On Vibration Analysis and Reduction for Damped Linear Systems

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von M. Sc. Jonas Denißen
geb. am 30.12.1983 in Berlin

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof. Dr. rer. nat. habil. Peter Benner
Prof. Dr. rer. nat. habil. Ludwig Kohaupt
Prof. Ph. D. Zoran Tomljanović

eingereicht am: 18.12.2018
Verteidigung am: 22.07.2019

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

ADI	Alternating Directions Implicit
ALE	Algebraic Lyapunov Equation
AMPL	A Mathematical Programing Language
BNB	Branch-&-Bound Algorithm
DAE	Differential Algebraic Equation
EAI	Ehrlich-Aberth Iteration
FEM	Finite Element Method
FFT	Fast Fourier Transformation
GEP	Generalized Eigenproblem
IC	Initial Condition
IVP	Initial Value Problem
JNF	Jordan Normal Form
KKT	Karush-Kuhn-Tucker conditions
LP	Linear Program
MATLAB	Matrix Laboratory
MILP	Mixed Integer Linear Program
MINLP	Mixed Integer Nonlinear Program
NLP	Nonlinear Program
NP	Nondeterministic Polynomial Time
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
QCLP	Quadratically Constrained Linear Program
QEP	Quadratic Eigenproblem
QR	QR Decomposition, Schur Decomposition
QZ	QZ Decomposition, generalized Schur Decomposition
RR-QR	Rank Revealing QR Decomposition
SEP	Standard Eigenproblem
SFM	Sign Function Method
SOS2	Special Ordered Set of Variables of Type 2
SVD	Singular Value Decomposition

List of Symbols

Basic Sets

$\mathbb{N}, \mathbb{N}_0$	set of natural numbers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, respectively
$\mathbb{Z}$	set of integers
$\mathbb{R}$	field of real numbers
$\mathbb{R}_+$	open set of nonnegative real numbers
$\mathbb{C}$	field of complex numbers
$\mathbb{C}_+, \mathbb{C}_-$	open sets of complex numbers with positive and negative real part, respectively
$\overline{S}$	closure of the set S
$\operatorname{span}(S)$	span is the intersection of all subspaces containing S, i.e., $\operatorname{span}(S) = \left\{ \sum_{i=1}^{k} \lambda_i v_i \middle\vert k \in \mathbb{N}, v_i \in S, \lambda_i \in \mathbb{F} \right\}$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$
$\mathcal{C}^k$	class of k-times differentiable functions such that $f, f', \ldots, f^{(k)}$ are continuous
$\mathcal{AC}^k$	class of k-times differentiable functions such that $f, f', \ldots, f^{(k)}$ are absolutely continuous
BV	class of functions of bounded variation
L_p	Lebesgue space
L_∞	Lebesgue space of functions that are bounded almost everywhere
L_p^k	Sobolev space
$\mathcal{C}^\omega$	class of analytic functions
$\mathscr{P}$	space of polynomials
$\mathscr{P}_m$	space of polynomials of degree at most m
$\mathcal{T}^k(\Omega_r)$	space of k-th order trigonometric splines w.r.t. nodes Ω_r
$\mathbb{R}[s], \mathbb{C}[s]$	ring of polynomials with coefficients in $\mathbb{R}$ and $\mathbb{C}$, respectively
$Gl_n(\mathbb{R}), Gl_n(\mathbb{C})$	group of invertible $n \times n$ matrices with entries in $\mathbb{R}$ and $\mathbb{C}$, respectively

Matrices and Vectors

I_n	$n \times n$ identity matrix
$0_{m\times n}$	$m \times n$ zero matrix
e_k	k-th unit vector, i.e., $e_k = [0 \cdots 0\, 1\, 0 \cdots 0]^T$
A^T	transpose of the matrix A
A^H	conjugate transpose of the matrix A, i.e., $A^H = \overline{A}^T$
A^{-1}	inverse of a matrix $A \in Gl_n(\mathbb{C})$
A^{-H}	inverse of a matrix $A^H \in Gl_n(\mathbb{C})$, $A^{-H} = (A^{-1})^H = \left(A^H\right)^{-1}$
$\mathrm{im}(A), \ker(A)$	image and kernel of a matrix $A \in \mathbb{C}^{m\times n}$, respectively
$\mathrm{rank}(A)$	rank of matrix $A \in \mathbb{C}^{m\times n}$
$\mathrm{tr}(A)$	trace of matrix $A \in \mathbb{C}^{n\times n}$, i.e., $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$
$\Lambda(A)$	spectrum of matrix $A \in \mathbb{C}^{n\times n}$, i.e., $\Lambda(A) = \{\lambda_1, \ldots, \lambda_r\}$, where $r \leq n$
$\det(A)$	determinant of $A \in \mathbb{C}^{n\times n}$, i.e., $\det(A) = \prod_{i=1}^{r} \lambda_i^{m_i}$, where m_i is the algebraic multiplicity of $\lambda_i \in \Lambda(A)$
$A > (\geq, <, \leq) B$	for two Hermitian matrices $A, B \in \mathbb{C}^{n\times n}$, the matrix $A - B$ is positive definite (positive semidefinite, negative definite, negative semidefinite)
$\mathrm{diag}(v_1, \ldots, v_k)$	diagonal matrix with $v_i \in \mathbb{C}$ for $i = 1, \ldots, k$, i.e., $A \in \mathbb{C}^{k\times k}$
$\mathrm{blockdiag}(A_1, \ldots, A_k)$	blockdiagonal matrix with $A_i \in \mathbb{C}^{m_i \times n_i}$ with $m_i, n_i \in \mathbb{N}_0$ for $i = 1, \ldots, k$, i.e., $A \in \mathbb{C}^{m\times n}$ with $m = \sum_{i=1}^{k} m_i$ and $n = \sum_{i=1}^{k} n_i$
$A \otimes B$	Kronecker product of $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$, i.e., $A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$
$\mathrm{vec}(A)$	vectorization of $A = [a_1, \ldots, a_n] \in \mathbb{C}^{m\times n}$, i.e., $\mathrm{vec}(A) = \left(a_1^H, \ldots, a_n^H\right)^H \in \mathbb{C}^{mn}$

1 Introduction

1.1 Motivation

A repetitive motion of a time-varying process is called *oscillation*. A mechanical oscillation is called *vibration*. The study of vibrations has been of human interest since musical instruments, such as whistles and drums, originated. Ever since, the vibration and its relation to sound have been subject to analytical examination. Even though the art of music was characterized by well-defined guidelines, those cannot be called scientific. The Greek mathematician and philosopher Pythagoras (582–507 B.C.) was probably the first to research the scientific grounds of musical sounds. He experimented with vibrating strings, working with the so-called monochord. Pythagoras discovered that if the same tension is applied to two strings, which have the same characteristics but different lengths, the shorter string generates a higher note. Especially, if the length of the longer string is exactly twice as long as the shorter string, then the shorter string generates a note which is an octave above the other. Although by the time of Pythagoras the notion of pitch was established, the correlation of pitch and frequency was not. The latter was not comprehended until the time of Galileo Galilei (1564–1642) and Marin Mersenne (1588–1648) [Rao07].
The phenomenon of vibration involves an alternating interchange of potential energy to kinetic energy and kinetic energy to potential energy. Hence, any vibrating system must have a component that stores potential energy and a component that stores kinetic energy. The components storing potential and kinetic energies are called a spring or an elastic element and a mass or an inertia element, respectively. The elastic element stores potential energy and supplies it to the inertia element as kinetic energy, and vice versa, in each cycle of motion.
Applications can be found in fields such as stability of mechanical structures, electrical network systems or quantum mechanical systems and vibrational examples include the displacement in strings, bars, shafts, beams, plates or membranes. Examples of vibrations are the periodic swinging motion of a simple pendulum, the sinusoidal vibration of an electric oscillator, or the random motion of a building due to gusts of wind. Famous examples of nonlinear problems are e.g. the van der Pol oscillator [Pol20] and the Duffing equation [Duf18].
We purposely restrict ourselves to a basic model leaving aside gyroscopic effects (i.e. we

do not take Coriolis inertia forces into account), free rigid body motion, singular mass matrices and nonlinearity. Studying each of these effects has its own right and leads to various phenomena such that the spectrum contains purely imaginary eigenvalues, a rotating reference frame or hidden additional algebraic constraints, respectively. In this thesis we consider linear damped oscillations. Our ultimate objective is to describe the time behavior of a damped linear system. The general solution of a damped linear system and therefore its time behavior can be described by the underlying spectral theory e.g. in [Lan02; TM01]. But there are some limitations of this approach, namely non-differentiability (in the classical sense) of the spectral abscissa and numerical computations effect the spectrum and the perturbations may destroy the stability of the system.

1.2 Structure of this Thesis

This thesis is structured as follows. In Chapter 2 we introduce the basic theory this work is based on. In particular, this includes an overview of spectral theory, specifically the theory of linear and quadratic eigenproblems and a short introduction to ordinary differential equations, which are essential for defining the general solution of a damped linear system. Furthermore, we review some approximation and interpolation properties of functions and classify basic optimization problems as needed in this thesis.
In Chapter 3 we study vibrations and relate them to certain norms. To analyze the time behavior of a vibrational system completely, all its components have to be considered individually. For large-scale systems such a detailed analysis is often not applicable, hence, all system components are combined to a single quantity — a *norm*. By this simplification a rough measure of the vibration behavior of the system is obtained. Local regularity for the norm of the solution can be shown but unfortunately classic differentiability is in general lost. By considering the time behavior of a vibrational system in certain norms, we obtain properties such as decoupling, filtering and monotonicity. Here, these norms are obtained by solving an algebraic Lyapunov equation or by considering the algebraic Lyapunov eigenvalue problem.
In Chapter 4 we relate the time behavior of a damped linear system to the solution's trace of an algebraic Lyapunov equation which represents the averaged system's energy. Since a vibration is an alternating interchange of potential energy to kinetic energy and vice versa, it is obvious that vibrations are reduced if the system's energy is minimized. In order to damp the system and reduce its energy, we consider external viscous dampers at fixed positions. Minimizing the averaged system's energy is classified as a Nonlinear Program (NLP) subject to a structured algebraic Lyapunov equation. Numerical methods, such as steepest descent and Newton's method, can be applied to solve the NLP to local optimality since it is sufficiently smooth. A minimum then fulfills the Karush-Kuhn-Tucker [Kar39; KT51] conditions. We show that the structure of the algebraic Lyapunov equation can be kept throughout a sign function iteration and hence, we derive the so-called structure exploiting sign function method. Furthermore, we derive an iterative eigenvalue algorithm based on the so-called Ehrlich-Aberth iteration, which takes the low rank of the viscous damping into account and hence, we can reduce the computational complexity significantly. The

results concerning the structure exploiting sign function method of this chapter have appeared in

> P. Benner and J. Denißen. Numerical solution to low rank perturbed Lyapunov equations by the sign function method. *Proc. Appl. Math. Mech.*, 16(1):723–724, 2016.

The iterative eigenvalue algorithm is described in

> P. Benner and J. Denißen. Ehrlich-Aberth iteration for vibrational systems. *Proceedings of ICoEV 2015*, pages 1540–1548, 2015.

While in Chapter 4 the positions of the viscous dampers were fixed, in Chapter 5 the problem of finding the optimal positions and their viscosities is considered. We show that the optimal positions can be computed in $\mathcal{O}(n^2)$ for sufficiently small viscosities, but unfortunately this result cannot be generalized to arbitrary viscosities. We present a new Mixed Integer Nonlinear Programming (MINLP) formulation of finding the optimal positions and their viscosities. This approach cannot be generalized to large-scale computations and hence, we discuss linearization strategies based on McCormick envelopes and piecewise linear functions. Moreover, we present a heuristic to find good damping positions. The heuristic is based on regulating sparsity of the viscosities by adding an ℓ_1-penalty term to the objective. Once, the damping positions are determined, the viscosities can be optimized by solving the Nonlinear Programming formulation of Chapter 4 with the help of the structure exploiting sign function method.
In Chapter 6 we study vibrations and their time behavior as a solution of time-periodic linear systems. We relate time-periodic linear systems to linear systems by the Floquet-Lyapunov transformation and therefore, results on the solution, such as two-sided rigorous bounds, decoupling, filtering and monotonicity in a newly defined norm can be generalized from linear systems. Moreover, the time behavior can be characterized by two-sided bounds for the Manhattan norm, the Euclidean norm and the maximum norm. Here, we use two different ideas in order to derive the two-sided bounds. While in the first method we approximate the solution of the time-periodic linear system by trigonometric splines and establish two-sided bounds on the quality of the approximation, the second method approximates the time-periodic linear system, which then turns out to be analytic. Hence, its solution can be represented as an infinite series. Depending on the smoothness of the time-periodic system, we formulate two-sided bounds, which incorporate the approximation error of the linear time-periodic system and the truncation error of the series representation. We show the order of convergence of the two-sided bounds to the solution of a linear time-periodic system depending on the smoothness of the linear time-periodic system. As a further result, the computational complexity of both methods has been derived. The results of this chapter have to some extent appeared in

> P. Benner, J. Denißen, and L. Kohaupt. Trigonometric spline and spectral bounds for the solution of linear time-periodic systems. *J. Appl. Math. Comput.*, 54(1):127–157, 2017.

Finally, in Chapter 7 we summarize the results of this thesis and discuss possible future research directions.

1.3 System Setup

In this thesis we perform a number of numerical experiments. These tests have been performed on an Ubuntu machine with Intel® Core™2 Duo CPU with 2.50GHz and 4 GB RAM. The algorithms have been implemented and tested in MATLAB® and Mixed Integer Nonlinear Programs (MINLPs) and Mixed Integer Linear Programs (MILPs) have been formulated in AMPL [FGK89]. The following software libraries and programs have been used:

- MATLAB version 8.3.0.532 (R2014a);
- AMPL version 20170207 [FGK89];
- BARON version 17.1.2 [TS05];
- MINOTAUR BNB 0.2 [Mah+11];
- IBM® ILOG® CPLEX® 12.7.0 [Ibm].

