Hankel-Norm Approximation of Descriptor Systems

An der Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg
zur Erlangung des akademischen Grades
Master of Science
angefertigte

Masterarbeit

vorgelegt von
Steffen Werner
geboren am 06.09.1992 in Stendal,
Studiengang Mathematik,
Studienrichtung Mathematik.

6. September 2016

Betreut am Institut für Analysis und Numerik von
Prof. Dr. rer. nat. Peter Benner

First of all, I would like to thank my advisor Prof. Dr. Peter Benner for the motivation of this thesis as well as all precious discussions and continuous support during the preparations.

Another word of thanks is addressed to Dr. Martin Stoll for the examination of this thesis and to Dr. Jens Saak for the support concerning the different test examples and the handling of the MATLAB M.E.S.S. toolbox.

Last I would like to thank all my colleagues at the Otto von Guericke University Magdeburg and at the Max Planck Institute for Dynamics of Complex Technical Systems Magdeburg for the continuous support and help during the entire processing time of this thesis.

Abbreviations and Notation

Abbreviations

ADI method	alternating implicit direction method
C-controllable	completely controllable
C-observable	completely observable
c-stable	continuous-time stable
GBT(SR)	generalized balanced truncation square-root method
GHNA	generalized Hankel-norm approximation method
LR-ADI method	low-rank alternating implicit direction method
RRQR decomposition	rank-revealing QR decomposition

Notation

$\mathbb{R}$	the field of the real numbers
$\mathbb{R}^- = (-\infty, 0)$	the negative real semi-axis
$j = \sqrt{-1}$	the imaginary unit
$j\mathbb{R}$	the imaginary axis
$\mathbb{C}$	the field of the complex numbers
$\mathrm{Re}(z)$	real part of $z \in \mathbb{C}$
$\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$	the open left half-plane
$\mathbb{F}^{n\times m}$	the space of matrices of real ($\mathbb{F} = \mathbb{R}$) or complex ($\mathbb{F} = \mathbb{C}$) matrices of size $n \times m$
A^T	the transpose of $A \in \mathbb{F}^{n\times m}$
A^{-1}	inverse of $A \in \mathbb{F}^{n\times n}$
$A^{-T} = (A^{-1})^T$	transposed inverse of $A \in \mathbb{F}^{n\times n}$
$A^H = \bar{A}^T$	conjugate transpose of $A \in \mathbb{F}^{n\times m}$
$A^\dagger$	the Moore-Penrose pseudoinverse of $A \in \mathbb{F}^{n\times m}$
$\mathrm{diag}(A_1, \ldots, A_k)$	block diagonal matrix with $A_j \in \mathbb{F}^{n_j\times n_j}$, $j = 1, \ldots, k$

$I_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$	the identity matrix of order n
$\det(A)$	the determinant of $A \in \mathbb{F}^{n\times n}$
$\mathrm{tr}(A)$	trace of the matrix $A \in \mathbb{F}^{n\times n}$
$\mathrm{rank}(A)$	the rank of $A \in \mathbb{F}^{n\times m}$
$\mathrm{Ker}(A) = \{x \in \mathbb{F}^m : Ax = 0\}$	the kernel (or null space) of $A \in \mathbb{F}^{n\times m}$
$\mathrm{Im}(A) = \{y \in \mathbb{F}^n : y = Ax,\ x \in \mathbb{F}^m\}$	the image (or range) of $A \in \mathbb{F}^{n\times m}$
$\Lambda(A) = \{\lambda \in \mathbb{C} : \det(\lambda I_n - A) = 0\}$	spectrum of $A \in \mathbb{F}^{n\times n}$
$\Lambda(A, E) = \{\lambda \in \mathbb{C} : \det(\lambda E - A) = 0\}$	spectrum of the matrix pencil $\lambda E - A$
$\Lambda_f(A, E) = \Lambda(A, E) \setminus \{\infty\}$	the finite spectrum of the matrix pencil $\lambda E - A$
$\pi(A)$	number of eigenvalues of $A \in \mathbb{F}^{n\times n}$ in the open right half-plane
$\nu(A)$	number of eigenvalues of $A \in \mathbb{F}^{n\times n}$ in the open left half-plane
$\delta(A)$	number of eigenvalues of $A \in \mathbb{F}^{n\times n}$ on the imaginary axis
$\mathrm{In}(A) = (\pi(A), \nu(A), \delta(A))$	the inertia of $A \in \mathbb{F}^{n\times n}$
$\sigma_1(A) \geq \ldots \geq \sigma_k(A) \geq 0$	singular values of $A \in \mathbb{F}^{n\times m}$
$\sigma_{\max}(A) = \sigma_1(A)$	the largest singular value of $A \in \mathbb{F}^{n\times m}$
$A_{k\to\infty} = \lim\limits_{k\to\infty} A_k$	limit of the sequence $A_k \in \mathbb{F}^{n\times m}$, $k = 0, 1, \ldots$
$\lvert\lvert.\rvert\rvert$	an arbitrary suitable norm, depending on the context
$\lvert\lvert A\rvert\rvert_2 = \sigma_{\max}(A)$	spectral norm of $A \in \mathbb{F}^{n\times m}$

Contents

List of Figures

List of Tables

List of Algorithms

1 The Hankel-Norm Approximation, a Method of Model Reduction

Today, many different real-world applications are modeled by systems of differential-algebraic equations. Chemical processes, electrical circuits and networks, or computational fluid dynamics are just a few examples. These models are used for simulations and the design of controllers since experiments can be very costly, time-consuming and expensive. In this thesis such systems of differential-algebraic equations have the form

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1.1}$$

with a singular E matrix. Such systems are called descriptor systems. Here the controls are used to influence the internal states x. The observed outputs of the system are given in y. The descriptor system (1.1) can be interpreted as a black box which gets an input and results in an output under certain rules.
Due to certain required properties, like an increased accuracy, the number of equations used to model the problem quickly enlarges. For this reason, the usage of the complete models often reaches the limits of computational resources like memory and computation time. The acquired data often contains a huge amount of unnecessary redundancies, especially for large-scale models . It would be beneficial to use an approximation of the original model by a system with a much smaller number of equations than the original one. This is the goal of model reduction.
Usually, systems of the form (1.1) have a small number of inputs and outputs but a much larger count of internal states, which have to be computed through the set of differential equations. Model reduction methods are then used to compute a new descriptor system of the form

$$\begin{aligned} \hat{E}\dot{\hat{x}}(t) &= \hat{A}\hat{x}(t) + \hat{B}u(t), \\ \hat{y}(t) &= \hat{C}\hat{x}(t) + \hat{D}u(t), \end{aligned} \tag{1.2}$$

where u are the same controls as in (1.1). The number of equations used in (1.2) shall be much smaller than in the original system (1.1). The reduction of the model is done in order to approximate the input-output behavior of (1.1), such that

$$||y - \hat{y}|| \leq tol \cdot ||u|| \tag{1.3}$$

holds for a given tolerance $tol > 0$ and all admissible inputs u. Beside this approximation, special properties of the original system should be preserved during the model reduction. Some of these properties are, for example, the stability or the passivity of the system.

For the task of model reduction, there exist many different methods and techniques. Most of them were developed for the much simpler standard case, where the E matrix in (1.1) is an identity matrix I_n. For regular E the descriptor system (1.1) can be reduced to the standard case by applying the inverse of E to the first equations of (1.1). The more complicated case occurs if E is a singular matrix. The introduction of spectral projectors has allowed the generalization of many known model reduction methods for descriptor systems.

A large number of model reduction methods is based on the computation of matrix equations. There, the solutions of matrix equations and the singular value decomposition are used to get measurements for a meaningful reduction of the original system. Different choices of matrix equations change the set of preserved properties of the original system. A survey of generalizations of some matrix equation based methods can be found in [11].

The input-output behavior of a descriptor system (1.1) can be described by a rational matrix-valued function, the transfer function. Another field of model reduction methods treats the interpolation of the system's transfer function. A detailed view on the generalization of interpolation based model reduction methods for descriptor systems can be found in [14].

As a last field of techniques, the proper orthogonal decomposition methods shall be mentioned. Here snapshots of the system's input-output behavior are computed to construct a reduced-order model. There exist different approaches, how to choose the snapshots. Based on spectral projectors, a generalization of proper orthogonal decomposition methods is presented in [18].

An often mentioned problem is the construction of a reduced-order model minimizing a certain system norm. A first approach in this direction was made by the introduction of the balanced truncation method and the corresponding error bound in the $\mathcal{H}_\infty$-norm. But in general, this method is not able to construct an optimal approximation. A refinement of the balanced truncation leads to another model reduction method which succeeds in finding a best approximation in the Hankel-norm. This Hankel-norm approximation method was based on the work of Adamjan, Arov and Krein about the approximation of the Hankel matrix. The main results of this work can be found in [1]. Further contributions to this theory were made by Glover in [13]. There, a characterization of all Hankel-norm approximations for standard linear systems is given using the theory of balanced realizations. As a result, an algorithm for the Hankel-norm approximation of standard systems was proposed.

Beside an exact error bound in the Hankel-norm, the Hankel-norm approximation can provide a better approximation behavior than other model reduction methods. Therefore, it would be beneficial to use it in case of descriptor systems (1.1). So far used approaches of the Hankel-norm approximation were designed only for the case of standard systems. In this thesis the Hankel-norm approximation method will be generalized to the descriptor system case. Therefor, the generalized concept of balanced realizations using spectral projectors will be considered to make use of the results from [13].

Starting form this presentation of the problem, in Chapter 2 some necessary basic tools from the linear algebra and the system and control theory are introduced. Most of the presented topics are considered in the framework of descriptor systems. In Chapter 3

the Hankel-norm approximation method for the standard case, introduced by Glover, is presented. Also, a special solution approach for descriptor systems, based on the Weierstrass canonical form, is presented there.
After clarifying all basics and the state of the art, the generalized Hankel-norm approximation is developed in Chapter 4 as an extension of the generalized balanced truncation method. Also, an approximated version of the generalized Hankel-norm approximation and the application on large-scale sparse systems will be considered here. In Chapter 5 a projection-free variant of the introduced method and, in this context, spectral projection based algorithms for the implementation of the generalized Hankel-norm approximation are shown.
The results of some numerical tests in MATLAB are displayed in Chapter 6. Details of two different dense implementations and a sparse implementation of the method are shown on different data examples. Finally, in Chapter 7 the results of this thesis are summarized and open points are outlined.

2 Mathematical Basics

For further discussions of model reduction methods, in this chapter necessary mathematical concepts are presented. After the revision of some basic linear algebra tools, a large bunch of system theoretical aspects and concepts will be introduced. From the beginning, the case of descriptor systems is considered.

One of the most used tools in the numerical linear algebra is the singular value decomposition. Given a matrix $A \in \mathbb{R}^{n \times m}$ there exist orthogonal matrices $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{m \times m}$ such that

$$A = U\Sigma V^T, \tag{2.1}$$

with Σ having the following form

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.2}$$

where $\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_k) \in \mathbb{R}^{k \times k}$ is a full-rank diagonal matrix. The matrix decomposition in (2.1) is called the singular value decomposition of the matrix A. The diagonal entries of Σ_1 in (2.2) are the singular values of A and ordered in a decreasing way $\sigma_1 \geq \ldots \geq \sigma_k > 0$. The columns of $U = [u_1, \ldots, u_n]$ are the left and the columns of $V = [v_1, \ldots, v_n]$ are the right singular value vectors of A, see [1]. Often only the non-zero singular values and the corresponding columns of the orthogonal matrices are needed. This economic version of the singular value decomposition is called skinny.

One important property of the singular value decomposition is given by the Schmidt-Eckart-Young-Mirsky theorem.

Theorem 2.1. *(See [1]). The best rank-r approximation of the matrix $A \in \mathbb{R}^{m \times n}$ is given by the formula*

$$A_k = \sum_{j=1}^{r} \sigma_j u_j v_j^T,$$

using the singular value decomposition in (2.1)*. The exact approximation error is then given by*

$$||A - A_k||_2 = \sigma_{r+1}(A),$$

where $||.||_2$ denotes the spectral norm and $\sigma_{r+1}(A)$ is the $(r+1)$-st singular value of the matrix A.

This construction principle is used, in a modified version, for model reduction methods. There are several other useful matrix decompositions. Another one, which will be needed in later discussions, is the QR decomposition. Given a matrix $A \in \mathbb{R}^{m \times n}$, there exists an orthogonal $Q \in \mathbb{R}^{m \times m}$ and an upper triangular $R \in \mathbb{R}^{m \times n}$, such that

$$A = QR. \tag{2.3}$$

This matrix decomposition is not unique. A special version of (2.3) is given by including a permutation matrix Π. This pivoted QR decomposition has the form

$$A\Pi = QR. \tag{2.4}$$

For example, the permutation matrix can be chosen, such that the diagonal elements r_{ii} of the upper triangular matrix R are ordered in a decreasing way. Another version can be used to determine the numerical rank of the matrix A. This one is called the rank-revealing QR decomposition (RRQR decomposition).

2.1 Descriptor Systems and Spectral Projectors

In this thesis linear dynamical systems with differential-algebraic equations are considered.

Definition 2.1. *Given a continuous-time linear time-invariant descriptor system*

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t),\\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.5}$$

with $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$ constant matrices. The functions $u(t) \in \mathbb{R}^m$ are the controls, $x(t) \in \mathbb{R}^n$ the internal states and $y(t) \in \mathbb{R}^p$ the outputs of the descriptor system (2.5)*. Then, the quintuple* (E, A, B, C, D) *is called a* realization of the descriptor system (2.5)*. The* order *of the descriptor system* (2.5) *is given by n.*

For a more accurate view on different aspects of descriptor systems, another powerful tool is needed. The Laplace transformation is mapping time domain functions onto functions in the frequency domain. For a time domain function $f(t)$ the Laplace transformation is given by

$$F(s) = \mathcal{L}(f(t)) = \int\limits_0^{+\infty} e^{-st} f(t)\,\mathrm{d}t, \tag{2.6}$$

with the complex parameter $s = \delta + j\omega \in \mathbb{C}$. From system theoretical background, j denotes here the imaginary unit $j = \sqrt{-1}$. For further evaluations of the frequency domain function, the parameter s is chosen as $\delta = 0$ and $\omega = 2\pi\nu$, where ν being the frequency. In [21] a more detailed view on the usage of the Laplace transformation can be found.

The linearity of the Laplace transformation is obvious, since it is defined by an integral. Also it can be shown that for the differential it holds

$$\mathcal{L}(\dot{x}(t)) = s\mathcal{L}(x(t)) + x(0) = sX(s) + x_0,$$

where X denotes the Laplace transform of $x(t)$ and $x_0 = x(0)$, see [21].

Now, the Laplace transformation can be applied to the descriptor system in (2.5). Thereby, the exact input-output behavior of the system can be described by the resulting equation

$$Y(s) = \big(C\,(sE - A)^{-1}\,B + D\big)\,U(s) + C\,(sE - A)^{-1}\,Ex_0, \tag{2.7}$$

where Y and U are the Laplace transforms of y and u, receptively. For simplicity, it is assumed that $Ex_0 = Ex(0) = 0$. Now, the form of (2.7) can be rewritten as

$$Y(s) = G(s) \cdot U(s),$$

with G a rational matrix-valued function. The input-output behavior is completely described by the function G.

Definition 2.2. *The rational matrix-valued function*

$$G(s) = C\,(sE - A)^{-1}\,B + D, \tag{2.8}$$

is called the transfer function of the descriptor system (2.5)*.*

Often the descriptor system and its realization are both associated with the corresponding transfer function.
The realization of a descriptor system is not unique. It exists an endless number of different realizations of a descriptor system with the same transfer function. Thus, they all have the same input-output behavior.

Definition 2.3. *Let (E, A, B, C, D) and $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ be two realizations of descriptor systems. The realizations are called* restricted system equivalent *if there exist non-singular matrices $W, T \in \mathbb{R}^{n\times n}$, such that*

$$(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = (WET, WAT, WB, CT, D). \tag{2.9}$$

The transformation between two equivalent systems is called generalized state space transformation*.*

With the generalized state space transformation (2.9), for transfer functions it holds

$$\begin{aligned}\tilde{G}(s) &= \tilde{C}\left(s\tilde{E} - \tilde{A}\right)^{-1}\tilde{B} + \tilde{D}\\ &= CTT^{-1}\,(sE - A)^{-1}\,W^{-1}W + D\\ &= G(s).\end{aligned}$$

Since the input-output behavior is described by the transfer function, both systems with different realizations have the same input-output behavior.
Next, some additional generalized concepts have to be introduced.

Definition 2.4. *Let $A, E \in \mathbb{R}^{n\times n}$. The pair of matrices A and E is called* matrix pencil *and further denoted by $\lambda E - A$.*

For such a matrix pencil, the *generalized spectrum* $\Lambda(A, E)$ is given as all $\lambda \in \mathbb{C}$ for which the characteristic polynomial $P(\lambda) = \det(\lambda E - A)$ vanishes. These values are the *eigenvalues of the matrix pencil* $\lambda E - A$.

Definition 2.5. *Let $\lambda E - A$ be a matrix pencil. The matrix pencil is called* regular *if there exists a $\lambda \in \mathbb{C}$, such that* $\det(\lambda E - A) \neq 0$. *Otherwise the matrix pencil is called* singular.

The regularity of a matrix pencil $\lambda E - A$ means that the numbers of finite and infinite eigenvalues are not endless.
Given a regular matrix pencil $\lambda E - A$, the *Weierstrass canonical form* can be introduced, see [23]. There are non-singular matrices W and T, such that

$$E = W \begin{bmatrix} I_{n_f} & 0 \\ 0 & N \end{bmatrix} T \quad \text{and} \quad A = W \begin{bmatrix} J & 0 \\ 0 & I_{n_\infty} \end{bmatrix} T. \tag{2.10}$$

The matrix $J \in \mathbb{C}^{n_f \times n_f}$ corresponds to the finite eigenvalues of the pencil $\lambda E - A$ and is in the Jordan canonical form. The dimension of the corresponding deflating subspace is given by n_f. Also, the matrix $N \in \mathbb{R}^{n_\infty \times n_\infty}$ is in Jordan canonical form with zeros on its diagonal. The number n_∞ is the dimension of the deflating subspace corresponding to the infinite eigenvalues of the matrix pencil $\lambda E - A$. The matrix N is nilpotent with the index ν. That means, it holds $N^{\nu-1} \neq 0$ and $N^\nu = 0$. If the matrix pencil $\lambda E - A$ is referred to a descriptor system (2.5), the index of nilpotency ν is also referred to as the index of the system (2.5).
Based on the Weierstrass canonical form, the $\mathbb{R}^n$ can be decomposed into two complementary deflating subspaces corresponding to the finite and infinite eigenvalues of the matrix pencil $\lambda E - A$. The two matrices

$$P_l = W \begin{bmatrix} I_{n_f} & 0 \\ 0 & 0 \end{bmatrix} W^{-1} \quad \text{and} \quad P_r = T^{-1} \begin{bmatrix} I_{n_f} & 0 \\ 0 & 0 \end{bmatrix} T, \tag{2.11}$$

where W and T are the transformation matrices from (2.10), are the spectral projectors onto the left and right deflating subspaces corresponding to the finite eigenvalues, respectively. Then the spectral projectors onto the left and right deflating subspaces corresponding to the infinite eigenvalues of the matrix pencil $\lambda E - A$ are given by

$$Q_l = I_n - P_l \quad \text{and} \quad Q_r = I_n - P_r, \tag{2.12}$$

see [8, 23]. Since the Weierstrass canonical form is difficult to compute, a more practicable representation of the spectral projectors can be found in [28]. First, a block-triangular form of $\lambda E - A$ is assumed, with

$$E = V \begin{bmatrix} E_f & E_u \\ 0 & E_\infty \end{bmatrix} U^T, \quad A = V \begin{bmatrix} A_f & A_u \\ 0 & A_\infty \end{bmatrix} U^T, \tag{2.13}$$

where U and V orthogonal, E_f nonsingular, E_∞ nilpotent with index ν and A_∞ nonsingular. Using this formulation, the spectral projectors are given by

$$P_l = V \begin{bmatrix} I_{n_f} & -Z \\ 0 & 0 \end{bmatrix} V^T \quad \text{and} \quad P_r = U \begin{bmatrix} I_{n_f} & -Y \\ 0 & 0 \end{bmatrix} U^T, \tag{2.14}$$

where Y and Z solve the generalized Sylvester equation

$$\begin{aligned} E_f Y - Z E_\infty &= -E_u, \\ A_f Y - Z A_\infty &= -A_u. \end{aligned}$$

Finally, the solution of descriptor systems is considered here. The following contents are based on [8].
From the Weierstrass canonical form (2.10) one can obtain the following Laurent expansion at infinity to get the generalized resolvent

$$(\lambda E - A)^{-1} = \sum_{k=-\infty}^{+\infty} F_k \lambda^{-k-1},$$

with the coefficients F_k of the form

$$F_k = \begin{cases} T^{-1} \begin{bmatrix} J^k & 0 \\ 0 & 0 \end{bmatrix} W^{-1}, & k = 0, 1, 2, \ldots, \\ \\ T^{-1} \begin{bmatrix} 0 & 0 \\ 0 & -N^{-k-1} \end{bmatrix} W^{-1}, & k = -1, -2, \ldots. \end{cases} \tag{2.15}$$

Now, the state variables x of the descriptor system (2.5) are transformed by the transformation matrices of the Weierstrass canonical form, such that

$$Tx(t) = \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix}, \tag{2.16}$$

with a partition according to the block structure of the Weierstrass canonical form (2.10). The matrices

$$W^{-1}B = \begin{bmatrix} B_f \\ B_\infty \end{bmatrix} \quad \text{and} \quad CT^{-1} = \begin{bmatrix} C_f, & C_\infty \end{bmatrix} \tag{2.17}$$

are partitioned similar to (2.16).
The descriptor system (2.5) is then transformed into

$$\begin{aligned} \dot{z}_1(t) &= J z_1(t) + B_f u(t), \\ N \dot{z}_2(t) &= z_2(t) + B_\infty u(t), \\ y(t) &= C_f z_1(t) + C_\infty z_2(t) + D u(t). \end{aligned}$$

This new system decouples into the *slow subsystem*

$$\begin{aligned} \dot{z}_1(t) &= J z_1(t) + B_f u(t), \\ y(t) &= C_f z_1(t), \end{aligned} \tag{2.18}$$

and the *fast subsystem*

$$\begin{aligned} N \dot{z}_2(t) &= z_2(t) + B_\infty u(t), \\ y(t) &= C_\infty z_2(t) + D u(t). \end{aligned} \tag{2.19}$$

The slow subsystem (2.18) is in the standard formulation and has a unique solution of the form

$$z_1(t) = e^{tJ} z_1^0 + \int\limits_0^t e^{(t-\tau)J} B_f u(\tau)\,\mathrm{d}\tau,$$

for any integrable input u and any initial value $z_1^0 \in \mathbb{R}^{n_f}$. For an input u, ν times continuously differentiable, the unique solution of the fast subsystem (2.19) is given by

$$z_2(t) = -\sum_{k=0}^{\nu-1} N^k B_\infty u^{(k)}(t), \tag{2.20}$$

where ν is the index of the descriptor system (2.5) and $u^{(k)}$ denotes the k-th derivative of the input function. It is necessary that the input function u is sufficiently smooth and the initial value z_2^0 satisfies

$$z_2^0 = -\sum_{k=0}^{\nu-1} N^k B_\infty u^{(k)}(0).$$

Considering this, the initial value x_0 of the descriptor system (2.5) has to be *consistent*, which means, it satisfies the condition

$$Q_r x_0 = \sum_{k=0}^{\nu-1} F_{-k-1} B u^{(k)}(0),$$

where Q_r is the spectral projector corresponding to the infinite eigenvalues of the matrix pencil $\lambda E - A$ and the matrices F_k are given in (2.15).
In consequence, the descriptor system (2.5) has a unique, continuously differentiable solution $x(t)$ of the form

$$x(t) = \mathcal{F}(t) E x_0 + \int\limits_0^t \mathcal{F}(t-\tau) B u(\tau)\,\mathrm{d}\tau + \sum_{k=0}^{\nu-1} F_{-k-1} B u^{(k)}(t),$$

if the matrix pencil $\lambda E - A$ is regular, the input u is ν times continuously differentiable and the initial value x_0 is consistent. The term $\mathcal{F}$ is a fundamental solution matrix of (2.5) given by

$$\mathcal{F}(t) = T^{-1} \begin{bmatrix} \mathrm{e}^{tJ} & 0 \\ 0 & 0 \end{bmatrix} W^{-1}. \tag{2.21}$$

In case of a non-consistent initial value x_0 or if the input u is not sufficiently smooth, the solution of the descriptor system (2.5) may have impulsive modes, see [8].

2.2 Controllability and Observability

In case of descriptor systems there exists a large number of different kinds of definitions for the controllability and observability of the system. A list of the different kinds is arranged in [11]. In this thesis, only the aspects of complete controllability and observability are considered.

Definition 2.6. *The descriptor system* (2.5) *is called*

(1) completely controllable (C-controllable) *if*

$$\operatorname{rank}[\alpha E - \beta A, B] = n \quad for\ all \quad (\alpha, \beta) \in \mathbb{C} \times \mathbb{C} \setminus \{(0,0)\}.$$

(2) completely observable (C-observable) *if*

$$\operatorname{rank}\left[\alpha E^T - \beta A^T, C^T\right] = n \quad for\ all \quad (\alpha, \beta) \in \mathbb{C} \times \mathbb{C} \setminus \{(0,0)\}.$$

The C-controllability of a system implies that for any given initial state $x_0 \in \mathbb{R}^n$ and final state $x_f \in \mathbb{R}^n$, there is an input u that transfers the state x_0 to the state x_f in finite time. On the other side, the C-observability of a system implies that if the output y is zero for all solutions x of the system with a zero input u, then this system has only the trivial solution $x \equiv 0$, see [8]. These interpretations are conform with the definitions of the controllability and observability of a dynamical system in [1]. A useful algebraic characterization is given in the following theorem.

Theorem 2.2. *(See [8]). Given a descriptor system* (2.5) *with a regular matrix pencil $\lambda E - A$.*

(1) The system (2.5) *is C-controllable if and only if* $\operatorname{rank}[\lambda E - A, B] = n$ *for all finite $\lambda \in \mathbb{C}$ and* $\operatorname{rank}[E, B] = n$.

(2) The system (2.5) *is C-observable if and only if* $\operatorname{rank}\left[\lambda E^T - A^T, C^T\right] = n$ *for all finite $\lambda \in \mathbb{C}$ and* $\operatorname{rank}\left[E^T, C^T\right] = n$.

Another important property of descriptor systems is the stability. For further observations on controllability and observability, the aspect of stability has to be introduced first.

Definition 2.7. *Given a descriptor system of the form* (2.5). *The system is called* asymptotically stable *if* $\lim_{t \to +\infty} x(t) = 0$ *for all solutions $x(t)$ of $E\dot{x}(t) = Ax(t)$.*

This definition is based on the practical interpretation of the systems stability. As in Theorem 2.2, an algebraic characterization using the realization of the descriptor system is more suitable.

Theorem 2.3. *(See [8]). Consider a descriptor system* (2.5) *with a regular matrix pencil $\lambda E - A$. The following statements are equivalent.*

(1) System (2.5) *is asymptotically stable.*

(2) All finite eigenvalues of the pencil $\lambda E - A$ lie in the open left half-plane.

(3) The projected generalized continuous-time Lyapunov equation

$$E^T XA + A^T XE + P_r^T QP_r = 0, \quad X = P_l^T XP_l$$

has a unique Hermitian, positive semidefinite solution X for every Hermitian, positive definite matrix Q.

