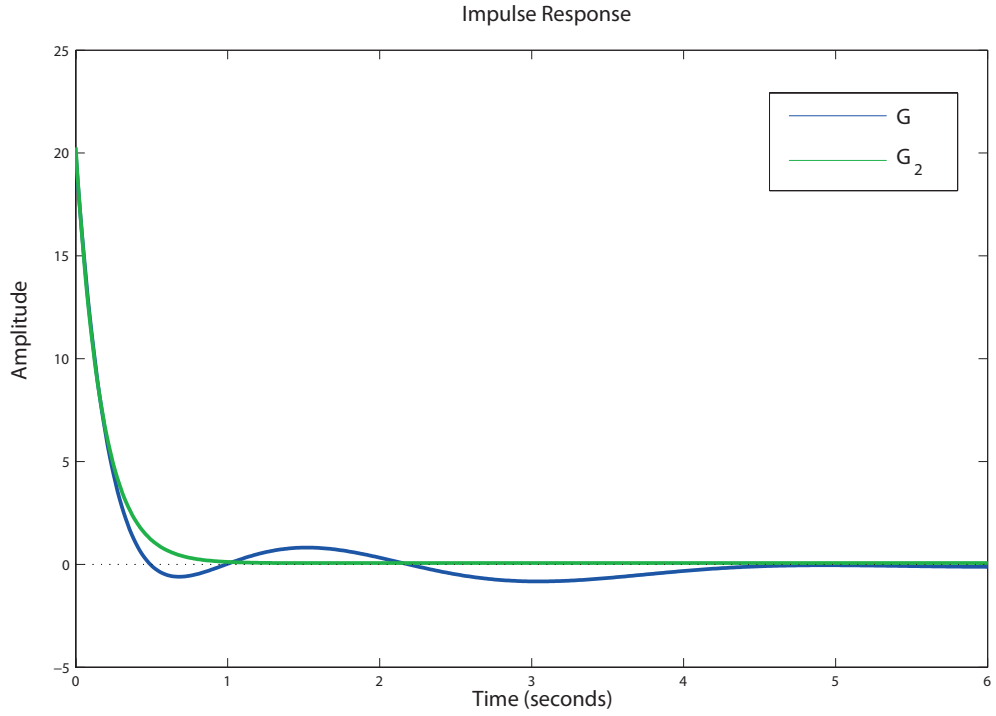

Figure 6.1.: Impulse response of a nonpositive system G with positve approximaton G_2

Example 6.2 (Nonpositive system)
Let us consider the following system

$$A := \begin{pmatrix} -2 & 0 & 2 & -4 \\ 0 & 0 & -1 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & -4 \end{pmatrix}, \quad b = c^T := \begin{pmatrix} -2 \\ 0 \\ 0 \\ -4 \end{pmatrix}.$$

The impulse response of the system can be seen in Figure 6.1 together with its positive

second-order approximation

$$A_2 := \begin{pmatrix} -5.77 & 0.31 \\ 0.31 & -0.05 \end{pmatrix}, \quad b_2 = c_2^T := \begin{pmatrix} 4.50 \\ 0 \end{pmatrix}.$$

However, in general we can say it is very unlikely to receive a positive approximation of a nonpositive system. This is reasonable, since in the same way as the preservation of the positivity, the maintenance of the nonpositivity is natural a requirement on its reduced model.

Observe, for $\tilde{A}$ in the proof to Theorem 6.2, we can also conclude, that $a_{ii} = 0$ for $i = 2, \ldots, k$. In case of a non-zero two-dimensional system $G(s)$, that has only one singular value, we can obtain a balanced realization

$$A = \begin{pmatrix} a_{11} & a_{12} \\ -a_{12} & 0 \end{pmatrix}, \quad b_1 = \begin{pmatrix} b_{11} \\ 0 \end{pmatrix}, \quad c_1 = \begin{pmatrix} b_{11} & 0 \end{pmatrix}.$$

If this was a positive system, it would hold $\|G\|_\infty = G(0) = -cA^{-1}b$. Since

$$A^{-1} = \frac{1}{a_{12}} \begin{pmatrix} 0 & -a_{12} \\ a_{12} & a_{11} \end{pmatrix}$$

it follows, that $G(0) = -cA^{-1}b = 0$, which is clearly a contradiction. This motivates the following conjecture, which has been proved only numerically so far.

Conjecture 6.1 (*Simple Hankel singular values*)

There does not exist a positive SISO-system which possesses a multiple largest Hankel singular value.

6.1. Symmetric Balanced Truncation Algorithm

By numerical experiments and intuition it can be observed, that especially in presents of many (dominant) real poles, such as for sparse systems, a higher dimension of the symmetric part can be expected. As mentioned in the introduction to Chapter 5, those systems occur very often in the context of discretized partial differential equation, which usually have a very large dimension and possess a symmetric A from the beginning. In order to use Symmetric Balanced Truncation, we have to reduce the system to dimension $m \approx 1000$, which allows us to apply Balanced Truncation. A natural desire of this pre-approximation should be, that the dimension of its balanced symmetric part is not decreased compared to the original balanced system, unless the error of both symmetric reduced systems is of the same quality.

In the end of Section 5.5 we said, that the Iterative Rational Krylov Algorithm performs comparable well as Balanced Truncation itself. Indeed, it turned out, during numerical experiments, that a pre-approximation via IRKA does not add any significant drawback in the context of Symmetric Balanced Truncation. Thus we can complete our method of Symmetric Balanced Truncation yielding the following algorithm.

Algorithm 6.1 (***Symmetric Balanced Truncation Algorithm***)

(i) For a given positive system of dimension n, choose m such that $m = \min\{n, 1000\}$ and apply IRKA starting with m random shifts. Denote the resulting approximation with G_m.

(ii) Compute a balanced realization (A_b, b_b, c_b) of G_m.

(iii) Compare the entries of b_b and c_b in order to identify the smallest k, where $b_{b_k} \neq c_{b_k}$.

(iv) If $k = 2$, perform the truncation of (A_b, b_b, c_b) to obtain a reduced system G_2 of order 2. Then apply Theorem 2.7.