2

Mathematical Preliminaries

In this chapter we introduce the main concepts this thesis is based on. First, we discuss general spectral theory and in particular the theory of linear and quadratic eigenproblems. Then we turn to classical differentiation and to function classes as they are needed for Chebyshev approximation and trigonometric spline interpolation. Moreover, we introduce ordinary differential equations, which are connected to damped linear systems. Furthermore, we review and classify some basic optimization problems as needed in this thesis.

2.1 Matrices and Eigenproblems

2.1.1 Linear Eigenproblems

Here we introduce some fundamentals for matrix pencils (E, A), i.e., first order matrix polynomials, with $E, A \in \mathbb{C}^{n\times n}$.

Definition 2.1.1. *The standard eigenvalue problem for a matrix $A \in \mathbb{C}^{n\times n}$ is the problem of finding a vector $v \in \mathbb{C}^n \setminus \{0\}$ and a scalar $\lambda \in \mathbb{C}$ that obey*

$$Av = \lambda v. \tag{2.1}$$

Definition 2.1.2. *The generalized eigenvalue problem for matrix pencils (E, A) is the problem of finding a vector $v \in \mathbb{C}^n \setminus \{0\}$ and a scalar $\lambda \in \mathbb{C}$ that obey*

$$Av = \lambda Ev, \tag{2.2}$$

where $E, A \in \mathbb{C}^{n\times n}$.

If $E \in Gl_n(\mathbb{C})$, then the standard eigenvalue problem (2.1) can be obtained by multiplying (2.2) by E^{-1}. However, in most situations it is preferable not to perform the inversion, since structural properties of E and A are lost and numerical errors are introduced. Obviously, the numerical errors depend on the condition of E. In general, it is better to solve the generalized eigenvalue problem as stated originally. If E is singular, then the pencil (E, A) is said to have one or more eigenvalues at infinity. The standard eigenvalue problem (2.1) is a specialization of the generalized eigenvalue problem (2.2).

We therefore introduce the basic theory for matrix pencils and implicitly capture the matrix case as well.

Definition 2.1.3. *We denote by*

$$p_{(E,A)}(\lambda) = \det(A - \lambda E). \tag{2.3}$$

the characteristic polynomial of the pencil (E, A).

By definition we obtain, that λ is an eigenvalue of the matrix pencil (E, A) if and only if it is a root of the characteristic polynomial $p_{(E,A)}(\lambda)$ and v is an eigenvector of the matrix pencil (E, A) if and only if it is in the kernel of $A-\lambda E$. Obviously, v is not unique since any multiple of v is an eigenvector as well. By the fundamental theorem of algebra it follows that the pencil (E, A) has exactly n eigenvalues (counting multiplicities). We define the algebraic multiplicity of an eigenvalue $\lambda \in \Lambda(E, A)$ as the number of times it occurs as a zero in $p_{(E,A)}(\cdot)$. We denote by the geometric multiplicity of an eigenvalue $\lambda \in \Lambda(E, A)$ the dimension of the corresponding eigenspace, i.e., $\dim(\ker(A-\lambda_i E)) > 0$. We recall that the geometric multiplicity is at most the algebraic multiplicity of the same eigenvalue $\lambda \in \Lambda(E, A)$.
The eigenvalues are an important tool for decomposing a matrix. Here, we do not consider a matrix pencil but a single matrix. We cite in this context the famous *Jordan Normal Form (JNF)*. Any matrix can be decomposed into its Jordan Normal Form. The decomposition is unique up to the ordering of the respective Jordan blocks.

Theorem 2.1.4 (Jordan Normal Form, e.g. [HJ85])**.** *Let* $r = |\Lambda(A)|$. *There exists* $U \in Gl_n(\mathbb{C})$ *such that*

$$UAU^{-1} = J = \text{blockdiag}(J_1, \ldots, J_m),$$

where $r \leq m \leq n$. *The Jordan blocks* J_i *for* $i = 1, \ldots, m$ *take the form*

$$J_i = J_i(\lambda_{k_i}) = \begin{bmatrix} \lambda_{k_i} & 1 & & 0 \\ & \lambda_{k_i} & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_{k_i} \end{bmatrix} \in \mathbb{C}^{n_i \times n_i}, \tag{2.4}$$

with corresponding $\lambda_{k_i} \in \Lambda(A)$, *where* $\sum_{i=1}^{m} n_i = n$.

Remark 2.1.5.

1. *A is called diagonalizable if it is similar to a diagonal matrix, i.e., if there exists* $S \in Gl_n(\mathbb{C})$ *such that* $S^{-1}AS$ *is a diagonal matrix.*

2. *A is diagonalizable or non-defective if and only if the algebraic and geometric multiplicity coincide for all eigenvalues* $\lambda \in \Lambda(A)$. *Otherwise, the matrix A is called defective.*

3. $J_i = J_i(\lambda_{k_i}) = \lambda_{k_i} I + N$ *with*

$$N = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & & 1 \\ 0 & & & 0 \end{bmatrix} \in \mathbb{R}^{n_i \times n_i}, \tag{2.5}$$

where the matrix N *is nilpotent, i.e.,* $N^{n_i-1} \neq 0$ *and* $N^{n_i} = 0$.

For a defective matrix A a *generalized eigenvector* v w.r.t. a defective eigenvalue λ can be defined. v is a nonzero vector satisfying

$$(A - \lambda I)^s v = 0,$$

where s is the algebraic multiplicity. The set of all generalized eigenvectors for a given eigenvalue λ form the generalized eigenspace for λ. In particular, for a given eigenvalue λ, eigenvectors and generalized eigenvectors $v_1, v_2, \ldots, v_s$ can be chosen such that they are linearly independent and satisfy

$$(A - \lambda I)v_k = \alpha_{k,1} v_1 + \ldots + \alpha_{k,k-1} v_{k-1}$$

for some coefficients $\alpha_{k,1}, \ldots, \alpha_{k,k-1}$ for $k = 1, \ldots, s$. We can choose the first vectors v_k to be eigenvectors for $k = 1, \ldots, \dim(\ker(A - \lambda I))$ and the remaining vectors v_k for $k = \dim(\ker(A - \lambda I)) + 1, \ldots, s$ as generalized eigenvectors. A particular choice of coefficients, which we use in Section 3.3, is $\alpha_{k,1} = \ldots = \alpha_{k,k-2} = 0$ and $\alpha_{k,k-1} = 1$, thus

$$Av_k = \lambda v_k + v_{k-1}\,, \qquad k = 1, \ldots, \mu(\lambda),$$

where $v_0 = 0$.
We conclude this section by a theorem on the simultaneous diagonalization of two matrices $A \in Gl_n(\mathbb{R})$ and $B \in \mathbb{R}^{n \times n}$. In general we cannot expect that A and B are simultaneously diagonalized by a similarity transformation, since this would yield that two arbitrary matrices A and B commute, i.e., $AB = BA$. Hence, we consider in the following theorem a congruence transformation.

Theorem 2.1.6 (Simultaneous diagonalization e.g. [HJ85])**.** *Let* $A \in Gl_n(\mathbb{R})$ *and* $B \in \mathbb{R}^{n \times n}$ *be symmetric matrices and* $A^{-1}B$ *be diagonalizable. Then there exists a* $S \in \mathbb{R}^{n \times n}$ *such that* $S^T AS$ *and* $S^T BS$ *are diagonal.*

Remark 2.1.7. *In Theorem 2.1.6 the matrices* A *and* B *can be interchanged, it is especially sufficient if either one of them is nonsingular, which e.g. can be guaranteed by positive definiteness of* A *or* B.

2.1.2 Quadratic Eigenproblems

So far we have considered the linear eigenproblem (standard eigenvalue and generalized eigenvalue problem) which is the simplest case of the more general polynomial eigenproblem. And here, we focus on the quadratic eigenproblem which often appears in mechanical and vibrating systems.

Definition 2.1.8. *The quadratic eigenvalue problem is the problem of finding a vector* $v \in \mathbb{C}^n \setminus \{0\}$ *and a scalar* $\lambda \in \mathbb{C}$ *that obey*

$$Q(\lambda)v = (A\lambda^2 + B\lambda + C)v = 0, \tag{2.6}$$

where $A, B, C \in \mathbb{C}^{n\times n}$.

In order to find eigenvalues and eigenvectors of $Q(\lambda)v$, one can transform the quadratic eigenproblem into an equivalent linear eigenproblem. Hence, the transformation is also called *linearization*.

Definition 2.1.9. *The generalized eigenvalue problem*

$$\begin{bmatrix} 0 & N \\ -C & -B \end{bmatrix} \begin{bmatrix} v \\ \lambda v \end{bmatrix} - \lambda \begin{bmatrix} N & 0 \\ 0 & A \end{bmatrix} \begin{bmatrix} v \\ \lambda v \end{bmatrix} = 0, \tag{2.7}$$

where $N \in Gl_n(\mathbb{C})$, *is called a linearization of* $Q(\lambda)v$.

A linearization is not unique. If it is possible, it is important to choose a linearization that respects structural properties of the quadratic eigenproblem 2.1.8.

2.2 Normed Vector Spaces

Definition 2.2.1. *Given a vector space* V *over a field* $\mathbb{C}$, *a norm on* V *is a function* $\|\cdot\| \colon V \to \mathbb{R}_+,\ x \mapsto \|x\|$ *that satisfies the following three axioms for all vectors* $x, y \in V$ *and all scalars* $\alpha \in \mathbb{C}$

$$\begin{array}{ll} \|\alpha\, x\| = |\alpha|\, \|x\| & \textit{(absolute homogeneity)}, \\ \|x + y\| \le \|x\| + \|y\| & \textit{(triangle inequality)}, \\ \|x\| \ge 0 \quad \textit{and} \quad \|x\| = 0 \Rightarrow x = \mathbf{0} & \textit{(positive definiteness)}. \end{array} \tag{2.8}$$

Definition 2.2.2. *Given a vector space* V *over a field* $\mathbb{C}$, *a scalar product or inner product is a map* $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{C}$ *that satisfies the following three axioms for all vectors* $x, y, z \in V$ *and all scalars* $\alpha \in \mathbb{C}$

$$\begin{array}{ll} \langle x, y\rangle = \overline{\langle y, x\rangle} & \textit{(conjugate symmetry)}, \\ \langle \alpha\, x, y\rangle = \alpha\, \langle x, y\rangle, \quad \langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle & \textit{(linearity)}, \\ \langle x, x\rangle \ge 0 \quad \textit{and} \quad \langle x, x\rangle = 0 \Rightarrow x = \mathbf{0} & \textit{(positive-definiteness)}. \end{array} \tag{2.9}$$

Example 2.2.3. *Let* $x \in \mathbb{C}^n$.

1. *Let* $p \ge 1$ *be a real number. We call*

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{2.10}$$

a p-norm which fulfills the axioms of a norm.

2. For $p = 1, 2$, we obtain the Manhattan norm and the Euclidean norm, respectively,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \tag{2.11}$$

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|}. \tag{2.12}$$

For $p = \infty$ we define the maximum norm as

$$\|x\|_\infty := \max_{i=1}^{n} |x_i|. \tag{2.13}$$

Let us derive the so-called energy norm. Inner product spaces have a naturally defined norm based upon the inner product of the space itself,

$$\|x\| = \sqrt{\langle x, x \rangle}.$$

Let $B \in \mathbb{C}^{n\times n}$ be a positive definite Hermitian matrix, i.e., $B > 0$ and $x, y \in \mathbb{C}^n$. Then a B scalar product can be defined as

$$\langle x, y \rangle_B := \langle x, By \rangle = y^H Bx$$

and the respective B energy norm is defined as

$$\|x\|_B = \sqrt{\langle x, x \rangle_B} = \sqrt{\langle x, Bx \rangle} = \sqrt{x^H Bx}. \tag{2.14}$$

Definition 2.2.4. *A matrix norm $\|\cdot\|$ is called sub-multiplicative if*

$$\|AB\| \leq \|A\|\|B\|$$

for all matrices $A, B \in \mathbb{C}^{n\times n}$.

Definition 2.2.5. *A matrix norm $\|\cdot\|_b$ on $\mathbb{C}^{n\times n}$ is called compatible with a vector norm $\|\cdot\|_a$ on $\mathbb{C}^n$ if*

$$\|Ax\|_a \leq \|A\|_b \|x\|_a$$

for all $A \in \mathbb{C}^{n\times n}, x \in \mathbb{C}^n$.

Definition 2.2.6. *The operator norm or induced norm w.r.t. the vector norm $\|\cdot\|$ is defined as*

$$\begin{aligned} \|A\| &= \max\{\|Ax\| : x \in \mathbb{C}^n \text{ with } \|x\| = 1\} \\ &= \sup\left\{ \frac{\|Ax\|}{\|x\|} : x \in \mathbb{C}^n \text{ with } x \neq 0 \right\}. \end{aligned}$$

Remark 2.2.7. *Induced norms are compatible by definition.*

Example 2.2.8. *Let $A \in \mathbb{C}^{m\times n}$. The operator norm corresponding to the p-norm for vectors is defined as*

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

In the case of $p = 1$ and $p = \infty$, the norms are

$$\begin{aligned} \|A\|_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|, \\ \|A\|_\infty &= \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|, \end{aligned}$$

which are the maximum absolute column and maximum absolute row sums, respectively.

Definition 2.2.9. *Two norms $\|\cdot\|$ and $|||\cdot|||$ of a vector space V are called equivalent if there exists $c, C > 0$ such that*

$$c|||x||| \leq \|x\| \leq C|||x|||$$

is fulfilled for all $x \in V$.

2.3 Differentiation and Function Classes

First, we start with the differentiability of a real valued function f. Consider the limit of a linear approximation of f at x_0:

$$f'(x_0) = \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

If this limit exists, f is said to be differentiable at x_0 and this limit is defined to be the derivative of the function f at x_0, i.e., $f'(x_0) = \lim_{h\to 0} \frac{f(x_0+h)-f(x_0)}{h}$.

Definition 2.3.1. *A function f is said to be of class $\mathcal{C}^k$, i.e., $f \in \mathcal{C}^k$, if the derivatives $f', f'', \ldots, f^{(k)}$ exist and are continuous.*

Definition 2.3.2. *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be Lipschitz continuous, if there exists a constant $L > 0$, such that*

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$$

for all $x_1, x_2 \in \mathbb{R}$.