From now on, the matrix pencil $\lambda E - A$ is called *continuous-time stable (c-stable)* if it is a regular matrix pencil and all the finite eigenvalues of $\lambda E - A$ have negative real parts. The infinite eigenvalues of $\lambda E - A$ do not affect the homogeneous system's behavior at infinity.
Now, the controllability and observability Gramians can be defined analogously to the standard system case [23].
Let $\lambda E - A$ be a c-stable matrix pencil. Then the following integrals exist

$$\mathcal{G}_{pc} = \int\limits_0^{+\infty} \mathcal{F}(t)BB^T\mathcal{F}^T(t)\,\mathrm{d}t, \quad \mathcal{G}_{po} = \int\limits_0^{+\infty} \mathcal{F}^T(t)C^TC\mathcal{F}(t)\,\mathrm{d}t,$$

with $\mathcal{F}$ is as in (2.21). The matrix $\mathcal{G}_{pc}$ is called the *proper controllability Gramian* and $\mathcal{G}_{po}$ the *proper observability Gramian* of the system (2.5), see [8, 23]. The *improper controllability Gramian $\mathcal{G}_{ic}$* and the *improper observability Gramian $\mathcal{G}_{io}$* of the system (2.5) are defined by

$$\mathcal{G}_{ic} = \sum_{k=-\nu}^{-1} F_kBB^TF_k^T, \quad \mathcal{G}_{io} = \sum_{k=-\nu}^{-1} F_k^TC^TCF_k,$$

respectively. The coefficients F_k are the matrices from (2.15).
Considering the standard system case $E = I_n$, the proper Gramians are the usual controllability and observability Gramians [1]. By applying the Parseval identity, the Gramians can be rewritten in the frequency domain in the following form

$$\mathcal{G}_{pc} = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} (j\omega E - A)^{-1}P_lBB^TP_l^T(-j\omega E - A)^{-T}\,\mathrm{d}\omega,$$

$$\mathcal{G}_{po} = \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} (-j\omega E - A)^{-T}P_r^TC^TCP_r(j\omega E - A)^{-1}\,\mathrm{d}\omega,$$

$$\mathcal{G}_{ic} = \frac{1}{2\pi}\int\limits_{0}^{2\pi} (\mathrm{e}^{j\omega}E - A)^{-1}Q_lBB^TQ_l^T(\mathrm{e}^{-j\omega}E - A)^{-T}\,\mathrm{d}\omega,$$

$$\mathcal{G}_{io} = \frac{1}{2\pi}\int\limits_{0}^{2\pi} (\mathrm{e}^{-j\omega}E - A)^{-T}P_r^TC^TCP_r(\mathrm{e}^{j\omega}E - A)^{-1}\,\mathrm{d}\omega.$$

In [23] it has been shown that the proper controllability Gramian $\mathcal{G}_{pc}$ and the proper observability Gramian $\mathcal{G}_{po}$ are the unique, positive semidefinite solutions of the following projected generalized continuous-time Lyapunov equations

$$E\mathcal{G}_{pc}A^T + A\mathcal{G}_{pc}E^T + P_lBB^TP_l^T = 0, \qquad \mathcal{G}_{pc} = P_r\mathcal{G}_{pc}P_r^T, \tag{2.22}$$
$$E^T\mathcal{G}_{po}A + A^T\mathcal{G}_{po}E + P_r^TC^TCP_r^T = 0, \qquad \mathcal{G}_{po} = P_l^T\mathcal{G}_{po}P_l. \tag{2.23}$$

Furthermore, the improper controllability Gramian $\mathcal{G}_{ic}$ and the improper observability Gramian $\mathcal{G}_{io}$ are the unique, positive semidefinite solutions of the projected generalized discrete-time Lyapunov equations

$$A\mathcal{G}_{ic}A^T - E\mathcal{G}_{ic}E^T - Q_lBB^TQ_l^T = 0, \qquad \mathcal{G}_{ic} = Q_r\mathcal{G}_{ic}Q_r^T, \tag{2.24}$$
$$A^T\mathcal{G}_{io}A - E^T\mathcal{G}_{io}E - Q_r^TC^TCQ_r = 0, \qquad \mathcal{G}_{io} = Q_l^T\mathcal{G}_{io}Q_l. \tag{2.25}$$

The Lyapunov equations (2.24) and (2.25) can be rewritten in the form

$$\begin{aligned} A\mathcal{G}_{ic}A^T - E\mathcal{G}_{ic}E^T - (I_n - P_l)BB^T(I_n - P_l)^T &= 0, \qquad & P_r\mathcal{G}_{ic}P_r^T &= 0,\\ A^T\mathcal{G}_{io}A - E^T\mathcal{G}_{io}E - (I_n - P_r)^TC^TC(I_n - P_r) &= 0, \qquad & P_l^T\mathcal{G}_{io}P_l &= 0, \end{aligned}$$

by the application of (2.12). With this formulation, only the spectral projectors corresponding to the finite eigenvalues of the matrix pencil $\lambda E - A$ are needed during computations of the Lyapunov equations.
As well as in standard system case, the Gramians can be used to define the Hankel singular values of the system. These are of great importance in the field of model reduction.
The proper controllability and observability Gramians $\mathcal{G}_{pc}$ and $\mathcal{G}_{po}$ as well as the improper controllability and observability Gramians $\mathcal{G}_{ic}$ and $\mathcal{G}_{io}$ are not system invariant. Using non-singular matrices W and T for the generalized state space transformation of the descriptor system (2.5), the proper system Gramians are transformed into $\tilde{\mathcal{G}}_{pc} = T^{-1}\mathcal{G}_{pc}T^{-T}$ and $\tilde{\mathcal{G}}_{po} = W^{-T}\mathcal{G}_{po}W^{-1}$, whereas the improper system Gramians are transformed into $\tilde{\mathcal{G}}_{ic} = T^{-1}\mathcal{G}_{ic}T^{-T}$ and $\tilde{\mathcal{G}}_{io} = W^{-T}\mathcal{G}_{io}W^{-1}$. It follows from

$$\begin{aligned} \tilde{\mathcal{G}}_{pc}\tilde{E}^T\tilde{\mathcal{G}}_{po}\tilde{E} &= T^{-1}\mathcal{G}_{pc}E^T\mathcal{G}_{po}ET,\\ \tilde{\mathcal{G}}_{ic}\tilde{A}^T\tilde{\mathcal{G}}_{io}\tilde{A} &= T^{-1}\mathcal{G}_{ic}A^T\mathcal{G}_{io}AT, \end{aligned}$$

that spectra of the matrix $\mathcal{G}_{pc}E^T\mathcal{G}_{po}E$ and $\mathcal{G}_{ic}A^T\mathcal{G}_{io}A$ do not change under generalized state space transformations. They are system invariant. These two matrices take the same role in the descriptor system case as the product of controllability and observability Gramian in the standard case [8]. In [26] it has been shown that the eigenvalues of these two matrices are real and non-negative.

Definition 2.8. *Let n_f and n_∞ be the dimensions of the deflating subspaces of the c-stable matrix pencil $\lambda E - A$ corresponding to the finite and infinite eigenvalues, respectively. The square roots of the n_f largest eigenvalues of the matrix $\mathcal{G}_{pc}E^T\mathcal{G}_{po}E$, ordered decreasingly and denoted by $\varsigma_1 \geq \ldots \geq \varsigma_{n_f} \geq 0$, are called the* proper Hankel singular values *of the descriptor system* (2.5)*.The square roots of the n_∞ largest eigenvalues of the matrix $\mathcal{G}_{ic}A^T\mathcal{G}_{io}A$, ordered decreasingly and denoted by $\theta_1 \geq \ldots \geq \theta_{n_\infty} \geq 0$, are called the* improper Hankel singular values *of the descriptor system* (2.5)*.*

The complete set of Hankel singular values of the descriptor system (2.5) is formed by

the sets of proper and improper Hankel singular values. In the standard case $E = I_n$, the proper Hankel singular values are the classical Hankel singular values.
It was mentioned before that the proper and improper controllability and observability Gramians are symmetric and positive semidefinite. Hence, there are Cholesky factorizations of the form

$$\begin{aligned} \mathcal{G}_{pc} &= R_p R_p^T, & \mathcal{G}_{po} &= L_p L_p^T, \\ \mathcal{G}_{ic} &= R_i R_i^T, & \mathcal{G}_{io} &= L_i L_i^T, \end{aligned} \tag{2.26}$$

where the matrices R_p, L_p, R_i and L_i are lower triangular. Another computation opportunity for the Hankel singular values is given by the usage of the factorized Gramians. The results are summarized in the following lemma.

Lemma 2.4. *(See [8]). Given a descriptor system* (2.5) *with a c-stable matrix pencil* $\lambda E - A$. *The proper Hankel singular values of the system* (2.5) *are the* n_f *largest singular values of the matrix* $L_p^T E R_p$ *and the improper Hankel singular values of the system* (2.5) *are the* n_∞ *largest singular values of the matrix* $L_i^T A R_i$, *with* R_p, L_p, R_i *and* L_i *the Cholesky factors from* (2.26).

2.3 Realizations of Descriptor Systems

It was already mentioned that the realization of a descriptor system is not unique. For example, let (E, A, B, C, D) be a realization of a descriptor system. Then another realization of the system with the same input-output behavior is given by

$$\left(\begin{bmatrix} E & 0 \\ 0 & \tilde{E} \end{bmatrix}, \begin{bmatrix} A & 0 \\ 0 & \tilde{A} \end{bmatrix}, \begin{bmatrix} B \\ \tilde{B} \end{bmatrix}, \begin{bmatrix} C & 0 \end{bmatrix}, D \right),$$

where $\tilde{E}, \tilde{A} \in \mathbb{R}^{k \times k}$ and $\tilde{B} \in \mathbb{R}^{k \times m}$ are matrices with arbitrary dimension k. It can be seen that the transfer function of a descriptor system (2.5) is invariant under the addition of uncontrollable and unobservable states.
It has been already seen that the transfer function is invariant under the generalized state space transformation. So, it is possible to obtain a realization of the descriptor system with desired properties. Systems can be transformed into special shapes to study the desired properties. One certain form is the Kalman decomposition. According to the previous section, the system can be split up into the four parts: controllable and observable, controllable but unobservable, uncontrollable but observable, and uncontrollable and unobservable. For any descriptor system (2.5) such a realization can be obtained, see [22].
A direct consequence of the Kalman decomposition and the invariance of the transfer function in terms of uncontrollable and unobservable states is that descriptor systems can be reduced to the controllable and observable part. So, descriptor systems (2.5) can be reduced to realizations of smaller order without changing the input-output behavior.

Definition 2.9. *A realization* $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$ *of a descriptor system* (2.5) *is called* minimal *if the order of the realization is the unique minimal number* $\hat{n} \geq 0$ *of states, necessary to describe the input-output behavior completely. The order of a minimal realization is called* McMillan degree.

In case of descriptor systems, the characterization of the minimal realization is more complicated than for standard systems. Therefor, an additional term for minimality has been introduced.
A minimal realization of the form $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, 0)$ is called *conditionally minimal*, see [22]. The term conditionally refers to the assumption $D = 0$. In this case, some characterizations are given in the following theorem.

Theorem 2.5. *(See [8]). Given a descriptor system* (2.5) *with a c-stable matrix pencil $\lambda E - A$. The following statements are equivalent:*

(1) The realization $(E, A, B, C, 0)$ is conditionally minimal.

(2) The descriptor system (2.5) *is C-controllable and C-observable.*

(3) The following rank conditions hold: $\operatorname{rank}(\mathcal{G}_{pc}) = \operatorname{rank}(\mathcal{G}_{po}) = \operatorname{rank}(\mathcal{G}_{pc} E^T \mathcal{G}_{po} E) = n_f$ *and* $\operatorname{rank}(\mathcal{G}_{ic}) = \operatorname{rank}(\mathcal{G}_{io}) = \operatorname{rank}(\mathcal{G}_{ic} A^T \mathcal{G}_{io} A) = n_\infty$ *hold.*

(4) The proper and improper Hankel singular values of (2.5) *are all positive.*

In general the feed-through term D does not have to be 0. In this case the characterization of a minimal realization differs from the above theorem.

Definition 2.10. *A realization (E, A, B, C, D) of the descriptor system* (2.5) *is called* deflated minimal *if the following conditions hold:*

(1) The realization is C-controllable and C-observable.

(2) The nilpotent matrix N in the Weierstrass canonical form (2.10) *of the pencil $\lambda E - A$ does not contain any Jordan blocks of index one.*

It can be shown that a deflated minimal realization of (2.5) has the same order as a minimal realization of (2.5), see [22].
The second condition in Definition 2.10 is equivalent to

$$A\operatorname{Ker}(E) \subseteq \operatorname{Im}(E),$$

see [8].
Later it is shown that it is not necessary to provide a deflated minimal realization. An appropriate assumption will be given by Theorem 2.5.
Beside these realizations, another important one is introduced in the next definition.

Definition 2.11. *A realization (E, A, B, C, D) of the descriptor system* (2.5) *is called* balanced *if*

$$\mathcal{G}_{pc} = \mathcal{G}_{po} = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{G}_{ic} = \mathcal{G}_{io} = \begin{bmatrix} 0 & 0 \\ 0 & \Theta \end{bmatrix},$$

where $\Sigma = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_{n_f})$ is a diagonal matrix containing the proper Hankel singular values and $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_{n_\infty})$ is a diagonal matrix containing the improper Hankel singular values of the system (2.5).

For a conditionally minimal realization of the descriptor system (2.5) with a c-stable matrix pencil $\lambda E - A$, it is possible to find a generalized state space transformation with non-singular matrices W_b and T_b, such that the transformed realization $(W_b E T_b, W_b A T_b, W_b B, C T_b, D)$ is balanced. A possible formulation of the matrices W_b and T_b is given by

$$\begin{aligned} W_b &= \left[L_p U_p \Sigma^{-\frac{1}{2}}, \quad L_i U_i \Theta^{-\frac{1}{2}} \right], \\ T_b &= \left[R_p V_p \Sigma^{-\frac{1}{2}}, \quad R_i V_i \Theta^{-\frac{1}{2}} \right], \end{aligned} \tag{2.27}$$

where R_p, R_i, L_p, and L_i are the lower triangular Cholesky factors from (2.26) and U_p, U_i, V_p, and V_i are the orthogonal matrices from the singular value decompositions of $L_p^T E R_p$ and $L_i^T A R_i$. The balanced realization of descriptor systems is not unique [8]. Note that for the transformation matrices in (2.27) it holds

$$E_b = W_b E T_b = \begin{bmatrix} I_{n_f} & 0 \\ 0 & E_\infty \end{bmatrix} \quad \text{and} \quad A_b = W_b A T_b = \begin{bmatrix} A_f & 0 \\ 0 & I_\infty \end{bmatrix}, \tag{2.28}$$

with E_∞ nilpotent and A_f nonsingular. The matrix pencil $\lambda E_b - A_b$ of the balanced realization resembles the Weierstrass canonical form (2.10).

2.4 System Norms for Descriptor Systems

A special property of transfer functions of descriptor systems (2.8) is, that even if G has no poles on the imaginary axis, the transfer function might be unbounded on $j\mathbb{R}$. This motivates the following definition [8, 30].

Definition 2.12. *The transfer function G is called* proper *if* $\lim\limits_{s \to +\infty} ||G(s)|| < +\infty$ *for any induced matrix norm* $||.||$ *and* improper *otherwise. If* $\lim\limits_{s \to +\infty} ||G(s)|| = 0$, *then G is called* strictly proper.

In order to define a norm for the transfer function of a descriptor system, some spaces have to be defined first [30].

Definition 2.13. *The Banach space of all $p \times m$ matrix-valued functions that are essentially bounded on $j\mathbb{R}$ is denoted by $\mathcal{L}_\infty^{p \times m}$. The rational subspace of $\mathcal{L}_\infty^{p \times m}$, termed by $\mathcal{RL}_\infty^{p \times m}$, consists of all proper and real rational $p \times m$ transfer functions with no poles on the imaginary axis.*

For these spaces, the corresponding norm can be defined as follows.

Definition 2.14. *For a matrix-valued function $F \in \mathcal{L}_\infty^{p \times m}$ the $\mathcal{L}_\infty$-norm is defined as*

$$||F||_{\mathcal{L}_\infty} = \text{ess} \sup_{\omega \in \mathbb{R}} \sigma_{\max}(F(i\omega)), \tag{2.29}$$

where $\sigma_{\max}(M)$ denotes the maximum singular value of the matrix M.

For proper transfer functions G of descriptor systems, i.e., $G \in \mathcal{RL}_\infty^{p \times m}$, the definition formula (2.29) simplifies to

$$||G||_{\mathcal{L}_\infty} = \sup_{\omega \in \mathbb{R}} \sigma_{\max}(G(i\omega)),$$

because G is continuous on the imaginary axis.
Remember the formulation of the approximation error in (1.3). Beside the norm for the transfer function G, also norms for the vector-valued input and output functions are needed.
In the time domain $[0, +\infty) \subseteq \mathbb{R}$ the $\mathcal{L}_2$-norm of a square integrable function $u(t) \in \mathbb{R}^m$ is given as the integral

$$||u||_{\mathcal{L}_2} = \left(\int\limits_0^{+\infty} u(t)^T u(t) \, \mathrm{d}t \right)^{\frac{1}{2}}.$$

By applying the Laplace transformation (2.6) on the $\mathcal{L}_2$-norm, the corresponding norm in the frequency domain is given by

$$||U||_{\mathcal{L}_2} = \left(\frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} U(-j\omega)^T U(j\omega) \, \mathrm{d}\omega \right)^{\frac{1}{2}}.$$

The definition of the $\mathcal{L}_2$-norm in the frequency domain can be used to derive an upper bound on the output Y of the descriptor system. It holds

$$\begin{aligned} ||Y||_{\mathcal{L}_2} &= ||GU||_{\mathcal{L}_2} \\ &= \left(\frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} ||G(j\omega)U(j\omega)||_2^2 \, \mathrm{d}\omega \right)^{\frac{1}{2}} \\ &\leq \left(\frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} \left[||G(j\omega)||_2 \, ||U(j\omega)||_2 \right]^2 \, \mathrm{d}\omega \right)^{\frac{1}{2}} \\ &\leq \sup_{\omega \in \mathbb{R}} ||G(j\omega)||_2 \left(\frac{1}{2\pi} \int\limits_{-\infty}^{+\infty} ||U(j\omega)||_2^2 \, \mathrm{d}\omega \right)^{\frac{1}{2}} \\ &= ||G||_{\mathcal{L}_\infty} \, ||U||_{\mathcal{L}_2}, \end{aligned} \tag{2.30}$$

where U is the input, G the transfer function and Y the output of the descriptor system (2.5) in the frequency domain.
Usually, only asymptotically stable systems are considered. Then, the space $\mathcal{H}_\infty$ is used, containing all proper transfer functions which are analytic and bounded in the open right half-plane. The space $\mathcal{H}_\infty$ is a closed subset of the space $\mathcal{L}_\infty$, see [30].
In further discussions, mainly the $\mathcal{H}_\infty$-norm is considered. It is given by

$$||F||_{\mathcal{H}_\infty} = \sup_{\mathrm{Re}(s) > 0} \sigma_{\max}(F(s)) = \sup_{\omega \in \mathbb{R}} \sigma_{\max}(F(j\omega)).$$

On the $\mathcal{H}_\infty$ space this definition is identical to the one of the $\mathcal{L}_\infty$-norm.
One of the goals of model reduction was to determine a bound on the approximation error as in (1.3). Using the Parseval identity and the obtained bound in (2.30), in the

time and frequency domain it holds

$$||y - \hat{y}||_{\mathcal{L}_2} \leq \left|\left|G - \hat{G}\right|\right|_{\mathcal{H}_\infty} ||u||_{\mathcal{L}_2},$$

where y is the original output, $\hat{y}$ is the approximated output, G is the transfer function of the original descriptor system, $\hat{G}$ is the transfer function of the reduced system and u is the input.
It can be shown that it holds

$$||G||_{\mathcal{H}_\infty} = \sup_{u \neq 0} \frac{||y||_{\mathcal{L}_2}}{||u||_{\mathcal{L}_2}}$$

in frequency and time domain, see [8, 30]. That is, the $\mathcal{H}_\infty$-norm of G is the ratio between the output and input energy of the descriptor system (2.5).
According to the decoupling of descriptor systems into the slow subsystem (2.18) and the fast subsystem (2.19), the transfer function (2.8) can be additively decomposed into

$$G(s) = G_{sp}(s) + P(s), \tag{2.31}$$

where G_{sp} is a strictly proper transfer function and P a polynomial one. With the block form of the Weierstrass canonical form (2.10) and the partition of the input and output matrices (2.17), the strictly proper part can be written in the form

$$G_{sp}(s) = C_f \left(sI_{n_f} - J\right)^{-1} B_f$$

and the polynomial part as

$$P(s) = C_\infty \left(sN - I_{n_\infty}\right)^{-1} B_\infty + D.$$

In the following, the feed-through term D is set to zero without loss of generality. For $D \neq 0$, a generalized descriptor system of the form

$$\begin{aligned} \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} \dot{\xi}(t) &= \begin{bmatrix} A & 0 \\ 0 & I_k \end{bmatrix} \xi(t) + \begin{bmatrix} B \\ D_2 \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} C & -D_1 \end{bmatrix} \xi(t) \end{aligned} \tag{2.32}$$

can be considered instead of (2.5). Here $D = D_1 D_2$ is a factorization of the feed-through term, for example, $D_1 = I_k$ and $D_2 = D$ can be chosen. This generalized system is equivalent to (2.5) in the sense that $x(t)$ is the solution of (2.5) with a given input u if and only if

$$\xi(t) = \begin{bmatrix} x(t) \\ D_2 u(t) \end{bmatrix}$$

satisfies the generalized system, see [25].
In this thesis, the problem of an optimal Hankel-norm approximation shall be considered. To introduce the Hankel-norm, the Sobolev spaces have to be considered first.

Definition 2.15. *Suppose $u \in \mathcal{L}_p(\Omega)$ and there are derivatives $\frac{\mathrm{d}^\alpha}{\mathrm{d}t^\alpha}u$ for a non-negative integer $\alpha \leq k$, such that*

$$\frac{\mathrm{d}^\alpha}{\mathrm{d}t^\alpha}u \in \mathcal{L}_p(\Omega)$$

holds for all $\alpha \leq k$. Then it is said that $u \in \mathcal{W}_p^k(\Omega)$, where $\mathcal{W}_p^k(\Omega)$ is called Sobolev space*.*

In the following, the assumed domain is $\Omega = (-\infty, 0]$. Also, $p = 2$ and $k = \nu - 1$ are set, with ν index of the descriptor system (2.5).
The *Hankel operator*

$$\mathcal{H} : \mathcal{W}_2^{\nu-1}(-\infty, 0] \to \mathcal{L}_2[0, +\infty)$$

is a mapping from past inputs $u_- : (-\infty, 0] \to \mathbb{R}^m$ to present and future system outputs $y_+ : (0, +\infty] \to \mathbb{R}^p$. It ignores the system response before the time 0.

Definition 2.16. *(See [12]). The* Hankel operator *$\mathcal{H}$ of the descriptor system* (2.5) *is defined as sum of the Hankel operators of the strictly proper part $\mathcal{H}_{sp}$ and the polynomial part $\mathcal{H}_p$ as*

$$\mathcal{H} = \mathcal{H}_{sp} + \mathcal{H}_p,$$

where the operators can be represented by

$$(\mathcal{H}_{sp}u)(t) = \int\limits_{-\infty}^{0} C_f \mathrm{e}^{J(t-\tau)} B_f u(\tau)\,\mathrm{d}\tau,$$

and

$$(\mathcal{H}_p u)(t) = -\sum_{k=0}^{\nu-1} C_\infty N^k B_\infty u^{(k)}(t),$$

with $t \geq 0$.

Now, the $\mathcal{L}_2$-norm can be used to measure the effect of the past inputs on future outputs.

Definition 2.17. *The Hankel-norm of a transfer function G is given by*

$$||G||_H = \sup_{u_- \in \mathcal{W}_2^{\nu-1}(-\infty,0]} \frac{||y_+||_{\mathcal{L}_2}}{||u_-||_{\mathcal{L}_2}}.$$

In case of standard systems, the Hankel-norm can be written as

$$||G||_H = \varsigma_{\max}(G), \tag{2.33}$$

where $\varsigma_{\max}(G)$ denotes the largest proper Hankel singular value of G, see [13]. It can be noted that the Hankel-norm is only a semi-norm on the Hardy space $\mathcal{H}_\infty$. It is easy to see that $||G||_H = 0$ does not imply $G \equiv 0$.

3 State of the Art

The problem of constructing an optimal Hankel-norm approximation was already considered for standard systems of the form

$$\begin{aligned}\dot{x}(t) &= Ax(t) + Bu(t),\\ y(t) &= Cx(t) + Du(t).\end{aligned} \tag{3.1}$$

In the next section, the basic theory of the Hankel-norm approximation for standard systems, introduced by Glover in [13], is summarized. For simplicity, the realization of the standard system (3.1) is written as (A, B, C, D). If the feed-through term D does not matter, the realization of (3.1) is reduced to (A, B, C).
A first approach on the generalized Hankel-norm approximation was already considered in [12]. A summary of the theory and the resulting algorithm is shown in the last section of this chapter.

3.1 Basic Functioning

The Hankel-norm approximation method introduced by Glover can be seen as an extension of the balanced truncation model reduction method. A detailed version of the following theoretical aspects can be found in [13].
First of all, the following definition about the position of the eigenvalues of a matrix will be needed for further discussions.

Definition 3.1. *The* inertia *of a general complex, square matrix A denoted as* $\mathrm{In}(A)$ *is the triple* $(\pi(A), \nu(A), \delta(A))$*, where*

(1) $\pi(A)$ is the number of eigenvalues of A in the open right half-plane,

(2) $\nu(A)$ is the number of eigenvalues of A in the open left half-plane,

(3) $\delta(A)$ is the number of eigenvalues of A on the imaginary axis.

The main idea of the Hankel-norm approximation is based on the relation between the Hankel singular values and the frequency response. Therefor, the following definition introduces para-conjugate unitary rational matrices, further known as all-pass transfer functions.

Definition 3.2. *A transfer function G is called* all-pass *if*

$$G(s)G^H(-\bar{s}) = I_p$$

holds for all $s \in \mathbb{C}$.

As for other system theoretical aspects before, it is useful to consider an algebraic characterization of all-pass transfer functions. Therefor, the next theorem is introduced.

Theorem 3.1. *(See [13]). Given a realization (A, B, C) with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{m\times n}$.*

(1) If (A, B, C) is controllable and observable, the following two statements are equivalent:

(a) There exists a D with $G(s)G^H(-\bar{s}) = \sigma^2 I_m$ for all $s \in \mathbb{C}$, where $G(s) = C(sI_n - A)B + D$.

(b) There exist P and Q, such that

(i) $P = P^H$, $Q = Q^H$,

(ii) $AP + PA^H + BB^H = 0$,

(iii) $A^HQ + QA + C^HC = 0$,

(iv) $PQ = \sigma^2 I_n$.

(2) Let part (1b) be satisfied. Then there exists a D with

$$\begin{aligned} D^H D &= \sigma^2 I_m, \\ D^H C + B^H Q &= 0, \\ DB^H + CP &= 0, \end{aligned}$$

and any such D will satisfy part (1a). Note that observability and controllability are not assumed.

First of all, note that this characterization does not need the assumption of a stable standard system. In case of a non-stable system, the Gramians P and Q are not defined. But the Lyapunov equations in part (1b(ii)) and (1b(iii)) still have solutions P and Q satisfying $PQ = \sigma^2 I_n$. Also, this theorem shows that the Hankel singular values of an all-pass transfer function are all equal to 1.

Another point to consider is the stability of the transfer function. Beside the Hardy space $\mathcal{H}_\infty$ of all proper transfer functions which are bounded and analytic in the right open half-plane, the following definition has to be used.

Definition 3.3. *The space of all transfer functions $G(s) : \mathbb{C} \to \mathbb{C}^{p\times m}$, which are bounded and analytic in the open left half-plane, is denoted by $\mathcal{H}_\infty^-$. Furthermore, the transfer function $G \in \mathcal{H}_\infty$ is called* stable *and the function $F \in \mathcal{H}_\infty^-$ is called* anti-stable*.*

One can show, for each transfer function $G \in \mathcal{H}_\infty$ it holds

$$||G||_H = \inf_{F\in\mathcal{H}_\infty^-} ||G - F||_{\mathcal{L}_\infty}. \tag{3.2}$$

This is, an approximation with the smallest possible $\mathcal{L}_\infty$ error of a stable transfer function can be made by an anti-stable one. Also, the smallest $\mathcal{L}_\infty$ error is given by the Hankel-norm of the transfer function G. An explicit construction of such an anti-stable F can be made by the application of the following theorem, assuming an already balanced realization.

Theorem 3.2. *(See [13]). Let the realization* (A, B, C, D) *with* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times n}$, *and* $D \in \mathbb{R}^{m\times m}$ *satisfy*

$$AP + PA^T + BB^T = 0, \tag{3.3}$$
$$A^TQ + QA + C^TC = 0, \tag{3.4}$$

for

$$P = P^T = diag(\Sigma_1, \sigma I_r), \tag{3.5}$$
$$Q = Q^T = diag(\Sigma_2, \sigma I_r), \tag{3.6}$$

with Σ_1 *and* Σ_2 *diagonal,* $\sigma \neq 0$ *and* $\delta\left(\Sigma_1\Sigma_2 - \sigma^2 I_{n-r}\right) = 0$. *Partition* (A, B, C) *conformally with* P *and* Q*, such that*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \tag{3.7}$$

and define

$$\tilde{A} = \Gamma^{-1}(\sigma^2 A_{11}^T + \Sigma_2 A_{11}\Sigma_1 - \sigma C_1^T U B_1^T), \tag{3.8}$$
$$\tilde{B} = \Gamma^{-1}(\Sigma_2 B_1 + \sigma C_1^T U), \tag{3.9}$$
$$\tilde{C} = C_1\Sigma_1 + \sigma U B_1^T, \tag{3.10}$$
$$\tilde{D} = D - \sigma U, \tag{3.11}$$

where U *is a unitary matrix satisfying*

$$B_2 = -C_2^T U \tag{3.12}$$

and

$$\Gamma = \Sigma_1\Sigma_2 - \sigma^2 I_{n-r}. \tag{3.13}$$

Also, define the error system of the form

$$A_e = \begin{bmatrix} A & 0 \\ 0 & \tilde{A} \end{bmatrix}, \quad B_e = \begin{bmatrix} B \\ \tilde{B} \end{bmatrix}, \quad C_e = \begin{bmatrix} C, & -\tilde{C} \end{bmatrix}, \quad D_e = D - \tilde{D}. \tag{3.14}$$

Then it holds

(1) (A_e, B_e, C_e) *satisfy the Lyapunov equations*

$$A_eP_e + P_eA_e^T + B_eB_e^T = 0, \tag{3.15}$$
$$A_e^TQ_e + Q_eA_e + C_e^TC_e = 0, \tag{3.16}$$

with

$$P_e = \begin{bmatrix} \Sigma_1 & 0 & I_{n-r} \\ 0 & \sigma I_r & 0 \\ I_{n-r} & 0 & \Sigma_2 \Gamma^{-1} \end{bmatrix}, \tag{3.17}$$

$$Q_e = \begin{bmatrix} \Sigma_2 & 0 & -\Gamma \\ 0 & \sigma I_r & 0 \\ -\Gamma & 0 & \Sigma_1 \Gamma \end{bmatrix}, \tag{3.18}$$

$$P_e Q_e = \sigma^2 I. \tag{3.19}$$

(2) Defining the error transfer function $E(s) = D_e + C_e \left(sI_{2n-r} - A_e\right)^{-1} B_e$. *Then it holds* $E(s)E^H(-\bar{s}) = \sigma^2 I_{2n-r}$.