(v) If Theorem 2.7 does not apply, perform the truncation of (A_b, b_b, c_b) to obtain a reduced symmetric system G_{k-1} of the order $k-1$. Then attain a positive realization of G_{k-1} with the help of Lanczos Iteration Algorithm.

Step (iv) is included, since not every second-order positive system needs to be symmetric after balancing, as seen in the next example.

Example 6.3 (Non-symmetric positive system)
Let a system (A, b, c) be given by

$$A := \begin{pmatrix} -9 & 5 \\ 5 & -10 \end{pmatrix}, \quad b := \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \quad c := \begin{pmatrix} 5 & 5 \end{pmatrix}.$$

By balancing this system we get

$$A := \begin{pmatrix} -4.37 & 1.01 \\ -1.01 & -14.63 \end{pmatrix}, \quad b := \begin{pmatrix} -3.90 \\ -0.45 \end{pmatrix}, \quad c := \begin{pmatrix} -3.90 & 0.45 \end{pmatrix}$$

For the sake of completeness, we should notice, if the Hankel Singular Values are close to each other, the symmetric part can be increased sometimes by a permutation of the balanced system states. However, so far we could not find an example, where a permutation made any significant difference.

6.2. Symmetric Balanced Truncation for MIMO-systems

In case of a positive MIMO-system Symmetric Balanced Truncation can usually not be applied. Of course one reasons is, that we use Lanczos Algorithm for the positive realization. But more important is, that Theorem 6.2 does not hold generally for MIMO-systems. Still, there are cases, where this procedure can be transferred.

Let us consider an n-dimensional positive MISO-system with balanced realization (A, B, c)

and $B = \begin{pmatrix} b_1 & \cdots & b_k \end{pmatrix}$. With $\Sigma = diag\left(\sigma_1, \ldots, \sigma_n\right)$ we can write the Lyapunov equation (3.27) as

$$A\Sigma + \Sigma A^T = BB^T = b_1 b_1^T + \cdots + b_k b_k^T.$$

By assuming, that $b_1 b_1^T = \cdots = b_k b_k^T$, we can conclude

$$A\Sigma + \Sigma A^T = k b_i b_i^T = (\sqrt{k} b_i)(\sqrt{k} b_i)^T,\ 1 \leq i \leq k \quad \text{and} \quad A^T\Sigma + \Sigma A = c^T c.$$

From Theorem 6.2 it follows then, that $|\sqrt{k} b_i| = \sqrt{k}|b_i| = |c^T|$ and $|A| = |A^T|$. Hence, if (A_r, B_r, c_r) denotes the reduced system of order r, with symmetric A_r, then Lanczos applied to (A_r, b_{r_i}, c_r) yields identical positive systems (A_p, b_{p_i}, c_p) for all $i = 1, \ldots, k$. Thus a positive realization of (A_r, B_r, c_r) can be obtained as (A, B_p, c_p), where $B_p = \begin{pmatrix} b_{p_1} & \cdots & b_{p_k} \end{pmatrix}$.

Observe, if we partition $A = \begin{pmatrix} A_{11} & * \\ * & * \end{pmatrix}$ and accordingly $b_i = \begin{pmatrix} b_{i1} \\ * \end{pmatrix}$ and $c = \begin{pmatrix} c_1 & * \end{pmatrix}$, it suffices to have $b_{11} b_{11}^T = \cdots = b_{k1} b_{k1}^T$ in order to get $|\sqrt{k} b_{i1}| = \sqrt{k}|b_{i1}| = |c_1^T|$ and $|A_{11}| = |A_{11}^T|$. Thus, we can proceed in the same way as before and observe, that it is not necessary to have a positive system with a B-matrix consisting of identical columns. Of course, the same applies in case of a SIMO-or MIMO-system, where we need to partition $C = \begin{pmatrix} c_1^T & \cdots & c_m^T \end{pmatrix}^T$ and assume $c_1 = \cdots = c_m$ or analogously for $c_i = \begin{pmatrix} c_{i1} & * \end{pmatrix}$, that $c_{11} c_{11}^T = \cdots = c_{m1} c_{m1}^T$.

7. Numerical Examples

In Chapter 3-6 we have studied many different model reduction approaches, that satisfy our aim, preservation of the positivity. All of them possess some theoretical advantages and limitations, which has influence on the quality of their approximations in the view of the relative $\mathcal{H}_\infty$-error. Based on some practical examples, this chapter will give a comparison of the quality among all the methods.

The examples presented here were run in MATLAB® Version 7.10.0 on a PC with with an Intel®Core™ i5-650 CPU 3.20 GHz. Moreover, YALMIP Version 3 and SeDuMi Version 1.3 were used for the optimizations with respect to the linear matrix inequalities. For methods which involve LMIs, we will not be able to show their performance for large-scale systems, due to the limits of computational power.

Throughout the examples we choose the following tolerances and maximal iterations:

- IRKA: $MAX_{iter} = 100$ and $TOL_\sigma = 1 \cdot 10^{-8}$.
- Generalized Balanced Truncation: $TOL_\alpha = 0.01$.
- ILMI I: $TOL_\alpha = TOL_\beta = TOL_{init} = 0.01$ and $Max_{iter} = 500$.
- ILMI II: $TOL_\delta = 1 \cdot 10^{-8}$ and $Max_{iter} = 1000$.

These levels turned out to be sufficient and do not add any significant disadvantage during the optimization of the shifts and the Generalized Hankel Singular Values. Furthermore, we distinguish between "Symmetric Balanced Truncation (IRKA)" using IRKA and "Symmetric Balanced Truncation" by direct balancing.

Of course, for ILMI I & II we cannot say if our tolerances are sufficient, but this is a general problem of these methods. Beside this, it should be noticed, for smaller tolerances and larger number of iterations, the algorithms become very time-consuming, which can also be considered as a limit of computational power. Under this consideration we will stop ILMI I after 1 hour of switching between its dual and its primal approach.