Definition 2.3.3. *A function f defined on $[a, b]$ is said to be of class $\mathcal{AC}[a, b]$, i.e., $f \in \mathcal{AC}[a, b]$, if it is absolutely continuous in $[a, b]$, i.e., for any ε, there exists $\delta > 0$ such that for all n and all $a \leq \underline{t}_1 \leq \bar{t}_1 \leq \underline{t}_2 \leq \bar{t}_2 \leq \ldots \leq \underline{t}_n \leq \bar{t}_n \leq b$ with $\sum_{i=1}^{n} |\bar{t}_i - \underline{t}_i| < \delta$:*

$$\sum_{i=1}^{n} \left| f(\bar{t}_i) - f(\underline{t}_i) \right| < \varepsilon.$$

Definition 2.3.4. *A function f is said to be of class $\mathcal{AC}^k$, i.e., $f \in \mathcal{AC}^k$, if the derivatives $f', f'', \ldots, f^{(k)}$ exist and are absolutely continuous.*

Definition 2.3.5. *A function $f : [a,b] \to \mathbb{R}$ is said to be of bounded variation, i.e., $f \in BV[a,b]$, if*

$$V := \sup_{p \in P} \sum_{i=0}^{n_p - 1} |f(x_{i+1}) - f(x_i)| < \infty,$$

where $P = \{p = \{x_0, \ldots, x_{n_p}\} : p \text{ is a partition of } [a,b]\}$.

Remark 2.3.6.

- *The continuity or absolute continuity of $f, f', \ldots, f^{(k-1)}$ is implied by differentiability for $f \in \mathcal{C}^k$ or $f \in \mathcal{AC}^k$, respectively.*
- *$\mathcal{C} = \mathcal{C}^0$ is the class of continuous functions.*
- *$\mathcal{AC} = \mathcal{AC}^0$ is the class of absolutely continuous functions.*
- *Obviously, any absolutely continuous function is continuous as well. Hence, $f \in \mathcal{AC}^k \Rightarrow f \in \mathcal{C}^k$.*
- *Any absolute continuous function is of bounded variation, i.e., $f \in \mathcal{AC}[a,b] \Rightarrow f \in BV[a,b]$, see e.g. [Roy88].*
- *The classes of functions are nested as follows:*

  $$f \in \mathcal{C}^1[a,b] \Rightarrow f \text{is Lipschitz continuous in } [a,b] \Rightarrow f \in \mathcal{AC}[a,b]$$
  $$\Rightarrow f \text{ is of bounded variation in } [a,b] \Rightarrow f \in \mathcal{C}[a,b],$$

  see e.g. [Roy88].

Definition 2.3.7. *A function f defined on $[a,b]$ is in the Lebesgue space $L_p[a,b]$, i.e., $f \in L_p[a,b]$, where $1 \le p \le \infty$,*

$$L_p[a,b] = \{f : f \text{ is measurable on } [a,b] \text{ and } \|f\|_p < \infty\},$$

where

$$\|f\|_{L_p[a,b]} = \|f\|_p = \left(\int_a^b |f(x)|^p \, \mathrm{d}x \right)^{1/p}, \quad \text{for } 1 \le p < \infty,$$

and

$$\|f\|_{L_\infty[a,b]} = \|f\|_\infty = \operatorname*{ess\,sup}_{x \in [a,b]} |f(x)|, \quad \text{for } p = \infty.$$

We define a certain subspace of the Lebesgue space L_p such that the functions posses $k-1$ smooth derivatives,

$$L_p^k[a,b] = \left\{ f : D^{k-1} f \in \mathcal{AC}[a,b] \text{ and } D^k f \in L_p[a,b] \right\} \tag{2.15}$$

with norm

$$\|f\|_{L_p^k[a,b]} = \sum_{j=0}^{k} \|D^j f\|_{L_p[a,b]}.$$

Remark 2.3.8.

- *$L_p[a,b]$ and $L_p^k[a,b]$ are normed linear spaces, in fact they are Banach spaces.*
- *$L_p^k[a,b]$ is a Sobolev space.*
- *The classes of smooth functions are nested as follows:*

$$\mathcal{C}^k[a,b] \subseteq L_\infty^k[a,b] \subseteq L_p^k[a,b] \subseteq L_1^k[a,b] \subseteq \mathcal{C}^{k-1}[a,b]$$

for all $1 \leq p \leq \infty$ and $k \in N$, see e.g. [Roy88; Sch81].

After defining the above function classes, let us now define Green's function $g_m(x;y)$:

$$g_m(x,y) := \begin{cases} \frac{(x-y)^{m-1}}{(m-1)!}, & x \geq y, \\ 0, & x < y, \end{cases} \tag{2.16}$$

where $m \in \mathbb{N}$. Green's function is associated with the Taylor series expansion in Theorem 2.3.9, in fact it is its kernel.

Theorem 2.3.9 (Taylor Series, e.g. in [Sch81])**.** *Let $f \in L_1^m[a,b]$. Then for all $x \in [a,b]$*

$$f(x) = \sum_{j=0}^{m-1} \frac{D^j f(a)(x-a)^j}{j!} + \int_a^b g_m(x;y) D^m f(y) \mathrm{d}y. \tag{2.17}$$

Moreover, there exists $\xi_x \in [a,b]$:

$$f(x) = \sum_{j=0}^{m-1} \frac{D^j f(a)(x-a)^j}{j!} + \frac{D^m f(\xi_x)(x-a)^m}{m!}.$$

We obtain from the Taylor series 2.3.9 the following corollary, which is given e.g. in [For10].

Corollary 2.3.10. *Let $f \in L_1^m[a,b]$. Then for all $x \in [a,b]$:*

$$f(x) = \sum_{j=0}^{m-1} \frac{D^j f(a)(x-a)^j}{j!} + o\left((x-a)^m\right).$$

Definition 2.3.11. *A function $f : D \to \mathbb{R}$ is said to be real analytic in D, i.e., $f \in \mathbb{C}^\omega(D)$, where D is an open set in $\mathbb{R}$, if $f \in \mathcal{C}^\infty(D)$ and for any $x_0 \in D$ the Taylor series*

$$\sum_{j=0}^{\infty} \frac{f^j(x_0)(x-x_0)^j}{j!}$$

converges to $f(x)$ for x in a neighborhood of x_0 pointwise.

By definition, it holds: $\mathcal{C}^\omega \subseteq \mathcal{C}^\infty$.
We conclude the topic Taylor series with two examples which will be used in Section 6.2.

Example 2.3.12. *The Taylor series of the trigonometric functions* sin *and* cos *are given as*

$$\begin{aligned}\sin(x) &= \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots \\ \cos(x) &= \sum_{n=0}^{\infty}(-1)^n\frac{x^{2n}}{(2n)!} = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \cdots\end{aligned}$$

for any $x \in \mathbb{R}$.

Definition 2.3.13. *A function* $f : \mathbb{C} \to \mathbb{C}$ *is said to be entire, if it is holomorphic (analytic) in the whole complex plane* $\mathbb{C}$*, i.e., it can be represented by a power series,*

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

for any $z \in \mathbb{C}$.

Remark 2.3.14. *Examples of analytic (entire) functions are: polynomials, the exponential function and trigonometric functions such as sine and cosine (its Taylor series are given in Example 2.3.12). The logarithmic function* $\log : \mathbb{C} \setminus \{0\} \to \mathbb{C}$ *is analytic as well.*

2.4 Approximation and Interpolation

Let V be a normed vector space over $\mathbb{R}$, $U \subset V$ a finite-dimensional subspace and $f \in V$. We consider the *best approximation* of f in U:

$$u \in U : \|u - f\| \leq \|v - f\| \qquad \forall v \in U. \tag{2.18}$$

It is well known that a solution to the best approximation problem exists as shown in the next theorem.

Theorem 2.4.1 (e.g. [Sch71])**.** *For every* $f \in V$ *there exists a solution* $u \in U$ *that satisfies the best approximation* (2.18)*.*

The proof of Theorem 2.4.1 is nonconstructive and in general the solution is not necessarily unique as it can be seen by the following example.

Example 2.4.2. *Let* $V = \mathbb{R}^2$, $\|\cdot\| = \|\cdot\|_\infty$, $U = \{v = (v_1, v_2) \in \mathbb{R}^2 : v_2 = 0\}$ *and* $f = (f_1, f_2)$*. Then every* $u = (u_1, 0) \in U$ *with* $u_1 \in [f_1 - |f_2|, f_1 + |f_2|]$ *is a solution to the best approximation, since*

$$\|u - f\|_\infty = \max\{|u_1 - f_1|, |f_2|\} = |f_2| = \min_{v \in U}\|v - f\|_\infty \quad \forall u_1 \in [f_1 - |f_2|, f_1 + |f_2|].$$

In Subsection 2.4.1 we consider the polynomial best approximation problem in the $\|\cdot\|_\infty$-norm. In Subsection 2.4.2 we relax the problem and it is not necessary to find the best solution but a "good" solution is sufficient. Hence, we consider the interpolation problem, where we focus on an interpolation with trigonometric splines.

2.4.1 Approximation by Chebyshev Polynomials

Let $\mathscr{P}_m$ be the vector space of real polynomials with degree at most m, i.e.

$$\mathscr{P}_m = \left\{ \sum_{i=0}^{m} a_i x^i,\ a_0, \ldots, a_m, x \in \mathbb{R} \right\}.$$

Clearly, $\mathscr{P}_m[a,b]$ is an $m+1$-dimensional subspace of $\mathcal{C}[a,b]$. Obviously, the monomials $\{1, t, t^2, \ldots, t^m\}$ are a basis of $\mathscr{P}_m$. Linear independence of the monomials can be checked by e.g. the Wronskian in Remark 2.5.5. Every $p \in \mathscr{P}_m$ can be expressed as $p(t) = \sum_{i=0}^{m} a_i t_i$, where $a_i \in \mathbb{R}$. Even though the monomials are linear independent, the Wronskian is often ill-conditioned. Hence, approximating a function by monomials may cause large numerical errors. We therefore consider a different basis, namely the Chebyshev basis, which we introduce in the following. We follow the presentation of Chebyshev projections based on [Tre13]. Any approximation can be used to replace the original function, but our focus is on Chebyshev polynomials due to Example 2.4.6. We consider the best approximation and choose as spaces $V = \mathcal{C}[a,b]$ and $U = \mathscr{P}_m$ in (2.18). Therefore,

$$p \in \mathscr{P}_m : \|p - f\| \leq \|q - f\| \qquad \forall q \in \mathscr{P}_m \tag{2.19}$$

for a given $f \in \mathcal{C}[a,b]$. Due to Theorem 2.4.1, a solution to (2.19) exists and it is unique, see e.g. [HH91].

First, let us introduce the Chebyshev polynomials of the first kind defined by the three term recurrence relation

$$T_{k+1}(t) = 2tT_k(t) - T_{k-1}(t),$$

where $T_0(t) = 1$, $T_1(t) = t$ for $k = 1, 2, 3, \ldots$

Chebyshev polynomials are orthogonal over the interval $[-1, 1]$:

$$\langle T_i, T_j \rangle_\omega := \int_{-1}^{1} T_i(t) T_j(t) \omega(t) \mathrm{d}t = \begin{cases} 0 & \text{for } i \neq j, \\ \pi & \text{for } i = j = 0, \\ \frac{\pi}{2} & \text{for } i = j \neq 0, \end{cases} \tag{2.20}$$

with the weight function $\omega(t) = \frac{1}{\sqrt{1-t^2}}$. In the following, we state results only for the interval $[-1, 1]$, but Chebyshev polynomials can be mapped to any interval by an affine variable transformation. A Lipschitz continuous f has a unique representation as a Chebyshev series [Tre13],

$$f(t) = \sum_{k=0}^{\infty} c_k T_k(t),$$

which is absolutely and uniformly convergent. The coefficients c_k are given by the orthogonality relationship (2.20),

$$c_0 = \frac{1}{\pi} \langle f, T_0 \rangle_\omega \qquad \text{and for } k > 0 : \qquad c_k = \frac{2}{\pi} \langle f, T_k \rangle_\omega. \tag{2.21}$$

The m-truncated Chebyshev series is defined as

$$f_m(t) := (P_m f)(t) := \sum_{k=0}^{m} c_k T_k(t). \tag{2.22}$$

Clearly, the Chebyshev polynomials T_k, $k = 0, 1, \ldots, m$, are a basis of $\mathscr{P}_m$. $P_m : \mathcal{C}[-1, 1] \to \mathscr{P}_m$ defined by (2.22) is a linear operator and it is also called Chebyshev projection since $P_m p = p$ for any $p \in \mathscr{P}_m$ and $P_m T_k = 0$ for $k > m$.
We recall the following two theorems given in [Tre08; Tre13] which are essential for the derivation of our spectral bounds in Section 6.3.

Theorem 2.4.3. *If f and its derivatives through $f^{(k-1)}$ are absolutely continuous on $[-1, 1]$ and if the k-th derivative $f^{(k)}$ is of bounded variation V for some $k \geq 1$, then for any $m > k$, the Chebyshev projection satisfies*

$$\|f - f_m\|_\infty \leq \frac{2V}{\pi k (m-k)^k}.$$

We conclude this section with a theorem when the function f and its derivatives are not only absolutely continuous and of bounded variation but if they are analytic. We therefore use the notion of a Bernstein ellipse $\mathcal{E}_\rho$.

Definition 2.4.4. *For $\rho > 1$ the Bernstein ellipse $\mathcal{E}_\rho$ is defined as*

$$\mathcal{E}_\rho := \left\{ \frac{re^{i\theta} + r^{-1}e^{-i\theta}}{2} \in \mathbb{C} : -\pi \leq \theta \leq \pi,\, 0 \leq r \leq \rho \right\}.$$

Since $re^{i\theta} + r^{-1}e^{-i\theta} = (\rho + \rho^{-1})\cos(\theta) + (\rho - \rho^{-1})i\sin(\theta)$ for $-\pi \leq \theta \leq \pi$, the boundary of the Bernstein ellipse $\partial\mathcal{E}_\rho$ can be written in parametric form as $\partial\mathcal{E}_\rho = \left\{ z \in \mathbb{C} : \frac{\mathrm{Re}(z)^2}{a_\rho^2} + \frac{\mathrm{Im}(z)^2}{b_\rho^2} = 1 \right\}$, where its semi-axes are $\frac{\rho+\rho^{-1}}{2}$ and $\frac{\rho-\rho^{-1}}{2}$ with foci at ± 1. Figure 2.1 shows Bernstein ellipses in the complex plane for $\rho = 1.1, 1.2, \ldots, 1.5$ as in [Tre13].