(3) If $\delta(A) = 0$ *then*

(a) $\delta(\tilde{A}) = 0$.

(b) If $\delta(\Sigma_1\Sigma_2) = 0$ *then*

$$In(\tilde{A}) = In(-\Sigma_1\Gamma) = In(-\Sigma_2\Gamma).$$

(c) If $P > 0$ *and* $Q > 0$*, then the McMillan degree of the stable part of* $(\tilde{A}, \tilde{B}, \tilde{C})$ *equals* $\pi(\Sigma_1\Gamma) = \pi(\Sigma_2\Gamma)$.

(d) If either

(i) $\Sigma_1\Gamma > 0$ *and* $\Sigma_2\Gamma > 0$ *or*

(ii) $\Sigma_1\Gamma < 0$ *and* $\Sigma_2\Gamma < 0$,

then $(\tilde{A}, \tilde{B}, \tilde{C})$ *is a minimal realization.*

The transformation formulas (3.8)-(3.11) are a direct result of the characterization of all-pass transfer functions in Theorem 3.1.
Let $\tilde{G}$ be the system constructed by the formulas (3.8)-(3.11). Then the normalized error transfer function of the form $\sigma^{-1}(G - \tilde{G})$ is all-pass. A detailed step-by-step derivation of the transformation formulas from the characterization of all-pass transfer functions can be found in chapter 8 in [1].
A direct consequence of Theorem 3.2 is the formula (3.2). Therefor, the number r is chosen, such that

$$\sigma_1 = \ldots = \sigma_r > \sigma_{r+1} \geq \sigma_{r+2} \geq \sigma_n > 0.$$

The resulting system $\tilde{G}$ is constructed, such that $\sigma_1^{-1}(G - \tilde{G})$ is all-pass, so

$$\left|\left|G - \tilde{G}\right|\right|_{\mathcal{L}_\infty} = \sigma_1.$$

From part (3) of Theorem 3.2 it follows that all poles of $\tilde{G}$ lie in the open left half-plane, so $\tilde{G}$ is anti-stable.
Next, the error of the transformed system has to be considered.

Lemma 3.3. *(See [13]). Given a stable transfer function G of dimensions $p \times m$ with Hankel singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} \geq \sigma_{r+2} \geq \ldots \geq \sigma_n > 0$. Then for all stable $\hat{G}$ and McMillan degree $\leq r$ it holds*

$$\left|\left|G - \hat{G}\right|\right|_H \geq \sigma_{r+1}(G).$$

Lemma 3.3 gives the lower bound on the Hankel-norm error for all stable systems of order r. Now, a certain system $\hat{G}$ has to be constructed to fulfill this lower error bound. This is the problem of the *optimal Hankel-norm approximation.*

Now, apply the transformation formulas (3.8)-(3.11) to a stable standard system G with the chosen Hankel singular value σ_{r+1}. From part (3) of Theorem 3.2 it follows that the resulting system $\tilde{G}$ has the form

$$\tilde{G} = \hat{G} + F, \tag{3.20}$$

with $\hat{G} \in \mathcal{H}_\infty$ and $F \in \mathcal{H}_\infty^-$. The corresponding error system $E = G - \hat{G} - F$ satisfies

$$E(s)E^H(-\bar{s}) = \sigma_{r+1} I_p. \tag{3.21}$$

From part (3) of Theorem 3.2 it follows that the system $\hat{G}$ is stable and has the order r. Also, it holds

$$\left|\left|G - \hat{G}\right|\right|_H = \sigma_{r+1}(G).$$

Hence, $\hat{G}$ is an optimal Hankel-norm approximation of G. Note that the realization of $\hat{G}$ is minimal.

Let k be the multiplicity of the Hankel singular value σ_{r+1}. With part (3) of Theorem 3.2 it follows that the system corresponding to the anti-stable transfer function $F(s)$ has the order $n - r - k$. From the formula (3.2), the additive decomposition of the transformed system (3.20) and the property of the error system (3.21), it follows that

$$\inf_{F \in \mathcal{H}_\infty^-, \hat{G}} \left|\left|G - \hat{G} - F\right|\right|_{\mathcal{L}_\infty} = \sigma_{r+1}(G).$$

This shows a relation between the Hankel singular values of a standard system (3.1) and an $\mathcal{L}_\infty$ optimization problem.

The formulas in Theorem 3.2 can be extended to non-square systems by using an explicit formulation for the unitary matrix in (3.12) with the condition $UU^H \leq I_p$. A commonly used example is

$$U = -C_2(B_2^T)^\dagger,$$

where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of M.

Note that the Hankel-norm approximation is not unique, since the Hankel singular values do not depend on the feed-through term D. So, the choice of $\hat{D} \in \mathbb{R}^{p \times m}$ is arbitrary. However, the $\mathcal{H}_\infty$ error depends on $\hat{D}$. Further characterizations of all optimal Hankel-norm approximations can be made by the formulas of the error system (3.14)-(3.19). In Algorithm 1 the complete Hankel-norm approximation method is summarized. Numerical tests for this algorithm can be found in [9].

```
Algorithm 1: Hankel-Norm Approximation for Standard Systems
Input: Stable realization $(A, B, C, D)$ of a standard systems (3.1)
Output: Realization of an optimal Hankel-norm approximation $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$
1: Compute a minimal balanced realization $(\breve{A}, \breve{B}, \breve{C}, D)$ of $(A, B, C, D)$ with
   Gramians
       $\mathcal{G}_c = \mathcal{G}_o = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_{n_{\min}})$,
   where $n_{\min}$ denotes the McMillan degree.
2: Choose a Hankel singular value $\varsigma_{r+1}$.
3: Permute the balanced realization $(\breve{A}, \breve{B}, \breve{C}, D)$, such that the Gramians have the
   form
       $\breve{\mathcal{G}}_c = \breve{\mathcal{G}}_o = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_r, \varsigma_{r+k+1}, \ldots, \varsigma_{n_{\min}}, \varsigma_{r+1} I_k)$
       $= \operatorname{diag}(\Sigma, \varsigma_{r+1} I_k)$,
   where $k$ is the multiplicity of the Hankel singular value $\varsigma_{r+1}$.
4: Partition the resulting permuted system according to the Gramians
       $\breve{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \breve{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \breve{C} = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$,
   where $A_{22} \in \mathbb{R}^{k \times k}$, $B_2 \in \mathbb{R}^{k \times m}$, and $C_2 \in \mathbb{R}^{p \times k}$.
5: Compute the transformation
       $\tilde{A} = \Gamma^{-1}(\sigma_{r+1}^2 A_{11}^T + \Sigma A_{11} \Sigma + \sigma_{r+1} C_1^T U B_1^T)$,
       $\tilde{B} = \Gamma^{-1}(\Sigma B_1 - \sigma_{r+1} C_1^T U)$,
       $\tilde{C} = C_1 \Sigma - \sigma_{r+1} U B_1^T$,
       $\tilde{D} = D + \sigma_{r+1} U$,
   with $U = (C_2^T)^\dagger B_2$ and $\Gamma = \Sigma^2 - \sigma_{r+1}^2 I_{n-k}$.
6: Compute the additive decomposition
       $\tilde{G} = \tilde{C}(sI_{n-k} - \tilde{A})^{-1}\tilde{B} + \tilde{D} = \hat{G}(s) + F(s)$,
   where $F$ is anti-stable and $\hat{G}$ is the stable Hankel-norm approximation with the
   realization $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$.
```

Let G be a standard system (3.1) and $\hat{G}$ an optimal Hankel-norm approximation computed by Algorithm 1. An $\mathcal{H}_\infty$ error bound is given by

$$\left|\left|G - \hat{G}\right|\right|_{\mathcal{H}_\infty} \leq 2(\sigma_{r+1} + \ldots + \sigma_n), \tag{3.22}$$

with σ_i the Hankel singular values of G, see [1]. This is the same error bound as for the balanced truncation method. It can be shown that there exists a D_0, such that

$$\left|\left|G - \hat{G} - D_0\right|\right|_{\mathcal{H}_\infty} \leq \sigma_{r+1} + \ldots + \sigma_n.$$

Still, there is no algorithm to compute such a D_0 for the Hankel-norm approximation. An alternative is given by an algorithm in [13]. There, a $\hat{D}_0$ is computed, such that

$$\left|\left|G - \hat{G} - \hat{D}_0\right|\right|_{\mathcal{H}_\infty} \leq \sigma_{r+1} + \mu_1 + \ldots + \mu_{n-r-k},$$

where $\mu_1, \ldots, \mu_{n-r-k}$ denote the Hankel singular values corresponding to the anti-stable system, computed by the transformation formulas (3.8)-(3.11).

3.2 Limits of Current Development

Still, there a several open points and questions concerning the Hankel-norm approximation method. One problem, announced before, was the construction of an appropriate feed-trough term D_0 for a more suitable $\mathcal{H}_\infty$ error bound. Another problem is the usage of the scaling matrix (3.13) in the transformations (3.8) and (3.9). Here, the typical numerical problems occur considering the division by small numbers. Especially for a small chosen Hankel singular value σ_{r+1}, the transformation becomes numerically unstable. A similar problem occurs if the balanced minimal realization contains too small Hankel singular values. Note that part (3) of Theorem 3.2 can be used as a criterion for numerical instability, since the numbers of anti-stable and stable poles of the resulting system are predetermined.

A further open point is the influence of the error accuracy used in partial computations. In Algorithm 1 there may be many steps needed to be computed up to a given accuracy tolerance. A special case is the application of the Hankel-norm approximation method on large-scale sparse systems. Another example is the use of iterative solvers during the computation of the minimal balanced realization and the additive decomposition of the transformed system.

The application of the generalized Hankel norm approximation method on large-scale sparse descriptor systems will be considered in the next chapter. The introduced theory can be used for the standard system case, too.

Until now, the Hankel-norm approximation method was only considered for standard systems of the form (3.1). But in practice, a version for descriptor systems

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned}$$

with a singular E matrix, would be beneficial. The problematic of the generalized Hankel-norm approximation is not completely new. A special solution approach for the generalized Hankel-norm approximation method is presented in the next section.

3.3 Display of a Special Solution Approach

The approach, presented in this section, is based on the work of Cao, Saltik and Weiland on the generalized Hankel-norm approximation in [12].
At first, a strong assumption has to be made. The descriptor system (2.5) must be transformed into the Weierstrass canonical form (2.10) with a partition according to the block structure (2.17). For simplicity, the feed-through term D is assumed as zero. For a non-zero term, the generalized descriptor system (2.32) can be used. The resulting system has the form

$$\begin{aligned} \begin{bmatrix} I_{n_f} & 0 \\ 0 & N \end{bmatrix} \dot{x}(t) &= \begin{bmatrix} J & 0 \\ 0 & I_{n_\infty} \end{bmatrix} x(t) + \begin{bmatrix} B_f \\ B_\infty \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} C_f, C_\infty \end{bmatrix} x(t). \end{aligned} \tag{3.23}$$

Moreover, the system is assumed to be c-stable and conditionally minimal.
Next, the system (3.23) is considered in its decoupled form with the slow subsystem

$$\begin{aligned} \dot{x}_f(t) &= Jx_f(t) + B_f u(t), \\ y_f(t) &= C_f x_f(t), \end{aligned} \tag{3.24}$$

and the fast subsystem

$$\begin{aligned} N\dot{x}_\infty(t) &= x_\infty(t) + B_\infty u(t), \\ y_\infty(t) &= C_\infty x_\infty(t). \end{aligned} \tag{3.25}$$

As the first case, let (3.23) be an index-1 descriptor system. Then the equations of the fast subsystem simplify to

$$\begin{aligned} 0 &= x_\infty(t) + B_\infty u(t), \\ y_\infty(t) &= C_\infty x_\infty(t), \end{aligned}$$

which can be rewritten as

$$y_\infty(t) = -C_\infty B_\infty u(t).$$

Finally, the complete descriptor system (3.23) simplifies to the form

$$\begin{aligned} \dot{x}_f(t) &= Jx_f(t) + B_f u(t), \\ y(t) &= C_f x_f(t) - C_\infty B_\infty u(t). \end{aligned} \tag{3.26}$$

So, the fast subsystem of an index-1 descriptor system is a static gain and the complete system can be written in standard form with the additional feed-trough term $-C_\infty B_\infty$. Now, a standard Hankel-norm approximation method can be used to reduce the system (3.26). For example, the method shown in Algorithm 1 would be an opportunity. The Hankel-norm approximation of (3.26) is the generalized Hankel-norm approximation of (3.23).
Next, let the index of the descriptor system be $\nu \geq 2$. For the reduction of the fast

subsystem (3.25), the block structure of N is considered

$$N = \begin{bmatrix} N_1 & & & \\ & N_2 & & \\ & & \ddots & \\ & & & N_l \end{bmatrix}, \tag{3.27}$$

where N_i are the Jordan blocks with zeros on the diagonal. Each block in (3.27) forms a new subsystem of the fast subsystem (3.25).

For SISO (single-input, single output) descriptor systems an explicit form of the minimal realization of the fast subsystem is given by the following theorem.

Theorem 3.4. *(See [12]). Given a fast SISO subsystem of the form* (3.25) *with index ν. If and only if the entire system* (3.23) *is C-controllable and C-observable a possible realization of* (3.25) *is given by*

$$\begin{aligned} \tilde{N}\dot{\tilde{x}}_\infty(t) &= \tilde{x}_\infty(t) + \tilde{B}_\infty u(t), \quad \tilde{x}_\infty(t_0) = \tilde{x}^0_\infty, \\ \tilde{y}_\infty(t) &= \tilde{C}_\infty \tilde{x}_\infty(t), \end{aligned} \tag{3.28}$$

where $\tilde{N} \in \mathbb{R}^{\nu\times\nu}$ is the largest block of (3.27)*, $\tilde{B}_\infty = \begin{bmatrix} 0, & 0, & \ldots, & 0, & -1 \end{bmatrix}^T \in \mathbb{R}^\nu$ and $\tilde{C}_\infty = -\begin{bmatrix} C_\infty N^{\nu-1} B_\infty, & \ldots, & C_\infty N^0 B_\infty \end{bmatrix} \in \mathbb{R}^\nu$.*

The construction of $\tilde{C}_\infty$ is done according to the solution of the fast subsystem (2.20). Let $P(s)$ be the transfer function of (3.25) and $\tilde{P}(s)$ the transfer function of (3.28). Then it holds $P(s) = \tilde{P}(s)$. A special feature of this method is the preservation of the index of the system.

The construction proposed in Theorem 3.4 can be extended to the MIMO (multi-input, multi-output) system case with m inputs and p outputs. Therefor, it is assumed that $\nu m \geq n_\infty$. Then the term $\tilde{B}_\infty$ is constructed by replacing the scalars by matrices of the form

$$\tilde{B}_\infty = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -I_m \end{bmatrix} \in \mathbb{R}^{\nu m \times m}.$$

The nilpotent matrix $\tilde{N}$ is no longer the largest block of (3.27) but

$$\tilde{N} = \begin{bmatrix} 0 & I_m & & 0 \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & I_m \\ 0 & & 0 & 0 \end{bmatrix} \in \mathbb{R}^{\nu m \times \nu m}.$$

The construction of the matrix $\tilde{C}_\infty$ does not change. The dimensions of the matrices depend on the number of inputs m, the index of the system ν and the number of states n_∞ of the fast subsystem (3.25). Let k be the dimension of $\tilde{B}_\infty$ and $\tilde{N}$, then it holds

$$k = \min(\nu m, n_\infty).$$

For the slow subsystem, a standard Hankel-norm approximation method can be applied. Again the method in Algorithm 1 would be possible. The complete reduced order model is then constructed by an additive decomposition of the Hankel-norm approximation of the slow subsystem and the constructed fast subsystem.
Let $\tilde{P}$ be the minimal fast subsystem and $\tilde{G}$ be the r-th order Hankel-norm approximation of (3.24). Then the generalized Hankel-norm approximation $\hat{G}$ of the descriptor system G is given by

$$\hat{G} = \tilde{G} + \tilde{P}.$$

Since $\hat{G}$ is a standard Hankel-norm approximation of G_{sp} and $P = \tilde{P}$ holds, all error bounds of the standard Hankel-norm approximation method can still be used.
That means, the reduced-order model fulfills

$$\begin{aligned} \left|\left|G - \hat{G}\right|\right|_H &= \left|\left|G_{sp} + P - \tilde{G} - \tilde{P}\right|\right|_H \\ &= \left|\left|G_{sp} - \tilde{G}\right|\right|_H \\ &= \varsigma_{r+1}(G), \end{aligned}$$

where $\varsigma_{r+1}(G)$ is the $(r+1)$-st proper Hankel singular value of G.
The complete generalized Hankel-norm approximation method based on the Weierstrass canonical form is summarized in Algorithm 2.
The main disadvantage of this solution approach is step 1 in Algorithm 2. The computation of the Weierstrass canonical form needs a high amount of computational effort. An efficient method for this computation is given in [16]. Even so, the descriptor system already has to be conditionally minimal, since both the construction of the fast subsystem as well as the use of Algorithm 1 for the computation of the standard Hankel-norm approximation assume the conditional minimality of the descriptor system. For an broad application of this method, the computation of the minimal realization should be adjusted to the computation of the Weierstrass canonical form.
Due to this problematic points, a more general and practicable method is presented in the next chapter.

Algorithm 2: Generalized Hankel-Norm Approximation using the Weierstrass Canonical Form

```
Input: Conditionally minimal realization $(E, A, B, C, 0)$, such that $\lambda E - A$ is
       c-stable
Output: Realization of Hankel-norm approximation $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$
Compute the Weierstrass canonical form and the transformed system
    $$\begin{bmatrix} I_{n_f} & 0 \\ 0 & N \end{bmatrix} \dot{x}(t) = \begin{bmatrix} J & 0 \\ 0 & I_{n_\infty} \end{bmatrix} x(t) + \begin{bmatrix} B_f \\ B_\infty \end{bmatrix} u(t), \quad y(t) = \begin{bmatrix} C_f, C_\infty \end{bmatrix} x(t),$$
from the realization $(E, A, B, C, 0)$ with index $\nu$.
Partition the resulting system into the realizations of the slow subsystem
$(I_{n_f}, J, B_f, C_f, 0)$ and the fast subsystem $(N, I_{n_\infty}, B_\infty, C_\infty, 0)$.
if $\nu = 1$ then
    Compute the $r$-th order Hankel-norm approximation $(I_r, \hat{A}, \hat{B}, \hat{C}, \hat{D})$ of the
    realization $(I_{n_f}, J, B_f, C_f, -C_\infty B_\infty)$.
else
    Get the number of states in the minimal fast subsystem
        $$k = \min(\nu m, n_\infty).$$
    if $\nu m < n_\infty$ then
        Compute the minimal realization $(\tilde{N}, I_k, \tilde{B}_\infty, \tilde{C}_\infty, 0)$ of the fast
        subsystem, where
            $$\tilde{N} = \begin{bmatrix} 0 & I_m & & 0 \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & I_m \\ 0 & & 0 & 0 \end{bmatrix} \in \mathbb{R}^{k \times k}, \quad \tilde{B}_\infty = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -I_m \end{bmatrix} \in \mathbb{R}^{k \times m}$$
        and $\tilde{C}_\infty = - \begin{bmatrix} C_\infty N^{\nu-1} B_\infty, \dots, C_\infty N^0 B_\infty \end{bmatrix} \in \mathbb{R}^{p \times k}$.
    else
        The fast subsystem $(N, I_{n_\infty}, B_\infty, C_\infty, 0) = (\tilde{N}, I_k, \tilde{B}_\infty, \tilde{C}_\infty, 0)$ is minimal.
    end
    Compute the $r$-th order Hankel-norm approximation $(I_r, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ of the
    realization $(I_{n_f}, J, B_f, C_f, 0)$.
    Construct the complete system $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$ with
        $$\hat{E} = \begin{bmatrix} I_r & 0 \\ 0 & \tilde{N} \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} \tilde{A} & 0 \\ 0 & I_k \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \tilde{B} \\ \tilde{B}_\infty \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} \tilde{C}, & \tilde{C}_\infty \end{bmatrix}, \quad \hat{D} = \tilde{D}.$$
end
```

4 The Generalized Hankel Norm Approximation

In contrast to the special approach presented before, the generalized Hankel-norm approximation method introduced in this chapter is not based on the Weierstrass canonical form. Here, the method is seen as an extension of the generalized balanced truncation. This is conform with the idea of the standard Hankel-norm approximation method introduced by Glover in [13]. For this purpose, several theoretical results of the generalized balanced truncation method will be presented as well as computational algorithms. On this basics, the generalized Hankel-norm approximation will be developed.

4.1 The Generalized Balanced Truncation

The following theoretical aspects concerning the generalized balanced truncation can be found in [8].

Main goal of the balanced truncation method is the computation of a balanced realization of the system and the truncation of undesired states at the same time. The idea of a balanced realization of a descriptor system (2.5) was already given in Definition 2.11. Uncontrollable and unobservable states can be associated with zero Hankel singular values. Since the input-output behavior of the system is invariant under the addition of unobservable and uncontrollable states, the zero Hankel singular values can be truncated without changing the system behavior.

Beside these unnecessary states, it would be beneficial to have a measurement for further states of the system with less influence on the input-output behavior. Therefor, the following theorem displays an energy interpretation of the proper controllability and observability Gramian.

Theorem 4.1. *(See [8, 26]). Consider a descriptor system of the form* (2.5)*. Assume that the matrix pencil $\lambda E - A$ is c-stable and the system is C-controllable. Let $\mathcal{G}_{pc}$ and $\mathcal{G}_{po}$ be the proper controllability and observability Gramian of* (2.5) *and let*

$$E_y := ||y||^2_{\mathcal{L}_2(\mathbb{R}_0^+)} = \int\limits_0^{+\infty} y(t)^T y(t)\,\mathrm{d}t, \quad E_u := ||u||^2_{\mathcal{L}_2(\mathbb{R}^-)} = \int\limits_{-\infty}^{0} u(t)^T u(t)\,\mathrm{d}t$$

be a future output energy and a past input energy, respectively. If $x_0 \in \operatorname{Im}(P_r)$ and $u(t) = 0$ for $t \geq 0$, then

$$E_y = x_0^T E^T G_{po} E x_0.$$

Furthermore, for $u_{\min}(t) = B^T \mathcal{F}(-t)\mathcal{G}_{pc}^{-} x_0$ *it holds*

$$E_{u_{\min}} = \min_{u \in \mathcal{L}_2(\mathbb{R}^-)} E_u = x_0^T \mathcal{G}_{pc}^{-} x_0,$$

where $\mathcal{F}$ *is the fundamental solution matrix from* (2.21) *and the matrix* $\mathcal{G}_{pc}^{-}$ *is a solution of the three matrix equations*

$$\mathcal{G}_{pc}\mathcal{G}_{pc}^{-}\mathcal{G}_{pc} = \mathcal{G}_{pc}, \qquad \mathcal{G}_{pc}^{-}\mathcal{G}_{pc}\mathcal{G}_{pc}^{-} = \mathcal{G}_{pc}^{-}, \qquad \left(\mathcal{G}_{pc}^{-}\right)^T = \mathcal{G}_{pc}^{-}.$$

This theorem implies that a large past input energy E_u is required to reach the state $x(0) = P_r x_0$ which lies in an invariant subspace of $\mathcal{G}_{pc}$ corresponding to its small non-zero eigenvalues from the state $x(-\infty) = 0$. On the other side, if x_0 is contained in an invariant subspace of the matrix $E^T\mathcal{G}_{po}E$ corresponding to its small non-zero eigenvalues, then the initial state $x(0) = x_0$ has a small effect on the future output energy E_y.

For a balanced realization of a descriptor system (2.5) it holds

$$\mathcal{G}_{pc} = E^T\mathcal{G}_{po}E = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}.$$

That is, the matrices $\mathcal{G}_{pc}$ and $E^T\mathcal{G}_{po}E$ have the same invariant subspaces. In this case, the states related to the small proper Hankel singular values are difficult to reach and observe at the same time. The truncation of such states essentially does not change the system properties, especially the input-output behavior.

This does not hold for the improper Hankel singular values. The truncation of small non-zero improper Hankel singular values may result in finite eigenvalues of the matrix pencil lying in the closed right half-plane [8]. Furthermore, the equations associated with the improper Hankel singular values describe constraints of the system. That means, these equations define a manifold in which the solution dynamics takes place. For this reason, a truncation of the equations corresponding to non-zero improper Hankel singular values can be identified by ignoring certain constraints of the system. Physically meaningless results may be expected.

Note that he number of the non-zero improper Hankel singular values of (2.5) is equal to $\operatorname{rank}(\mathcal{G}_{ic}A^T\mathcal{G}_{io}A)$. This number can in turn be bounded by

$$\operatorname{rank}(\mathcal{G}_{ic}A^T\mathcal{G}_{io}A) \leq \min\left(\nu m, \nu p, n_\infty\right), \tag{4.1}$$

with ν, the index of the system (2.5), m, the number of inputs, p, the number of outputs, and n_∞, the dimension of the deflating subspace of $\lambda E - A$ corresponding to the infinite eigenvalues. This bound shows that if the numbers of inputs and outputs multiplied by the index ν are much smaller than the dimension n_∞, the order of the descriptor system (2.5) can be reduced significantly by the truncation of states corresponding to zero improper Hankel singular values.

Now, let (E, A, B, C, D) be a realization (not necessary minimal) of the descriptor system (2.5) and let the matrix pencil $\lambda E - A$ be c-stable. With the Cholesky factorizations

(2.26) let

$$\begin{aligned} L_p^T E R_p &= \begin{bmatrix} U_1, & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \quad \text{and} \\ L_i^T A R_i &= U_3 \Theta_3 V_3^T \end{aligned}$$

be the skinny singular value decompositions of the matrices $L_p^T E R_p$ and $L_i^T A R_i$. The matrices $\begin{bmatrix} U_1, & U_2 \end{bmatrix}$, $\begin{bmatrix} V_1, & V_2 \end{bmatrix}$, U_3 and V_3 have orthonormal columns. The two matrices $\Sigma_1 = \mathrm{diag}(\varsigma_1, \ldots, \varsigma_{l_f})$ and $\Sigma_2 = \mathrm{diag}(\varsigma_{l_f+1}, \ldots, \varsigma_{r_f})$ with $\varsigma_1 \geq \ldots \geq \varsigma_{l_f} > \varsigma_{l_f+1} \geq \ldots \geq \varsigma_{r_f} > 0$ contain the non-zero proper Hankel singular values of the descriptor system (2.5). The number of non-zero proper Hankel singular values is given by $r_f = \mathrm{rank}(L_p^T E R_p) \leq n_f$. Similarly, the non-zero improper Hankel singular values of the descriptor system (2.5) are given in the matrix $\Theta_3 = \mathrm{diag}(\theta_1, \ldots, \theta_{l_\infty})$ by $\theta_1 \geq \ldots \geq \theta_{l_\infty} > 0$, where the number of non-zero improper Hankel singular values is $l_\infty = \mathrm{rank}(L_i^T A R_i) \leq n_\infty$. The partition by the number l_f is chosen, such that the proper Hankel singular values corresponding to undesired states are contained in Σ_2.

Now, a balanced realization $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D}) = (W_l^T E T_l, W_l^T A T_l, W_l^T B, C T_l, D)$ of the descriptor system (2.5) can be computed by the application of the two matrices

$$\begin{aligned} W_l &= \begin{bmatrix} L_p U_1 \Sigma_1^{-\frac{1}{2}}, & L_i U_3 \Theta_3^{-\frac{1}{2}} \end{bmatrix} \in \mathbb{R}^{n \times l} \quad \text{and} \\ T_l &= \begin{bmatrix} R_p V_1 \Sigma_1^{-\frac{1}{2}}, & R_i V_3 \Theta_3^{-\frac{1}{2}} \end{bmatrix} \in \mathbb{R}^{n \times l} \end{aligned}$$

for a generalized state space transformation. The resulting realization has the order $l = l_f + l_\infty$.

The transformation using the matrices W_l and T_l can be seen as a truncated version of the transformation in (2.28). Therefore, the resulting matrices

$$\hat{E} = W_l^T E T_l = \begin{bmatrix} I_{l_f} & 0 \\ 0 & E_\infty \end{bmatrix} \quad \text{and} \quad \hat{A} = W_l^T A T_l = \begin{bmatrix} A_f & 0 \\ 0 & I_{l_\infty} \end{bmatrix} \tag{4.2}$$

resemble a truncated version of the Weierstrass canonical form (2.10).

The balanced truncation method using the matrices W_l and T_l for the transformation of the system is known as the square-root method. The complete generalized balanced truncation square-root method is summarized in Algorithm 3.