Furthermore, ILMI I & II require a prescribed error bound. We start with a relative error of 0.1, because in the view of Theorem 3.3, everything above this border would not justify to use such numerically expensive methods. For the same reason, we will always decrease the order of the prescribed error bound in case of an increasing dimension of the approximation. For example, if we attain a first order reduced system with an error of 0.04, then we will prescribe an error bound of 0.01 for the second order approximation.

For the sake of fairness we include step (iv) of the Symmetric Balanced Truncation Algorithm into Arnoldi/Lanczos and Biorthogonalization.

7.1. Water reservoirs

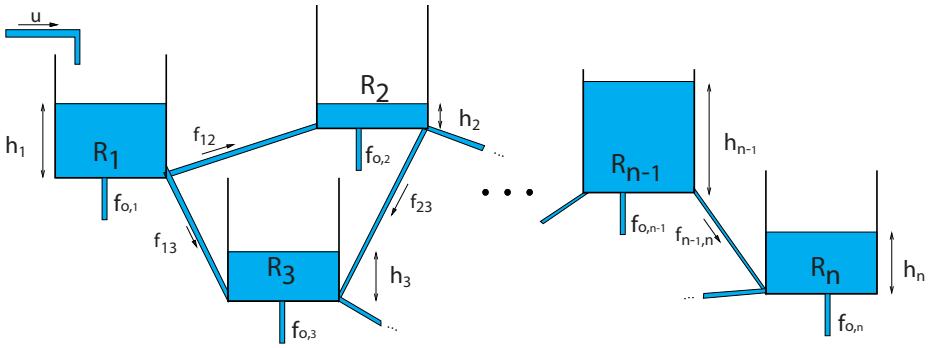

Figure 7.1.: System of n water reservoirs

The example of n connected water reservoirs as schematically shown in Figure 7.1, was presented in [22], in order to show the properties of *Generalized Balanced Truncation*. For simplicity all the reservoirs $R_1, \ldots, R_n$ are assumed to be located on the same level, i.e the connection between two water reservoirs is always horizontal. We denote with a_i and h_i the base area and the fill level of the reservoir R_i, respectively. Moreover, let R_i and R_j be connected by a pipe of diameter $d_{ij} = d_{ji} \geq 0$, then the direct flow f_{ij} from R_i to R_j is assumed to be linearly dependent of the pressure difference on both ends. We consider the external inflow to reservoir R_1 as the input of the system. The output is given by the sum of all outflows $f_{o,i}$ of R_i through a pipe with diameter $d_{o,i}$. With the help of Pascal's law we can describe the system flows by

$$f_{ij}(t) = d_{ij}^2 \cdot k \cdot (h_i(t) - h_j(t)) \quad \text{and} \quad f_{o,i}(t) = d_{o,i}^2 \cdot k \cdot (h_i(t) - h_j(t)),$$

where k is a constant representing gravity as well as viscosity and density of the medium. Thus, the fill level h_i of R_i follows the differential equation

$$\dot{h}_i = \frac{k}{a_i} \left(-d_{o,i}^2 h_i(t) + \sum_{j=1}^{n} d_{ij}^2 (h_j(t) - h_i(t)) \right) + \frac{1}{a_i} \delta_{1i} u(t),$$

where δ_{1i} stands for the Delta-Kronecker symbol, i.e. $\delta_{1i} = 1$ if and only if $i = 1$ and zero otherwise. Writing this equations as linear state-space system results in a SISO-system (A, b, c) with $b = \begin{pmatrix} \frac{1}{a_1} & 0 & \cdots & 0 \end{pmatrix}^T$, $c = k \begin{pmatrix} d_{o,1}^2 & \cdots & d_{o,n}^2 \end{pmatrix}$ and a symmetric A-matrix with entries

$$a_{ij} := \frac{k}{a_i} \begin{cases} -d_{o,i}^2 - \sum_{m=1}^{n} d_{im}^2, & i = j \\ d_{ij}^2 & i \neq j, \end{cases} \quad \text{with} \quad d_{ii} := 0.$$

In [22] the system was supposed to consist of two substructures, each with five reservoirs. In both substructures each reservoir is assumed be connected to every other reservoir by a pipe of diameter 1, i.e $d_{ij} = 1$ for $i \neq i$, $1 \leq i, j \leq 5$ and $6 \leq i, j \leq 10$, respectively. The connection of the substructures is given by a pipe of diameter $d_{1,10} = d_{10,1} = 0.2$, between reservoir 1 and 10. Moreover, for simplicity we set $a_i = 1$ and $k = 1$.

For a reduced order of 5 an error bound of 0.06 was given in [22] by applying Generalized Balanced Truncation. This coincides with our results and can be compared with the relative $\mathcal{H}_\infty$-errors of the other methods, given in the following tabular.

Order	1	4	5
Generalized Balanced Truncation	0.80	0.51	0.01
Symmetric Balanced Truncation (IRKA)	0	-	-
Symmetric Balanced Truncation	0	-	-
ILMI I	$7.65 \cdot 10^{-5}$	-	-
ILMI II	0.02	-	-
Arnoldi/Lanczos	0.80	0	-
Biorthogonalization	0	-	-

The results show, that the system is actually of first order and thus by Theorem 3.3 Symmetric Balanced Truncation returns a first-order positive approximation without causing an error. The same holds for Biorthogonalization because of its minimal realization property. Similarly, Arnoldi/Lanczos has removed the uncontrollable states. Notice, by Theorem 5.3 it follows, that if we apply Arnoldi/Lanczos to the transposed of its approximation, we also end up with a positive minimal system of first order.

With this knowledge, even the fifth order reduced system resulting from Generalized Balanced Truncation must be considered as a poor approximation. By inheriting the difficulties of truncating an unbalanced system, Generalized Balanced Truncation cannot perform any better. For our other examples this will become even more significant.

Although, ILMI I & II are numerically very expensive, in case of such a small system they terminated within half an hour. We observe, ILMI I performs a great deal better than ILMI II. Unfortunately, in both cases, we could not find a realization of order 4 or 5 with a significantly smaller error.