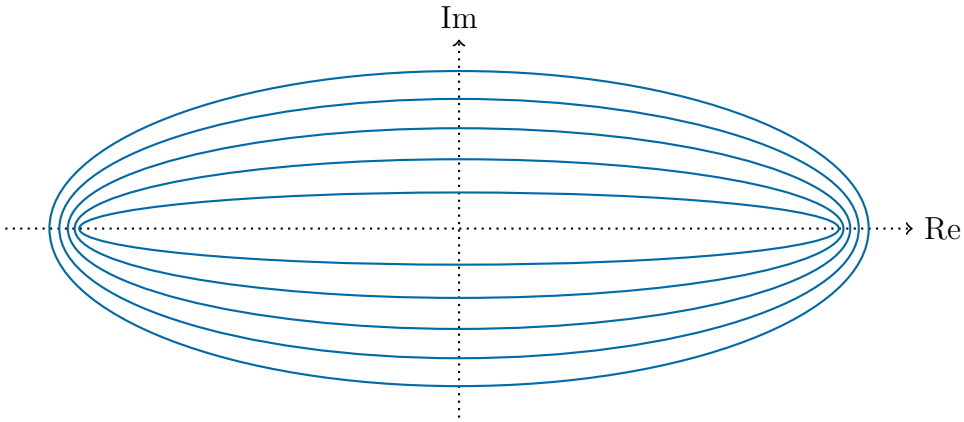


Figure 2.1: Bernstein ellipses $\partial\mathcal{E}_\rho$ for $\rho = 1.1, 1.2, \ldots, 1.5$.

Theorem 2.4.5. *If f is analytic in $[-1,1]$ and analytically continuable to the open Bernstein ellipse $\mathcal{E}_\rho$, where it satisfies $|f(t)| \leq M$ for some M, then for each $m \geq 0$ its Chebyshev projection satisfies*

$$\|f - f_m\|_\infty \leq \frac{2M\rho^{-m}}{\rho - 1}.$$

2.4.2 Interpolation by Trigonometric Splines

Let $f \in \mathcal{C}[a,b]$ be fixed. In this section we do not consider the best possible approximation of f as in Section 2.4.1, but a sufficiently good one. Therefore, we consider the *interpolation problem*:

$$u_r \in U_r : \quad u_r(t_i) = f(t_i) \qquad \forall i = 0, \ldots, r, \tag{2.23}$$

where $U_r \subset \mathcal{C}[a,b]$ for $r = 1, 2, \ldots$ are subspaces of ansatz functions. Nodes

$$\Omega_r = \{t_0, \ldots, t_r : a \leq t_0 < t_1 < \cdots < t_r \leq b\} \tag{2.24}$$

have to be chosen suitably. The choice of the nodes is very important since the interpolation error can be arbitrarily large as it can be seen by the next example.

Example 2.4.6 (Runge phenomenon, e.g. [Tre13])**.** *Interpolating the Runge function $f(x) = \frac{1}{1+25x^2}$ at equidistant points t_i between -1 and 1 such that: $t_i = \frac{2i}{m} - 1$ for $i \in \{0, 1, \ldots, m\}$ with a polynomial of degree at most m, i.e., $p_m(x) \in \mathscr{P}_m[-1,1]$, yields*

$$\lim_{m\to\infty} \left(\max_{-1\leq x\leq 1} |f(x) - p_m(x)| \right) = +\infty.$$

Interpolating the Runge function $f(x) = \frac{1}{1+25x^2}$ between -1 and 1 with a polynomial of degree at most m, i.e., $p_m(x) \in \mathscr{P}_m[-1,1]$, at the Chebyshev nodes t_i, which are the roots of T_m (Chebyshev polynomial of the first kind), yields

$$\lim_{m\to\infty} \left(\max_{-1\leq x\leq 1} |f(x) - p_m(x)| \right) = 0.$$

By the Runge phenomenon we have seen that interpolation at equidistant points can be arbitrarily bad, but it even holds for any sequence of nodes (2.24), that there exists an $f \in \mathcal{C}[a,b]$ such that $\liminf_{m\to\infty} \|f - p_m\| = \infty$ [Fab14].
Instead of polynomial interpolation, we consider trigonometric interpolation, and hence U_r in (2.23) is chosen to be the space of trigonometric functions, i.e.,

$$U_r = \{1, \cos(jt), \sin(jt), j = 1, \ldots, r\}.$$

Trigonometric functions are periodic, which is a sought property for the derivation of bounds for time-periodic systems in Section 6.2. Due to the linear independence of the trigonometric function $1, \cos(t), \sin(t), \ldots, \cos(rt), \sin(rt)$, we obtain a unique solution $u_r \in U_r$ to the interpolation problem (2.23) for any $f \in \mathcal{C}[a,b]$, e.g. in [DH08]. But in

order to avoid Runge type phenomena we want to consider *piecewise* ansatz functions, which in the interpolation setting are called *splines*. Here, we consider trigonometric splines. By (2.23) the spline is not uniquely determined, but we have some freedom to choose the spline, which is used by adding more regularity to the spline w.r.t. the nodes Ω_r in (2.24). Hence, we further demand that the spline fulfills

$$u_r \in U_r : \quad u_r^{(j)}(t_i) = f^{(j)}(t_i) \qquad \forall i = 0, \dots, r, j = 1, \dots, k, \tag{2.25}$$

where k is the order of the spline.

For brevity we choose in the following $r+1$ equidistant nodes in Ω_r defined in (2.24). Therefore, the nodes t_i are given as $t_i = ih$ for $i = 0, 1, \dots, r$ with $h = \frac{T}{r}$ in the interval $[0, T]$. Trigonometric splines have to fulfill some restriction on the step length h, which in the case of quadratic and cubic splines will be stated explicitly in following sections.

A convenient way to express trigonometric splines is to use trigonometric basis splines (abbreviated as trigonometric B-splines), which will be used in the following. A basis spline is a spline function that has minimal support, see Figures 2.2 and 2.3. A recursive definition of trigonometric splines is given in [LW79], but if the order k of the spline is small, it is easier to use the exact formulas given in equation (2.26) for quadratic splines and in equation (2.27) for cubic splines.

Quadratic Trigonometric Splines

Quadratic trigonometric splines $S_i^2(t)$ are defined by

$$S_i^2(t) = \theta_2 \begin{cases} \sin^2\left(\frac{t-t_i}{2}\right) & \text{if } t \in [t_i, t_{i+1}), \\ \sin\left(\frac{t-t_i}{2}\right)\sin\left(\frac{t_{i+2}-t}{2}\right) + \sin\left(\frac{t_{i+3}-t}{2}\right)\sin\left(\frac{t-t_{i+1}}{2}\right) & \text{if } t \in [t_{i+1}, t_{i+2}), \\ \sin^2\left(\frac{t_{i+3}-t}{2}\right) & \text{if } t \in [t_{i+2}, t_{i+3}), \\ 0 & \text{if } t \notin [t_i, t_{i+3}], \end{cases} \tag{2.26}$$

with $\theta_2 = \frac{1}{\sin(h)\sin\left(\frac{h}{2}\right)}$, see e.g. [Nik93; Nik04; Sch81]. A quadratic trigonometric spline $S_i(t)$ is shown in Figure 2.2. First, as it can be seen in Figure 2.2, for any inner subinterval $[t_i, t_{i+1}]$ with $1 < i \leq r$, the spline $S_i(t)$ is fully described. For the intervals $[t_0, t_1]$ and $[t_1, t_2]$, artificial intervals $[t_{-2}, t_{-1}]$ and $[t_{-1}, t_0]$ have to be included in the definition of $S_i^2(t)$ such that the restriction to the respective subinterval is still a linear combination of the functions $1, \cos(t)$ and $\sin(t)$. If we denote by $\mathcal{T}^2(\Omega_r)$ the space of quadratic trigonometric splines in $[0, T]$ w.r.t. the nodes Ω_r, then $\mathcal{T}^2(\Omega_r) = \operatorname{span}\{S_i^2\}_{i=-2}^{r-1}$. Hence, every quadratic trigonometric spline can be expressed in the form $\sum_{i=-2}^{r-1} \alpha_i S_i^2(t)$. For representing a quadratic trigonometric spline, the summation index i runs from -2 to $r-1$, which does not represent the number of nodes, but the number of intervals $[t_i, t_{i+1}]$ for $i = -2, \dots, r-1$ which includes the aforementioned artificial intervals. The coefficients α_i are unknown and have to be determined. The steplength h has to be chosen sufficiently small, i.e., $h < \frac{2\pi}{3}$ for quadratic trigonometric splines.

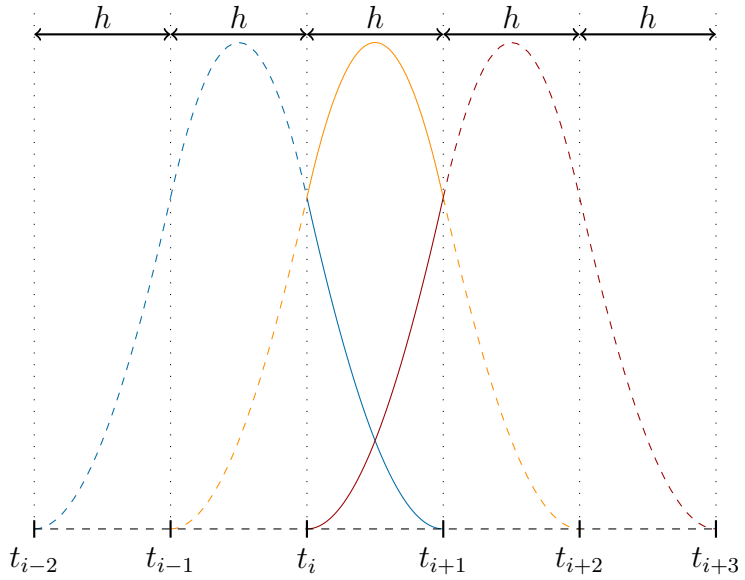


Figure 2.2: Quadratic trigonometric splines at equidistant nodes.

Cubic Trigonometric Splines

Cubic trigonometric splines $S_i^3(t)$ are defined by

$$S_i^3(t) = \theta_3 \begin{cases} \sin^3\left(\frac{t-t_i}{2}\right) & \text{if } t \in [t_i, t_{i+1}), \\ \sin^2\left(\frac{t-t_i}{2}\right)\sin\left(\frac{t_{i+2}-t}{2}\right) + \sin^2\left(\frac{t-t_{i+1}}{2}\right)\sin\left(\frac{t_{i+4}-t}{2}\right) & \\ \quad + \sin\left(\frac{t-t_i}{2}\right)\sin\left(\frac{t_{i+3}-t}{2}\right)\sin\left(\frac{t-t_{i+1}}{2}\right) & \text{if } t \in [t_{i+1}, t_{i+2}), \\ \sin^2\left(\frac{t_{i+3}-t}{2}\right)\sin\left(\frac{t-t_i}{2}\right) + \sin^2\left(\frac{t_{i+4}-t}{2}\right)\sin\left(\frac{t-t_{i+2}}{2}\right) & \\ \quad + \sin\left(\frac{t_{i+4}-t}{2}\right)\sin\left(\frac{t-t_{i+1}}{2}\right)\sin\left(\frac{t_{i+3}-t}{2}\right) & \text{if } t \in [t_{i+2}, t_{i+3}), \\ \sin^3\left(\frac{t_{i+4}-t}{2}\right) & \text{if } t \in [t_{i+3}, t_{i+4}], \\ 0 & \text{otherwise,} \end{cases} \tag{2.27}$$

where $\theta_3 = \frac{1}{\sin\left(\frac{3h}{2}\right)\sin(h)\sin\left(\frac{h}{2}\right)}$, see e.g. in [Sch81]. Artificial intervals $[t_{-3}, t_{-2}]$, $[t_{-2}, t_{-1}]$ and $[t_{-1}, t_0]$ have to be included in the definition of $S_i^3(t)$ such that the restriction to the respective subinterval is still a linear combination of the functions $\sin\left(\frac{t}{2}\right)$, $\cos\left(\frac{t}{2}\right)$, $\sin\left(\frac{3t}{2}\right)$, $\cos\left(\frac{3t}{2}\right)$ as in the previous section. If we denote by $\mathcal{T}^3(\Omega_r)$ the space of cubic trigonometric splines in $[0, T]$ w.r.t. the nodes Ω_r, then $\mathcal{T}^3(\Omega_r) = \operatorname{span}\{S_i^3\}_{i=-3}^{r-1}$. Hence, every cubic trigonometric spline can be expressed in the form $\sum_{i=-3}^{r-1} \alpha_i S_i^3(t)$. For representing a cubic trigonometric spline, the summation index i runs from -3 to $r-1$ which does not represent the number of nodes, but the number of intervals $[t_i, t_{i+1}]$ for $i = -3, \ldots, r-1$ which includes the aforementioned artificial intervals. The coefficients α_i are unknown and have to be determined. The steplength h has to be chosen sufficiently small, i.e., $h < \frac{\pi}{2}$ for cubic trigonometric splines.

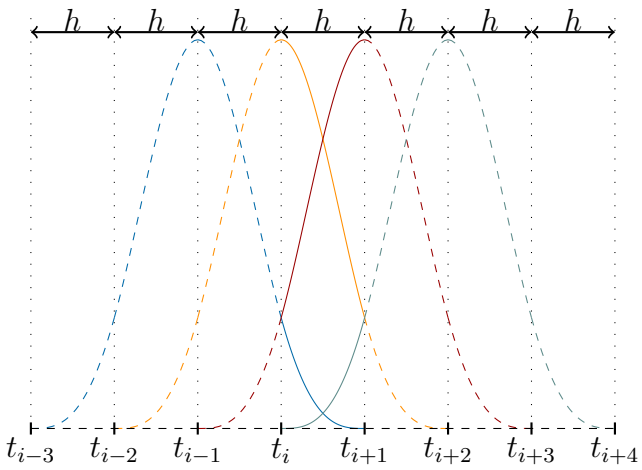


Figure 2.3: Cubic trigonometric splines at equidistant nodes.