The computation of the balanced truncation can be interpreted as a generalized state space transformation with the transformation matrices W and T of the form

$$(WET, WAT, WB, CT, D) = \left(\begin{bmatrix} E_f & 0 \\ 0 & E_\infty \end{bmatrix}, \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix}, \begin{bmatrix} B_f \\ B_\infty \end{bmatrix}, \begin{bmatrix} C_f, & C_\infty \end{bmatrix}, D \right), \tag{4.3}$$

where the matrix pencil $\lambda E_f - A_f$ contains all the finite eigenvalues and $\lambda E_\infty - A_\infty$ contains the infinite eigenvalues of the original matrix pencil $\lambda E - A$.

The orders of the decoupled subsystems $(E_f, A_f, B_f, C_f, 0)$ and $(E_\infty, A_\infty, B_\infty, C_\infty, D)$ are then reduced separately, where E_f and A_∞ are both non-singular.

```
Algorithm 3: Generalized Balanced Truncation Square-Root Method
Input: Realization $(E, A, B, C, D)$, such that $\lambda E - A$ is c-stable
Output: Reduced-order balanced realization $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$
1: Compute the Cholesky factors $R_p$ and $L_p$ of the proper controllability Gramian
   $\mathcal{G}_{pc} = R_p R_p^T$ and the proper observability Gramian $\mathcal{G}_{po} = L_p L_p^T$ that satisfy
   $$\begin{aligned} E\mathcal{G}_{pc}A^T + A\mathcal{G}_{pc}E^T + P_lBB^TP_l^T &= 0, & \mathcal{G}_{pc} &= P_r\mathcal{G}_{pc}P_r^T, \\ E^T\mathcal{G}_{po}A + A^T\mathcal{G}_{po}E + P_r^TC^TCP_r^T &= 0, & \mathcal{G}_{po} &= P_l^T\mathcal{G}_{po}P_l. \end{aligned}$$
2: Compute the Cholesky factors $R_i$ and $L_i$ of the improper controllability
   Gramian $\mathcal{G}_{ic} = R_i R_i^T$ and the improper observability Gramian $\mathcal{G}_{io} = L_i L_i^T$ that
   satisfy
   $$\begin{aligned} A\mathcal{G}_{ic}A^T - E\mathcal{G}_{ic}E^T - (I_n - P_l)BB^T(I_n - P_l)^T &= 0, & P_r\mathcal{G}_{ic}P_r^T &= 0, \\ A^T\mathcal{G}_{io}A - E^T\mathcal{G}_{io}E - (I_n - P_r)^TC^TC(I_n - P_r) &= 0, & P_l^T\mathcal{G}_{io}P_l &= 0. \end{aligned}$$
3: Compute the skinny singular value decomposition
   $$L_p^TER_p = \begin{bmatrix} U_1, & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix},$$
   where $\begin{bmatrix} U_1, & U_2 \end{bmatrix}$ and $\begin{bmatrix} V_1, & V_2 \end{bmatrix}$ have orthonormal columns, $\Sigma_1 = \operatorname{diag}(\varsigma_1, \dots, \varsigma_{l_f})$
   and $\Sigma_2 = \operatorname{diag}(\varsigma_{l_f+1}, \dots, \varsigma_{r_f})$ with the proper non-zero Hankel singular values of
   the system and $r_f = \operatorname{rank}(L_p^TER_p)$.
4: Compute the skinny singular value decomposition
   $$L_i^TAR_i = U_3\Theta_3V_3^T,$$
   where $U_3$ and $V_3$ have orthonormal columns, $\Theta_3 = \operatorname{diag}(\theta_1, \dots, \theta_{l_\infty})$ with the
   improper non-zero Hankel singular values and $l_\infty = \operatorname{rank}(L_i^TAR_i)$.
5: Compute the transformation matrices
   $$W_l = \left[ L_pU_1\Sigma_1^{-\frac{1}{2}}, \quad L_iU_3\Theta_3^{-\frac{1}{2}} \right], \quad T_l = \left[ R_pV_1\Sigma_1^{-\frac{1}{2}}, \quad R_iV_3\Theta_3^{-\frac{1}{2}} \right].$$
6: Compute the reduced-order model
   $$(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D}) = (W_l^TET_l, W_l^TAT_l, W_l^TB, CT_l, D).$$
```

The resulting realization of the reduced-order model $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$ is minimal, c-stable and, due to the choice of the transformation matrices W_l and T_l, balanced. According to the decoupled form (4.3) of the descriptor system (2.5), the transfer function can be additively decomposed in the form $G = G_{sp} + P$, where the strictly proper part is given by

$$G_{sp}(s) = C_f\left(sE_f - A_f\right)^{-1}B_f$$

and the polynomial part by

$$P(s) = C_\infty \left(sE_\infty - A_\infty\right)^{-1} B_\infty + D.$$

As mentioned above, the reduced-order model is computed by model reduction of the subsystems separately. So, the additive form of the transfer function does not change. The transfer function of the reduced-order model has the form $\hat{G} = \hat{G}_{sp} + \hat{P}$, where the strictly proper can be written as

$$\hat{G}_{sp}(s) = \hat{C}_f \left(s\hat{E}_f - \hat{A}_f\right)^{-1} \hat{B}_f$$

and the polynomial part has the form

$$\hat{P}(s) = \hat{C}_\infty \left(s\hat{E}_\infty - \hat{A}_\infty\right)^{-1} \hat{B}_\infty + \hat{D}.$$

These are the transfer functions of the two decoupled reduced-order subsystems.
Since the matrix E_f is non-singular, the classical balanced truncation method can be applied to the realization $(E_f, A_f, B_f, C_f, 0)$. So, the $\mathcal{H}_\infty$ error bound of the classical balanced truncation method holds for this subsystem

$$\left|\left|G_{sp} - \hat{G}_{sp}\right|\right|_{\mathcal{H}_\infty} \leq 2(\varsigma_{l_f+1} + \ldots + \varsigma_{r_f}), \tag{4.4}$$

see [13].
In contrast to the reduction of the proper part, the reduction of the subsystem corresponding to the polynomial part $(E_\infty, A_\infty, B_\infty, C_\infty, D)$ can be interpreted as classical balanced truncation of the discrete-time system

$$A_\infty \zeta_{k+1} = E_\infty \zeta_k + B_\infty \eta_k, \tag{4.5}$$
$$\omega_k = C_\infty \zeta_k + D\eta_k, \tag{4.6}$$

with the inputs η_k, the outputs ω_k, the internal states ζ_k and the non-singular matrix A_∞. The classical Hankel singular values of this system are the improper Hankel singular values of the original descriptor system (2.5).
As mentioned before, for the polynomial part only states corresponding to zero improper Hankel singular values can be truncated. As a result, the equality $P = \hat{P}$ holds for the polynomial parts of the original and the reduced subsystem.
The index of the reduced-order model is equal to $\deg(P) + 1$, where $\deg(P)$ denotes the degree of the polynomial transfer function P. Equivalent to this number is the multiplicity of the pole at infinity of the transfer function G, see [8].
Hence, the error system can be written in the form

$$\begin{aligned} G - \hat{G} &= G_{sp} + P - \hat{G}_{sp} - \hat{P} \\ &= G_{sp} - \hat{G}_{sp}. \end{aligned}$$

That means, the error system's transfer function is strictly proper and the following error bound holds

$$\left|\left|G - \hat{G}\right|\right|_{\mathcal{H}_\infty} \leq 2(\varsigma_{l_f+1} + \ldots + \varsigma_{r_f}), \tag{4.7}$$

since (4.4) can be used for the strictly proper parts of the transfer functions.

This error bound is one of the main advantages of the generalized balanced truncation method compared to, for example, moment matching methods. Therewith, the order of the strictly proper part can be chosen dynamically depending on the desired approximation error.

A problem of the balanced truncation square-root method are the possible ill-conditioned transformation matrices. This case can occur for highly unbalanced descriptor systems, if quite small Hankel singular values are involved in the computation or if the angle between the subspaces of the matrix pencil $\lambda E - A$ corresponding to the finite and infinite eigenvalues is too small. The prevention of an accuracy loss in the reduced-order model can be achieved by changing the construction of the transformation matrices.

The generalized balanced truncation balancing-free square-root method can be obtained by the computation of two QR decompositions of the form

$$Q_R R_0 = \begin{bmatrix} R_p V_1, & R_i V_3 \end{bmatrix} \quad \text{and} \quad Q_L L_0 = \begin{bmatrix} L_p U_1, & L_i U_3 \end{bmatrix},$$

where the matrices R_p, V_1, R_i, V_3, L_p, U_1, L_i, U_3 are computed by the steps 1-4 in Algorithm 3. The matrices Q_R and Q_L have orthonormal columns and are used as transformation matrices. The realization of the reduced-order model is then given by

$$(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D}) = (Q_L^T E Q_R, Q_L^T A Q_R, Q_L^T B, C Q_R, D).$$

Both generalized balanced truncation methods are formally equivalent in the sense that in exact arithmetic, they return the same transfer function. Also, the reduced-order model obtained by the balancing-free square-root method is minimal, c-stable and satisfies the same error bound in the $\mathcal{H}_\infty$-norm. Since the transformation matrices Q_L and Q_R have orthonormal columns, they may be significantly less sensitive to perturbations than the projection matrices W_l and T_l computed by the square-root method. It is possible to get a resulting system resembling the form (4.3) by separating the QR decompositions into the parts corresponding to the proper and improper Hankel singular values, respectively. Note that the realization obtained by the balancing-free square-root method is in general not balanced. Thus, this method will not further be considered.

The complete algorithm for the generalized balanced truncation balancing-free square-root method can be found in [8].

4.2 The Generalized Hankel-Norm Approximation

After considering all basics of systems theory, the classical Hankel-norm approximation and the generalized balanced truncation, the generalized Hankel-norm approximation will finally be introduced in this section. In contrast to the special approach based on the Weierstrass canonical form, the method shown in this section is an extension of the generalized balanced truncation square-root method.

The basic idea of the generalized Hankel-norm approximation is the decoupling of the descriptor system (2.5) and the individual reductions of the two resulting subsystems. As seen before, the descriptor system is decomposed in its slow subsystem

$$\begin{aligned} E_f\dot{x}_1(t) &= A_f x_1(t) + B_f u(t), \\ y_1(t) &= C_f x_1(t), \end{aligned} \tag{4.8}$$

with non-singular E_f, and the fast subsystem

$$\begin{aligned} E_\infty\dot{x}_2(t) &= A_\infty x_2(t) + B_\infty u(t), \\ y_2(t) &= C_\infty x_2(t) + Du(t), \end{aligned} \tag{4.9}$$

where E_∞ is nilpotent with index ν.

Since the fast subsystem (4.9) corresponds to the constraints of the solution dynamics, only the states corresponding to the zero improper Hankel singular values can be truncated. There is no sense to further consider this part for the Hankel-norm approximation.

Now, the slow subsystem (4.8) is considered. Since the matrix E_f is regular, a standard system can be obtained by applying the inverse of E_f to the first equations of (4.8). Then, the classical Hankel-norm approximation method can be applied. In fact, this transformation to the standard form is made by the generalized balanced truncation square-root method.

The first step of the Hankel-norm approximation method in Algorithm 1 is the construction of a minimal balanced realization. For descriptor systems this can be made by the generalized balanced truncation square-root method. Since this method resembles the Weierstrass canonical form (4.2), the decoupled slow subsystem (4.8) simplifies to the form

$$\begin{aligned} \dot{x}_1(t) &= A_f x_1(t) + B_f u(t), \\ y_1(t) &= C_f x_1(t), \end{aligned} \tag{4.10}$$

This resulting slow subsystem is in standard form. Through the use of the generalized balanced truncation, the subsystem (4.10) is minimal and balanced.

Note that the generalized balanced truncation balancing-free square root method is not suited, since the resulting realization is usually not balanced.

Let $(I_r, A_h, B_h, C_h, D_h)$ be the resulting r-th order Hankel-norm approximation obtained by Algorithm 1 and $(\hat{E}_\infty, I_{l_\infty}, \hat{B}_\infty, \hat{C}_\infty, D)$ the balanced fast subsystem, obtained by the truncation of the zero improper Hankel singular values. Then the complete Hankel-norm approximation of the descriptor system (2.5) is given by the coupled descriptor system

$$\begin{aligned} \begin{bmatrix} I_r & 0 \\ 0 & \hat{E}_\infty \end{bmatrix} \dot{z}(t) &= \begin{bmatrix} A_h & 0 \\ 0 & I_{l_\infty} \end{bmatrix} z(t) + \begin{bmatrix} B_h \\ \hat{B}_\infty \end{bmatrix} u(t), \\ \hat{y}(t) &= \begin{bmatrix} C_h, \hat{C}_\infty \end{bmatrix} z(t) + (D + D_h)u(t). \end{aligned}$$

The summarized method can be found in Algorithm 4.

Now, the question of the approximation error arises.

Algorithm 4: Generalized Hankel-Norm Approximation

```
Input: Realization $(E, A, B, C, D)$, such that $\lambda E - A$ is c-stable
Output: Realization of Hankel-norm approximation $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$
1: Compute the Cholesky factors $R_p$ and $L_p$ of the proper controllability Gramian
   $\mathcal{G}_{pc} = R_p R_p^T$ and the proper observability Gramian $\mathcal{G}_{po} = L_p L_p^T$ that satisfy
       $E\mathcal{G}_{pc}A^T + A\mathcal{G}_{pc}E^T + P_l B B^T P_l^T = 0, \qquad \mathcal{G}_{pc} = P_r \mathcal{G}_{pc} P_r^T,$
       $E^T\mathcal{G}_{po}A + A^T\mathcal{G}_{po}E + P_r^T C^T C P_r^T = 0, \qquad \mathcal{G}_{po} = P_l^T \mathcal{G}_{po} P_l.$
2: Compute the Cholesky factors $R_i$ and $L_i$ of the improper controllability
   Gramian $\mathcal{G}_{ic} = R_i R_i^T$ and the improper observability Gramian $\mathcal{G}_{io} = L_i L_i^T$ that
   satisfy
       $A\mathcal{G}_{ic}A^T - E\mathcal{G}_{ic}E^T - (I_n - P_l)BB^T(I_n - P_l)^T = 0, \qquad P_r \mathcal{G}_{ic} P_r^T = 0,$
       $A^T\mathcal{G}_{io}A - E^T\mathcal{G}_{io}E - (I_n - P_r)^T C^T C(I_n - P_r) = 0, \qquad P_l^T \mathcal{G}_{io} P_l = 0.$
3: Compute the skinny singular value decomposition
       $L_p^T E R_p = U_1 \Sigma V_1^T,$
   where $U_1$ and $V_1$ have orthonormal columns, $\Sigma = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_{l_f})$ with the
   proper non-zero Hankel singular values of the system and $l_f = \operatorname{rank}(L_p^T E R_p)$.
4: Compute the skinny singular value decomposition
       $L_i^T A R_i = U_2 \Theta V_2^T,$
   where $U_2$ and $V_2$ have orthonormal columns, $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_{l_\infty})$ with the
   improper non-zero Hankel singular values and $l_\infty = \operatorname{rank}(L_i^T A R_i)$.
5: Compute the projection matrices
       $W_{l_f} = L_p U_1 \Sigma^{-\frac{1}{2}}, \quad W_{l_\infty} = L_i U_2 \Theta^{-\frac{1}{2}},$
       $T_{l_f} = R_p V_1 \Sigma^{-\frac{1}{2}}, \quad T_{l_\infty} = R_i V_2 \Theta^{-\frac{1}{2}}.$
6: Compute the minimal balanced standard realization of the slow subsystem
       $(I_{l_f}, A_f, B_f, C_f, 0) = (W_{l_f}^T E T_{l_f}, W_{l_f}^T A T_{l_f}, W_{l_f}^T B, C T_{l_f}, 0).$
7: Choose the proper Hankel singular value $\varsigma_{r+1}$.
8: Permute the standard realization $(I_{l_f}, A_f, B_f, C_f, 0)$, such that the proper
   system Gramians are
       $\check{\mathcal{G}}_{pc} = \check{\mathcal{G}}_{po} = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_r, \varsigma_{r+k+1}, \ldots, \varsigma_{l_f}, \varsigma_{r+1} I_k)$
       $= \operatorname{diag}(\check{\Sigma}, \varsigma_{r+1} I_k),$
   where $k$ is the multiplicity of the Hankel singular value $\varsigma_{r+1}$.
```

```
9: Partition the resulting permuted system according to the proper Gramians
       $$\check{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \check{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \check{C} = \begin{bmatrix} C_1 & C_2 \end{bmatrix},$$
   where $A_{22} \in \mathbb{R}^{k \times k}$, $B_2 \in \mathbb{R}^{k \times m}$, and $C_2 \in \mathbb{R}^{p \times k}$.
10: Compute the transformation
       $$\tilde{A} = \Gamma^{-1}(\varsigma_{r+1}^2 A_{11}^T + \check{\Sigma} A_{11} \check{\Sigma} + \varsigma_{r+1} C_1^T U B_1^T),$$
       $$\tilde{B} = \Gamma^{-1}(\check{\Sigma} B_1 - \varsigma_{r+1} C_1^T U),$$
       $$\tilde{C} = C_1 \check{\Sigma} - \varsigma_{r+1} U B_1^T,$$
       $$\tilde{D} = \varsigma_{r+1} U,$$
   with $U = (C_2^T)^{\dagger} B_2$ and $\Gamma = \check{\Sigma}^2 - \varsigma_{r+1}^2 I_{l_f - k}$.
11: Compute the additive decomposition
       $$\tilde{G} = \tilde{C}(sI_{l_f-k} - \tilde{A})^{-1}\tilde{B} + \tilde{D} = G_h(s) + F(s),$$
   where $F(s)$ is anti-stable and $G_h(s)$ is the stable Hankel-norm approximation
   with the realization $(I_r, A_h, B_h, C_h, D_h)$.
12: Compute the minimal realization of the fast subsystem
       $$(E_\infty, I_{l_\infty}, B_\infty, C_\infty, D) = (W_{l_\infty}^T E T_{l_\infty}, W_{l_\infty}^T A T_{l_\infty}, W_{l_\infty}^T B, C T_{l_\infty}, D).$$
13: Additive construction of the resulting system
       $$(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D}) = \left( \begin{bmatrix} I_r & 0 \\ 0 & E_\infty \end{bmatrix}, \begin{bmatrix} A_h & 0 \\ 0 & I_{l_\infty} \end{bmatrix}, \begin{bmatrix} B_h \\ B_\infty \end{bmatrix}, \begin{bmatrix} C_h, & C_\infty \end{bmatrix}, D + D_h \right).$$
```

Therefore, let $G = G_{sp} + P$ be the original transfer function with its strictly proper part G_{sp} and its polynomial part P. Assume $\hat{G} = G_h + F$ is the transfer function resulting from the transformation formulas in Theorem 3.2 for the system G_{sp}, where G_h is the stable, F the anti-stable part and the $(r+1)$-st proper Hankel singular value ς_{r+1} of G was chosen. Let $\hat{P}$ be the conditionally minimal realization of the polynomial transfer function P, which means $\hat{P} = P$. Now, Consider the error system of the form

$$\begin{aligned} E &= G - \hat{G} - \hat{P} \\ &= G_{sp} + P - \hat{G} - \hat{P} \\ &= G_{sp} - \hat{G} \\ &= G_{sp} - G_h - F. \end{aligned}$$

Since G_h is the Hankel-norm approximation of G_{sp} and F is the corresponding anti-stable transfer function, part (2) of Theorem 3.2 can be used and it holds

$$E(s)E^H(-\bar{s}) = \varsigma_{r+1}^2 I_r.$$

So, the scaled error transfer function is all-pass. That is the reason, the error bounds proposed in section 3.1 still hold for the descriptor system case. For the Hankel-norm the error is given by

$$\left|\left|G - G_h - \hat{P}\right|\right|_H = \varsigma_{r+1}(G), \tag{4.11}$$

where $\varsigma_{r+1}(G)$ is the $(r+1)$-th proper Hankel singular value of the original descriptor system G. Hence, the method in Algorithm 4 is a Hankel-norm approximation for descriptor systems.

With the same approach as for the Hankel-norm, the error bound in the $\mathcal{H}_\infty$-norm is given by

$$\left|\left|G - \hat{G}\right|\right|_{\mathcal{H}_\infty} \leq 2 \sum_{k=r+1}^{l_f} \varsigma_k(G).$$

Finally, the following theorem can be summarized.

Theorem 4.2. *Let G be a c-stable descriptor system of the form* (2.5)*. In exact arithmetic, the reduced-order descriptor system $\hat{G}$ obtained from Algorithm 4 has the following properties:*

(1) The realization of $\hat{G}$ is conditionally minimal, c-stable and resembles the Weierstrass canonical form.

(2) The error in the Hankel-norm is given by

$$\left|\left|G - \hat{G}\right|\right|_H = \varsigma_{r+1}(G).$$

(3) The following $\mathcal{H}_\infty$ error bound holds

$$\left|\left|G - \hat{G}\right|\right|_{\mathcal{H}_\infty} \leq 2 \sum_{k=r+1}^{l_f} \varsigma_{r+1}(G).$$

4.3 The Approximated Hankel-Norm Approximation

The generalized Hankel-norm approximation quickly becomes a numerically unstable method. This problem arises because of a small chosen proper Hankel singular value or a large McMillan degree with many small proper Hankel singular values of the slow subsystem. In this section the case of a large McMillan degree is considered.

First, step 10 in Algorithm 4 is regarded. This is the transformation formula from Theorem 3.2 used for the slow subsystem of (2.5). In this step the permuted system matrices $\check{A}$ and $\check{B}$ are both scaled by the diagonal matrix $\Gamma^{-1} = \left(\check{\Sigma}^2 - \varsigma_{r+1}^2 I_{l_f - k}\right)^{-1}$ during the transformation. Quite small proper Hankel singular values in $\check{\Sigma}$ can lead to large errors in the transformed system. There, the choice of a small ς_{r+1} Hankel singular value as approximation error is as problematic as the use of the remaining small proper Hankel singular values in $\check{\Sigma}$.

Another point to consider is the computation of the balanced system by the generalized balanced truncation square-root method. Here, the transformation matrices W_{l_f} and T_{l_f} get also ill-conditioned for small proper Hankel singular values. The resulting minimal balanced realization would loose a certain amount of accuracy up to this point [8].

Beside these accuracy related problems, also the computational costs should be mentioned. The resulting transformed unstable system has the order equal to the McMillan degree of the slow subsystem (4.8) reduced by the multiplicity of the chosen proper Hankel singular value. So for a large McMillan degree and well distributed Hankel singular values the resulting system may have a large order. The problem is that this system has to be decomposed in the stable and anti-stable parts and this becomes very costly for larger orders.

All these problems could be eliminated by allowing the computation of a smaller balanced realization of the slow subsystem than the McMillan degree. Obviously, this results in an additional error which hurts the exact error bound of the Hankel-norm approximation. The question that arises is, how much is the resulting error disturbed. Therefore, a special result from [13] for standard systems is used.

Let G_b be a r-th order balanced realization of the original standard system G. Then it holds

$$||G - G_b||_H \leq 2 \sum_{k=r+1}^{n} \varsigma_k.$$

As in the previous section shown, this result can be applied to the descriptor system case by the decoupling of the slow and fast subsystems.

Now, let $G = G_{sp} + P$ be the original descriptor system (2.5), G_b a balanced realization of the strictly proper part G_{sp} with order $n_b \leq n_f$ and let $\hat{G} = G_h + \hat{P}$ be the r-th order Hankel-norm approximation of the balanced system G_b with $r \leq n_b$.

Then for the error in the Hankel-norm it holds

$$\begin{aligned}
\left|\left|G - \hat{G}\right|\right|_H &= \left|\left|G_{sp} + P - G_h + \hat{P}\right|\right|_H \\
&= ||G_{sp} - G_h||_H \\
&= ||G_{sp} - G_b + G_b - G_h||_H \\
&\leq ||G_b - G_h||_H + ||G_{sp} - G_b||_H \\
&\leq \varsigma_{r+1}(G_b) + 2 \sum_{k=n_b+1}^{n_f} \varsigma_k(G),
\end{aligned}$$

with n_f the dimension of the deflating subspace corresponding to the finite eigenvalues of the matrix pencil $\lambda E - A$.

Since G_b is a balanced realization of order n_b, only n_b non-zero proper Hankel singular values remain for further computations. If $r = n_b$ is chosen, the value $\varsigma_{n_b+1}(G_b)$ is set to zero. For G_b the balanced realization of G, it holds

$$\varsigma_k(G_b) = \varsigma_k(G)$$

for $k = 1, \dots, n_b$.
Now, the Hankel-norm error can be rewritten as

$$\left|\left|G - \hat{G}\right|\right|_H \leq \varsigma_{r+1}(G) + 2 \sum_{k=n_b+1}^{n_f} \varsigma_k(G). \tag{4.12}$$

Obviously, in case of G_b is a minimal balanced realization of the strictly proper part G_{sp}, the error formula (4.12) simplifies to

$$\left|\left|G - \hat{G}\right|\right|_H \leq \varsigma_{r+1}(G).$$

With the result of Lemma 3.3, equality holds. This is conform with the error of the exact generalized Hankel-norm approximation (4.11).
Using the same approach for the error bound in the $\mathcal{H}_\infty$-norm, one obtains

$$\left|\left|G - \hat{G}\right|\right|_{\mathcal{H}_\infty} \leq 2 \sum_{k=r+1}^{n_b} \varsigma_k(G_b) + 2 \sum_{k=n_b+1}^{n_f} \varsigma_k(G) = 2 \sum_{k=r+1}^{n_f} \varsigma_k(G),$$

which is identical to the original $\mathcal{H}_\infty$ error bound for the generalized Hankel-norm approximation.
Now, the remaining open point is the implementation of this approximated version of the generalized Hankel-norm approximation. Therefor, consider step 3 of Algorithm 4. There, the skinny singular value decomposition has to be changed in the way used for the balanced truncation method. That means, the singular value decomposition has the form

$$L_p^T E R_p = \begin{bmatrix} U_1, & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix},$$

where $\begin{bmatrix} U_1, & U_2 \end{bmatrix}$ and $\begin{bmatrix} V_1, & V_2 \end{bmatrix}$ have orthonormal columns, $\Sigma_1 = \operatorname{diag}(\varsigma_1, \dots, \varsigma_{n_b})$ and $\Sigma_2 = \operatorname{diag}(\varsigma_{n_b+1}, \dots, \varsigma_{l_f})$ are diagonal with the proper non-zero Hankel singular values of the system and $l_f = \operatorname{rank}(L_p^T E R_p)$. The partition is chosen, such that Σ_1 contains the proper Hankel singular values which shall be used for further computations and Σ_2 the Hankel singular values which are no longer needed. The matrices U_1, Σ_1, and V_1 are then used for the computation of the balanced realization G_b and hence, for further transformations of the generalized Hankel-norm approximation method.
This approximated method take the advantage of the generalized balanced truncation method in form of the adaptive choice of the order. The order n_b of the balanced realization G_b might be chosen as

$$2 \sum_{k=n_b+1}^{n_f} \varsigma_k(G) \ll \varsigma_{r+1}(G).$$

Then the resulting Hankel-norm error can be seen as

$$\left|\left|G - \hat{G}\right|\right|_H \approx \varsigma_{r+1}(G).$$

This approach can be seen as a numerically disturbed version of the exact generalized Hankel-norm approximation method. So, the additional error is not worse than the unavoidable round-off errors.
Applying this approximated generalized Hankel-norm approximation is definitely more stable than the original one. Still, there is no numerical analysis of the exact or the approximated Hankel-norm approximation method.
Beside the accuracy problem, the case of large McMillan degrees can be tackled by this approximated method. Memory resources can be saved for the transformation matrices W_{l_f} and T_{l_f} as well as for the resulting balanced realization. In general, the computational costs after step 3 of Algorithm 4 are reduced, since further computations are made on smaller system matrices. Especially, the block diagonalization of the transformed system in step 10 needs a computational effort of $\mathcal{O}((l_f - k)^3)$ which clearly reduces for smaller l_f.