The first-order property of this system is clearly a result of the strong connection and the homogeneity within the substructures. Therefore an increased dimension (amount of reservoirs) does not change the order. Instead, let us consider a slight modification of the outflows, i.e. we set $d_{o,i} = 0.1 \cdot i$. This system is not of first order any longer and thus we expect larger errors, as shown in the next tabular.

Order	1	2	5
Generalized Balanced Truncation	1.00	0.98	0.08
Symmetric Balanced Truncation (IRKA)	0.02	$1.99 \cdot 10^{-3}$	-
Symmetric Balanced Truncation	0.02	$1.99 \cdot 10^{-3}$	-
ILMI I	0.08	-	-
ILMI II	-	-	-
Arnoldi/Lanczos	1.00	$1.68 \cdot 10^{-2}$	-
Biorthogonalization	-	-	-

For Biorthogonalization the first order model is already unstable and hence, this method is not applicable. In contrast, Symmetric Balanced Truncation led to the best results and there is no difference between IRKA and Balanced Truncation, though the system is not just reduced to minimality.

ILMI II could not return a model with an error, that is smaller than 0.1. On the other hand, ILMI I gives a quite good first order reduced system, but again no second order approximation with a smaller error could be found. Beside this, for Arnolid as well as for Symmetric Balanced Truncation, the highest achievable order for a positive reduced system is restricted to 2.

As a last point let us consider the same system with $n = 250$. In this case the optimizations for ILMI I & II take such a long time, that the first optimization already requires hours. Thus, we are left with Balanced Truncation and the Krylov subspaces methods, which perform as follows.

Order	1	2	100
Generalized Balanced Truncation	1.00	0.99	0.45
Symmetric Balanced Truncation (IRKA)	0.13	$1.51 \cdot 10^{-3}$	-
Symmetric Balanced Truncation	0.13	$1.51 \cdot 10^{-3}$	-
Arnoldi/Lanczos	1.00	0.15	-
Biorthogonalization	-	-	-

The important consequence of the results in this section is, that among all the methods, only Symmetric Balanced Truncation is robust with respect to an increasing order.
The reason why we could still apply Generalized Balanced Truncation for such a medium scale system with 250 states is, that for a system with n states, we only need to optimize $2n$ variables with $2n$ rows. As seen in the end of Chapter 4, this gives a complexity of $\mathcal{O}((2n)^2(2n)^{2.5} + (2n)^{3.5})$, which is already extremely high. In case of the ILMI-approaches this becomes even worse because of the large amount of rows, which leads to the exceeding of computational power.

7.2. Compartmental Networks

Compartmental networks built a general class of systems that consist of a finite number of homogeneous subsystems (compartments), which all interact with each other and their environment.[14] Representatively, the interaction of two compartments is schematically

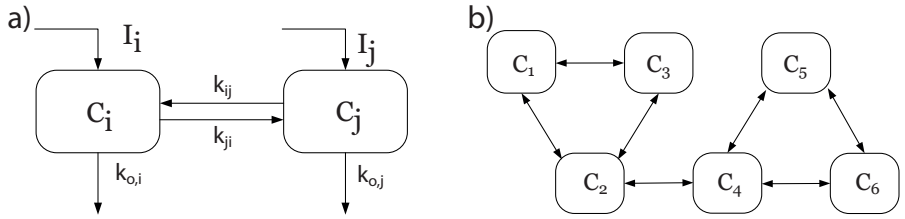

Figure 7.2.: a) Segment of a compartmental network. b) Compartmental network of 6 components

given in Figure 7.2 a). Observe, the systems in Section 7.1 are classical examples for compartmental networks. Generally, these networks can be described as follows [14]: if $x_i(t)$ denotes the mass, which compartment i is using at time t, then we denote with $k_{ij}x_j$ the mass flow from compartment j to i and with $k_{o,i}x_i$ the sum of all outflows of compartment i. Further, the external inflow of compartment i is given by $I_i = \sum_{j=1}^{m} b_{ij}u_j(t)$ where $u_j(t)$ represents the j-th input source. As for the water reservoir example, a compartmental network consisting of n compartments can be described by the linear differential equation

$$\dot{x}_i(t) = -k_{o,i}x_i(t) + \sum_{j\neq i}^{n}[k_{ij}x_j(t) - k_{ji}x_i(t)] + \sum_{j=1}^{m} b_{ij}u_j(t) \quad \text{for } i = 1,\dots,n.$$

The state-space representation of this system is given by $A = [a_{ij}]_{n\times n}$ with

$$a_{ij} = \begin{cases} -k_{o,i} - \sum_{j\neq i}^{n} k_{ji}, & i = j \\ k_{ij}, & i \neq j \end{cases},$$

and $B = [b_{ij}]_{n\times m}$. In Figure 7.2 a) we can observe the difference to the water reservoir examples: we do not assume any longer, that the influence of compartment i to j is mutual and thus the system does not need to be symmetric. In [14] the following system, consisting of the 6 compartments as shown in Figure 7.2 b), has been used to demonstrate the ILMI I approach

$$A := \begin{pmatrix} -1.5 & 0.6 & 1.0 & 0 & 0 & 0 \\ 0.3 & -1.9 & 0.2 & 0 & 0 & 0 \\ 0.2 & 0.5 & -2.7 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & -3 & 0.6 & 0.5 \\ 0 & 0 & 0.4 & -1.6 & 0.3 & 0 \\ 0 & 0 & 0 & 0.6 & 0.5 & -1.6 \end{pmatrix}, \quad B := \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$
$$C := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Observe, this is a MIMO-system and thus only Generalized Balanced Truncation, Balanced Truncation to first order and the ILMI-approaches can be applied. A comparison of these methods gives

Order	1	2	3
Generalized Balanced Truncation	0.78	0.26	0.06
Balanced Truncation to first order	0.01	-	-
ILMI I	0.01	-	-
ILMI II	-	-	-

Hence, as before for the SISO-case, the best results arise from ILMI I and Balanced Truncation. In contrast ILMI II could not find any solution smaller than 0.1. Let us see how these results transfer to the SISO-case. We transform this non-symmetric example to a SISO-system by assuming, that compartment C_1 and C_2 share the same input source, i.e. $B := \begin{pmatrix}1 & 1 & 0 & 0 & 0 & 0\end{pmatrix}^T$. The results for this non-symmetric example are summarized in the following tabular.