2.5 Ordinary Differential Equations

Ordinary differential equations (ODEs) are well-established to describe scientific processes such as the motion of a particle, displacements in strings, bars, shafts, beams, plates or membranes. Here, we only give a very short introduction to them and cite only the most important results, which will be used throughout this thesis. For many more classes of ODEs and their theory we refer to [Wal90; CC97]. A very practical introduction to ODEs and their occurrence is given in [Col90].

Let $f : \Omega \subset \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R} \to \mathbb{R}$ be a function, which depends on the state x and its derivatives $x', x'', \ldots, x^{(k-1)}$. Then an equation of the form

$$x^{(k)} = f(t, x, x', \ldots, x^{(k-1)}), \tag{2.28}$$

is called an explicit ordinary differential equation (ODE). The highest derivative occurring in (2.28) is called the order of the ODE, i.e, the ODE (2.28) is of order k. We classify this ODE depending on the function f. If f can be written as a linear combination of the derivatives of x, then the ODE (2.28) is said to be a linear ODE and otherwise it is a nonlinear ODE. A function $x : [a, b] \subseteq \mathbb{R} \to \mathbb{R}$ is called a solution of (2.28), if x is k-times differentiable on I, and fulfills $x^{(k)} = f(t, x, x', \ldots, x^{(k-1)})$ for any $t \in I$. In the following we show existence and uniqueness of a solution, the solution or the solution's structure, respectively, when it is possible. Any ODE of order k in (2.28) can be reduced to a coupled system of k ODEs of first order [Col90]. Reducing the order of an ODE has a practical and theoretical impact. Firstly, solutions to higher order ODEs can be found with the order reduction, e.g. in Chapter 4 a quadratic system of ODEs is linearized to a first order system. Secondly, higher order ODEs can theoretically be analyzed since they can be rewritten as a system of first order ODEs, as e.g. in Section 2.5.2. In the following we restrict the theoretical analysis to first order ODEs since by the above considerations corresponding results for higher order ODEs can be drawn.

2.5.1 Nonlinear Ordinary Differential Equations

Any ordinary differential equation of the form (2.28) with arbitrary order can be reduced to a first order system of ODEs [Col90]. Hence, in this section we state existence and uniqueness of a solution to a first order system of ODEs

$$x' = f(t, x), \tag{2.29}$$

where $f : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{C}^n$ and Ω is an open set in $\mathbb{R} \times \mathbb{R}^n$. The ODE (2.29) with some initial condition $x(t_0) = x_0 \in \mathbb{C}^n$ is called *initial value problem (IVP)* or *Cauchy problem.*

Theorem 2.5.1 (local Picard-Lindelöf e.g. in [CL55])**.** *Let the initial value problem (IVP) be given as*

$$x' = f(t, x), \quad x(t_0) = x_0.$$

Suppose $f(t, x)$ is Lipschitz-continuous in x and continuous in t. Then, for a sufficiently small $\varepsilon > 0$ there exists a unique solution $x(t)$ to the initial value problem on $[t_0 - \varepsilon, t_0 + \varepsilon]$.

Gronwall's lemma [Gro19] can be used to prove uniqueness of a solution to an initial value problem in the Picard-Lindelöf Theorem 2.5.1. Here, we cite an integral version of Gronwall's lemma by R. Bellman [Bel43], which is given e.g. in [Wal70].

Lemma 2.5.2 (Gronwall's lemma, integral version e.g. in [Wal70])**.** *Let $g : [a, b] \mapsto \mathbb{R}$ and $\beta : [a, b] \mapsto \mathbb{R}$ be continuous, $\alpha : [a, b] \mapsto \mathbb{R}$ be integrable on $[a, b]$ and $\beta(t) \geq 0$. Assume g satisfies*

$$g(t) \leq \alpha(t) + \int_a^t \beta(s) g(s) \mathrm{d}s, \qquad t \in [a, b].$$

Then

$$g(t) \leq \alpha(t) + \int_a^t \alpha(s)\beta(s) \exp\left(\int_s^t \beta(r)\,\mathrm{d}r\right) \mathrm{d}s, \qquad \forall t \in [a, b].$$

Furthermore, if α is non-decreasing and $\beta > 0$ is constant, then

$$g(t) \leq \alpha(t) e^{\beta(t-a)}, \qquad \forall t \in [a, b].$$

We will use Lemma 2.5.2 to derive bounds on the solution of time-periodic systems in Section 6.3.

2.5.2 Linear Ordinary Differential Equations

In this section we consider linear ODEs, i.e., the function $f : \Omega \subset \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R} \to \mathbb{C}$ in (2.28) can be written as a linear combination of x and its derivatives. Hence, the ODE (2.28) can be written as

$$\mathcal{L}_k x = g, \tag{2.30}$$

where $\mathcal{L}_k$ is a linear differential operator of order k, i.e.,

$$\begin{aligned}\mathcal{L}_k x &= \frac{\mathrm{d}^k x}{\mathrm{d}t^k} + a_{k-1}(t)\frac{\mathrm{d}^{k-1}x}{\mathrm{d}t^{k-1}} + \cdots + a_1(t)\frac{\mathrm{d}x}{\mathrm{d}t} + a_0(t)x \\ &= \frac{\mathrm{d}^k x}{\mathrm{d}t^k} + \sum_{j=0}^{k-1} a_j(t)\frac{\mathrm{d}^j x}{\mathrm{d}t^j}. \end{aligned} \tag{2.31}$$

The coefficients $a_0, \ldots, a_{k-1}$ are allowed to depend on t, but they should be continuous, i.e., we assume that $a_j \in \mathcal{C}[a,b]$ for $j = 0, \ldots, k-1$. First, we introduce some basic notation for linear homogeneous differential equations (2.30). If $g = 0$ in (2.30), then the linear differential equation is called *homogeneous* and otherwise *inhomogeneous*. One of the main differences between linear and nonlinear homogeneous ODEs is that the solutions of a linear ODE form a vector space, see e.g. in [CC97].

Definition 2.5.3 (Null space of linear differential operator)**.** *The null space of a linear differential operator $\mathcal{L}_k$ is defined as*

$$N_{\mathcal{L}_k} := \left\{ x \in L_1^k[a,b] : \mathcal{L}_k x(t) = 0\,, t \in [a,b] \right\},$$

where the Sobolev space $L_1^k[a,b]$ is defined in (2.15)*.*

Definition 2.5.4 (Fundamental solution)**.** *A set of k linear independent functions $\{x_1(t), \ldots, x_k(t)\} \in L_1^{k-1}[a,b]$ that are solutions to the homogeneous linear differential equation $\mathcal{L}_k x = 0$ for $t \in [a,b]$ is called a fundamental solution.*

A fundamental solution spans the null space $N_{\mathcal{L}_k}$ of the linear differential operator $\mathcal{L}_k$. Linear independence of functions can be checked e.g. by the Wronskian in Remark 2.5.5

Remark 2.5.5. *Let $x_1(t), \ldots, x_k(t) \in \mathcal{C}^{k-1}[a,b]$. Then the Wronskian is defined as*

$$W(x_1, \ldots, x_k)(t) = \begin{vmatrix} x_1(t) & x_2(t) & \cdots & x_k(t) \\ x_1'(t) & x_2'(t) & \cdots & x_k'(t) \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(k-1)}(t) & x_2^{(k-1)}(t) & \cdots & x_k^{(k-1)}(x) \end{vmatrix}, \qquad t \in [a,b]. \tag{2.32}$$

$x_1(t), \ldots, x_k(t)$ are linear independent if and only if $W(x_1, \ldots, x_k)(t) \not\equiv 0$.

We have introduced a linear differential operator $\mathcal{L}$, its null space $N_{\mathcal{L}}$ and a fundamental solution of $\mathcal{L}$. Now we can cite basic results for trigonometric splines in Lemma 2.5.6 and Remark 2.5.7, which are needed in the derivation of two-sided bounds by trigonometric splines in Chapter 6.

Lemma 2.5.6 (e.g. [Sch81])**.** *Trigonometric splines are $\mathcal{L}$-splines, where the $\mathcal{L}$ corresponds to a certain linear differential operator.*

Remark 2.5.7.

- *Trigonometric splines are functions in $\mathcal{C}^{n-1}[a,b]$ so that the restriction of them in every subinterval $[t_i, t_{i+1}]$ is a linear combination of functions of the null space $N_{\mathcal{L}}$ (see Definition 2.5.3) of the corresponding linear differential operator $\mathcal{L}$, [Sch81].*

- *$\mathcal{L}$-splines and hence, trigonometric splines by Lemma 2.5.6, fulfill an extended Taylor formula [Sch81].*

We conclude this introduction to linear (homogeneous) differential equations by two examples which are used for trigonometric splines in Section 2.4.2 and 6.2.

Example 2.5.8. *Quadratic trigonometric splines defined in* (2.26) *are $\mathcal{L}$-splines by Lemma 2.5.6, where the linear differential operator is given as $\mathcal{L}_3 \cdot = \frac{\mathrm{d}^3}{\mathrm{d}t^3} \cdot + \frac{\mathrm{d}}{\mathrm{d}t} \cdot$. $x_1(t) = 1, x_2(t) = \cos(t)$ and $x_3(t) = \sin(t)$ are solutions to $\mathcal{L}_3 x = 0$. $\{x_1(t), x_2(t), x_3(t)\}$ is a fundamental solution to $\mathcal{L}_3 x = 0$, since*

$$W(x_1, x_2, x_3)(t) = \begin{vmatrix} 1 & \cos(t) & \sin(t) \\ 0 & -\sin(t) & \cos(t) \\ 0 & -\cos(t) & -\sin(t) \end{vmatrix} = \sin^2(t) + \cos^2(t) = 1 \not\equiv 0.$$

Example 2.5.9. *Cubic trigonometric splines defined in* (2.27) *are $\mathcal{L}$-splines by Lemma 2.5.6, where the linear differential operator is given as $\mathcal{L}_4 \cdot = \frac{\mathrm{d}^4}{\mathrm{d}t^4} \cdot + \frac{5}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2} \cdot + \frac{9}{16} \cdot$. $x_1(t) = \sin\left(\frac{t}{2}\right)$, $x_2(t) = \cos\left(\frac{t}{2}\right)$, $x_3(t) = \sin\left(\frac{3t}{2}\right)$ and $x_4(t) = \cos\left(\frac{3t}{2}\right)$ are solutions to $\mathcal{L}_4 x = 0$ for $t \in [0, T]$. $\{x_1(t), x_2(t), x_3(t), x_4(t)\}$ is a fundamental solution, since there exists $t \in [0, T]$ such that the Wronskian is nonzero, e.g.*

$$W(x_1, x_2, x_3, x_4)(0) = \begin{vmatrix} 0 & 1 & 0 & 1 \\ 0.5 & 0 & 1.5 & 0 \\ 0 & -0.25 & 0 & -2.25 \\ -0.125 & 0 & -3.375 & 0 \end{vmatrix} = 3 \neq 0.$$

More on splines can be found in [DB01], on trigonometric splines in [Sch81], on interpolation with trigonometric splines in [Sch64].
The linear ordinary differential equation (2.30) can be transformed into a first order system of ODEs

$$x' = A(t)x + b(t), \tag{2.33}$$

where

$$A(t) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0(t) & -a_1(t) & -a_2(t) & \cdots & -a_{k-1}(t) \end{bmatrix}, \quad b(t) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ g(t) \end{bmatrix}.$$

First, we want to cite the solution's structure for (2.33), which is given, e.g., [CC97]. Suppose x_p is a particular solution to the non-homogeneous first order ODE system (2.33) and x_h is the general solution to the homogeneous first order ODE system $x' = A(t)x$ for $t \in I$. Then $x(t) = x_\mathrm{h}(t) + x_\mathrm{p}(t)$ is the general solution to the non-homogeneous first order ODE system (2.33). Hence, it is worth to investigate homogeneous first order ODE systems, which is being done in the following section. For finding a particular solution to a non-homogeneous first order ODE systems we refer to e.g. [CL55; CC97].

2.5.3 First Order System of Ordinary Differential Equations

In this section we consider a homogeneous first order system of ODEs, which is also known as a homogeneous linear system of ordinary differential equations,

$$x' = A(t)x, \qquad \text{for } t \in I, \tag{2.34}$$

where $A : I \subset \mathbb{R} \to \mathbb{C}^{n\times n}$, $I = [t_0, t_{\text{end}}]$ and an initial condition (IC) $x(t_0) = x_0$ are given. The question of existence and uniqueness of a solution to nonlinear ordinary differential equations has already been answered by Theorem 2.5.1, where $f(t, x)$ has to be Lipschitz-continuous in x and continuous in t. For a homogeneous linear ODE of first order, the function $f(t, x)$ in (2.29) can be expressed as $f(t, x) = A(t)x$. If the function $A : I \subset \mathbb{R} \to \mathbb{C}^{n\times n}$ is continuous, then obviously $f(t, x) = A(t)x$ is Lipschitz-continuous in x. Hence, we summarize in Theorem 2.5.10 the conditions of existence and uniqueness of a solution to a linear ordinary differential equation. Theorem 2.5.10 turns out to be a global existence and uniqueness result, see e.g. [CC97].

Theorem 2.5.10 ([CC97])**.** *Let $A : I \subset \mathbb{R} \to \mathbb{C}^{n\times n}$ be continuous. For any initial condition $x(t_0) = x_0$ there exists a unique solution $x(t)$ of the ODE* (2.34)*.*

We introduce the fundamental matrix for the first order ODE system (2.34), which will be used for definition of the solution to the ODE (2.34) and later on in this section for the solution's structure for time-periodic matrix functions $A : I \subset \mathbb{R} \to \mathbb{C}^{n\times n}$.