4.4 Application to Sparse Systems

A frequently appearing case in practice is the model reduction of large-scale sparse descriptor systems. There, at least the matrices A and E from the descriptor system (2.5) are in a large-scale sparse form, i.e., the dimension n is large, the matrices A and E can be stored using $\mathcal{O}(n)$ memory and the matrix-vector multiplication can be computed with $\mathcal{O}(n)$ effort. Often these matrices are the result of the discretization of partial differential equations.
Considering Algorithm 4 with respect to the sparse structure of the matrices, one can observe that after the steps 6 and 12 both, the fast and slow subsystems, have to be stored dense, due to the use of dense transformation matrices. Also, it is not possible to use the Cholesky factors for the computation of the balanced realization because they are dense and of dimensions $n \times n$. For the usage on sparse data, the projected Lyapunov equations in the first two steps of Algorithm 4 have to be solved in sparse form.
It has been observed that the eigenvalues of the symmetric positive semidefinite solutions of the Lyapunov equations with low-rank right-hand sides generally decay rapidly. Such solutions may be well approximated by low-rank matrices [8]. The same result holds for the projected Lyapunov equations [29]. For example, consider the Lyapunov equation (2.22). If it is possible to find a matrix $Z \in \mathbb{R}^{n\times k}$ with much smaller number of columns k, such that ZZ^T is an approximated solution of (2.22), Z is referred to as the low-rank Cholesky factor of $\mathcal{G}_{pc}$ which is the solution of the projected continuous-time Lyapunov equation (2.22).
Since the matrix pencil $\lambda E - A$ is assumed as c-stable, the matrix A is invertible. So, the projected generalized continuous-time Lyapunov equation (2.22) can be rewritten as standard projected Lyapunov equation in the form

$$(A^{-1}E)X + X(A^{-1}E)^T = -P_r A^{-1} B B^T A^{-T} P_r^T, \quad X = P_r X P_r^T, \tag{4.13}$$

with $X = \mathcal{G}_{pc}$. The equation (4.17) can be solved by the alternating implicit direction

(ADI) method. The iteration scheme of this method is given by

$$\begin{aligned} X_k = &(A^{-1}E + \tau_k I_n)^{-1}(A^{-1}E - \bar{\tau}_k I_n)X_{k-1}(A^{-1}E - \tau_k I_n)^T(A^{-1}E + \bar{\tau}_k I_n)^{-T} \\ &-2\mathrm{Re}(\tau_k)(A^{-1}E + \tau_k I_n)^{-1}P_r A^{-1}BB^T A^{-T}P_r^T(A^{-1}E + \bar{\tau}_k I_n)^{-T}, \end{aligned} \tag{4.14}$$

with the initial solution $X_0 = 0$ and the complex shift parameters $\{\tau_1, \ldots, \tau_{k_q}\} \subset \mathbb{C}^-$, see [29]. The second equation in (4.13) follows from

$$\begin{aligned} P_r(A^{-1}E - \bar{\tau}_k I_n) &= (A^{-1}E - \bar{\tau}_k I_n)P_r, \\ P_r(A^{-1}E + \tau_k I_n)^{-1} &= (A^{-1}E + \tau_k I_n)^{-1}P_r. \end{aligned}$$

Now, the iteration scheme (4.14) can be rewritten in the form

$$\begin{aligned} X_k = &(E + \tau_k A)^{-1}(E - \bar{\tau}_k A)X_{k-1}(E - \tau_k A)^T(E + \bar{\tau}_k A)^{-T} \\ &-2\mathrm{Re}(\tau_k)(E + \tau_k A)^{-1}P_l BB^T P_l^T(E + \bar{\tau}_k A)^{-T}. \end{aligned} \tag{4.15}$$

It can be shown that the iteration (4.15) converges to the solution X of (4.13) and consequently, the iteration converges to the solution $\mathcal{G}_{pc}$ of (2.22), see [29].
The formulation (4.15) can be exploited for the low-rank Cholesky factorization $X = ZZ^T$. Therefor, let $X_k = Z_k Z_k^H$ be the new iteration variable. The iteration (4.15) can be written in the form

$$Z_k = \left[\sqrt{-2\mathrm{Re}(\tau_k)}(E + \tau_k A)^{-1}P_l B,\ (E + \tau_k A)^{-1}(E - \bar{\tau}_k A)Z_{k-1}\right].$$

By explicitly setting the term Z_{k-1} the final iteration has the form

$$Z_k = \left[B_0, F_{k-1}B_0, F_{k-2}F_{k-1}B_0, \ldots, F_1 \cdots F_{k-1}B_0\right], \tag{4.16}$$

where $B_0 = \sqrt{-2\mathrm{Re}(\tau_k)}(E + \tau_k A)^{-1}P_l B$ and

$$F_j = \sqrt{\frac{\mathrm{Re}(\tau_k)}{\mathrm{Re}(\tau_{k-1})}}\left(I_n - (\bar{\tau}_{k-1} + \tau_k)(E + \tau_k A)^{-1} A\right).$$

In the literature, this method is known as the low-rank ADI (LR-ADI) method. The introduced method is summarized in Algorithm 5. For simplicity it is assumed, that the resulting update matrices are real. In general, the proposed method has to be adjusted for the different cases of complex shifts. A more detailed view on this method can be found in [8, 29].
If all finite eigenvalues of the matrix pencil $\lambda E - A$ lie in the open left half-plane, then Z_k converges to the solution factor of (2.22). The rate of convergence strongly depends on the choice of the shift parameters $\tau_1, \ldots, \tau_q$. The optimal ADI shifts satisfy the generalized ADI minimax problem

$$\{\tau_1, \ldots, \tau_{k_q}\} = \underset{\tau_1,\ldots,\tau_{k_q} \in \mathbb{C}^-}{\operatorname{argmin}} \max_{t \in \Lambda_f(A,E)} \frac{|(1 - \bar{\tau}_1) \cdot \ \ldots \ \cdot (1 - \bar{\tau}_q)|}{|(1 + \tau_1) \cdot \ \ldots \ \cdot (1 + \tau_q)|},$$

where $\Lambda_f(A, E)$ denotes the finite spectrum of the regular matrix pencil $\lambda E - A$, see

[29].

Algorithm 5: Generalized LR-ADI Method for Projected Continuous-Time Lyapunov Equations

```
Input: Matrices $A, E, P_l \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and shift parameters $\tau_1, \ldots, \tau_q \in \mathbb{C}^-$
Output: Low-rank Cholesky factor $Z_k$ of the solution $\mathcal{G}_{pc} \approx Z_k Z_k^T$ of (2.22)
1: $Z^{(1)} = \sqrt{-2\mathrm{Re}(\tau_1)}\,(E + \tau_1 A)^{-1} P_l B, \quad Z_1 = Z^{(1)}$
2: for $k = 2, 3, \ldots$ do
3:     $Z^{(k)} = \sqrt{\frac{\mathrm{Re}(\tau_k)}{\mathrm{Re}(\tau_{k-1})}}\,\big(I_n - (\bar{\tau}_{k-1} + \tau_k)(E + \tau_k A)^{-1} A\big)\, Z^{(k-1)}$
4:     $Z_k = \big[Z_{k-1},\ \ Z^{(k)}\big]$
5: end
```

In exact arithmetic, the matrices Z_k satisfy the condition $Z_k = P_r Z_k$. Hence, the second equation in (2.22) is fulfilled for the approximated solution $Z_k Z_k^T$. However, in finite precision arithmetic a drift-off effect may occur. In this case, the update matrices $Z^{(k)}$ need to be reprojected onto the image of P_r by pre-multiplying $Z^{(k)}$ with P_r. In order to limit the computational costs, it is beneficial to restrict this additional projection to every second or third iteration step [29]. To avoid complex operations, caused by the choice of the shifts, the formulation of the LR-ADI should be adjusted, see [7].

For the second projected generalized continuous-time Lyapunov equation (2.23), the transposed method can be used with the matrices $C \in \mathbb{R}^{p\times n}$ and $P_r \in \mathbb{R}^{n\times n}$. In this case the reprojection uses the projector P_l.

As well as for the continuous-time Lyapunov equations (2.22) and (2.22), the projected generalized discrete-time Lyapunov equations (2.24) and (2.25) have to be solved in sparse form. From the upper bound on the number of non-zero improper Hankel singular values (4.1), it can be seen that the terms νm and νp provide an upper bound on the size of the low-rank factorizations for the improper controllability and observability Gramian.

A first method for the computation of this projected generalized discrete-time Lyapunov equations can be constructed by a LR-ADI method for generalized discrete-time Lyapunov equations, see [6]. For the projected version of the algorithm the same adaptations as in the continuous-time case have to be made. In contrast to the normal formulation of the discrete-time Lyapunov equations, the sign of the constant term is negative, so the roles of the A and E matrices have to be swapped. This leads to a drawback of the ADI method used in the projected case. Since the projected generalized discrete-time Lyapunov equations (2.24) and (2.25) relate to the deflating subspaces corresponding to the infinite eigenvalues of the matrix pencil $\lambda E - A$, this spectrum is used in most shift computation methods. But the roles of the A and E matrices have been swapped, so the considered spectrum completely consists of zero eigenvalues. Such shifts cannot be used for the discrete version of the LR-ADI method [6]. Still, the use of this method is possible by a cyclic approach of the LR-ADI method with small non-zero chosen shifts.

However, for this special case of projected generalized discrete-time Lyapunov equations there exists a much more efficient computation method. Next, the projected generalized discrete-time Lyapunov equation (2.24) is considered. As mentioned before, the matrix pencil $\lambda E - A$ is assumed as c-stable and thus, the matrix A is invertible. Then, the equation (2.24) is equivalent to the projected discrete-time Lyapunov equation of the form

$$\mathcal{G}_{ic} - (A^{-1}E)\mathcal{G}_{ic}(A^{-1}E)^T = Q_r A^{-1} B B^T A^{-T} Q_r^T, \quad \mathcal{G}_{ic} = Q_r \mathcal{G}_{ic} Q_r^T. \tag{4.17}$$

Note that Q_r can now be seen as the spectral projector onto the subspace of $A^{-1}E$ corresponding to the zero eigenvalues. In this case $Q_r A^{-1} E = A^{-1} E Q_r$ is nilpotent with index ν. So, this index is identical to the index of the matrix pencil $\lambda E - A$. The unique solution of the new projected Lyapunov equation (4.17) is given by

$$\mathcal{G}_{ic} = \sum_{k=0}^{\nu-1} (A^{-1}E)^k Q_r A^{-1} B B^T A^{-T} Q_r^T \left((A^{-1}E)^T\right)^k,$$

see [29].
Thus, the full-rank Cholesky factor Z of the solution $\mathcal{G}_{ic} = ZZ^T$ of (4.17) can be written in the form

$$Z = \left[Q_r A^{-1} B, A^{-1} E Q_r A^{-1} B, \ldots, (A^{-1}E)^{\nu-1} Q_r A^{-1} B\right].$$

The computation of this full-rank factor is made by the Smith method, displayed in Algorithm 6.

Algorithm 6: Generalized Smith Method for Projected Discrete-Time Lyapunov Equations

Input: Matrices $A, E, Q_r \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$
Output: Full-rank Cholesky factor Z_k of the solution $\mathcal{G}_{ic} = Z_k Z_k^T$ of (2.22)

```
1: $Z^{(1)} = Q_r A^{-1} B, \quad Z_1 = Z^{(1)}$
2: for $k = 2, 3, \ldots, \nu$ do
3:     $Z^{(k)} = A^{-1} E Z^{(k-1)}$
4:     $Z_k = \left[Z_{k-1}, \;\; Z^{(k)}\right]$
5: end
```

In contrast to a LR-ADI method, this procedure makes an exact construction of the Cholesky factor by the exact number of iteration steps. That is why, the Smith method does not need any computed shift parameters for convergence acceleration.
If the index of the system (2.5) is unknown, the Algorithm 6 can be stopped when $\left|\left|Z^{(k)}\right|\right|$ or $\left|\left|Z^{(k)}\right|\right| / \left|\left|Z_k\right|\right|$ is small enough for a certain matrix norm $||\cdot||$.
As in the continuous-time case, the constructed matrices Z_k are affected by a drift-off effect. To avoid this, the columns of Z_k have to be reprojected onto the image of Q_r, so $Z^{(k)}$ should be pre-multiplied by the matrix Q_r. As before, the second projected generalized discrete-time Lyapunov equation (2.25) can be solved with the same method using the transposed matrices of A and E, the spectral projector Q_l and the matrix C.
Back to the generalized Hankel-norm approximation method, the steps 3-13 of Algorithm 4 stay the same as before. Only the computation methods for the solution of

the Lyapunov equations change for the sparse case. In consequence, the dimensions of the used Cholesky factors have been changed. It can be noted that computation of low-rank Cholesky factors for the slow subsystem is an implicit use of the theory in the previous section.

5 Implementation in the MORLAB Toolbox

The Model Order Reduction LABoratory (MORLAB) toolbox is a collection of MATLAB routines for the model reduction of dense continuous-time linear systems, consisting of modal and balancing related model reduction methods. Furthermore, algorithms for the partial stabilization of systems based on Lyapunov and Bernoulli equations are given. For the solution of the corresponding matrix equations spectral projection methods like the matrix sign function and the disk function are used. This whole chapter deals with the implementation of the generalized Hankel-norm approximation method in the MORLAB toolbox by the usage of spectral projection methods [2].

5.1 The Projection-Free Generalized Hankel-Norm Approximation

In the first steps of the generalized Hankel-norm approximation, two balanced minimal realizations are computed with the decoupling of the system into a slow and a fast subsystem as an additional result. Therefor, the Cholesky factors of the projected generalized Lyapunov equations (2.22)-(2.25) are needed. Considering the dense case, the required spectral projectors are not generally constructed in an explicit way, so they have to be computed. One possible way would be (2.14).

In [24, 26] it was proposed to use a decoupling of the descriptor system (2.5) rather than the computation of the spectral projectors for the generalized balanced truncation method. The algorithm mentioned there was based on the generalized Schur decomposition. The *generalized Schur form* of a regular matrix pencil $\lambda E - A$ is given as

$$E = V \begin{bmatrix} E_f & E_u \\ 0 & E_\infty \end{bmatrix} U^T, \quad A = V \begin{bmatrix} A_f & A_u \\ 0 & A_\infty \end{bmatrix} U^T, \tag{5.1}$$

where V and U are orthogonal. The matrix pencil $\lambda E_f - A_f$ is quasi-triangular and contains all the finite eigenvalues of $\lambda E - A$. On the other hand, the matrix pencil $\lambda E_\infty - A_\infty$ is triangular and contains infinite eigenvalues.

To compute the block diagonalization corresponding to the deflating subspaces of finite and infinite eigenvalues, the generalized Sylvester equation

$$\begin{aligned} E_f Y - Z E_\infty &= -E_u, \\ A_f Y - Z A_\infty &= -A_u \end{aligned} \tag{5.2}$$

has to be solved. This Sylvester equation has a unique solution, since the matrix pencils $\lambda E_f - A_f$ and $\lambda E_\infty - A_\infty$ have no common eigenvalues.

Now, these matrices can be used for the generalized state space transformation of the descriptor system (2.5). The resulting descriptor system has the form

$$\begin{bmatrix} E_f & 0 \\ 0 & E_\infty \end{bmatrix} \dot{x}(t) = \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix} x(t) + \begin{bmatrix} B_f \\ B_\infty \end{bmatrix} u(t),$$
$$y(t) = \begin{bmatrix} C_f, C_\infty \end{bmatrix} x(t) + Du(t),$$

where the remaining matrices are constructed as

$$V^T B = \begin{bmatrix} B_u \\ B_\infty \end{bmatrix}, \qquad B_f = B_u - ZB_\infty,$$
$$CU = \begin{bmatrix} C_f, C_u \end{bmatrix}, \quad C_\infty = C_f Y + C_u.$$

Obviously, this realization of the descriptor system (2.5) decouples into the slow and fast subsystems. It was mentioned before that the generalized balanced truncation method can be seen as an individual reduction of the slow and fast subsystems, respectively. Concerning this, the projected generalized continuous-time Lyapunov equations (2.22) and (2.23) corresponding to the proper system Gramians can be replaced by the two generalized continuous-time Lyapunov equations

$$E_f X_{pc} A_f^T + A_f X_{pc} E_f^T + B_f B_f^T = 0, \tag{5.3}$$
$$E_f^T X_{po} A_f + A_f^T X_{po} E_f + C_f^T C_f = 0, \tag{5.4}$$

for the slow subsystem $(E_f, A_f, B_f, C_f, 0)$. The same approach can be used for the projected generalized discrete-time Lyapunov equations (2.24) and (2.25) corresponding to the improper system Gramians. These two equations are replaced by the generalized discrete-time Lyapunov equations

$$A_\infty X_{ic} A_\infty^T - E_\infty X_{ic} E_\infty^T - B_\infty B_\infty^T = 0, \tag{5.5}$$
$$A_\infty^T X_{io} A_\infty - E_\infty^T X_{io} E_\infty - C_\infty^T C_\infty = 0, \tag{5.6}$$

for the fast subsystem $(E_\infty, A_\infty, B_\infty, C_\infty, D)$, see [24]. It is possible to reconstruct all system Gramians and their Cholesky factorizations of the original descriptor system (2.5) from the solutions of the new Lyapunov equations (5.3)-(5.6), see [26].
Now, the projection-free Lyapunov equations (5.3)-(5.6) can be used for two separated balanced truncation approaches. First, the slow subsystem $(E_f, A_f, B_f, C_f, 0)$ is considered. For the matrix $L_f^T E_f R_f$, the skinny singular decomposition of the form

$$L_f^T E_f R_f = U_1 \Sigma V_1^T$$

is computed, where $X_{pc} = R_f R_f^T$ satisfies (5.3) and $X_{po} = L_f L_f^T$ satisfies (5.4). The matrices U_1 and V_1 consist of orthonormal columns and the matrix $\Sigma = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_{l_f})$ contains the Hankel singular values of the slow subsystem, with $l_f = \operatorname{rank}(L_f^T E_f R_f) \leq n_f$. These singular values are exactly the non-zero proper Hankel singular values of the original descriptor system (2.5).
Finally, the slow subsystem has to be reduced. Therefor, the two transformation ma-

trices

$$W_{l_f} = L_f U_1 \Sigma^{-\frac{1}{2}} \quad \text{and} \quad T_{l_f} = R_f V_1 \Sigma^{-\frac{1}{2}}$$

are used. The resulting realization is minimal and balanced, so it can be used for the classical Hankel norm approximation method in Algorithm 1.

For the fast subsystem, it was mentioned before that the reduction of this subsystem is equivalent to the balanced truncation method applied to a discrete-time system of the form (4.5). Hence, the skinny singular value decomposition for the matrix $L_\infty^T A_\infty R_\infty$ has to be computed in the form

$$L_\infty^T A_\infty R_\infty = U_2 \Theta V_2^T,$$

where $X_{ic} = R_\infty R_\infty^T$ and $X_{io} = L_\infty L_\infty^T$ satisfy the equations (5.5) and (5.6), respectively. The matrices U_2 and V_2 have orthonormal columns and the matrix $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_{l_\infty})$ contains the improper non-zero Hankel singular values of the original descriptor system (2.5), with $l_\infty = \operatorname{rank}(L_\infty^T A_\infty R_\infty)$.

All further computation steps of the generalized Hankel-norm approximation method stay the same as before. The new resulting method uses implicitly the spectral projectors corresponding to the deflating subspaces of the finite and infinite eigenvalues by the computation of the block diagonalization with respect to the eigenvalues. But the projectors are no longer explicitly involved in the computation. The complete projection-free Hankel-norm approximation method is summarized in Algorithm 7.

In the beginning, the generalized Schur form was mentioned as bases for the block diagonalization of the matrix pencil $\lambda E - A$. An advantage of this method is the exploitation of the resulting quasi-triangular structure during the following computation steps. For example, the generalized Schur-Hammarling method can be used for the computation of the continuous-time Lyapunov equations (5.3) and (5.4) which takes advantage of the already computed quasi-triangular form of the matrices. The introduction of this method and the corresponding perturbation theory for projected generalized continuous-time Lyapunov equations can be found in [24].

An alternative to the Schur based approach for the block diagonalization as well as for the computation of the matrix equations is given by the application of spectral projection methods. In the following sections, the spectral projection methods needed to compute all necessary steps of Algorithm 7 will be considered.

5.2 Additive Decomposition of Descriptor Systems

To understand the spectral projection based methods specified in this and the following sections, the fundamental theory of spectral projectors is denoted here. Until now, only the spectral projectors of the matrix pencil $\lambda E - A$ corresponding to the deflating subspaces of the finite eigenvalues (2.11) and the infinite eigenvalues (2.12) were stated. The general case is given by the following definitions.

Definition 5.1. *A matrix $P \in \mathbb{R}^{n \times n}$ is a* projector *onto the subspace $\mathcal{S} \subset \mathbb{R}^n$ if* $\operatorname{range}(P) = \mathcal{S}$ *and $P^2 = P$.*

Algorithm 7: Projection-Free Generalized Hankel-Norm Approximation

```
Input: Realization $(E, A, B, C, D)$, such that $\lambda E - A$ is c-stable
Output: Realization of Hankel-norm approximation $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D})$
1: Compute the realization in block diagonal form, such that
       $\left( \begin{bmatrix} E_f & 0 \\ 0 & E_\infty \end{bmatrix}, \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix}, \begin{bmatrix} B_f \\ B_\infty \end{bmatrix}, \begin{bmatrix} C_f & C_\infty \end{bmatrix}, D \right).$
2: Compute the Cholesky factors $R_f$ and $L_f$ of the solutions $X_{pc} = R_f R_f^T$ and
   $X_{po} = L_f L_f^T$ that satisfy
       $E_f X_{pc} A_f^T + A_f X_{pc} E_f^T + B_f B_f^T = 0,$
       $E_f^T X_{po} A_f + A_f^T X_{po} E_f + C_f^T C_f = 0.$
3: Compute the Cholesky factors $R_\infty$ and $L_\infty$ of the solutions $X_{ic} = R_\infty R_\infty^T$ and
   $X_{io} = L_\infty L_\infty^T$ that satisfy
       $A_\infty X_{ic} A_\infty^T - E_\infty X_{ic} E_\infty^T - B_\infty B_\infty^T = 0,$
       $A_\infty^T X_{io} A_\infty - E_\infty^T X_{io} E_\infty - C_\infty^T C_\infty = 0.$
4: Compute the skinny singular value decomposition
       $L_f^T E_f R_f = U_1 \Sigma V_1^T,$
   where $U_1$ and $V_1$ have orthonormal columns, $\Sigma = \operatorname{diag}(\varsigma_1, \ldots, \varsigma_{l_f})$ with the
   proper non-zero Hankel singular values of the system and $l_f = \operatorname{rank}(L_f^T E_f R_f)$.
5: Compute the skinny singular value decomposition
       $L_\infty^T A_\infty R_\infty = U_2 \Theta V_2^T,$
   where $U_2$ and $V_2$ have orthonormal columns, $\Theta = \operatorname{diag}(\theta_1, \ldots, \theta_{l_\infty})$ with the
   improper non-zero Hankel singular values and $l_\infty = \operatorname{rank}(L_\infty^T A_\infty R_\infty)$.
6: Compute the projection matrices
       $W_{l_f} = L_f U_1 \Sigma^{-\frac{1}{2}}, \quad W_{l_\infty} = L_\infty U_2 \Theta^{-\frac{1}{2}},$
       $T_{l_f} = R_f V_1 \Sigma^{-\frac{1}{2}}, \quad T_{l_\infty} = R_\infty V_2 \Theta^{-\frac{1}{2}}.$
7: Compute the minimal standard realization of the slow subsystem
       $(I_{l_f}, A_b, B_b, C_b, 0) = (W_{l_f}^T E_f T_{l_f}, W_{l_f}^T A_f T_{l_f}, W_{l_f}^T B_f, C_f T_{l_f}, 0).$
8: Choose the proper Hankel singular value $\varsigma_{r+1}$.
```

```
9: Permute the balanced standard realization $(I_{l_f}, A_b, B_b, C_b, 0)$, such that the proper system Gramians are
   $\check{\mathcal{G}}_{pc} = \check{\mathcal{G}}_{po} = \operatorname{diag}(\varsigma_1, \dots, \varsigma_r, \varsigma_{r+k+1}, \dots, \varsigma_{l_f}, \varsigma_{r+1} I_k) = \operatorname{diag}(\check{\Sigma}, \varsigma_{r+1} I_k),$
   where $k$ is the multiplicity of the Hankel singular value $\varsigma_{r+1}$.
10: Partition the resulting permuted system according to the proper Gramians
   $\check{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \check{B} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad \check{C} = \begin{bmatrix} C_1 & C_2 \end{bmatrix},$
   where $A_{22} \in \mathbb{R}^{k \times k}$, $B_2 \in \mathbb{R}^{k \times m}$, and $C_2 \in \mathbb{R}^{p \times k}$.
11: Compute the transformation
   $\tilde{A} = \Gamma^{-1}(\varsigma_{r+1}^2 A_{11}^T + \check{\Sigma} A_{11} \check{\Sigma} + \varsigma_{r+1} C_1^T U B_1^T),$
   $\tilde{B} = \Gamma^{-1}(\check{\Sigma} B_1 - \varsigma_{r+1} C_1^T U),$
   $\tilde{C} = C_1 \check{\Sigma} - \varsigma_{r+1} U B_1^T,$
   $\tilde{D} = \varsigma_{r+1} U,$
   with $U = (C_2^T)^\dagger B_2$ and $\Gamma = \check{\Sigma}^2 - \varsigma_{r+1}^2 I_{l_f - k}$.
12: Compute the additive decomposition
   $\tilde{G} = \tilde{C}(sI_{n_f - k} - \tilde{A})^{-1} \tilde{B} + \tilde{D} = G_h(s) + F(s),$
   where $F(s)$ is anti-stable and $G_h(s)$ is the stable Hankel-norm approximation with the realization $(I_r, A_h, B_h, C_h, D_h)$.
13: Compute the minimal realization of the fast subsystem
   $(E_i, I_{l_\infty}, B_i, C_i, D) = (W_{l_\infty}^T E_\infty T_{l_\infty}, W_{l_\infty}^T A_\infty T_{l_\infty}, W_{l_\infty}^T B_\infty, C_\infty T_{l_\infty}, D).$
14: Additive construction of the resulting system
   $(\hat{E}, \hat{A}, \hat{B}, \hat{C}, \hat{D}) = \left( \begin{bmatrix} I_r & 0 \\ 0 & E_i \end{bmatrix}, \begin{bmatrix} A_h & 0 \\ 0 & I_{l_\infty} \end{bmatrix}, \begin{bmatrix} B_h \\ B_i \end{bmatrix}, \begin{bmatrix} C_h, & C_i \end{bmatrix}, D + D_h \right).$
```

Definition 5.2. *Let $Y, X \in \mathbb{R}^{n \times n}$ be a regular matrix pencil with $\Lambda(Y, X) = \Lambda_1 \cup \Lambda_2$, where $\Lambda_1 \cap \Lambda_2 = \emptyset$, and let $\mathcal{S}_1$ be the (right) deflating subspace of the matrix pencil $\lambda X - Y$ corresponding to Λ_1. Then a projector onto $\mathcal{S}_1$ is called a* spectral projector.

The first problem which will be considered, is the decoupling of the descriptor system (2.5). Therefore, a matrix pencil $\lambda X - Y$ has to be block diagonalized. This will be done in two steps. First, the matrix pencil will be transformed into a block triangular form and then, the block diagonal form is computed.

Let $\lambda X - Y$ and $\mathcal{S}_1$ be as in Definition 5.2. Given a spectral projector P onto $\mathcal{S}_1$, the orthonormal basis of the corresponding right deflating subspace $\mathcal{S}_1$ can be computed using the following way.
Let

$$P = VR_1\Pi_1^T, \quad R_1 = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix},$$

with $R_{11} \in \mathbb{R}^{k\times k}$, be a QR decomposition of P with column pivoting. Here, Π_1 is a permutation matrix. Then, an orthonormal basis of $\mathcal{S}_1$ is given by the first k columns of $V = \begin{bmatrix} V_1, & V_2 \end{bmatrix}$, see [4].
The basis of the left deflating subspace can be obtained by the first k columns of U, where

$$UR_2\Pi_2^T = U \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \Pi_2^T = \begin{bmatrix} YV_1, & XV_1 \end{bmatrix}$$

is a QR decomposition of the matrix $\begin{bmatrix} YV_1, & XV_1 \end{bmatrix}$ with column pivoting, if $\Lambda_1 = \Lambda(Y, X) \cap \Gamma_1$, where Γ_1 is the stability region of the matrix pencil $\lambda X - Y$ and the matrix pencil has no eigenvalues on the boundary $\partial\Gamma_1$, see [4].
Then it holds

$$U^T(\lambda X - Y)V = \lambda \begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix} - \begin{bmatrix} Y_{11} & Y_{12} \\ 0 & Y_{22} \end{bmatrix},$$

where $\Lambda(Y_{11}, X_{11}) = \Lambda_1$ and $\Lambda(Y_{22}, X_{22}) = \Lambda_2$.
Now, a method for the computation of such a spectral projector is needed.
Consider the Weierstrass canonical form

$$\lambda X - Y = W \begin{bmatrix} \lambda I_k - J^0 & 0 \\ 0 & \lambda N - J^\infty \end{bmatrix} T,$$

where J^0 contains the Jordan blocks corresponding to the eigenvalues of $\lambda E - A$ inside the unit circle, and J^∞ contains the Jordan blocks corresponding to the eigenvalues outside the unit circle.
Then, the *right matrix pencil disk function* is defined for the regular matrix pencil $\lambda X - Y$ as

$$\operatorname{disk}(Y, X) = T^{-1} \left(\lambda \begin{bmatrix} 0 & 0 \\ 0 & I_{n_\infty} \end{bmatrix} - \begin{bmatrix} I_{n_f} & 0 \\ 0 & 0 \end{bmatrix} \right) T = \lambda \mathcal{P}^\infty - \mathcal{P}^0.$$

The matrix $\mathcal{P}^0$ defines a skew projection onto the right deflating subspace corresponding to the eigenvalues of $\lambda X - Y$ inside the unit circle. On the other hand, the matrix $\mathcal{P}^\infty$ is a skew projection onto the right deflating subspace corresponding to the eigenvalues of $\lambda X - Y$ outside the unit circle.
Other splittings than the unit circle in the complex can be computed by applying a suitable conformal mapping to the matrix pencil $\lambda X - Y$.
The method for the computation of the right matrix disk function can be found in [4]. The needed parts of this method are summarized in Algorithm 8. In the literature, this algorithm is referred to as the disk function method or the inverse free method.