Order	1	2	3
Generalized Balanced Truncation	0.69	0.24	0.06
Symmetric Balanced Truncation (IRKA)	$1.70 \cdot 10^{-3}$	$7.41 \cdot 10^{-4}$	-
Symmetric Balanced Truncation	$1.70 \cdot 10^{-3}$	$7.41 \cdot 10^{-4}$	-
ILMI I	$8.18 \cdot 10^{-3}$	-	-
ILMI II	-	-	-
Arnoldi/Lanczos	0.26	$1.89 \cdot 10^{-3}$	-
Biorthogonalization	0.03	$5.19 \cdot 10^{-3}$	-

ILMI II could not find any solution, though the dimension of this system is comparably low. This shows, together with the the forgone examples, that ILMI II is only of theoretical interest, but performs not sufficiently well in practice. Another interesting fact of this example is, that ILMI I performs well for finding a first order system, but by trying to find a higher order approximations, below the error of the first order system, ILMI I ran into numerical problems.

For the other methods we can observe, the quality of their approximations remained almost the same compared with those of the water reservoir example.

7.3. Heat Equation

The heat equation is an important partial differential equation, which is given in the plain as

$$\dot{T} = \triangle T = \frac{\partial^2}{\partial x^2} T + \frac{\partial^2}{\partial y^2} T. \tag{7.1}$$

By discretizing this system with the help of finite differences [26] and numbering the discretization points as shown schematically in Figure 7.3, we can write

$$\triangle T_{ij} \approx -\frac{1}{h^2}(4T_{ij} - T_{i+1,j} - T_{i,j+1} - T_{j-1,j} - T_{i,j-1}),$$

where h denotes the grid step. If we consider the temperature on the grid boundaries to be steered by four different inputs, then we can write $\dot{T}_{ij} = \triangle T_{ij}$ as a linear positive system with input $u_1, \ldots, u_4$. For this purpose let us consider the unit-square, discretized

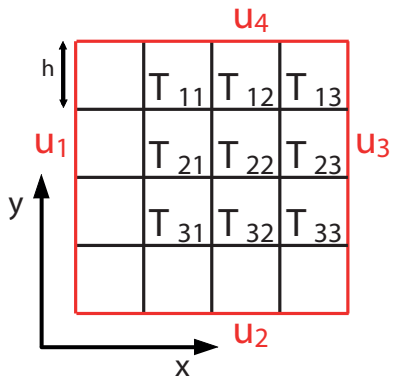

Figure 7.3.: Discretized heat equation on a quadratic plate

with $(n+2)^2$ points. Since the boundary points are taken by the inputs, our state variable consists of the n^2 inner points as shown in Figure 7.3. Then we can define our system matrices

$$A := \begin{pmatrix} P & I & 0 & \cdots & \\ I & P & I & 0\cdots & \\ & \ddots & \ddots & \ddots & \\ \cdots & 0 & I & P & I \\ & \cdots & 0 & -I & P \end{pmatrix} \in \mathbb{R}^{n^2\times n^2} \quad \text{with} \quad P := \begin{pmatrix} -4 & 1 & 0 & \cdots & \\ 1 & -4 & 1 & 0\cdots & \\ & \ddots & \ddots & \ddots & \\ \cdots & 0 & 1 & -4 & 1 \\ & \cdots & 0 & 1 & -4 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

and $B := [b_{ij}] \in \mathbb{R}^{n^2\times 4}$, where

$$b_{i1} := \begin{cases} 1, & \text{for } i = 1,\dots,n \\ 0, & \text{otherwise} \end{cases}, \quad b_{i2} := \begin{cases} 1, & \text{for } i = n, 2n,\dots,n^2 \\ 0, & \text{otherwise} \end{cases},$$

$$b_{i3} := \begin{cases} 1, & \text{for } i = n(n-1)+1,\dots,n^2 \\ 0, & \text{otherwise} \end{cases}, \quad b_{i4} := \begin{cases} 1, & \text{for } i = 1, n+1,\dots,n(n-1)+1 \\ 0, & \text{otherwise} \end{cases}.$$

By setting $x := \begin{pmatrix} T_{11} & \dots & T_{n1} & T_{12} & \dots & T_{n2} & \dots & T_{1n} & \dots & T_{nn} \end{pmatrix}^T \in \mathbb{R}^{n^2}$, we have discretized (7.1) as

$$\dot{x} \approx \frac{1}{h^2}Ax + \frac{1}{h^2}Bu \quad \text{with} \quad u := \begin{pmatrix} u_1 & \cdots & u_4 \end{pmatrix}^T \in \mathbb{R}^4.$$

As the output of the system we want to consider the average temperature, i.e.

$$y = \frac{1}{n}CT, \quad \text{with} \quad C := \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{1\times n}$$

Notice, the smaller h the better the approximation of the heat equation. Thus, in case of a good approximation, the system (A, B, C) will be become large-scale since $h = \frac{1}{n+1}$.

Let us start the model reduction comparison with a "bad" approximation, due to the LMI problems for high orders. In particular we choose $n = 3$ and set $u_2 = u_3 = u_4 = 0$, which yields a SISO state-space system of dimension 9. Then our discussed methods perform as follows

Order	1	2	3	5	8
Generalized Balanced Truncation	0.88	0.70	0.49	0.31	0.07
Symmetric Balanced Truncation (IRKA)	0.02	$2.73 \cdot 10^{-5}$	0	-	-
Symmetric Balanced Truncation	0.02	$2.73 \cdot 10^{-5}$	0	-	-
ILMI I	$9.74 \cdot 10^{-3}$	-	-	-	-
ILMI II	-	-	-	-	-
Arnoldi/Lanczos	0.50	0.23	0.08	0	-
Biorthogonalization	0.20	$7.94 \cdot 10^{-3}$	0	-	-

Considering, that we only measure the the average temperature, it is not surprising, that we cannot observe all the states and thus the state-space representation is not minimal. Symmetric Balanced Truncations as well as Biorthogonalization end up in a positive minimal realization. In contrast, Generalized Balanced Truncation gives for an order of $n - 1$ a worse result than reduction to first order.