Definition 2.5.11 (Fundamental matrix)**.** *Let $\{x_1, \ldots, x_n\}$ a fundamental solution to* (2.34)*. Then*

$$\Phi(t) := \big[x_1(t), \cdots, x_n(t)\big] \in \mathbb{C}^{n\times n}$$

is called a fundamental matrix of the first order ODE system (2.34)*.*

$\Phi(t)$ is called principal fundamental matrix if $\Phi(t)$ is a fundamental matrix and $\Phi(t_0) = I$. A principal fundamental matrix can be constructed from a fundamental matrix $\Phi(t)$ by $\Phi(t)\Phi^{-1}(t_0)$. Therefore, from now on when we use the term fundamental matrix, we mean in general its principal fundamental matrix. We summarize the following properties of a fundamental matrix in the following remark, which is given e.g. in [MS76]:

Remark 2.5.12.

- *The solution to the first order ODE system* (2.34) *with $x(t_0) = x_0$ is given as*

$$x(t) = \Phi(t)x_0. \tag{2.35}$$

- *The fundamental matrix $\Phi(t)$ solves the matrix-valued first order system*

$$\Phi'(t) = A(t)\Phi(t). \tag{2.36}$$

- *A fundamental matrix fulfills*

$$\Phi(t_2) = \Phi(t_2)\Phi(t_1) \tag{2.37}$$

  for $t_1, t_2 \geq t_0$.

- *Liouville's formula:*

$$\det \Phi(t) = \exp \int_{t_0}^{t} \operatorname{tr}(A(\tau))\mathrm{d}\tau, \tag{2.38}$$

where $\det \Phi(t)$ *is called Wronskian, see Remark 2.5.5.*

In (2.35) we have defined the solution to the ODE (2.34). Now, we want to know the conditions when a solution is analytic. The answer is given by the following theorem.

Theorem 2.5.13 (e.g. in [CC97])**.** *Let* $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ *be analytic at* $\tau \in \mathbb{R}$*, where* $\rho > 0$ *is its radius of convergence, i.e.,*

$$A(t) = \sum_{k=0}^{\infty} A_k (t-\tau)^k,$$

where $|t-\tau| < \rho$ *and* $A_k \in \mathbb{C}^{n\times n}$ *for* $k = 0, 1, \ldots$*. Given any* x_0*, there exist an analytic solution* x *of* (2.34) *with the same convergence radius* $\rho > 0$ *satisfying* $x(\tau) = x_0$*, i.e.,* x *has a power series representation*

$$x(t) = \sum_{k=0}^{\infty} c_k (t-\tau)^k,$$

where $|t-\tau| < \rho$ *and* $c_k \in \mathbb{C}^n$ *for* $k = 0, 1, \ldots$*.*

We give some remarks for the power series representation of the analytic solution x of (2.34).

Remark 2.5.14.

- *The matrix function* $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ *is continuous since it is analytic and hence, by Theorem 2.5.10 the first order ODE system* (2.34) *has a solution, which is unique.*
- *We have* $c_0 = x_0$ *and* $c_k \in \mathbb{C}^n$ *can be determined uniquely in terms of* c_0 *by substituting the series into the first order ODE system* (2.34)*.*

We want to conclude the investigation on a general time-variant matrix by the following assumption. If the matrix function $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ commutes for all times, i.e.,

$$A(t_1)A(t_2) = A(t_2)A(t_1)$$

for any $t_1, t_2 \geq t_0$. Then the solution $x(t)$ of the first order ODE system (2.34) with the initial condition $x(t_0) = x_0$ is given as

$$x(t) = e^{\int_{t_0}^{t} A(s)\,\mathrm{d}s} x_0.$$

In the following we distinguish two different cases for the matrix function $A : \mathbb{R} \to \mathbb{C}^{n\times n}$:

1. A is time-invariant, i.e., it is constant over time: $A \equiv const$ and

2. A is time-periodic, i.e., it has some periodicity $t_p > 0$ such that $A(t) = A(t + t_p)$ for any $t \geq t_0$,

We show in the following the solution or the solution's structure for first order ODE system (2.34), where A is time-invariant and time-periodic, respectively.

Time-Invariant

We start the investigation on the first order ODE system (2.34), where the matrix function $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ is time-invariant, i.e., $A := A(t) \equiv const$,

$$x' = Ax, \qquad \text{for } t \in I, \tag{2.39}$$

where $A \in \mathbb{C}^{n\times n}$ and $x(t_0) = x_0 \in \mathbb{C}^n$. We follow the presentation of a flow operator, e.g. given in [BD08].

Theorem 2.5.15 ([BD08])**.** *The flow operator Φ^t of (2.39) is given by*

$$\Phi(t) = e^{(t-t_0)A}. \tag{2.40}$$

The series

$$e^{tA} := \sum_{k=0}^{\infty} \frac{1}{k!}(tA)^k \tag{2.41}$$

converges uniformly on finite time intervals $[a, b]$ and the solution of the IVP (2.39) is given by

$$x(t) = \Phi(t)x_0 = e^{(t-t_0)A}x_0, \tag{2.42}$$

e.g. in [Wal90; CC97].

Remark 2.5.16. *The matrix exponential (2.41) has the following properties:*

$$\begin{aligned}
&(i)\ e^{t(TAT^{-1})} = Te^{tA}T^{-1}\,,\ \forall T \in Gl_n(\mathbb{C}),\\
&(ii)\ e^{t(A+B)} = e^{tA}e^{tB}\,,\ \forall B \in \mathbb{C}^{n\times n} \text{ with } AB = BA,\\
&(iii)\ A = \text{blockdiag}(A_1, \ldots, A_m) \Rightarrow e^{tA} = \text{blockdiag}(e^{tA_1}, \ldots, e^{tA_m}),\\
&(iv)\ e^{\alpha I} = e^{\alpha} I\,,\ \alpha \in \mathbb{C}.
\end{aligned}$$

By Theorem 2.1.4, we can decompose the matrix A in Jordan Normal Form, i.e., $UAU^{-1} = J$, where J consists of the Jordan blocks. By using the property (i) of Remark 2.5.16 for the matrix exponential, we obtain $Ue^{At}U^{-1} = e^{Jt}$ and obtain the following stability result which e.g. is given in [CC97].

Remark 2.5.17. *Let $\Lambda(A)$ denote the spectrum of A.*

- *If $\text{Re}(\lambda) < 0$ for all $\lambda \in \Lambda(A)$, then the solution $x(t)$ of (2.39) is called asymptotically stable, i.e., $\|x(t)\| \to 0$ as $t \to \infty$ for any $x_0 \in \mathbb{C}^n$.*
- *If $\text{Re}(\lambda) \leq 0$ for all $\lambda \in \Lambda(A)$ and the algebraic multiplicity and geometric multiplicity of the eigenvalue λ coincide for all $\lambda \in \Lambda(A)$ with $\text{Re}(\lambda) = 0$, then the solution $x(t)$ of (2.39) is called stable, i.e., for $x_0 \in \mathbb{C}^n$ there exists $C > 0$ such that $\|x(t)\| \leq C$ as $t \to \infty$.*

- *If* $\mathrm{Re}(\lambda_i) > 0$ *for any* $i = 1, \ldots, n$*, then the solution* $x(t)$ *of* (2.39) *is called unstable, i.e., there exists* $x_0 \in \mathbb{C}^n$ *such that* $\|x(t)\| \to \infty$ *as* $t \to \infty$*.*

Time-Periodic

Now, we assume that the matrix function $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ is time-periodic, i.e., there exists some $t_p > 0$ such that $A(t) = A(t + t_p)$ for any $t \in \mathbb{R}$. The time-periodic linear initial value problem (IVP) is then given as

$$\begin{array}{lcll} x' & = & A(t)x, & \text{for } t \in I, \\ A(t) & = & A(t+t_p), & \text{for } t \in \mathbb{R}, \\ x(t_0) & = & x_0, & \end{array} \tag{2.43}$$

where $x(t_0) = x_0 \in \mathbb{C}^n$ denotes the initial condition (IC) and the matrix function $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ is periodic with periodicity $t_p > 0$, i.e., $A(t) = A(t + t_p)$ for any $t \in \mathbb{R}$. The following theorem was initially given for a single time-periodic ordinary differential equation in [Flo83], here we cite the matrix version from e.g. [MS76].

Theorem 2.5.18 (Floquet's Theorem [Flo83])**.** *Let* $\Phi(t)$ *be a fundamental matrix of* (2.43)*. Then for all* $t \in \mathbb{R}$*:*

$$\Phi(t + t_p) = \Phi(t)C, \tag{2.44}$$

where $C = \Phi(t_0 + t_p) \in Gl_n(\mathbb{C})$ *is a constant nonsingular matrix. In addition, for a matrix* L *such that*

$$e^{Lt_p} = \Phi(t_0 + t_p), \tag{2.45}$$

there is a periodic matrix function $t \mapsto Z(t)$ *such that*

$$\Phi(t) = Z(t)e^{L(t-t_0)}, \quad \forall t \geq t_0. \tag{2.46}$$

We give some remarks about Floquet's theory.

Remark 2.5.19.

- L *is in general not unique. It can be given by as* $L = \frac{1}{t_p} \ln \Phi(t_0 + t_p)$*, where* $\Phi(t_0 + t_p)$ *is the monodromy matrix.*
- *The eigenvalues of* e^{Lt_p} *are called the characteristic multipliers of the system* (2.43)*.*
- $\mu \in \mathbb{C}$ *is called Floquet exponent if* $e^{\mu t_p}$ *is a characteristic multiplier of the system* (2.43)*. Floquet exponents are not unique, since* $e^{\mu t_p + 2\pi i k}$ *is a Floquet exponent as well for* $k \in \mathbb{Z}$*.*
- *Let* $\mu \in \mathbb{C}$ *be a Floquet exponent. Then its real part, i.e.,* $\mathrm{Re}(\mu)$*, is called Lyapunov exponent.*
- *The solution of the IVP* (2.43) *can be given in terms of Floquet's Theorem 2.5.18 as*

$$x(t) = \Phi(t)x_0 = Z(t)e^{L(t-t_0)}x_0.$$

- *$Z(t)$ has full rank for all $t \geq t_0$, i.e., $Z(t) \in Gl_n(\mathbb{C})$ for all $t \geq t_0$ since equation (2.46) yields $Z(t) = \Phi(t)e^{-L(t-t_0)}$, where the columns of the fundamental matrix $\Phi(t)$ are per definition linearly independent and therefore $\Phi(t)$ has full rank.*

- *$Z(t)$ gives a coordinate transformation $x(t) = Z(t)y(t)$, the so-called Floquet-Lyapunov transformation, which transforms (2.43) into a constant linear system $\dot{y} = Ly$ and initial condition $y(t_0) = x_0$. The solution to $\dot{y} = Ly$ has been given in Section 2.5.2 as $y(t) = e^{L(t-t_0)}x_0$.*

- *If all Lyapunov exponents are negative, then the solution $x(t)$ is asymptotically stable. This result is obtained by the coordinate transformation and Remark 2.5.17.*

- *Knowledge of $\Phi(t)$ for all $t \in [t_0, t_0 + t_p]$ is sufficient for the knowledge of $\Phi(t)$ for any $t \in \mathbb{R}$, due to the semigroup property of the solution:*
  1. *determine L as $L = \frac{1}{t_p} \ln \Phi(t_0 + t_p)$ by the first remark,*
  2. *determine $Z(t)$ in $[t_0, t_0 + t_p]$ by equation (2.46),*
  3. *$Z(t)$ is known for any $t \in \mathbb{R}$ since it is periodic.*

2.5.4 Second Order System of Ordinary Differential Equations

Now, we consider a second order system of ordinary differential equations, where the coefficient matrices are time-invariant, i.e., we consider the initial value problem (IVP)

$$A\frac{\mathrm{d}^2x}{\mathrm{d}t^2} + B\frac{\mathrm{d}x}{\mathrm{d}t} + Cx = g(t), \tag{2.47}$$

where $A, B, C \in \mathbb{C}^{n\times n}$ and an initial condition (IC) $x(t_0) = x_0$ and $x'(t_0) = x_0'$ is given. A first order linear system of dimension $2n$ can be obtained from (2.47) by

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} N & 0 \\ 0 & A \end{bmatrix}\begin{bmatrix} x(t) \\ x'(t) \end{bmatrix} = \begin{bmatrix} 0 & N \\ -C & -B \end{bmatrix}\begin{bmatrix} x(t) \\ x'(t) \end{bmatrix}, \tag{2.48}$$

for any $N \in Gl_n(\mathbb{C})$.

Remark 2.5.20. *With the separation hypothesis $x(t) = e^{\lambda t}v$ in (2.47) we obtain the corresponding quadratic eigenproblem 2.1.8. With the ansatz $x(t) = e^{\lambda t}v$ in (2.48) we obtain the linearization 2.1.9 of the quadratic eigenproblem.*

By Remark 2.5.20 there is a close connection between a quadratic ordinary differential equation and a corresponding quadratic eigenproblem. (2.48) is in fact a first order linear system of ODEs with time-invariant coefficients which is treated in Section 2.5.2. Its solution is given by the flow operator in Theorem 2.5.15 with the solution (2.42):

$$x(t) = \Phi(t)\begin{bmatrix} x_0 \\ x_0' \end{bmatrix} = e^{(t-t_0)L}\begin{bmatrix} x_0 \\ x_0' \end{bmatrix},$$

where L is given as

$$L = \begin{bmatrix} N^{-1} & 0 \\ 0 & A^{-1} \end{bmatrix} \begin{bmatrix} 0 & N \\ -C & -B \end{bmatrix} = \begin{bmatrix} 0 & I \\ -A^{-1}C & -A^{-1}B \end{bmatrix},$$

and $N \in Gl_n(\mathbb{R})$. By Theorem 2.1.4 the matrix L can be decomposed into Jordan normal form and with the properties of the matrix exponential (2.5.16), we obtain that the general solution of the quadratic ordinary differential equation (2.47) can be expressed by Jordan pairs or more precisely as

$$x(t) = Ve^{Jt}c, \tag{2.49}$$

where (V, J) is a Jordan pair [GLR09; LT85] and $c \in \mathbb{C}^{2n}$ is a vector of constants. Here, the Jordan pair consist of the Jordan matrix J of size $2n \times 2n$ that contains the eigenvalues and their multiplicities such that $J = \operatorname{diag}(J_1, \ldots, J_m)$, where each $J_i \in \mathbb{C}^{n_i \times n_i}$ is a Jordan block and $\sum_{i=1}^{m} n_i = 2n$. The matrix $V = [v_1, \ldots, v_{2n}]$ is of size $n \times 2n$ and contains the corresponding Jordan chains v_i for $i = 1, \ldots, 2n$. If the eigenvalues are semi-simple, then the set of Jordan chains and the set of eigenvectors span the same space. The vector of constants c can be determined by a given initial conditions $x(t_0) = x_0$ and $x'(t_0) = x'_0$ for the quadratic ordinary differential equation (2.47).