Algorithm 8: Disk Function Method

```
Input: Matrix pencil $\lambda X - Y$ with no eigenvalues on the unit disk
Output: $Y_k$ and $X_k$ with null spaces corresponding to eigenvalues of $\lambda X - Y$
        inside and outside the unit circle
$Y_0 = Y, \quad X_0 = X$
while not converged do
    Compute the $QR$ factorization
        $$\begin{bmatrix} X_k \\ -Y_k \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_k \\ 0 \end{bmatrix}.$$
    $Y_{k+1} = Q_{12}^T Y_k$
    $X_{k+1} = Q_{22}^T X_k$
end
```

Using this algorithm the matrix pencil disk function can be constructed by

$$\operatorname{disk}(Y, X) = (Y_k + X_k)^{-1} (\lambda Y_k - X_k),$$

with $k \to +\infty$. In this case, the skew projections are given by

$$\mathcal{P}^0 = (Y_k + X_k)^{-1} X_k \quad \text{and} \quad \mathcal{P}^\infty = (Y_k + X_k)^{-1} Y_k, \tag{5.7}$$

for $k \to +\infty$. From this, the transformation matrices for the block triangular form can be computed as mentioned above.

A more efficient approach can be used without the explicit computation of the disk function as well as the skew projections. Therefor, a suitable subspace extraction method was proposed in [4]. A slightly modified version of this method can be found in Algorithm 9. The original algorithm uses a RRQR decomposition for the determination of the numerical rank. Here, the rank-revealing QR decomposition was replaced by the singular value decomposition because on modern computers it can be used more efficiently.

Until now, the deflating subspaces corresponding to the eigenvalues of $\lambda X - Y$ inside and outside the unit circle were considered. Now, the methods shall be used for the separation of the finite and infinite eigenvalues of the matrix pencil $\lambda E - A$. Therefor, the disk function method has to be applied to the matrix pencil $\lambda(\alpha A) - E$, with

$$\frac{1}{\alpha} > \max\{|\lambda| : \lambda \in \Lambda(A, E) \setminus \{\infty\}\},$$

and using the subspace extraction method for the original matrix pencil $\lambda E - A$. The scalar α is used to scale the spectrum of $\lambda E - A$. The finite eigenvalues of $\lambda E - A$ correspond to the eigenvalues of $\lambda(\alpha A) - E$ outside the unit circle and the infinite eigenvalues of $\lambda E - A$ correspond to the eigenvalues of $\lambda(\alpha A) - E$ inside the unit circle. The estimation of α is still an unsolved problem [10].

Algorithm 9: Subspace Extraction for the Disk Function Method

```
Input: Matrix pencil $\lambda X - Y$ with no eigenvalues on the unit disk, $Y_k$ as
       computed by Algorithm 8, $\tau$ tolerance for rank detection
Output: Orthogonal matrices $Q, Z \in \mathbb{R}^{n \times n}$ sucht that
       $Q^T(\lambda X - Y)Z = \lambda \begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix} - \begin{bmatrix} Y_{11} & Y_{12} \\ 0 & Y_{22} \end{bmatrix},$
       with $\Lambda(Y_{11}, X_{11}) \subset \{z \in \mathbb{C} : |z| < 1\}$, $\Lambda(Y_{22}, X_{22}) \subset \{z \in \mathbb{C} : |z| > 1\}$
1: Compute the singular value decomposition
       $Y_k = \begin{bmatrix} U_1, & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix},$
   with $V_1 \in \mathbb{R}^{n \times r}$ and $V_2 \in \mathbb{R}^{n \times (n-r)}$. The size of the partitioning $r$ is based on
   the tolerance $\tau$.
2: Set $Z = [V_2, V_1]$.
3: Compute the $QR$ decomposition with column pivoting
       $QR\Pi_Q^R = \begin{bmatrix} YV_2, & XV_2 \end{bmatrix}.$
```

Using the obtained transformation matrices Q and Z from the subspace extraction, the block triangular form of the matrix pencil $\lambda E - A$ can be computed as

$$Q^T(\lambda E - A)Z = \lambda \begin{bmatrix} E_f & E_u \\ 0 & E_\infty \end{bmatrix} - \begin{bmatrix} A_f & A_u \\ 0 & A_\infty \end{bmatrix}. \tag{5.8}$$

An approach for the block diagonalization is the solution of the generalized Sylvester equation (5.2). But since there is no spectral projection method for the generalized Sylvester equation with singular coefficients, standard methods, like the generalized Schur method, have to be used.

A more suitable approach for the block diagonalization can be developed from [15]. It is possible to get a block diagonalizing equivalence transformation from two opposed block triangularizations. Consider the following generalization of Theorem 4.1 in [15].

Theorem 5.1. *Let $\Gamma \subset \mathbb{C}$ be a region in the complex plane which contains n_1 eigenvalues of the matrix pencil $\lambda E - A$. Let $Q, Z \in \mathbb{R}^{n \times n}$ be orthogonal matrices that transform the matrix pencil $\lambda E - A$ into the upper block triangular form:*

$$Q^T(\lambda E - A)Z = \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix} (\lambda E - A) \begin{bmatrix} Z_1, Z_2 \end{bmatrix} = \begin{bmatrix} \lambda E_{11}^{(1)} - A_{11}^{(1)} & \lambda E_{12}^{(1)} - A_{12}^{(1)} \\ 0 & \lambda E_{22}^{(1)} - A_{22}^{(1)} \end{bmatrix}, \tag{5.9}$$

with $\Lambda(A_{11}^{(1)}, E_{11}^{(1)}) \subseteq \Gamma$ and $\Lambda(A_{11}^{(1)}, E_{11}^{(1)}) \cap \Lambda(A_{22}^{(1)}, E_{22}^{(1)}) = \emptyset$. Similarly, let $U, V \in \mathbb{R}^{n \times n}$ be orthogonal matrices that transform the matrix pencil $\lambda E - A$ into the upper block *triangular form:*

$$U^T(\lambda E - A)V = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} (\lambda E - A) \begin{bmatrix} V_1, V_2 \end{bmatrix} = \begin{bmatrix} \lambda E_{11}^{(2)} - A_{11}^{(2)} & \lambda E_{12}^{(2)} - A_{12}^{(2)} \\ 0 & \lambda E_{22}^{(2)} - A_{22}^{(2)} \end{bmatrix}, \tag{5.10}$$

with $\Lambda(A_{22}^{(2)}, E_{22}^{(2)}) \subseteq \Gamma$ and $\Lambda(A_{11}^{(2)}, E_{11}^{(2)}) \cap \Lambda(A_{22}^{(2)}, E_{22}^{(2)}) = \emptyset$. Then

$$X = \begin{bmatrix} U_2, & Q_2 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} Z_1, & V_1 \end{bmatrix} \tag{5.11}$$

are transformation matrices, such that $X^T(\lambda E - A)Y$ has a block diagonal structure where the upper block contains the n_1 eigenvalues lying inside Γ and the lower block has the remaining $n - n_1$ eigenvalues of $\lambda E - A$ outside of Γ.

Proof.
First, consider the transformed matrix pencil

$$X^T(\lambda E - A)Y = \begin{bmatrix} U_2^T(\lambda E - A)Z_1 & U_2^T(\lambda E - A)V_1 \\ Q_2^T(\lambda E - A)Z_1 & Q_2^T(\lambda E - A)V_1 \end{bmatrix}.$$

From (5.9) one can obtain

$$(\lambda E - A)Z_1 = Q_1(\lambda E_{11}^{(1)} - A_{11}^{(1)}),$$

such that

$$U_2^T(\lambda E - A)Z_1 = U_2^T Q_1(\lambda E_{11}^{(1)} - A_{11}^{(1)}).$$

Using (5.9) and (5.1), the transformed matrix can be rewritten as

$$X^T(\lambda E - A)Y = \begin{bmatrix} U_2^T Q_1(\lambda E_{11}^{(1)} - A_{11}^{(1)}) & U_2^T U_1(\lambda E_{11}^{(2)} - A_{11}^{(2)}) \\ Q_2^T Q_1(\lambda E_{11}^{(1)} - A_{11}^{(1)}) & Q_2^T U_1(\lambda E_{11}^{(2)} - A_{11}^{(2)}) \end{bmatrix}.$$

Since Q and U are orthogonal matrices, the matrix products $U_2^T U_1$ and $Q_2^T Q_1$ are both zero. So, the transformed matrices have block diagonal structure. The desired spectral properties of the diagonal blocks follows from the fact that the upper left block is an equivalence transformation of $\lambda E_{11}^{(1)} - A_{11}^{(1)}$ and the lower right block is an equivalence transformation of $\lambda E_{11}^{(2)} - A_{11}^{(2)}$. □

Note that by construction of the transformation matrices (5.11), the block columns of X and Y have orthonormal bases which ensure transformation matrices with an optimal condition number.
In [15] the block triangular form is achieved by applying the QZ algorithm, such that the matrix pencil is first transformed into the generalized Schur form. Then the computation of the opposed generalized Schur form is made by a reordering of the eigenvalues. By the use of the disk function method, it was possible to obtain orthogonal matrices Q and Z to get the block triangularization (5.8). For the opposed block triangularization it would be possible to use the disk function method on the matrix pencil $\lambda E - \alpha A$ and again the subspace extraction method to get orthogonal matrices U and V, such

that

$$U^T(\lambda E - A)V = \lambda \begin{bmatrix} E_\infty & E_v \\ 0 & E_f \end{bmatrix} - \begin{bmatrix} A_\infty & A_v \\ 0 & A_f \end{bmatrix}. \tag{5.12}$$

This additional use of the disk function method is not necessary. By the construction of the skew projection matrices (5.7), one can see that Algorithm 8 already computes the matrices corresponding to the eigenvalues of $\lambda X - Y$ inside the unit circle as well as the matrix corresponding to the eigenvalues outside the unit circle. So, the transformation matrices from (5.12) can be obtained be applying the subspace extraction to the matrix X_k resulting from Algorithm 8.

It should be noted that since the dimensions of the deflating subspaces were determined in the first use of the subspace extraction method, the second application does not need a rank determination anymore. So, it is possible to replace the singular value decomposition by a QR decomposition with column pivoting.

The resulting method for the additive decomposition of descriptor systems into the slow and fast subsystems is summarized in Algorithm 10.

Until now, it was always assumed that the matrix pencil $\lambda E - A$ is c-stable. But using the additive decomposition, the methods also can be applied to unstable descriptor systems.

Algorithm 10: Additive Decomposition of Descriptor Systems

Input: Realization (E, A, B, C, D), such that $\lambda E - A$ is a regular matrix pencil

Output: Block diagonalized realization, such that

$$(\hat{E}, \hat{A}, \hat{B}, \hat{C}, D) = \left(\begin{bmatrix} E_f & 0 \\ 0 & E_\infty \end{bmatrix}, \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix}, \begin{bmatrix} B_f \\ B_\infty \end{bmatrix}, \begin{bmatrix} C_f, & C_\infty \end{bmatrix}, D \right),$$

where $(E_f, A_f, B_f, C_f, 0)$ is the slow and $(E_\infty, A_\infty, B_\infty, C_\infty, D)$ is the fast subsystem.

1: Compute A_k and E_k by applying the disk function method (Algorithm 8) to the matrix pencil $\lambda(\alpha A) - E$.

2: Compute the transformation matrices $Q = \begin{bmatrix} Q_1, & Q_2 \end{bmatrix}$ and $Z = \begin{bmatrix} Z_1, & Z_2 \end{bmatrix}$ by applying the subspace extraction method (Algorithm 9) on A_k and the matrix pencil $\lambda E - A$.

3: Compute the transformation matrices $U = \begin{bmatrix} U_1, & U_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1, & V_2 \end{bmatrix}$ by applying the subspace extraction method (Algorithm 9) on E_k and the matrix pencil $\lambda E - A$.

4: Compute the block diagonalized matrices

$$\begin{bmatrix} E_f & 0 \\ 0 & E_\infty \end{bmatrix} = \begin{bmatrix} U_2^T E Z_1 & 0 \\ 0 & Q_2^T E V_1 \end{bmatrix}, \quad \begin{bmatrix} A_f & 0 \\ 0 & A_\infty \end{bmatrix} = \begin{bmatrix} U_2^T A Z_1 & 0 \\ 0 & Q_2^T A V_1 \end{bmatrix}.$$

5: Compute the effect on B and C

$$\begin{bmatrix} B_f \\ B_\infty \end{bmatrix} = \begin{bmatrix} U_2^T \\ Q_2^T \end{bmatrix} B, \quad \begin{bmatrix} C_f, & C_\infty \end{bmatrix} = C \begin{bmatrix} Z_1, & V_1 \end{bmatrix}.$$

To do so, the additive decomposition of the system has to be computed, such that $G(s) = G_-(s) + G_+(s)$, with $G_-(s) = C_-(sE_- - A_-)^{-1}B_- + D$ and $G_+(s) = C_+(sE_+ - A_+)^{-1}B_+ + D$. Here, the matrix pencil $\lambda E_- - A_-$ is c-stable and all the eigenvalues of the pencil $\lambda E_+ - A_+$ are finite and have non-negative real parts. Then, the reduced-order system $\hat{G}_-(s) = \hat{C}_-(s\hat{E}_- - \hat{A}_-)^{-1}\hat{B}_- + \hat{D}$ is determined by applying the model reduction methods to the c-stable subsystem $G_-(s)$. Finally, the reduced-order approximation of the complete system $G(s)$ is given by $\hat{G}(s) = \hat{G}_-(s) + G_+(s)$, where $G_+(s)$ is included unmodified. Such an additive decomposition of a descriptor system (2.5) can be computed by using a modified version of Algorithm 10. Therefor, the disk function method has to be used on the matrix pencil $\lambda(A - E) - (A + E)$ and the subspace extraction on the original matrix pencil $\lambda E - A$.

5.3 Solving Lyapunov Equations

After the additive decomposition of the descriptor system (2.5), now the Cholesky factors of the continuous-time and discrete-time generalized Lyapunov equations in step 2 and 3 of Algorithm 7 have to be computed. First, the continuous-time generalized Lyapunov equations corresponding to the proper Hankel singular values are considered. Let $Y \in \mathbb{R}^{n\times n}$ be a matrix with no eigenvalues on the imaginary axis. The Jordan canonical form of Y is given by

$$Y = S \begin{bmatrix} J^- & 0 \\ 0 & J^+ \end{bmatrix} S^{-1},$$

where $J^- \in \mathbb{C}^{k\times k}$ contains the Jordan blocks corresponding to eigenvalues in the left open half-plane and $J^+ \in \mathbb{C}^{(n-k)\times(n-k)}$ contains the Jordan blocks corresponding to eigenvalues in the right open half-plane. Then, the *matrix sign function* of Y is defined by

$$\operatorname{sign}(Y) = S \begin{bmatrix} -I_k & 0 \\ 0 & I_{n-k} \end{bmatrix} S^{-1}. \tag{5.13}$$

Note that the matrix sign function of Y is unique [5].
A generalization of this matrix sign function of the matrix pencil $\lambda X - Y$, where both matrices X and Y are non-singular and no eigenvalues of $\lambda X - Y$ lie on the imaginary axis, is given in [4, 5]. There, a Newton iteration of the form

$$Y_{k+1} = \frac{1}{2}\left(\frac{1}{c_k}Y_k + c_k X Y_k^{-1} X\right), \quad Y_0 = Y, \tag{5.14}$$

is considered. The scalar c_k is chosen in order to accelerate the Newton iteration. There are different possibilities how to choose the acceleration parameter c_k. Some of them are given in [8]. Now, consider the generalized Lyapunov equation of the form

$$EZA^T + AZE^T + Q = 0, \tag{5.15}$$

where $A, E, Z, Q \in \mathbb{R}^{n\times n}$ and $Q^T = Q$. Assuming $\lambda E - A$ to be c-stable and both matrices to be non-singular, the Newton iteration in (5.14) can be applied to the matrix pencil

$$\lambda X - Y = \lambda \begin{bmatrix} E^T & 0 \\ 0 & E \end{bmatrix} - \begin{bmatrix} A^T & 0 \\ Q & -A \end{bmatrix}. \tag{5.16}$$

Then, the iteration converges to $Y_{k\to\infty} = \lim\limits_{k\to\infty} Y_k$, with

$$Y_{k\to\infty} = \begin{bmatrix} -E^T & 0 \\ 2EZE^T & E \end{bmatrix}. \tag{5.17}$$

Here, Z is the solution of (5.15). Using the iteration (5.14) and the block structure of (5.16), one can define the following iteration:

$$\begin{aligned} A_{k+1} &= \frac{1}{2}\left(\frac{1}{c_k}A_k + c_k E^T A_k^{-1} E^T\right), & A_0 &= A^T, \\ Q_{k+1} &= \frac{1}{2}\left(\frac{1}{c_k}Q_k + c_k E A_k^{-1} Q_k A_k^{-T} E^T\right), & Q_0 &= Q. \end{aligned}$$

Let the limits of the iteration be denoted by

$$A_{k\to\infty} = \lim_{k\to\infty} A_k, \quad Q_{k\to\infty} = \lim_{k\to\infty} Q_k,$$

the resulting solution for the complete matrix pencil (5.16) can be constructed in the form

$$Y_{k\to\infty} = \begin{bmatrix} A_{k\to\infty}^T & 0 \\ Q_{k\to\infty} & -A_{k\to\infty} \end{bmatrix}. \tag{5.18}$$

Hence, with (5.17) the solution of the Lyapunov equation (5.15) is given by

$$Z = E^{-1} Q_{k\to\infty} E^{-T}.$$

Now, the symmetric Q term is replaced by BB^T. So, the iteration method can be applied to the generalized continuous-time Lyapunov equation of the form (5.3). The second continuous-time Lyapunov equation (5.4) can be solved by the application of the transposed method. Since both iterations only differ in the construction of the solution, it is possible to compute the solutions of both Lyapunov equations at the same time with only one iterate A_k.

Since the matrix Q appears in the considered cases in a symmetric factored form, all iterates Q_k are also in a symmetric factored form. In the case $Q = FF^T$, the iteration of Q_k can be replaced by

$$F_{k+1} = \frac{1}{\sqrt{2c_k}}\left[F_k, \quad c_k E A_k^{-1} F_k\right], \quad F_0 = F.$$

From (5.17) and (5.18) one obtains $A_{k\to\infty} = -E$. This suggests a stopping criterion of the form

$$||A_k + E|| \leq tol \cdot ||E||,$$

with an appropriate matrix norm $||.||$ and a user defined tolerance *tol*. The complete resulting method is summarized in Algorithm 11.

Algorithm 11: Sign Function Method for Generalized Dual Lyapunov Equations

Input: Matrices $A_f, E_f \in \mathbb{R}^{n\times n}$, $B_f \in \mathbb{R}^{n\times m}$, $C_f \in \mathbb{R}^{p\times n}$ from (5.3) and (5.4)

Output: Low-rank Cholesky factors $X_{pc} = RR^T$ and $X_{po} = L^TL$ of (5.3) and (5.4)

```
1: $A_0 = A_f, \quad R_0 = B_f, \quad L_0 = C_f$
2: while $||A_k + E_f|| > tol \cdot ||E_f||$ do
3:     Compute scaling factor $c_k$.
4:     $R_{k+1} = \frac{1}{\sqrt{2c_k}} \left[R_k, \quad c_k E_f A_k^{-1} R_k\right]$
5:     $L_{k+1} = \frac{1}{\sqrt{2c_k}} \begin{bmatrix} L_k \\ c_k L_k A_k^{-1} E_f \end{bmatrix}$
6:     $A_{k+1} = \frac{1}{2c_k} \left(A_k + c_k^2 E_f A_k^{-1} E_f\right)$
7: end
8: $R = \frac{1}{\sqrt{2}} E_f^{-1} R_k$
9: $L = \frac{1}{\sqrt{2}} L_k E_f^{-1}$
```

Note that the workspace for storing the iterates L_k and R_k in Algorithm 11 doubles at each step. This can be avoided by using a column compression method for R_k and a row compression method for L_k with a suitable tolerance criterion. This can be done, for example, by the singular value decomposition or the QR decomposition of L_k and R_k. Using this compression, the costs of solving the systems of linear equation with the matrix E_f in the last two iteration steps can be reduced.

Hence, it remains to solve the generalized discrete-time Lyapunov equations (5.5) and (5.6). As announced before, there is no spectral projection method for the solution of the generalized discrete-time Lyapunov equations with singular coefficient matrices. But it is possible to construct the exact solution by a Smith method.

Since the matrix pencil $\lambda E - A$ was assumed as c-stable, the matrix A is invertible. Especially, this holds for the additive decomposition, such that A_f and A_∞ are both invertible, too. Analogously to (4.17) the generalized discrete-time Lyapunov equation (5.5) is equivalent to the discrete-time Lyapunov equation of the form

$$X_{ic} - A_\infty^{-1} E_\infty X_{ic} (A_\infty^{-1} E_\infty)^T = A_\infty^{-1} B_\infty B_\infty^T A_\infty^{-T},$$

where the matrix $A_\infty^{-1} E_\infty$ has the same index of nilpotency ν as the complete descriptor system (2.5).

That's why the unique solution of (5.5) is given by

$$X_{ic} = \sum_{k=0}^{\nu-1} (A_\infty^{-1} E_\infty)^k A_\infty^{-1} B_\infty B_\infty^T A_\infty^{-T} ((A_\infty^{-1} E_\infty)^T)^k,$$

and the full-rank Cholesky factor of X_{ic} is constructed as

$$Z = \left[A_\infty^{-1}B_\infty, \quad A_\infty^{-1}E_\infty A_\infty^{-1}B_\infty, \quad \dots, \quad (A_\infty^{-1}E_\infty)^{\nu-1}A_\infty^{-1}B_\infty\right].$$

The resulting Smith method can be found in Algorithm 12. It is the projection-free equivalent version of Algorithm 6. Because of this, the same criterion for convergence as for Algorithm 6 can be used here.

Algorithm 12: Projection-Free Generalized Smith Method

```
Input: Matrices $A_\infty, E_\infty \in \mathbb{R}^{n_\infty \times n_\infty}$, $B_\infty \in \mathbb{R}^{n_\infty \times m}$ from (5.5)
Output: Full-rank Cholesky factor $X_{ic} = Z_k Z_k^T$ of (5.5)
1: $Z^{(1)} = A_\infty^{-1}B_\infty, \quad Z_1 = Z^{(1)}$
2: for $k = 2, 3, \dots, \nu$ do
3:     $Z^{(k)} = A_\infty^{-1}E_\infty Z^{(k-1)}$
4:     $Z_k = \left[Z_{k-1}, \quad Z^{(k)}\right]$
5: end
```

In general, the Algorithm 12 is numerically unstable, since it can be seen as a power method by applying the matrix $A_\infty^{-1}E_\infty$ to the iterate $Z^{(k)}$ in each iteration step. In practice, the index of descriptor systems is usually less than or equal to 3. So, Algorithm 12 needs only a few steps to compute the desired solution factor and the instability is of small influence. A possible way to stabilize this method would be an orthogonalization during the iteration steps.

5.4 Additive Decomposition of Standard Systems

The last undiscussed step in the projection-free Hankel-norm approximation method is the computation of an additive decomposition of the standard system in step 12 of Algorithm 7. Here, it is necessary to separate the stable Hankel-norm approximation G_h from the anti-stable part F.

Let $(I_{n-k}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ be the standard realization resulting from step 11 of Algorithm 7. A first approach can be made by using the disk function based decomposition method from Section 5.2. There, it was claimed that the decomposition into the stable and anti-stable part can be made by applying the disk function to the matrix pencil $\lambda(A - E) - (A + E)$. In case of a standard system, this would be the matrix pencil of the form $\lambda(\tilde{A} - I_{n_f-k}) - (\tilde{A} + I_{n_f-k})$. Even so, the application of Theorem (5.1) for the standard case is not recommended, since the product of two different orthogonal matrices does not preserve the standard form, i.e., $\hat{E} = U_2^T I_{n_f-k} Z_1 \neq I_r$.

A more practicable approach is the usage of the standard matrix sign function. The definition of the matrix sign function for a matrix $Y \in \mathbb{R}^{n \times n}$, having no eigenvalues on the imaginary axis, was given in (5.13). In standard case, the generalized sign function iteration (5.14) simplifies to

$$Y_{k+1} = \frac{1}{2}\left(\frac{1}{c_k}Y_k + c_k Y_k^{-1}\right), \quad Y_0 = Y, \tag{5.19}$$

where the scalars c_k are chosen again to accelerate the convergence of the iteration [8]. From the matrix sign function, projectors onto the stable and anti-stable invariant subspaces can be constructed. The projector onto the subspace corresponding to the eigenvalues of Y in the open left half-plane is given by

$$\mathcal{P}^{-} = \frac{1}{2}\left(I_n - \operatorname{sign}(Y)\right).$$

On the other hand, the projector onto the anti-stable deflating subspace is given by

$$\mathcal{P}^{+} = \frac{1}{2}\left(I_n + \operatorname{sign}(Y)\right).$$

Note that P^- and P^+ are not orthogonal projectors, but skew projectors along the Y-invariant subspace [8].

Let p^- be the number of eigenvalues of Y with negative real part and p^+ the number of eigenvalues with positive real part. It can be shown that

$$p^- = \frac{1}{2}(n + \operatorname{tr}(\operatorname{sign}(Y))) \quad \text{and} \quad p^+ = \frac{1}{2}(n - \operatorname{tr}(\operatorname{sign}(Y))) \tag{5.20}$$

hold, where $\operatorname{tr}(M)$ denotes the trace of the matrix M [8].

As in Section 5.2, it is possible to obtain orthogonal transformation matrices for the block triangularization of Y from the spectral projectors by applying a pivoted QR decomposition. Since the dimensions of the deflating subspaces are given by (5.20), there is no rank-revealing method required. So, with a pivoted QR decomposition of the form

$$QR\Pi^T = I_n - \operatorname{sign}(Y),$$

the following block triangularization can be computed

$$Q^TYQ = \begin{bmatrix} Y^- & Y_u \\ 0 & Y^+ \end{bmatrix}, \tag{5.21}$$

where Y^- contains the eigenvalues of Y with negative real part and Y^+ the eigenvalues with positive real part.

At this point, there are two different possibilities for the block diagonalization of the matrix Y. The first one would be the application of Theorem 5.1 by computing a second pivoted QR decomposition of the matrix $I_n + \operatorname{sign}(Y)$. As mentioned before, this is not desired, since it destroys the standard form of the system. The second way is the computation of a stable Sylvester equation of the form

$$Y^-Z - ZY^+ + Y_u = 0, \tag{5.22}$$

for the solution Z. Hence, the block triangular matrix (5.21) can be block diagonalized by

$$\begin{bmatrix} Y^- & 0 \\ 0 & Y^+ \end{bmatrix} = \begin{bmatrix} I_{p^-} & -Z \\ 0 & I_{p^+} \end{bmatrix} \begin{bmatrix} Y^- & Y_u \\ 0 & Y^+ \end{bmatrix} \begin{bmatrix} I_{p^-} & Z \\ 0 & I_{p^+} \end{bmatrix}.$$

Since the matrices used in the Sylvester equation (5.22) have no eigenvalues on the

imaginary axis, a spectral projection method based on the sign function can be used to compute the solution.
Consider a Sylvester equation of the form

$$AZ + ZB + C = 0, \tag{5.23}$$

with $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{n \times m}$. Then, a unique solution Z exists if $\alpha + \beta \neq 0$ for $\alpha \in \Lambda(A)$ and $\beta \in \Lambda(B)$, which is in fact fulfilled if the eigenvalues of A and B lie in the same half-plane [3]. Since the method shall be applied to solve Sylvester equations of the form (5.22), it can be assumed that all eigenvalues of A and B lie in the open left half-plane. Analogously to the sign function based solver for dual Lyapunov equations in section 5.3, here the matrix sign function is applied to a block matrix of the form

$$H = \begin{bmatrix} A & C \\ 0 & -B \end{bmatrix}.$$

By utilizing the block structure of H, an iterative scheme for the three matrices A, B and C can be developed, see [3].
Since all the eigenvalues of A and B lie in the open left half-plane, it holds

$$A_{k\to\infty} = -I_n, B_{k\to\infty} = -I_m.$$

So, the same stopping criterion as for Algorithm 11 can be used for the iteration matrices A_k and B_k at the same time, i.e.,

$$\max\{||A_k + I_n||, ||B_k + I_m||\} > tol,$$

with a suitable matrix norm $||.||$ and a given tolerance tol.
The solution of (5.23) can be obtained by $X = \frac{1}{2}C_{k\to\infty}$.
The resulting method is summarized in Algorithm 13.

Algorithm 13: Sign Function Method for Stable Sylvester Equations

```
Input: Matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$, $C \in \mathbb{R}^{n\times m}$ from (5.23)
Output: Solution $Z$ of the Sylvester equation (5.23)
1: $A_0 = A, \quad B_0 = B, \quad C_0 = C$
2: while $\max\{||A_k + I_n||, ||B_k + I_m||\} > tol$ do
3:     Compute scaling factor $c_k$.
4:     $A_{k+1} = \frac{1}{2}\left(\frac{1}{c_k}A_k + c_k A_k^{-1}\right)$
5:     $B_{k+1} = \frac{1}{2}\left(\frac{1}{c_k}B_k + c_k B_k^{-1}\right)$
6:     $C_{k+1} = \frac{1}{2}\left(\frac{1}{c_k}C_k + c_k A_k^{-1} C_k B_k^{-1}\right)$
7: end
8: $Z = -\frac{1}{2}C_k$
```

5.5 Notes and Open Points of the MORLAB Toolbox

There are several reasons, the routines in the MORLAB toolbox are criticized for. The most prominent one is the computation of an explicit inverse in each step of the sign function iteration methods. Even so, since the matrix sign function is undefined for matrices with purely imaginary eigenvalues it suffers from numerical problems in the presence of eigenvalues close to the imaginary axis. The same problem occurs for the disk function method with eigenvalues close to the unit circle. Fortunately, there are many applications with descriptor system, where the eigenvalues are suitable apart from the imaginary axis. Without the numerical problems, the sign and disk function methods solve problems which are usually better conditioned than the problems solved by the (generalized) Schur approach. This is because, in the Schur approach a quasi-upper triangular matrix is computed while the sign and disk function based methods compute only the block structures. In case of stable matrices, the condition number of $\operatorname{sign}(Z)$ is one and hence, the computation of itself is a well-conditioned problem. Therefore, the results computed by the spectral projection methods often are more accurate than those obtained by the Schur-type algorithms [8]. A second advantage of the spectral projection based methods is that the algorithms basically consist only of basic linear algebra subroutines. Reviewing the announced algorithms in this chapter, only subroutines like solving systems of linear equations, computing an inverse, the QR decomposition and the singular value decomposition are used. Hence, the spectral projection methods are well suited for parallel computations in contrast to the Schur based methods.