In this example, ILMI I gives for a first order approximations the best result. However, as in the previously discussed examples, if we want to improve this error by increasing the order, the optimization runs into numerical problems. Further notice, the application of Lanczos to its transposed fifth order system does not result in a positive system this time.

Let us consider the same system with $n = 10$, i.e a state-space-dimension of 100. Then we attain

Order	1	9	15	18
Generalized Balanced Truncation	0.99	0.85	0.74	0.31
Symmetric Balanced Truncation (IRKA)	0.15	$2.52 \cdot 10^{-13}$	-	-
Symmetric Balanced Truncation	0.66	$2.52 \cdot 10^{-13}$	0	-
Arnoldi/Lanczos	0.81	0.14	$3.60 \cdot 10^{-4}$	$5.85 \cdot 10^{-6}$
Biorthogonalization	0.5	$1.04 \cdot 10^{-4}$	0	-

Since direct balancing performs perfectly well, this indicates, that this system always possesses a symmetric minimal realization. As in all the examples, there is no difference between IRKA and direct balancing, which shows, that IRKA is indeed a reliable precursor. This is especially important in the context of large-scale systems, what we will consider now.

In case of $n = 30$, we have to deal with 900 states and Generalized Balanced Truncation is not applicable any longer. Thus we are left with the Krylov subspaces methods and Symmetric Balanced Truncation, which give

Order	1	11	45	55
Symmetric Balanced Truncation (IRKA)	0.36	$8.73 \cdot 10^{-11}$	-	-
Symmetric Balanced Truncation	0.16	$2.37 \cdot 10^{-10}$	0	-
Arnoldi/Lanczos	0.93	0.53	$2.54 \cdot 10^{-4}$	$5.85 \cdot 10^{-6}$
Biorthogonalization	0.88	$1.04 \cdot 10^{-4}$	$6.32 \cdot 10^{-12}$	0

Biorthognalization works in the same way as before, but it is remarkable how much better Symmetric Balanced Truncation performs.

Notice, from a practical perspective there is no difference in which side of the plate is used as the single-input and thus for all sides we get the same results. Moreover, since $CB = \begin{pmatrix} n & \cdots & n \end{pmatrix}$, for all the four possible SISO-systems we even get identical approximations by applying Biorthogonalization. It is easy to see, that this extends to Arnoldi and Symmetric Balanced Truncation. Thus, we are in the situation of Section 6.2, which allows us to apply those methods to the full MISO-system (A, B, C). If (A_r, b_r, c_r) denotes the approximation of one of the four SISO-systems, then the reduced MISO-system can be given by (A_r, B_r, c_r), with $B_r := \begin{pmatrix} b_r & b_r & b_r & b_r \end{pmatrix}$. Such an approach is not applicable for the SISO-systems attained by ILMI I. Consequently, obtaining a MISO-approximation becomes even more expensive due to an in increase of the number of variables. The results of MISO-system approximations for $n = 3$ is summarized in the following tabular.

Order	1	2	3	5	8
Generalized Balanced Truncation	0.95	0.89	0.84	0.69	0.33
Symmetric Balanced Truncation (IRKA)	0.02	$2.74 \cdot 10^{-5}$	0	-	-
Symmetric Balanced Truncation	0.02	$2.74 \cdot 10^{-5}$	0	-	-
ILMI I	$9.74 \cdot 10^{-3}$	-	-	-	-
ILMI II	-	-	-	-	-
Arnoldi/Lanczos	0.50	0.23	0.08	0	-
Biorthogonalization	0.20	$7.94 \cdot 10^{-3}$	0	-	-

Except for Generalized Balanced Truncation, all the methods returned the same error as for the SISO-case. This result can be extended to higher order systems and we conclude, in case of a MIMO-system the drawbacks of Generalized Balanced Truncation show up even more.

Conclusions and Open Problems

All positivity preserving model order reduction methods, that were found by the author till this day, have been discussed in this thesis. As a consequence of high numerical complexity and generally poor approximation properties, we could observe, that none of these methods are applicable to high-dimensional systems. This restricts these methods to systems, where we usually do not see the need to reduce them.

Basically, all the discussed LMI-approaches share the same problem, that they cannot take advantage of well-established methods, such as Balanced Truncation and Krylov subspaces methods. The main reason for this lies in the difficulty, that the established methods mostly do not return a positive approximation, even though the reduced system might be positively realizable.

To this end, it has been shown, that in case of a symmetric system, the Krylov subspaces methods can be considered a positive realization algorithm. Moreover, a new symmetry characterization of balanced SISO-systems has been presented. Combined with the Krylov subspace methods, this led to the applicability of Balanced Truncation to positive SISO-systems. Based on its good approximation properties, this method outperforms the LMI-approaches in most cases. Additionally, we motivated to use IRKA for a pre-approximation in order to make our new approach applicable to large-scale systems.

As a consequence of these results the positive realizability of the Krylov subspaces methods has been extended and we found a way to replace SISO-systems by symmetric approximations. This could be of great interest e.g. in the context of system analysis. Beside this, we discovered a new necessary condition for the positive realizability of an arbitrary transfer function, which avoids the consideration of the impulse response.

In Section 6.2 we transferred the symmetry approach to a certain class of MIMO-systems. Nevertheless, a full analysis is still missing here. Also the use of other pre-approximation methods as well as the consequences for time-varying systems, non-linear systems, etc. have not been investigated so far.