2.6 Optimization Problems

In this section we introduce the most basic optimization framework, which is needed throughout this thesis. We do not give any kind of introduction here but only classify problems, namely linear and nonlinear with continuous and discrete variables. For more on the theory we list some books and survey papers but this list is not at all exhaustive. For the theory of linear and integer programming we refer to [BT97; Sch86] and for more details on linear and nonlinear programming to [LY15]. [Bel+13] is a good survey on nonlinear integer programming and a comparison of the respective solvers can be found in [BV10]. In this context the pioneering work of [Coo71; Kar72] on complexity classes and their reduction should be mentioned. But now let us start defining some optimization problems that are used in this thesis.

A Nonlinear Program (NLP) is defined as follows,

$$\begin{array}{lll} \min & f(x) & \\ \text{s.t.} & g_i(x) \leq 0 & \forall i \in I, \\ & x \in X, & \end{array} \tag{2.50}$$

where at least the objective function (cost function) $f(\cdot)$ or some constraint functions $g_i(x)$ are nonlinear for some $i \in I$, where I is some given index set and $x \in X \subseteq \mathbb{R}^{n_x}$ are continuous variables for a bounded polyhedral set X.

If some of the variables have to fulfill some integer constraints, then we refer to a Mixed

Integer Nonlinear Program (MINLP),

$$\begin{array}{lll} \min & f(x,y) & \\ \text{s.t.} & g_i(x,y) \leq 0 & \forall i \in I, \\ & x \in X, & \\ & y \in \mathbb{Z}^{n_y}, & \end{array} \tag{2.51}$$

where $X \subseteq \mathbb{R}^{n_x}$ is a bounded polyhedral set. Here, we specify that the variables $y \in \mathbb{Z}^{n_y}$ are integers. This class includes Binary Nonlinear Programs, where the integer variables are binaries, i.e., $y \in \{0,1\}^{n_y}$. Problem (2.51) is an NP-hard combinatorial problem, since it includes a Mixed Integer Linear Program (MILP), see e.g. [KM78]. It is even worse, non-convex integer optimization problems are in general undecidable [Jer73]. An example of a quadratically constrained integer program is shown in [Jer73] such that no computing device can compute the optimum for all problems in this class. In the remainder of this thesis, we concentrate on the case where (2.51) is decidable, which we can achieve either by ensuring that the set X is compact or by assuming that the problem functions are convex.
If the objective and all constraint functions of (2.50) are linear, then we refer to a Linear Program (LP),

$$\begin{array}{lll} \min & c^T x & \\ \text{s.t.} & a_i x \leq b_i & \forall i \in I, \\ & x \in X, & \end{array} \tag{2.52}$$

where I is some given index set and $X \subseteq \mathbb{R}^{n_x}$. Linear Programs can be solved efficiently (in polynomial time) e.g. by the ellipsoidal method or interior point methods [BT97]. As above, if some of the variables have to fulfill some integer constraints, then we refer to a Mixed Integer Linear Program (MILP),

$$\begin{array}{lll} \min & c_1^T x + c_2^T y & \\ \text{s.t.} & a_i^T x + b_i^T y \leq d_i & \forall i \in I, \\ & x \in X, & \\ & y \in \mathbb{Z}^{n_y}, & \end{array} \tag{2.53}$$

where $X \subseteq \mathbb{R}^{n_x}$ and I is some given index set. The variables $y \in \mathbb{Z}^{n_y}$ in (2.53) are specified as integers. Integer programming is in general NP-hard, e.g. in [Sch86].

3

Vibrations and Norms

For a time-varying process, a repetitive motion of some measure about a central value (e.g. an equilibrium point) or between two or more different states is called *oscillation*. For a mechanical oscillation we use instead the term *vibration*, i.e., a vibration underlies a mechanical system. Vibrations occur in many systems and familiar examples of vibration include a swinging pendulum or the random motion of a building due to gusts of wind. Further examples are given in Chapter 1. Vibrations are often an unwanted behavior for the underlying system since they may produce friction and heat, which results in material stress and fatigue. Hence, it is important to measure and reduce vibrations.
First, we introduce the notion of a vibration in Section 3.1. In Section 3.2 we investigate the differentiability of a general time-varying function in certain norms. We apply these results to vibrations in Section 3.3 and connect them to the algebraic Lyapunov equation in order to obtain a norm in which the vibration behaves monotonic. Finally, in Section 3.4 we introduce the vibration reduction problems for viscous dampers, which we will consider in this thesis. Overall, this chapter provides a basis for vibration reduction by viscous dampers in Chapter 4 and 5 and for the time behavior of vibrations in certain norms, which are generalized in Chapter 6.

3.1 Vibrations

Let us start this section with the definition of a vibrational system.

Definition 3.1.1. *Let $M, C, K \in \mathbb{R}^{n\times n}$ be symmetric and real $n \times n$ matrices. Furthermore, let M, K be positive definite, i.e., $M, K > 0$, the pencil (M, K) be regular and C be positive semidefinite, i.e., $C \geq 0$. Then the second order ordinary differential equation*

$$Mx'' + Cx' + Kx = g(t) \tag{3.1}$$

is called a vibrational system. Here, M, C and K are called mass, damping and stiffness matrix, respectively, and $g(t)$ is a time-dependent external force vector.

As described in Section 2.5.4, the homogeneous vibrational system (3.1), where $g = 0$,

corresponds to a quadratic eigenproblem (QEP),

$$Q(\lambda)v = \left(M\lambda^2 + C\lambda + K\right) v = 0 \tag{3.2}$$

for eigenvalues λ and eigenvectors v. Let (λ_k, v_k) for $k = 1, \ldots, 2n$ be eigenpairs of the QEP (3.2). In the following we assume for notational simplicity that the eigenvalues λ_k for $k = 1, \ldots, 2n$ are non-defective. The case for defective eigenvalues is considered in Section 2.5.4. In general λ_k and v_k of the QEP (3.2) are complex, i.e., $\lambda_k = d_k + i\omega_k$, where $d_k, \omega_k \in \mathbb{R}$. Then $\{x_1(t), \ldots, x_{2n}(t)\}$ is a fundamental solution to the homogeneous vibrating system $Mx'' + Cx' + Kx = 0$, where

$$x_k(t) = e^{\lambda_k t} = e^{d_k t} \left(\cos(t\omega_k)\,\mathrm{Re}(v_k) - \sin(t\omega_k)\,\mathrm{Im}(v_k)\right) \tag{3.3}$$

is a basic solution to the vibrational system for $k = 1, \ldots, 2n$. $x_k(t)$ describes how the system vibrates between the two configurations given as $\mathrm{Re}(v_k)$ and $\mathrm{Im}(v_k)$. Furthermore, it shows that d_k and ω_k correspond to damping and circular frequency, respectively. Since the solutions of linear homogeneous differential equations form a vector space, the general solution to $Mx'' + Cx' + Kx = 0$ is obtained as

$$x(t) = \sum_{k=1}^{2n} c_k e^{\lambda_k t} = \sum_{k=1}^{2n} c_k e^{d_k t} \left(\cos(t\omega_k)\,\mathrm{Re}(v_k) - \sin(t\omega_k)\,\mathrm{Im}(v_k)\right), \tag{3.4}$$

where $c_k \in \mathbb{R}$ for $k = 1, \ldots, 2n$. If there exists $\omega_k \neq 0$, where $k = 1, \ldots, 2n$, the vibrational system has one or more frequencies that it vibrates at once it has been disturbed. When the forcing frequency of $g(t)$ is close to a frequency ω_k, where $k = 1, \ldots, 2n$, the amplitude of the vibration may get extremely high. This phenomenon is called resonance and often it is an unwanted and very harmful behavior since it leads to material stress.

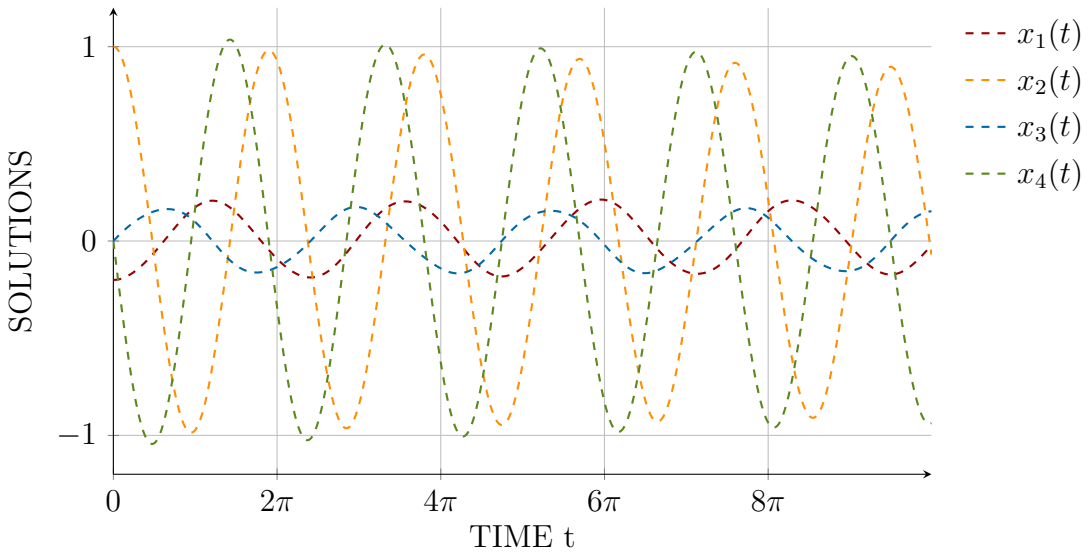


Figure 3.1: Solutions for a Jeffcott rotor on an anisotropic shaft for $t \in [0, 10\pi]$.

A basic solution x_k in (3.3) is called asymptotically stable if $x_k(t) \to 0$ as $t \to \infty$. $x_k(t)$

is asymptotically stable if and only if $d_k < 0$, i.e., $\lambda_k \in \mathbb{C}_-$, compare Remark 2.5.17. A basic solution x_k in (3.3) is called stable if there exists $C > 0$ such that $|x(t)| \leq C$ for any $t \in \mathbb{R}$. $x(t)$ is stable and not asymptotically stable if and only if $d_k = 0$, i.e., $\lambda_k = i\omega_k$ is purely imaginary, compare Remark 2.5.17. Then the basic solution x_k is given as

$$x_k(t) = \cos(t\omega_k)\,\mathrm{Re}(v_k) - \sin(t\omega_k)\,\mathrm{Im}(v_k), \tag{3.5}$$

which has no damping part and theoretically oscillates forever. Obviously, asymptotic stability implies stability, i.e., if $x_k(t)$ is asymptotically stable then it is stable as well. A general solution $x(t)$ is called (asymptotically) stable if and only if $x_k(t)$ for $k = 1, \ldots, 2n$ is (asymptotically) stable. In Figure 3.1 the solutions $x_k(t)$ for $k = 1, \ldots, 4$ of a Jeffcott rotor on an anisotropic shaft are shown. The Jeffcott rotor is investigated in more detail in [All09] and it serves as an example in Chapter 6. In this thesis we do not investigate a single vibration as given in equation (3.3), but all vibrations at once, i.e., we investigate the time behavior and how all vibrations can be reduced at once for $x(t)$, where $x(t)$ is given in equation (2.49) and (3.4) for defective and non-defective eigenvalues, respectively, as the solution to a vibrational system (3.1). Therefore, we consider $x(t)$ in a norm, i.e., $\|x(t)\|$ and investigate its transient behavior by two-sided bounds, i.e., $f_\ell(t) \leq \|x(t)\| \leq f_u(t)$ for all $t \geq t_0$ in Chapter 6. The most well-known upper bound depends on the spectral abscissa α, see e.g. [Koh02], where the spectral abscissa is defined as the maximal real part of the eigenvalues of the QEP (3.2),

$$\alpha = \max\{\mathrm{Re}\,\lambda : \lambda \in \Lambda\}. \tag{3.6}$$

The spectral abscissa gives the asymptotic rate of the vibrational system (3.1). Then the upper bound based on the spectral abscissa can be defined as $f_u(t) = Ce^{\alpha t}$ [Koh02], where an optimal constant $C > 0$ can be determined by the differential calculus of norms, which is introduced in the following section, i.e., we investigate $\|f(t)\|$, where $f : \mathbb{R} \to \mathbb{C}^n$ is a general time-varying function in various norms $\|\cdot\|$. In general $\|f(t)\|$ cannot be classically differentiated as it can be seen by the following simple example.