Beside the projection-free version of the generalized Hankel-norm approximation and the announced equation solvers, the MORLAB toolbox contains several other algorithms, e.g., a projection-free version of the balanced truncation square-root method and the balanced truncation balancing-free square-root method. Also, there are other balancing related methods corresponding to other matrix equations than the continuous-time Lyapunov equations, e.g., the balanced stochastic truncation, the bounded-real and positive-real balanced truncation, the linear-quadratic-Gaussian balanced truncation, and the $\mathcal{H}_\infty$ balanced truncation. Corresponding to these model reduction methods, the toolbox provides solvers for non-factored continuous-time algebraic Lyapunov equations, continuous-time and $\mathcal{H}_\infty$ algebraic Riccati equations. Also, the toolbox contains methods for the partial stabilization of systems based on continuous-time algebraic Lyapunov equations or algebraic Bernoulli equations.

The next step in the implementation of the MORLAB toolbox is the application of the presented theoretical aspects of the improper system Gramians for other model reduction methods. So far, the additive decomposition of the descriptor system (2.5) was made without any following reduction of the fast subsystem. But the truncation of zero improper Hankel-singular values can be applied to other matrix equation based model reduction methods as it is shown in [11].

6 Numerical Results

After introducing all relevant algorithms and methods, in this section some numerical examples are presented to illustrate the effectiveness of the described methods. All implementations of the introduced methods were made in MATLAB.
First, the projection-free generalized Hankel-norm approximation method is presented on two different medium size data examples. Beside the MORLAB implementation of this method, also a version based on the Schur form was implemented using SLICOT subroutines for the solution of the required matrix equations. SLICOT is the Subroutine Library In systems and COntrol Theory and contains the most important algorithms in system and control theory [17].
For the presentation of the generalized Hankel-norm approximation method with spectral projectors, two sparse data examples were chosen. For the implementation of the sparse method, subroutines from the M-M.E.S.S. were modified for the application with spectral projectors and used for the computation of the projected generalized continuous-time Lyapunov equations (2.22) and (2.23). The M-M.E.S.S. is the MATLAB version of the Matrix Equation Sparse Solver toolbox containing efficient subroutines for sparse matrix equations [19]. For the projected generalized discrete-time Lyapunov equations (2.24) and (2.25), Algorithm 6 has been implemented in MATLAB. For the remaining computation steps, MORLAB subroutines are used.

6.1 A First Index-1 Test Example

To get a first view on the basic behavior of the projection-free Hankel-norm approximation method, a small, dense index-1 descriptor system has been constructed.
Therefor, the matrices were chosen as follows

$$E = Q^T \begin{bmatrix} I_{190} & 0 \\ 0 & 0 \end{bmatrix} Q \quad \text{and} \quad A = Q^T \begin{bmatrix} -1 & & 0 \\ & \ddots & \\ 0 & & -200 \end{bmatrix} Q,$$

where $Q \in \mathbb{R}^{200 \times 200}$ is a random orthogonal matrix, such that the matrix pencil $\lambda E - A$ has equally distributed eigenvalues. The resulting system has 190 finite and 10 infinite states. The input term $B \in \mathbb{R}^{200 \times 3}$ is a dense matrix with random entries and the output term has the form

$$C = \begin{bmatrix} I_2 & 0 & \cdots & 0 \end{bmatrix},$$

such that the first 2 states of the system are of interest. The feed-through term $D \in \mathbb{R}^{2 \times 3}$ is chosen at random. So, the system has multiple inputs and outputs and the matrix pencil $\lambda E - A$ is c-stable with equally distributed eigenvalues and without any special structure. Via the construction of A and E, the system is of index 1.

To this problem, the two implementations of the projection-free generalized Hankel-norm approximation are applied. The convergence histories of the different applied spectral projection based methods in the MORLAB implementation can be found in the tables A.1-A.4 in Appendix A.1. The convergence of the disk function and the matrix sign function method was evaluated by measuring the relative and absolute change of the iteration matrices. The quadratic convergence of the Newton-based methods can be seen after a short warm-up phase in the tables A.2-A.4. Even so, all of the iterations converge in a small number of required steps to a reliable accuracy.
Since the additive decomposition of the system and the computation of the Gramian's factors belong to the computation of the balanced realization, the convergence histories in the tables A.1 and A.2 are identical for each chosen order.
As mentioned in Section 4.3, it is not recommended to compute the real minimal realization by only truncating the zero proper Hankel singular values. Therefore, the tolerance

$$10 \cdot \log n \cdot \epsilon \approx 1.177 \cdot 10^{-14},$$

with $\epsilon \approx 2.22 \cdot 10^{-16}$ the machine precision, was chosen. So, all proper Hankel singular values smaller than this tolerance multiplied with the largest proper Hankel singular value were truncated. The computed proper Hankel singular values and the tolerance for the minimal realization are displayed in Figure 6.1. The resulting balanced slow subsystem is of order 40 with an additional error of $1.7702 \cdot 10^{-15}$ in the Hankel- and $\mathcal{H}_\infty$-norm. This error is neglectable small compared to the later chosen proper Hankel singular values. Considering the infinite states of the system, the projection-free Smith method provides exact two non-zero improper Hankel singular values

$$\theta_1 = 1.042 \cdot 10^{-3} \quad \text{and} \quad \theta_2 = 8.968 \cdot 10^{-4}.$$

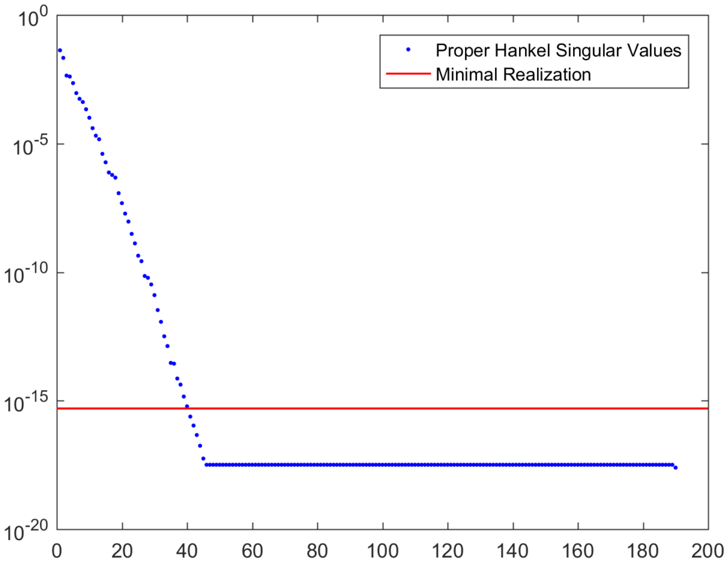

Figure 6.1: Computed proper Hankel singular values and tolerance bound for the balanced realization, example in Section 6.1

The resulting reduced-order fast subsystem is of order 2.
For a more differentiated view on the approximation behavior of the Hankel-norm approximation, reduced-order models of order 4 and 10 were computed.
In Figure 6.2 the absolute error in the 2-induced matrix norm is shown. Beside the realizations computed by the generalized Hankel-norm approximation method (GHNA), also realizations of the same order were computed by the generalized balanced truncation square-root method (GBT(SR)).

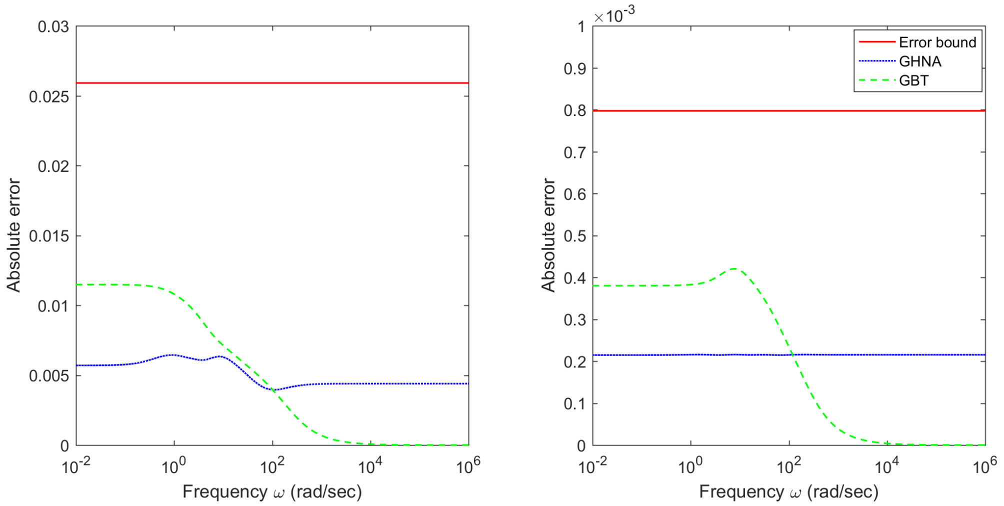

Figure 6.2: Comparison of absolute errors from GHNA and GBT(SR) reduced systems of order 4 (left) and 10 (right), example in Section 6.1

In the right figure the reduced-order model is of order 4 which means the system has $r = 2$ finite and $l_\infty = 2$ infinite states. The chosen proper Hankel singular value for the generalized Hankel-norm approximation is $\varsigma_3 = 4.395764 \cdot 10^{-3}$. Since the additional error is neglectable small compared to ς_3, the Hankel-norm error of the reduced-order model is given by ς_3. The error of the Hankel-norm approximation system seems to approach the chosen proper Hankel singular value ζ_3 of the system. This effect is even stronger noticeable in the right figure. There, both reduced-order models are of order 10 ($r = 8$, $l_\infty = 2$) and the chosen proper Hankel singular value for the generalized Hankel-norm approximation is $\varsigma_9 = 2.1534 \cdot 10^{-4}$. The $\mathcal{H}_\infty$ error of the Hankel-norm approximation seems to be nearly equal to the chosen proper Hankel singular value. This effect can be ascribed to the construction of the reduced-order system based on a scaled all-pass error function (3.21). So, if the influence of the anti-stable part in the transformation becomes smaller, the $\mathcal{H}_\infty$ error of the generalized Hankel-norm approximation begins to approach the chosen Hankel singular value. It has to be noted that this is only a practical observation in the absolute sense. If one would enlarge the error plot, the error function still fluctuates on a relatively high scale around the chosen proper Hankel singular value.
Another observation can be made by comparing the Hankel-norm approximation with the balanced truncation of the same order. The $\mathcal{H}_\infty$ error of the Hankel-norm approximation is smaller than for the balanced truncation. Different practical tests have shown that this observation is often correct. Considering Figure 6.2, it has to be said that the error of the Hankel-norm approximation is smaller in the beginning but after

a while larger than the error of the balanced truncation.
For such small orders as in Figure 6.2, there is no noticeable difference in the comparison of the two implemented versions of the GHNA. Therefore, only the MORLAB version was plotted.
As an advantage of the MORLAB implementation, in contrast to to the Schur-based SLICOT version, the higher accuracy was mentioned. In Figure 6.3 the errors of the reduced-order systems of order 25 ($r = 23$, $l_\infty = 2$) resulting from both implementations of the GHNA are plotted.

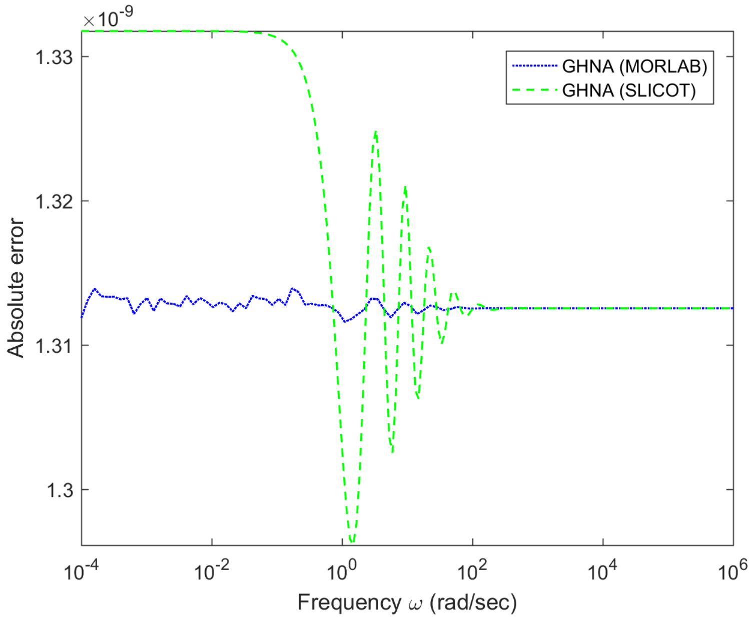

Figure 6.3: Comparison of absolut errors from the MORLAB and SLICOT versions of the GHNA, example in Section 6.1

At the beginning, the error of the SLICOT version is much larger and fluctuates more than the error of MORLAB version before converging to the chosen proper Hankel singular value. The computation of the Hankel-norm of both error systems shows that the error of the SLICOT version is slightly larger than the chosen Hankel singular value while the MORLAB version still serves this condition.

6.2 Semidiscretized Stokes Equation

As a second example, a semidiscretization of the following Stokes equation is considered. Stokes equations describe the flow of fluids at very low velocities without convection and coincide with the linearization of Navier-Stokes equations around the zero-state. In their classical formulation, a Stokes control system for incompressible fluids can be written as

$$
\begin{aligned}
\frac{\partial v}{\partial t} &= \nu \Delta v - \nabla p + Bu, && \text{in } (0,T] \times \Omega,\\
0 &= \operatorname{div}(v) && \text{in } (0,T] \times \Omega,\\
v &= 0 && \text{on } (0,T] \times \partial\Omega,\\
v &= v_0 && \text{in } \{0\} \times \Omega,\\
y &= Cv && \text{in } (0,T] \times \Xi.
\end{aligned}
$$

Here, Ω is a bounded domain in $\mathbb{R}^d$, $d \geq 2$, filled by an incompressible fluid. The constant scalar ν is the viscosity of the fluid, the vector field $v : [0,T] \times \Omega \to \mathbb{R}^d$ describes the velocity of the fluid with an initial state $v_0 \in \mathbb{R}^d$, and the scalar field $p : [0,T] \times \Omega \to \mathbb{R}$ describes the pressure. The term B maps the inputs u into the set of volume forces $f : [0,T] \times \Omega \to \mathbb{R}^d$ and the C term maps the velocity field to the outputs on the observed domain Ξ, see [20].

The spatial discretization of the Stokes equation system above by the finite volume method on a uniform staggered grid leads to a descriptor system of the form

$$
\begin{aligned}
\dot{v}_h(t) &= A_{11} v_h(t) + A_{12} p_h(t) + B_1 u(t),\\
0 &= A_{12}^T v_h(t) + B_2 u(t),\\
y(t) &= C_1 v_h(t) + C_2 p_h(t),
\end{aligned}
\tag{6.1}
$$

where v_h and p_h are the semidiscretized vectors of velocity and pressure, respectively, see [8]. There, A_{11} is the discrete Laplace operator, $-A_{12}$ the discrete gradient operator, and $-A_{21}$ the discrete divergence operator. Due to the non-uniqueness of the pressure, the matrix A_{12} has a rank defect one. In this case, a full-rank matrix A_{12} is obtained by discarding the last column. In the following, the data is already suitably constructed with A_{12} full-rank and $A_{21} = A_{12}^T$. In this case, both matrices A_{12} and A_{21} have full-rank and the overall system is of index $\nu = 2$.

The generation of data is based on the test configuration 3.3 in [20]. That is, the spatial discretization is made on a unit square domain $\Omega = [0,1] \times [0,1]$. For generating the input term B the active control of the flow will be assumed to be restricted to the smaller rectangular domain $\Omega_c = [0.1, 0.9] \times [0.1, 0.3]$. Similarly, the output term C is generated by restricting the observation of the system to the domain $\Omega_m = [0.4, 0.6] \times [0.4, 0.9]$. The resulting order of the full order system is given by $n = n_v + n_p - 1$, with n_v the degrees of freedom in the velocity component and n_p the degrees of freedom in the pressure component.

6.2.1 A Small Dense Problem

As a second application of the projection-free generalized Hankel-norm approximation method, a data set for a medium size system was generated. Therefor, the Stokes equation was discretized on Ω by a uniform staggered grid with 20×20 points. This leads to a problem of order $n = 1159$. The dimensions of the deflating subspaces of the matrix pencil $\lambda E - A$ corresponding to the finite and infinite eigenvalues are $n_f = 361$ and $n_\infty = 798$, respectively.

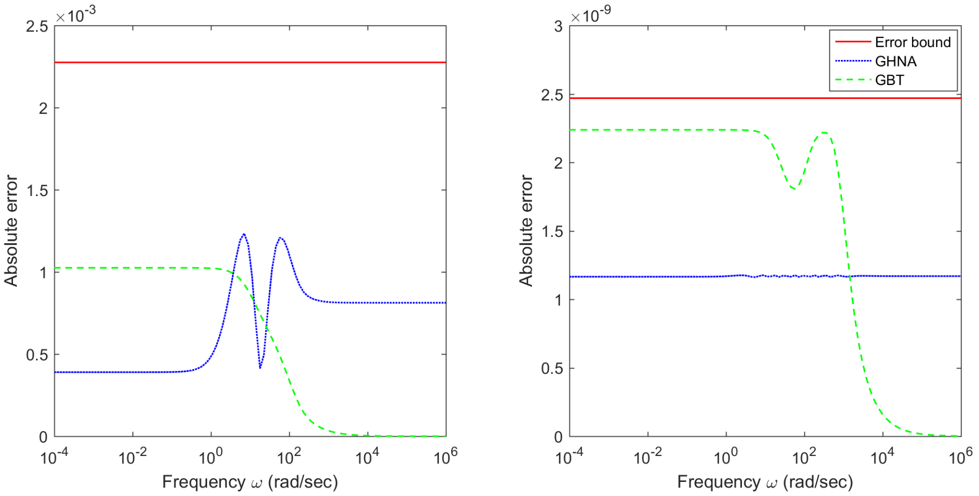

Figure 6.4: Comparison of absolute errors from GHNA and GBT(SR) reduced models of order 2 (left) and 10 (right), example in Section 6.2.1

As in the previous section, the convergence histories of the iterative methods used in the MORLAB implementations of the GHNA and GBT(SR) can be found in the tables A.5-A.8 in Appendix A.2. There, a similar convergence behavior compared to the previous example can be seen. In contrast to before, the column compression during the sign function iteration method for solving the dual Lyapunov equations truncated more than 100 columns of the Cholesky factors which correspond to nearly zero proper Hankel singular values. For the minimal balanced realization the same tolerance formula as before. The obtained balanced realization is of order 17. Considering the fast subsystem, the Smith method computed only one non-zero improper Hankel singular value $\theta_1 = 4.4792 \cdot 10^{-16}$. This improper Hankel singular value might be small, but since it corresponds to the constraints defining the manifold for the solution dynamics, it should not be truncated.

As before, two error plots are displayed for reduced-order models computed by the GHNA and the GBT(SR) are given in Figure 6.4 with the corresponding $\mathcal{H}_\infty$ error bound. There, the left figure shows the reduced-order models of order 2 ($r = 1,\ l_\infty = 1$). The chosen proper Hankel singular value for the GHNA was $\varsigma_2 = 8.1240 \cdot 10^{-4}$. This is an example for a Hankel-norm approximation with a larger $\mathcal{H}_\infty$ error than the balanced truncation of the same order. In contrast, the right figure shows again the behavior seen before. There the order of the approximated models is 10 ($r = 9,\ l_\infty = 1$) with a chosen proper Hankel singular value of $\varsigma_{10} = 1.1694 \cdot 10^{-9}$. The error plot of the GHNA is again approaching the chosen proper Hankel singular value and compared to the balanced truncation, the $\mathcal{H}_\infty$ error is smaller.

Compared to the MORLAB implementation, the SLICOT version of the GHNA can only be used to compute small orders for this example. Numerical disturbances starting to occur at order 8 ($r = 7,\ l_\infty = 1$) and for further increasing orders the SLICOT based GHNA becomes more and more unstable. In Figure 6.5 the error plot of both versions is shown for the reduced system of order 11 ($r = 10,\ l_\infty = 1$). The plot shows that the reduced-order model computed by the SLICOT is not a Hankel-norm approximation anymore.

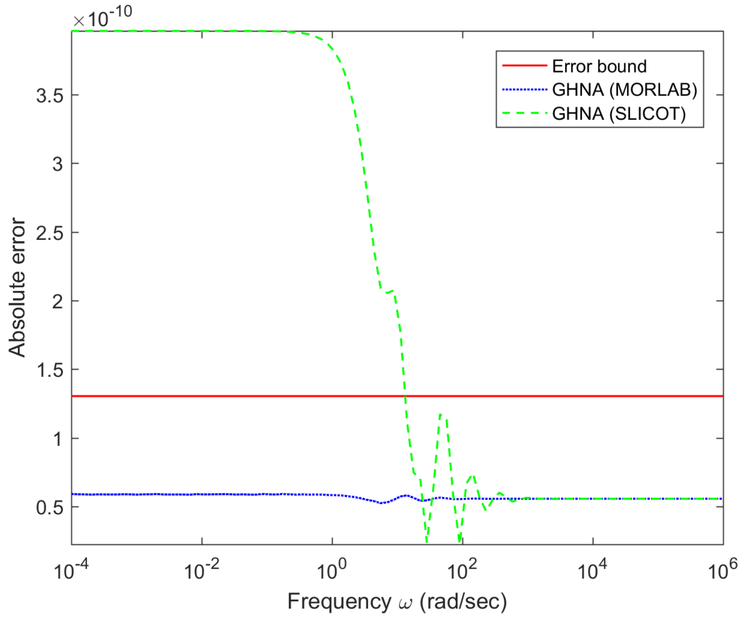

Figure 6.5: Comparison of absolut errors from the MORLAB and SLICOT versions of the GHNA, example in Section 6.2.1

Especially, the $\mathcal{H}_\infty$ error bound does not hold for the computed result anymore. On the other hand, the system obtained by the spectral projection methods still fulfills all conditions. So, the MORLAB version provides to be the more accurate and stable implementation of the GHNA on this example.

6.2.2 A Large Sparse Problem

The dimensions of the discretized descriptor system (6.1) quickly enlarge for a more accurate grid on Ω. But the computed matrices have only a small number of entries unequal to zero. So, it is more practicable to make use of the special large-scale sparse structure of the system than using the projection-free method. For this purpose, the sparse implementations of the GHNA and the GBT(SR) are used for this example. First, note that for the usage of the sparse implementations the spectral projectors onto the left and right deflating subspaces of the matrix pencil $\lambda E - A$ corresponding to the finite eigenvalues are needed, where

$$E = \begin{bmatrix} I_{n_f} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & 0 \end{bmatrix}$$

are sparse and have a special block structure. This block structure can be exploited to explicitly construct the spectral projectors P_l and P_r. It holds

$$P_l = \begin{bmatrix} \Pi & -\Pi A_{11} A_{12} (A_{12}^T A_{12})^{-1} \\ 0 & 0 \end{bmatrix}, \quad P_r = \begin{bmatrix} \Pi & 0 \\ -(A_{12}^T A_{12})^{-1} A_{12}^T A_{11} \Pi & 0, \end{bmatrix},$$

where $\Pi = I_{n_f} - A_{12}(A_{12}^T A_{12})^{-1} A_{12}^T$ is the orthogonal projector onto $\mathrm{Ker}(A_{12}^T)$ along $\mathrm{Im}(A_{12})$, see [27]. A collection of more general construction formulas for matrix pencils of different indices and forms can be found in [29].

For the spatial discretization of the Stokes equation a uniform staggered grid of the size 80 × 80 was used on Ω. As result, the generated full order system (6.1) has the order $n = 19039$, where the matrix pencil $\lambda E - A$ has $n_f = 6241$ finite eigenvalues and $n_\infty = 12798$ infinite ones. Considering the upper bound for the size of the reduced fast subsystem (4.1), one can see that this system can be massively reduced because $\nu m = \nu p = 2$.

It was mentioned before that for the computation of the projected generalized continuous-time Lyapunov equations (2.22) and (2.23) routines from the M-M.E.S.S. toolbox were used. The applied LR-ADI method converged after 51 iteration steps for both low-rank factors $R_{51} \in \mathbb{R}^{n \times 51}$ and $L_{51} \in \mathbb{R}^{n \times 51}$, where $\mathcal{G}_{pc} \approx R_{51}R_{51}^T$ and $\mathcal{G}_{po} \approx L_{51}L_{51}^T$ approximate the proper controllability and observability Gramian, respectively. To prevent the drift-off effect, in every fourth iteration step the update matrix was projected back onto the corresponding deflating subspace. Since this additional projection is only done each fourth iteration step, the computational overhead has its limits. The shifts used in the LR-ADI method were computed during the iteration by using the previous computed low-rank solution factor for a projection. For shift methods based on the eigenvalues of the matrix pencil $\lambda E - A$, the shift computation has to be restricted to the deflating subspace corresponding to the finite eigenvalues of $\lambda E - A$. The convergence histories of both LR-ADI methods is plotted in Figure A.1 in the Appendix A.3. The implemented Smith method for the projected generalized discrete-time Lyapunov equations (2.24) and (2.25) converged after one iteration step. So, the fast subsystem has only one non-zero improper Hankel singular value $\theta_1 = 5.3046 \cdot 10^{-18}$.

For this example of the semidiscretized Stokes equation, note that the computation of the low-rank factors corresponding to the proper Gramians can be computed using much smaller continuous-time Lyapunov equations as well as there exists an explicit construction for the low-rank factors corresponding to the improper Gramians using the block structure of the descriptor system, see [27].

The balanced minimal realization of the slow subsystem was determined by using the same tolerance formula as before. The realization is of order 21. For the remaining additive decomposition of the transformed slow subsystem into its stable and anti-stable parts, MORLAB subroutines has been used. For this example, no further comparison between the SLICOT and MORLAB implementation versions is done since the application of these subroutines is restricted to the additive decomposition of a small dense subsystem. The convergence histories of the two used MORLAB subroutines can be seen in the tables A.9 and A.10 in Appendix A.3. As in the previous examples, the quadratic convergence of the sign function based methods can be seen.

The semidiscretized Stokes system (6.1) was approximated by reduced-order systems of order $l = 5$ ($r = 4$, $l_\infty = 1$) computed by the GHNA and the GBT(SR). So, for the generalized Hankel-norm approximation the proper Hankel singular value $\varsigma_5 = 1.8370 \cdot 10^{-6}$ was chosen. The error plot for this example is shown in Figure 6.6. There, the same approximation behavior as for the dense methods in the previous sections can be seen for the GHNA and the GBT(SR).

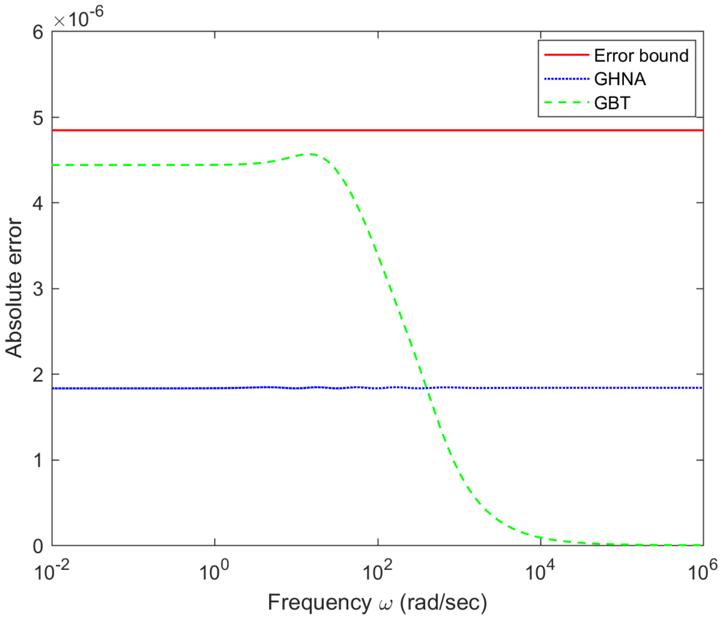

Figure 6.6: Comparison of absolute errors from GHNA and GBT(SR) reduced-order models of order 5, example in Section 6.2.2

6.3 Constraint Damped Mass-Spring System

As last numerical example, a damped mass-spring system with a holonomic constraint is considered, see Figure 6.7.