A. Appendix

A.1. Cones

Definition A.1 (*Cone*)

Let $X \subseteq \mathbb{R}^n$, then the set

$$C(X) := \{y \mid y = \alpha x, \alpha \geq 0, x \in X\}$$

is called the cone hull of the set X and X is said to be a cone if and only if $X = C(X)$.[6]

Definition A.2 (*Dual Cone*)

The dual of a set $X \subseteq \mathbb{R}^n$ is defined by

$$X^* := \{y \mid \langle y, x \rangle \geq 0 \ \forall x \in X\}.$$

If X is a cone, we call X^* its dual cone.[2]

Definition A.3 (*Convex Cone*)

A cone X is called convex if it contains the line segment between any two points of it, i.e.

$$x_1, x_2 \in X \ \Rightarrow \ \alpha x_2 + (1 - \alpha) x_1 \in X, \ 0 \leq \alpha \leq 1,$$

or equivalently by cone defintion

$$x_1, x_2 \in X \ \Rightarrow \ \alpha x_1 + \beta x_2 \in X, \alpha, \beta \geq 0.[2]$$

Remark: If $X \subset \mathbb{R}^n$, then the *smallest convex cone containing* X consists of all finite nonnegative linear combinations of elements of X.

The dual set is defined by the scalar product and hence by the continuity and linearity of the scalar product, we get it the following result.

Lemma A.1

For every set X, its dual X^* is a closed convex cone.[2]

Definition A.4 (*Pointed and Solid Convex Cone*)

For a convex cone X we say it is

- *pointed if $X \cap -X = \{0\}$,*
- *solid if the interior of X, $\mathring{X} \neq \emptyset$.[2]*

Definition A.5 (*Polyhedral Cone*)

A cone $X \subset \mathbb{R}^n$ is called polyhedral, if it is finitely generated, i.e. if

$$X = B\mathbb{R}^k_+$$

for some natural number k and an $n \times k$ matrix B.[2]

A polyhedral cone is in a manner of speaking, a cone with a finite number of edges. Thus the cone consists by definition of nonnegative linear combinations of finite number and is therefore automatically convex. It also has to be closed because of the continuity of the linear mapping B.

Lemma A.2

Every polyhedral cone is closed and convex.[2]

Definition A.6 (*Proper Cone*)

A closed, pointed, solid convex cone is called proper cone.[2]

Lemma A.3

In $\mathbb{R}^2$ every closed proper cone is polyhedral.[1]

Proof: Let X be a closed convex cone in $\mathbb{R}^2$. Since X is pointed, the maximum angular between two vectors of X must be strictly smaller than π. By taking exactly the two vectors of X with maximum angular, the area enclosed by their convex combinations must lie in X and by assumption of the maximum angular there cannot be any point in X outside this area. ■

A direct consequence of the well-known Seperation Theorem in functional analysis [25] is the following result.

Lemma A.4 (*Seperation Theorem for convex cones*)

Let $X \subset \mathbb{R}^n$ be a closed convex cone and $x_0 \in \mathbb{R}^n \setminus X$. Then there exists $x' \in \mathbb{R}^n$ such that

$$\langle x_0, x' \rangle < 0 \quad \text{and} \quad \langle x, x' \rangle \geq 0 \ \forall x \in X.$$

By the help of this lemma it will be easy to prove the following important theorem about the dual of a dual cone.

Theorem A.1

X is a closed convex cone if and only if $X = X^{**} := (X^*)^*$.[29]

Proof: ▶ Sufficiency: If $X^{**} = X$ then by Lemma A.1 X is a closed and convex cone.

▶ Necessity: $X^* = \{y \mid \langle x, y \rangle \geq 0 \ \forall x \in X\} \ \Rightarrow \ \langle y, x \rangle \geq 0 \ \forall x \in X$ and $\forall y \in X^*$.
Since $X^{**} = \{z \mid \langle y, z \rangle \geq 0 \ \forall y \in X^*\}$ it is clear that $X \subset X^{**}$.
Suppose there exists a $z_0 \in X^{**} \setminus X$, then by Lemma A.4 there is a vector x', such that

$$\langle z_0, x' \rangle < 0 \quad \text{and} \quad \langle x, x' \rangle \geq 0 \ \forall x \in X.$$

Hence x' must be an element of X^* and therefore $\langle z_0, x' \rangle < 0$ contradicts the definition of X^{**}. ■

Observe, if X was not closed, X^{**} would be equal to the smallest closed convex cone including X. Next we want to characterize a closed convex cone by its dual.

Theorem A.2 (*Duality Theorem*)

A closed convex cone $X \subset \mathbb{R}^n$ is pointed if and only if X^* is solid. [29]

Proof: ▶ Sufficiency: Suppose X is not pointed, then $\exists \tilde{x} \in X : \tilde{x} \in X \cap -X$, i.e. $\tilde{x}$ and $-\tilde{x}$ are both elements of X and hence $\alpha\tilde{x} \in X, \ \forall \alpha \in \mathbb{R}$ by definition of a cone.
By definition of the dual cone, $\langle y, \alpha\tilde{x} \rangle \geq 0, \forall y \in X^*$.

If X^* is solid, then $\exists \epsilon > 0, \ x_0 \in X^* : B_\epsilon(x_0) \subset X^*$ and therefore it is possible to find a $y \in B_\epsilon(x_0) : \langle y, \tilde{x} \rangle \neq 0$. Choosing $\alpha = -\text{sign}(\langle y, \tilde{x} \rangle)$ contradicts the condition $\langle y, \alpha\tilde{x} \rangle \geq 0$.

▶ Necessity: Assume X^* is not solid and let $\{x_1, \ldots, x_k\}$ denote k linear independent unit vectors of it.
If we could choose $k = n$, then $z = \sum_{i=0}^{n} x_i$ is an inner point X^* :
Let $1 > \epsilon > 0$ and $y \in B_\epsilon(z)$. Since $\{x_1, \ldots, x_k\}$ spans the whole $\mathbb{R}^n$ we can write $y = z + \sum_{i=0}^{n} \alpha_i x_i$ with $\sum_{i=0}^{n} \alpha_i^2 < \epsilon^2 < 1$. Thus $|\alpha_i| < 1$ and therefore y is a nonnegative linear

combination with coefficients $(1 - \alpha_i) \geq 0$.

Hence, $k < n$ and it is possible to find a vector $p \in \mathbb{R}^n$ that is orthogonal to the set $\{x_1, \ldots, x_k\}$ and $\langle p, x \rangle = 0 \; \forall x \in \pm X^*$. That implies $p \in X^{**} \cap -X^{**}$ which is by Theorem A.1 a contradiction to X being pointed. ■

Bibliography

[1] Berman, A., Neumann, M. and Stern, R. J. [1989], *Nonnegative Matrices in Dynamical Systems*, John Wiley & Sons.

[2] Berman, A. and Plemmons, R. J. [1979], *Nonnegative Matrices in the Mathematical Sciences*, Academic Press.

[3] Cheney, W. and Kincaid, D. [2010], *Linear Algebra: Theory and Applications*, 2nd edn, Jones & Bartlett Learning, pp. 533–534.

[4] Damm, T. [2009], *Lecture Notes: Numerical Methods in Control*, TU Kaiserslautern.

[5] Farina, L. [1995], Necessary conditions for positive realizability of continuous-time linear systems, *in* 'System & Controll Letters', number 25, pp. 121–124.

[6] Farina, L. and Rinaldi, S. [2000], *Positive Linear Systems: Theory and Applications*, John Wiley & Sons.

[7] Feng, J. et al. [2010], Internal positivity preserved model reduction, *in* 'International Journal of Control', Vol. 83, Taylor & Francis, pp. 575–585.

[8] Gathmann, A. [2006], *Vorlesungsskript: Einführung in die Funktionentheorie*, Fachbereich Mathematik, TU Kaiserslautern, p. 58.

[9] Ghaoui, L. E., Oustry, F. and AitRami, M. [1997], A cone complementarity linearization algorithm for static output-feedback and related problems, *in* 'IEEE Transactions On Automatic Control', Vol. 42, pp. 1171 – 1176.

[10] Glad, T. and Ljung, L. [2000], *Control Theory: Multivariable and Nonlinear Methods*, Taylor & Francis, p. 69.

[11] Gugercin, S. and Antoulas, A. C. [2003], An $\mathcal{H}_2$ error expression for the lanczos procedure, *in* 'Proceedings of the 42nd IEEE Conference on Decision and Control'.

[12] Gugercin, S., Antoulas, A. C. and Beattie, C. [2008], $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems, *in* 'SIAM Journal on Matrix Analysis and Applications', Vol. 3, pp. 609–638.

[13] Johansson, R. [2010], *System Modelling & Identification, Course Textbook*, KFS AB, Lund University.

[14] Li, P. et al. [2011], Positivity-preserving $\mathcal{H}_\infty$ model reduction for positive systems, *in* 'Automatica', Vol. 47, pp. 1504 – 1511.

[15] Liu, W., Sreeram, V. and Teo, K. [1998], Model reduction for state-space symmetric systems, *in* 'Systems & Control Letters', number 34, ELSEVIER, pp. 209–215.

[16] Luenberger, D. G. [1979], *Introduction to Dynamic Systems: Theory, Models & Applications*, John Wiley & Sons.

[17] Mayer, C. [2008], *Vorlesungsmitschrift: Einführung in die Funktionalanalysis*, Arbeitsgruppe Geomathematik, Fachbereich Mathematik, TU Kaiserslautern, pp. 18–19 and 41–42.

[18] Meyer, C. D. [2001], *Matrix Analysis and Applied Linear Algebra Book and Solutions Manual*, SIAM, chapter 8, pp. 670 – 678.

[19] Ohta, Y., Maeda, H. and Kodama, S. [1984], Reachability, observability, and realizability of continuous-time positive systems, *in* 'SIAM J. Control and Optimization', Vol. 22, pp. 171–180.

[20] Peaucelle, D. et al. [2002], *User's Guide for SeDuMi Interface 1.04*, p. 4.

[21] Rao, P. R. [2008], *Signals And Systems*, Tata McGraw-Hill, pp. 151–152.

[22] Reis, T. and Virnik, E. [2009], Positivity preserving balanced truncation for descriptor systems, *in* 'SIAM Journal on Control and Optimization', Vol. 48, pp. 2600 – 2619.

[23] Safonov, M. G. and Chiang, R. Y. [1989], A schur method for balanced-truncation model reduction, *in* 'IEEE Transactions on Automatic Control', Vol. 34, pp. 729–733.

[24] Sandberg, H. and Rantzer, A. [2004], Balanced truncation of linear time-varying systems, *in* 'IEEE Transactions on Automatic Control', Vol. 49.

[25] Functional Analysis & Stochastic Analysis Group [2008], *Lecture notes: Functional Analysis*, Department of Mathematics, TU Kaiserslautern, pp. 22–25.

[26] Technomathematics Group [2009], *Lecture notes: Numerical Methods for Partial Differential Equations I: Elliptic and Parabolic Equations*, Department of Mathematics, TU Kaiserslautern, chapter 3-4 and 7.

[27] Trefethen, L. N. and Bau, D. [1997], *Numerical Linear Algebra*, SIAM, chapter 32-39.

[28] Zerz, E. [2005], *Lecture Notes: Introduction to System and Control Theory*, pp. 63–65.

[29] Zhang, S. [2010], *Lecture notes: SEG5660: Conic Optimization and Applications*, The Chinese University of Hong Kong, chapter 1-2.

[30] Zhou, K. and Doyle, J. C. [1999], *Essentials of Robust Control*, Prentice Hall.

Hereby I declare that I am the only author of this work and that no sources other than those listed have been used in this work.

Kaiserslautern, November 29, 2012

_____

Christian Grußler

Acknowledgement

It is a pleasure to thank those who made this work possible, my supervisors Tobias Damm and Anders Rantzer. I am very grateful for receiving this topic from them but also for their help, advice and encouragement throughout this work and beyond that. This thesis would not have been possible without the immense contributions and patience of Tobias Damm. I would like to thank my family, particularly my mother and grandmother, for their invaluable support during my whole life and education. My profound gratitude also goes to my Mathematics teacher Helmut Blauth for promoting my interest in Mathematics and encouraging me to study it. I am very thankful for the support of all my friends, but notably I would like to mention Jochen Kall for his help and advice while writing this thesis. Finally, I would like to show my appreciation to all those, who have not been mentioned here, but were important to me in one way or another in helping me reach this far.

Thank you all a lot.