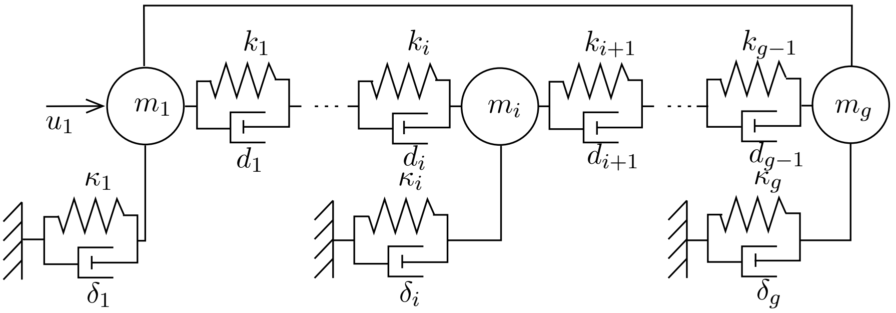

Figure 6.7: A damped mass-spring system with a holonomic constraint

The i-th mass of weight m_i is connected to the $(i+1)$-st mass by a spring and a damper with constants k_i and d_i, respectively, and also to the ground by a spring and a damper with constants κ_i and δ_i, respectively. Additionally, the first mass is connected to the last one by a rigid bar and it is influenced by the control $u(t)$. The vibration of this system is described by a descriptor system of the form

$$
\begin{aligned}
\dot{p}(t) &= v(t),\\
M\dot{v}(t) &= Kp(t) + Dv(t) - G^T\lambda(t) + B_2u(t),\\
0 &= Gp(t),\\
y(t) &= C_1p(t),
\end{aligned}
\tag{6.2}
$$

where $p(t) \in \mathbb{R}^g$ is the position vector, $v(t) \in \mathbb{R}^g$ is the velocity vector, $\lambda(t) \in \mathbb{R}$ is the Lagrange multiplier, $M = \operatorname{diag}(m_1, \ldots, m_g)$ is the mass matrix, D and K are the tridiagonal damping and stiffness matrices, and $G = [1, 0, \ldots, 0, -1] \in \mathbb{R}^{1\times g}$ is the constraint matrix. The active part of the input term is given as $B_1 = e_1$ and the active part of the output term as $C_1 = [e_1, e_2, e_{g-1}]^T$, where e_i denotes the i-th column of the identity matrix I_g. The descriptor system (6.2) is of index $\nu = 3$ and the projectors P_l and P_r can be explicitly constructed by

$$
\begin{aligned}
P_l &= \begin{bmatrix} \Pi_1 & 0 & -\Pi_1 M^{-1} D G_1 \\ -\Pi_1^T D(I_g - \Pi_1) & \Pi_1^T & -\Pi_1^T(K + D\Pi_1 M^{-1} D)G_1 \\ 0 & 0 & 0 \end{bmatrix},\\
P_r &= \begin{bmatrix} \Pi_1 & 0 & 0 \\ -\Pi_1 M^{-1} D(I_g - \Pi_1) & \Pi_1 & 0 \\ G_1^T(K\Pi_1 - D\Pi_1 M^{-1} D(I_g - \Pi_1)) & G_1^T D\Pi_1 & 0 \end{bmatrix},
\end{aligned}
$$

where $G_1 = M^{-1}G^T(GM^{-1}G^T)^{-1}$ and $\Pi_1 = I_g - G_1G$ is a projection onto $\operatorname{Ker}(G)$ along $\operatorname{im}(M^{-1}G^T)$, see [8, 29].

For the construction of the data, it was assumed that $m_1 = \ldots = m_g = 100$ as well as

$$
\begin{aligned}
k_1 = \ldots = k_{g-1} = \kappa_2 = \ldots = \kappa_{g-1} = 2, &\quad \kappa_1 = \kappa_g = 4,\\
d_1 = \ldots = d_{g-1} = \delta_2 = \ldots = \delta_{g-1} = 2, &\quad \delta_1 = \delta_g = 10.
\end{aligned}
$$

For $g = 6000$ the resulting descriptor system has the order $n = 12001$ with $m = 1$ inputs and $p = 3$ outputs. The dimensions of the deflating subspaces of the matrix pencil corresponding to the finite and infinite eigenvalues are $n_f = 11998$ and $n_\infty = 3$, respectively.

Since the resulting descriptor system has a large-scale sparse structure, the sparse implementations of the GHNA and the GBT(SR) are used again. Therefore, in Figure A.2 in Appendix A.4 the convergence histories of the LR-ADI method is shown for the two computed low-rank factors $R_{32} \in \mathbb{R}^{n\times 32}$ and $L_{38} \in \mathbb{R}^{n\times 114}$. As in the previous example, the shift parameters are computed during the iteration by projection and in every fourth iteration step the update matrix is reprojected. The resulting balanced minimal realization of the slow subsystem is of order 26. The computed improper Hankel singular values of the system are all zero. This implies that the transfer function $G(s)$ of the full order model is proper and the reduced-order models do not contain any fast subsystem anymore. The convergence histories of the iterative methods used in for the additive decomposition of the transformed slow subsystem in the GHNA can be found in the tables A.11 and A.12 in Appendix A.4.

The descriptor system (6.2) is approximated by standard systems of order 10 ($r =$

10, $l_\infty = 0$) computed by the GHNA and GBT(SR). For the GHNA the proper Hankel singular value $\varsigma_6 = 0.0013$ was chosen. In Figure 6.8 the magnitude of the $(3, 1)$ components of the full order model as well as the reduced-order models obtained by the GHNA and the GBT(SR) are shown. For this small order of the reduced-order systems, it is possible to see slight differences in this magnitude plot. In the beginning, the generalized Hankel-norm approximation approximates the original system better while after the peak the generalized balanced truncation provides a better approximation of the original system. Note that for higher orders no differences are visible anymore in the magnitude plots. For the same frequency range the error plot is given in Figure 6.9.

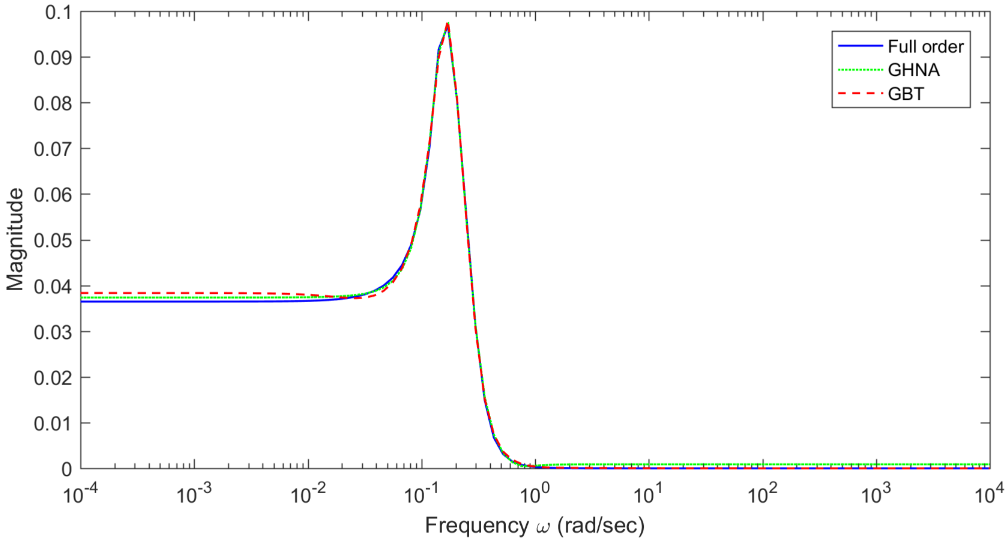

Figure 6.8: Magnitude plots of $G_{31}(j\omega)$ for the full order and the reduced-order models obtained by GHNA and the GBT(SR), example in Section 6.3

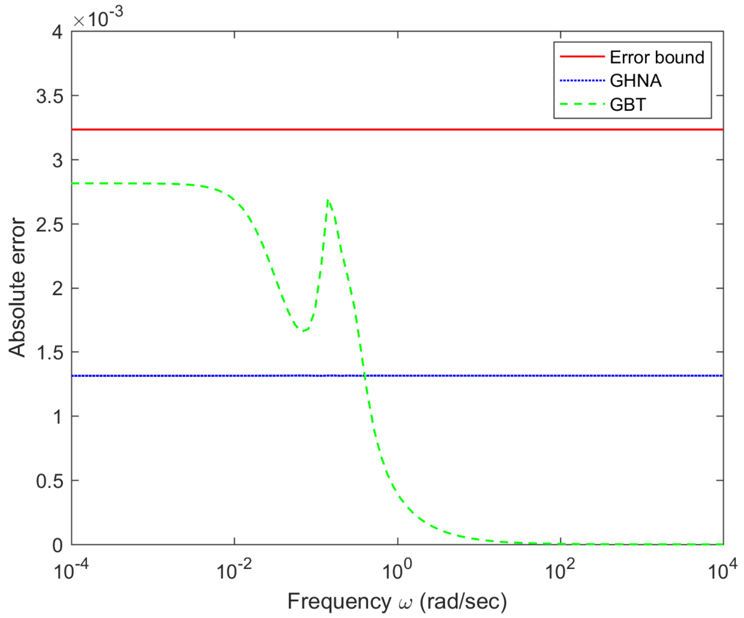

Figure 6.9: Comparison of absolute errors from GHNA and GBT(SR) reduced-order models of order 10, example in Section 6.3

7 Conclusions and Outlook

In this thesis several theoretical aspects and methods for the computation of the generalized Hankel-norm approximation for continuous-time linear time-invariant descriptor systems have been presented.

The Hankel-norm approximation for standard systems was introduced as an extension of the generalized balanced truncation square-root method. As a result, an algorithm for the computation of the generalized Hankel-norm approximation, using the spectral projectors corresponding to the finite and infinite eigenvalues of the matrix pencil of the descriptor system, has been developed. For the application of this method on descriptor systems with large McMillan degrees, an approximated version of the introduced method was considered. A new error bound for the resulting method was shown in the Hankel-norm and the $\mathcal{H}_\infty$-norm. According to this, the generalized Hankel-norm approximation has been considered for the case of large-scale sparse systems. Suitable algorithms for an implementation of this case were given.

For the usage of the generalized Hankel-norm approximation on medium size dense systems, a projection-free version of the method has been developed. Matrix sign function and disk function based methods were introduced as a version of implementation for several main steps in the projection-free generalized Hankel-norm approximation. These results have been implemented in the MORLAB toolbox [2]. This implementation was compared with another one, based on the generalized Schur decomposition and SLICOT subroutines, on two different numerical examples. As it turned out, the MORLAB implementation of the generalized Hankel-norm approximation was the more stable and accurate implementation.

Additionally, an implementation of the generalized Hankel-norm approximation for large-scale sparse systems has been implemented. Therefor, the M-M.E.S.S. toolbox was extended for the usage of spectral projectors. This implementation of the generalized Hankel-norm approximation method was tested on two large-scale sparse data examples with reliable results.

Still, there are many open problems and questions which have to be analyzed in further research. The method, presented in this thesis, was based on the generalized balanced truncation square-root method. A second approach would consider the generalization of the Theorem 3.2. Since this transformation formula is based on the characterization of all-pass transfer functions, this concept has to be generalized to the case of descriptor systems.

Another open question, which has to be considered, is the numerical stability of the method. There are some steps in the introduced algorithm for which the generalized Hankel-norm approximation can become quickly unstable. That has been shown by the numerical examples in Chapter 6. Especially, the numerical stability of the transformation formula in Theorem 3.2 strongly depends on the size of the minimal realization and the chosen proper Hankel singular value. If the minimal realization has many small proper Hankel singular values, the transformation becomes numerically unstable. In Section 4.3 it has been shown that small proper Hankel singular values can be truncated to receive a smaller minimal realization. The problem is to obtain an order for the minimal realization which avoid a large additional error and still is small enough that the method is numerically stable. A similar problem exists for the chosen proper Hankel singular value. If this value is too small, the algorithm becomes numerically unstable. It is unknown how small the proper Hankel singular value can be chosen before the method becomes unstable.

Bibliography

[1] A. C. Antoulas. *Approximation of large-scale dynamical systems.* Society for Industrial and Applied Mathematics, Philadelphia, 2005.

[2] P. Benner. MORLAB - Model Order Reduction Laboratory. MATLAB Toolbox. `http://www.mpi-magdeburg.mpg.de/1657682/morlab` (Online; accessed August 28th, 2016).

[3] P. Benner. Factorized solution of Sylvester equations with application in control. In *Proceedings of the Sixteenth International Symposium on: Mathematical Theory of Network and Systems, MTNS 2004*, pages 1–10, Leuven, Belgium, July 2004.

[4] P. Benner. Partial stabilization of descriptor systems using spectral projectors. In P. Van Dooren, P. S. Bhattacharyya, H. R. Chan, V. Olshevsky, and A. Routray, editors, *Numerical Linear Algebra in Signals, Systems and Control*, chapter 3, pages 55–76. Springer Netherlands, Dordrecht, 2011.

[5] P. Benner, J. M. Claver, and E. S. Quintana-Ortí. Efficient solution of coupled Lyapunov equations via matrix sign function iteration. In A. Dourado et al., editors, *Proc. 3^{rd} Portuguese Conf. on Automatic Control CONTROLO'98*, pages 205–210, Coimbra, 1998.

[6] P. Benner and P. Kürschner. Computing real low-rank solutions of Sylvester equations by the factored ADI method. Preprint on webpage at `https://www2.mpi-magdeburg.mpg.de/preprints/2013/MPIMD13-05.pdf`, July 2013.

[7] P. Benner, P. Kürschner, and J. Saak. Real versions of low-rank ADI methods with complex shifts. Preprint on webpage at `https://www2.mpi-magdeburg.mpg.de/preprints/2012/MPIMD12-11.pdf`, May 2012.

[8] P. Benner, V. Mehrmann, and D. C. Sorensen. *Dimension reduction of large-scale systems.* Springer-Verlag, Berlin, Heidelberg, New York, 2005.

[9] P. Benner, E. S. Quintana-Ortí, and G. Quintana-Ortí. Computing optimal Hankel norm approximations of large-scale systems. In *2004 43rd IEEE Conference on Decision and Control (CDC)*, volume 3, pages 3078–3083, Atlantis, Paradise Island, Bahamas, December 2004. Institute of Electrical and Electronics Engineers.

[10] P. Benner, E. S. Quintana-Ortí, and G. Quintana-Ortí. Parallel model reduction of large-scale linear descriptor systems via balanced truncation. In M. Daydé, J. Dongarra, V. Hernández, and J. Palma, editors, *High Performance Computing for Computational Science - VECPAR 2004: 6th International Conference, Valencia, Spain, June 28-30, 2004, Revised Selected and Invited Papers*, pages 340–353. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[11] P. Benner and T. Stykel. Model order reduction for differential-algebraic equations: a survey. Preprint on webpage at `https://www2.mpi-magdeburg.mpg.de/preprints/2015/MPIMD15-19.pdf`, November 2015.

[12] X. Cao, B. Saltik, and S. Weiland. Hankel model reduction for descriptor systems. In *2015 IEEE 54th Annual Conference on Decision and Control (CDC)*, pages 4668–4673, Osaka, Japan, December 2015. Institute of Electrical and Electronics Engineers.

[13] K. Glover. All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. *International Journal of Control*, 39(6):1115–1193, 1984.

[14] S. Gugercin, T. Stykel, and S. Wyatt. Model reduction of descriptor systems by interpolatory projection methods. *SIAM Journal on Scientific Computing*, 35(5):B1010–B1033, 2013.

[15] B. Kågström and P. Van Dooren. A generalized state-space approach for the additive decomposition of a transfer matrix. *Int. J. Numerical Linear Algebra with Applications*, 1(2):165–181, 1992.

[16] G. Kalogeropoulos, M. Mitrouli, A. Pantelous, and D. Triantafyllou. The Weierstrass canonical form of a regular matrix pencil: Numerical issues and computational techniques. In S. Margenov, L. G. Vulkov, and J. Waśniewski, editors, *Numerical Analysis and Its Applications: 4th International Conference, NAA 2008, Lozenetz, Bulgaria, June 16-20, 2008. Revised Selected Papers*, pages 322–329. Springer Berlin Heidelberg, 2009.

[17] Niconet e.V. SLICOT Basic System and Control Toolbox. MATLAB Toolbox. `http://slicot.org/` (Online; accessed August 8th, 2016).

[18] R. Romijn, S. Weiland, and W. Marquardt. Proper orthogonal decomposition for model reduction of linear differential-algebraic equation systems. Preprint of the 18th IFAC World Congress, Milano (Italy), September 2011.

[19] J. Saak, M. Köhler, and P. Benner. M-M.E.S.S.-1.0.1 – The Matrix Equations Sparse Solvers library. DOI:10.5281/zenodo.50575, April 2016. see also:`www.mpi-magdeburg.mpg.de/projects/mess`.

[20] M. Schmidt. *Systematic discretization of input/output maps and other contributiuons to the control of distributed parameter systems.* PhD thesis, Berlin University of Technology, May 2007.

[21] W. Skolaut, editor. *Maschinenbau - Ein Lehrbuch für das ganze Bachelor-Studium.* Springer-Verlag, Berlin, Heidelberg, 2014.

[22] V. I. Sokolov. *Contributions to the minimal realization problem for descriptor systems.* PhD thesis, Chemnitz University of Technology, 2006.

[23] T. Stykel. *Analysis and numerical solution of generalized Lyapunov equations.* PhD thesis, Berlin University of Technology, June 2002.

[24] T. Stykel. Numerical solution and perturbation theory for generalized Lyapunov equations. *Linear Algebra and its Applications*, 349:155–185, 2002.

[25] T. Stykel. Input-output invariants for descriptor systems. Preprint on webpage at `https://scwww.math.uni-augsburg.de/~stykel/Publications/PIMS-03-1.pdf`, February 2003.

[26] T. Stykel. Gramian based model reduction for descriptor systems. *Mathematics of Control, Signals, and Systems*, 16:297–319, 2004.

[27] T. Stykel. Balanced truncation model reduction for semidiscretized Stokes equation. *Linear Algebra and its Applications*, 415:262–289, 2006.

[28] T. Stykel. Solving projected generalized Lyapunov equations using SLICOT. In *2006 IEEE Conference on Computer Aided Control System Design, 2006 IEEE International Conference on Control Applications, 2006 IEEE International Symposium on Intelligent Control*, pages 14–18, October 2006.

[29] T. Stykel. Low-rank iterative methods for projected generalized Lyapunov equations. *Electronic Transactions on Numerical Analysis*, 30:187–202, 2008.

[30] M. Voigt. $\mathcal{L}_\infty$*-norm computation for descriptor systems.* Diploma thesis, Chemnitz University of Technology, 2010.

Appendix: Convergence Histories of the Numerical Examples

The convergence histories of the iterative MORLAB and M-M.E.S.S. subroutines, used for the computations on the displayed numerical examples in Chapter 6, are presented in form of tables and figures in this appendix.

A.1 Convergence Histories of the Index-1 Text Example

The following tables belong to the computation of the generalized Hankel-norm approximation and the generalized balanced truncation on the constructed index-1 example in Section 6.1.

Table A.1: Convergence history of the disk function method, example in Section 6.1

Iteration Step	Absolute Change	Relative Change
1	9.620254e+00	1.115906e+00
2	3.994510e−08	4.633453e−09
3	1.840975e−14	2.135449e−15

Table A.2: Convergence history of the dual Lyapunov equation sign function solver, example in Section 6.1

Iteration Step	Absolute Error	Relative Error
1	2.300651e+01	1.669068e+00
2	4.159987e+00	3.017972e−01
3	1.259797e+00	9.139525e−02
4	3.257941e−01	2.363559e−02
5	3.910992e−02	2.837332e−03
6	7.358685e−04	5.338551e−05
7	2.705521e−07	1.962791e−08
8	4.997945e−14	3.625890e−15
9	3.007648e−14	2.181977e−15

The tables A.3 and A.4 are used to compute the additive decomposition of the stable and anti-stable system. Here the computation of the Hankel-norm approximation of the order 4 is shown. The transformed system has $n_s = 2$ stable poles and $n_u = 37$ anti-stable poles.

Table A.3: Convergence history of the sign function iteration, example in Section 6.1

Iteration Step	Absolute Change	Relative Change
1	6.570649e+02	2.810483e+01
2	1.797430e+01	2.026950e+00
3	2.711054e+00	3.495354e−01
4	5.928985e−01	7.646071e−02
5	7.652252e−02	9.864063e−03
6	1.621295e−03	2.089895e−04
7	7.321225e−07	9.437264e−08
8	1.487571e−13	1.917521e−14
9	1.783243e−15	2.298651e−16

Table A.4: Convergence history of the Sylvester equation sign function solver, example in Section 6.1

Iteration Step	Absolute Error	Relative Error
1	1.372674e+01	7.537494e−01
2	1.983747e+00	1.326800e−01
3	4.710526e−01	3.228748e−02
4	5.998742e−02	4.111733e−03
5	1.249820e−03	8.566674e−05
6	5.641439e−07	3.866826e−08
7	1.149899e−13	7.881785e−15
8	3.524449e−27	2.415772e−28

A.2 Convergence Histories of the Small Stokes Example

The tables in this section show the convergence histories of the iterative methods used for the small Stokes example in Section 6.2.1.

Table A.5: Convergence history of the disk function method, example in Section 6.2.1

Iteration Step	Absolute Change	Relative Change
1	1.000000e+00	1.000000e+00
2	3.696554e+00	1.302256e+00
3	3.898837e+00	1.200364e+00
4	8.374486e−15	2.578315e−15

Table A.6: Convergence history of the dual Lyapunov equation sign function solver, example in Section 6.2.1

Iteration Step	Absolute Error	Relative Error
1	1.481392e+01	2.710221e+00
2	1.698613e+00	3.107628e−01
3	2.943649e−01	5.385434e−02
4	2.785851e−02	5.096740e−03
5	3.773928e−04	6.904438e−05
6	7.118580e−08	1.302351e−08
7	5.980939e−15	1.094218e−15
8	5.347254e−15	9.782853e−16

The following tables show the computation of the additive decomposition for generalized Hankel-norm approximation is of order 2. The transformed system has $n_s = 1$ stable pole and $n_u = 15$ anti-stable poles.

Table A.7: Convergence history of the sign function iteration, example in Section 6.2.1

Iteration Step	Absolute Change	Relative Change
1	2.736134e+03	1.221279e+02
2	1.425108e+01	1.471830e+00
3	2.498954e+00	3.129493e−01
4	3.242142e−01	4.051723e−02
5	1.209384e−02	1.511326e−03
6	1.913715e−05	2.391505e−06
7	3.275834e−11	4.093699e−12
8	9.743300e−16	1.217587e−16

Table A.8: Convergence history of the Sylvester equation sign function solver, example in Section 6.2.1

Iteration Step	Absolute Error	Relative Error
1	1.333112e+01	1.802437e−01
2	2.238004e+00	9.625141e−03
3	3.242938e−01	1.204074e−04
4	1.207087e−02	4.481807e−06
5	1.913706e−05	7.105422e−09
6	3.275773e−11	1.216266e−14
7	1.928930e−22	7.161947e−26

A.3 Convergence Histories of the Large-Scale Stokes Example

In this appendix the convergence histories of the iterative methods used for the large-scale stokes example in Section 6.2.2 are shown. The normalized residual norms of R_k and L_k denote numerical values of the form

$$\frac{||ER_kR_k^TA^T + AR_kR_k^TE^T + P_lBB^TP_l^T||}{||P_lBB^TP_l^T||}$$

where A, E, B, and P_l are the matrices from the projected generalized continuous-time Lyapunov equation (2.22). An analog formulation is used for the low-rank factors L_k. Here, the Frobenius norm is used as matrix norm.

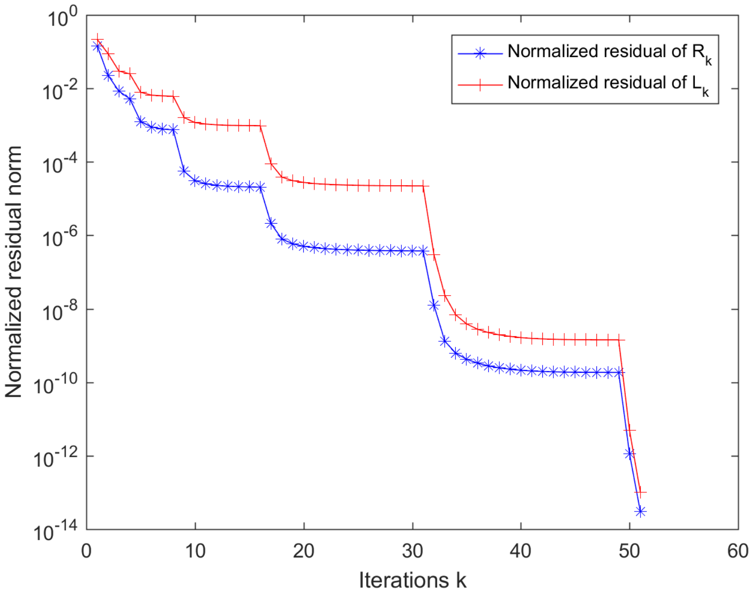

Figure A.1: Convergence histories of the normalized residual norms of the low-rank factors $\mathcal{G}_{pc} \approx R_kR_k^T$ and $\mathcal{G}_{po} \approx L_kL_k^T$, example in Section 6.2.2

The following two tables describe the convergence histories of the additive decomposition of the transformed system for Hankel-norm approximation of the order 5. The transformed system has $n_s = 4$ stable poles and $n_u = 16$ unstable poles.

Table A.9: Convergence history of the sign function iteration, example in Section 6.2.2

Iteration Step	Absolute Change	Relative Change
1	1.532838e+04	6.881858e+01
2	2.138810e+02	5.416575e+00
3	2.905646e+01	1.270019e+00
4	6.249697e+00	2.978575e−01
5	2.486120e+00	1.184752e−01
6	1.115318e+00	5.328100e−02
7	2.130972e−01	1.017866e−02
8	9.630014e−03	4.599808e−04
9	8.512319e−06	4.065938e−07
10	1.091815e−11	5.215089e−13
11	4.952740e−15	2.365693e−16

Table A.10: Convergence history of the Sylvester equation sign function solver, example in Section 6.2.2

Iteration Step	Absolute Error	Relative Error
1	8.038769e+01	1.226552e−02
2	2.201734e+01	3.359396e−03
3	8.519506e+00	1.299903e−03
4	3.435850e+00	5.242405e−04
5	1.225533e+00	1.869914e−04
6	2.070073e−01	3.158508e−05
7	9.608322e−03	1.466034e−06
8	8.428610e−06	1.286034e−09
9	1.081253e−11	1.649771e−15

A.4 Convergence Histories of the Constraint Damped Mass-Spring Example

In this section the convergence histories of the iterative methods used for the constraint damped mass-spring example in Section 6.3 are displayed. For the explanation of the normalized residual norms see the previous appendix.

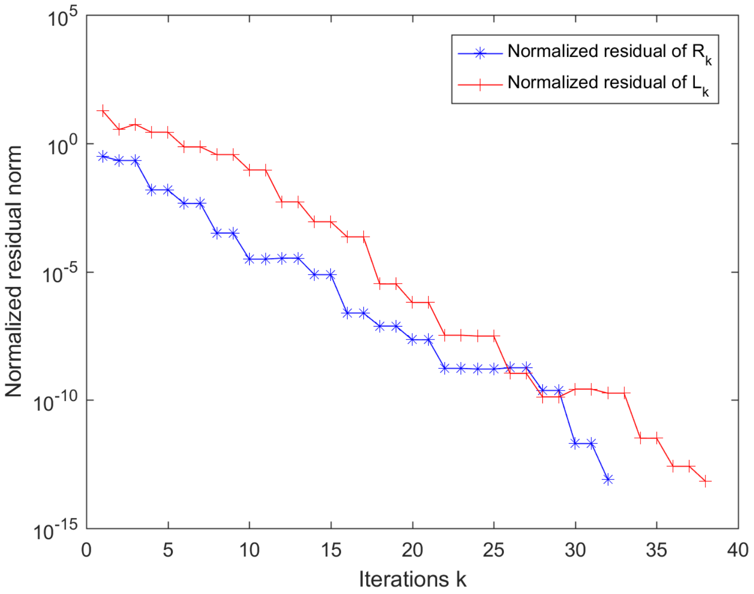

Figure A.2: Convergence histories of the normalized residual norms of the low-rank factors $\mathcal{G}_{pc} \approx R_k R_k^T$ and $\mathcal{G}_{po} \approx L_k L_k^T$, example in Section 6.3

The following two tables show the convergence histories of the matrix sign function method and the Sylvester equation solver for the generalized Hankel-norm approximation of order 5, where the transformed system has $n_s = 5$ stable poles and $n_u = 20$ unstable poles.

Table A.11: Convergence history of the sign function iteration, example in Section 6.2.2

Iteration Step	Absolute Change	Relative Change
1	9.444578e+00	8.587927e−01
2	1.009273e+01	8.109110e−01
3	1.094279e+01	1.783576e+00
4	1.010800e+00	1.681469e−01
5	5.109452e−02	8.490514e−03
6	1.605096e−04	2.667220e−05
7	6.294943e−09	1.046043e−09
8	8.556132e−16	1.421790e−16

Table A.12: Convergence history of the Sylvester equation sign function solver, example in Section 6.2.2

Iteration Step	Absolute Error	Relative Error
1	1.040199e+01	5.140414e+00
2	9.490775e+00	4.690113e+00
3	8.610184e−01	4.254947e−01
4	3.953860e−02	2.636415e−02
5	1.546096e−04	1.473669e−04
6	6.228503e−09	5.936727e−09
7	1.166690e−17	1.112036e−17
8	1.550940e−36	1.478286e−36

Statutory Declaration

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Magdeburg, September 6, 2016