Example 3.1.2. *The absolute value of the linear function t is obviously not classically differentiable at $t = 0$. The absolute value coincides with the maximum norm and the Manhattan norm for $n = 1$.*

By the above example, we see that not every norm can be classically differentiated everywhere. But at the exceptional point $t = 0$, the absolute value function has a left derivative, $D_-|t| = -1$ and a right derivative, $D_+|t| = +1$, which do not coincide. Obviously, if $x(t)$ is asymptotically stable, then the solution and the upper bound $f_u(t)$ converges to zero, i.e., $\|x(t)\|, f_u(t) \to 0$ as $t \to \infty$. Here, we want to classify this behavior more precisely. Vibrations $x(t)$ behave different in different norms and as a motivating example we return to the time behavior of the Jeffcott rotor on an anisotropic shaft and show the solution $x(t)$ in various norms in Figure 3.2. We are interested in choosing a norm such that it fulfills certain properties such as monotonic convergence, decoupling and filtering which will be defined later on in this chapter. We analyze the behavior of $x(t)$ in various p-norms, where $\|x(t)\|_p := (\sum_{i=1}^n |x_i(t)|^p)^{\frac{1}{p}}$ for $p \in [1, \infty)$ has been introduced in (2.10). $\|x\|_p$ for $0 < p < 1$ is often called a norm

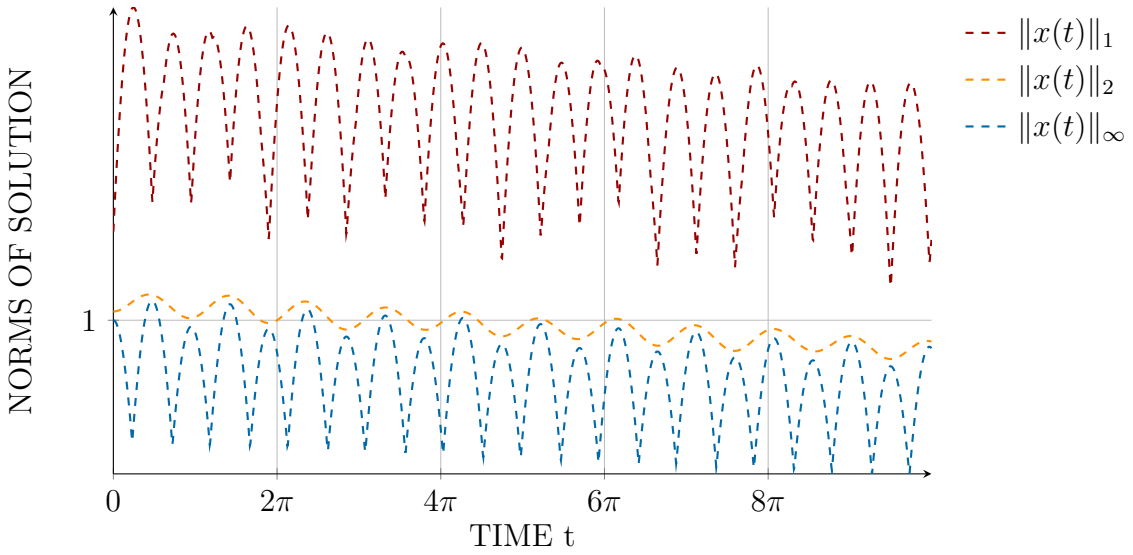


Figure 3.2: Time behavior for a Jeffcott rotor on an anisotropic shaft for $t \in [0, 10\pi]$.

but actually it is not since it is not subadditive. In Figure 3.3 unit "circles" of $\|\cdot\|_p$ for $p = \frac{1}{2}, 1, 2, \infty$ are visualized.

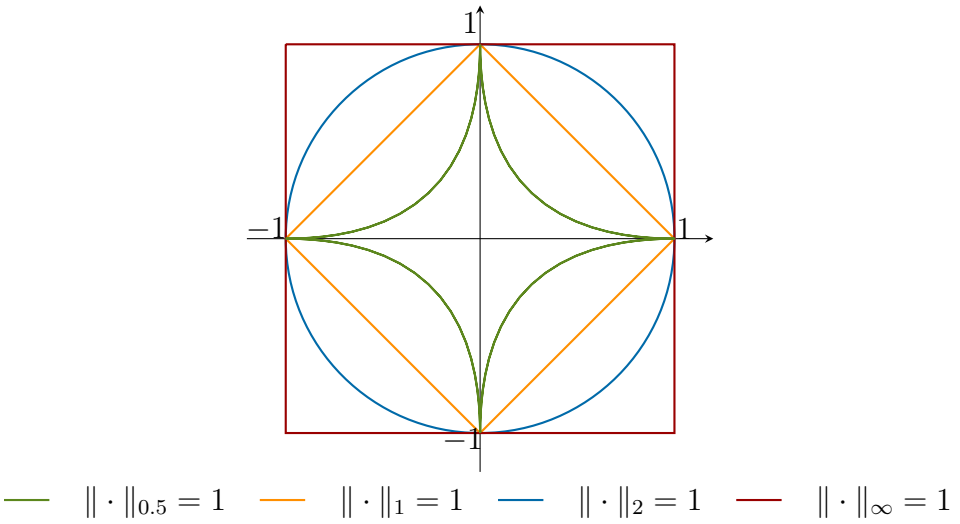


Figure 3.3: Unit "circles" of $\|\cdot\|_p$ for $p = \frac{1}{2}, 1, 2, \infty$.

3.2 Differential Calculus of Norms

In this section, we investigate the differentiability of the norms $\|f(t)\|_p$ for $p \in [1, \infty]$ w.r.t. a complex-valued vector function $f : \mathbb{R} \to \mathbb{C}^n$. $\|f(t)\|_p$ cannot be classically differentiated everywhere as seen in Example 3.1.2. A further example for non-classic differentiability is $\|x(t)\|_1$ and $\|x(t)\|_\infty$ in Figure 3.2. Obviously, if a function can be differentiated, then the left and right derivative coincide with the usual derivative. In the latter, the results we obtain should be carried over to vibrational systems (3.1) that

evolve in time, hence we restrict ourselves to right derivatives and derive expressions for right derivatives of $\|\cdot\|_p$ for $p \in [1,\infty]$. First, we are interested how $\|f(t)\|_p$ behaves locally and therefore, we cite a regularity lemma from [Koh02]. We consider the mapping $t \mapsto \|f(t)\|_p$ for a sufficiently smooth vector function f.

Lemma 3.2.1 ([Koh02]). *Let $f : \mathbb{R} \to \mathbb{C}^n$ be an n-dimensional complex-valued vector function that is m times continuously differentiable, i.e., $f \in \mathcal{C}^m(\mathbb{R}, \mathbb{C}^n)$, and $p \in [1,\infty)$. For every $t_0 \in \mathbb{R}$ there exists a number $\delta > 0$ and a function $\widetilde{f} : t \mapsto \widetilde{f}(t)$, which is real and m times continuously differentiable on $[t_0, t_0+\delta]$, i.e., $\widetilde{f} \in \mathcal{C}^m([t_0, t_0+\delta], \mathbb{R})$, such that*

$$\widetilde{f}(t) = \|f(t)\|_p$$

for every $t \in [t_0, t_0+\delta]$.

Lemma 3.2.2 ([Koh02]). *Let $f : \mathbb{R} \to \mathbb{C}^n$ be an n-dimensional complex-valued vector function that is m times continuously differentiable, i.e., $f \in \mathcal{C}^m(\mathbb{R}, \mathbb{C}^n)$. Assume that each two components of f are either identical or they intersect each other at most finitely often near t_0. Then there exists a number $\delta > 0$ and a function $\widetilde{f} : t \mapsto \widetilde{f}(t)$, which is real and m times continuously differentiable on $[t_0, t_0+\delta]$, i.e., $\widetilde{f} \in \mathcal{C}^m([t_0, t_0+\delta], \mathbb{R})$, such that*

$$\widetilde{f}(t) = \|f(t)\|_\infty$$

for every $t \in [t_0, t_0+\delta]$.

Let $f \in \mathcal{C}^m(\mathbb{R}, \mathbb{C}^n)$, then all right derivatives $D_+^k \|f(t)\|_p$ for $k \leq m$ exist for $p \in [1,\infty)$ due to Lemma 3.2.1. For $p = \infty$ we have to further assume that any two components of f are either identical or their intersection is finite (see Lemma 3.2.2). This assumption may be difficult to prove, but if f is analytic for $t \geq t_0$ or in a neighborhood of t_0, then this assumption is fulfilled for the solution $x(t)$ of the linear first order ordinary differential equation $x' = Ax$ [Koh02]. By Lemma 3.2.1 and 3.2.2 formulae for the right derivatives can be derived and are stated for $k = 1, 2$ and $p = 1, 2, \infty$. In the following let us assume that f is at least twice continuously differentiable, i.e., $f \in \mathcal{C}^m(\mathbb{R}, \mathbb{C}^n)$, where $m \geq 2$. By the Taylor series in Theorem 2.3.9 and Corollary 2.3.10 it follows that

$$f(t) = f(t_0) + Df(t_0)(t-t_0) + D^2 f(t_0)\frac{(t-t_0)^2}{2!} + o((t-t_0)^2),$$

where $t \geq t_0$. We start with the general case where $p \in [1,\infty)$ and in the latter we state

expressions for $p = 1, 2, \infty$. First, we define the following functionals for $i \in \{1, \ldots, n\}$:

$$
\begin{aligned}
S_i^{(0)} &:= |f_i(t_0)|\,, \\
S_i^{(1)} &:= \begin{cases} \frac{\operatorname{Re}(f_i(t_0))\operatorname{Re}(Df_i(t_0))+\operatorname{Im}(f_i(t_0))\operatorname{Im}(Df_i(t_0))}{|f_i(t_0)|}, & f_i(t_0) \neq 0, \\ |Df_i(t_0)|\,, & f_i(t_0) = 0, \end{cases} \\
S_i^{(2)} &:= \begin{cases} \frac{|Df_i(t_0)|^2+\operatorname{Re}(f_i(t_0))\operatorname{Re}(D^2f_i(t_0))+\operatorname{Im}(f_i(t_0))\operatorname{Im}(D^2f_i(t_0))}{|f_i(t_0)|} \\ -\frac{[\operatorname{Re}(f_i(t_0))\operatorname{Re}(Df_i(t_0))+\operatorname{Im}(f_i(t_0))+\operatorname{Im}(f_i(t_0))]^2}{|f_i(t_0)|^3}, & f_i(t_0) \neq 0, \\ \frac{\operatorname{Re}(Df_i(t_0))\operatorname{Re}(D^2f_i(t_0))+\operatorname{Im}(Df_i(t_0))\operatorname{Im}(D^2f_i(t_0))}{|Df_i(t_0)|}, & f_i(t_0) = 0, Df_i(t_0) \neq 0, \\ |D^2f_i(t_0)|\,, & f_i(t_0) = 0, Df_i(t_0) = 0. \end{cases}
\end{aligned}
$$

We start with the case $p = \infty$ and define the following index sets recursively,

$$
\begin{aligned}
I_{-1} &:= \{1, \ldots, n\}\,, \\
I_0 &:= \left\{ i_0 \in I_{-1} : S_{i_0}^{(0)} = \max_{i \in I_{-1}} S_i^{(0)} \right\}, \\
I_1 &:= \left\{ i_1 \in I_0 : S_{i_1}^{(1)} = \max_{i \in I_0} S_i^{(1)} \right\},
\end{aligned}
$$

and with this we cite the following theorem from [Koh02].

Theorem 3.2.3 ([Koh02])**.** *Let $f : \mathbb{R} \to \mathbb{C}^n$ be an n-dimensional vector function that is twice continuously differentiable, i.e., $f \in \mathcal{C}^2(\mathbb{R}, \mathbb{C}^n)$ and let $t_0 \in \mathbb{R}$. Suppose additionally that each two components of f are either identical or they intersect each other at most finitely often near t_0. Then,*

$$
\begin{aligned}
\|f(t_0)\|_\infty &= \max_{i \in I_{-1}} S_i^{(0)}, \\
D_+^1 \|f(t_0)\|_\infty &= \max_{i \in I_0} S_i^{(1)}, \\
D_+^2 \|f(t_0)\|_\infty &= \max_{i \in I_1} S_i^{(2)}.
\end{aligned}
$$

For the more general case $p \in [1,\infty)$ we define the following functionals:

$$
\begin{aligned}
S^{(0,p)} &:= \left(\sum_{i=1}^{n} \left(S_i^{(0)} \right)^p \right)^{1/p}, \\
S^{(1,p)} &:= \begin{cases} \frac{\sum_{i=1}^{n} \left(S_i^{(0)} \right)^{p-1} S_i^{(1)}}{\left(S^{(0,p)} \right)^{p-1}}, & S^{(0,p)} \neq 0, \\ \left(\sum_{i=1}^{n} \left(S_i^{(1)} \right)^p \right)^{1/p}, & S^{(0,p)} = 0, \end{cases} \\
S^{(2,p)} &:= \begin{cases} \frac{\sum_{i=1}^{n} \left(S_i^{(0)} \right)^{p-1} S_i^{(2)} + (p-1) \sum_{i=1}^{n} \left(S_i^{(0)} \right)^{p-2} \left(S_i^{(1)} \right)^2}{\left(S^{(0,p)} \right)^{p-1}} & \\ + \frac{(1-p) \left(\sum_{i=1}^{n} \left(S_i^{(0)} \right)^{p-1} S_i^{(1)} \right)^2}{\left(S^{(0,p)} \right)^{2p-1}}, & S^{(0,p)} \neq 0, \\ \frac{\sum_{i=1}^{n} \left(S_i^{(1)} \right)^{p-1} S_i^{(2)}}{\left(S^{(1,p)} \right)^{p-1}}, & S^{(0,p)} = 0, S^{(1,p)} \neq 0, \\ \left(\sum_{i=1}^{n} \left(S_i^{(2)} \right)^p \right)^{1/p}, & S^{(0,p)} = S^{(1,p)} = 0. \end{cases}
\end{aligned}
$$

Theorem 3.2.4 ([Koh02]). *Let $f : \mathbb{R} \to \mathbb{C}^n$ be an n-dimensional vector function that is twice continuously differentiable, i.e., $f \in \mathcal{C}^2(\mathbb{R}, \mathbb{C}^n)$ and let $t_0 \in \mathbb{R}$ and $p \in [1,\infty)$. Then*

$$
\begin{aligned}
\|f(t_0)\|_p &= S^{(0,p)}, \\
D_+^1 \|f(t_0)\|_p &= S^{(1,p)}, \\
D_+^2 \|f(t_0)\|_p &= S^{(2,p)}.
\end{aligned}
$$

By Theorem 3.2.4 we have derived expressions for $\|f(t_0)\|_p$ and its first two right derivatives $D_+^1 \|f(t_0)\|_p$ and $D_+^2 \|f(t_0)\|_p$ for $p \in [1,\infty]$. The results that we have obtained in this section can easily be transferred to vibrations $x(t)$, which are solutions to a vibrational system (3.1), since $x(t)$ is obviously twice continuously differentiable or even analytic.

Let us finally state the first derivative of the energy norm and the scalar product, which will be used in Section 3.3.

Example 3.2.5. *Let $B \in \mathbb{C}^{n \times n}$ be a positive definite Hermitian matrix, i.e., $B > 0$ and $f : \mathbb{R} \to \mathbb{C}^n$ be differentiable. Then the B energy norm is well-defined in* (2.14) *and the first derivative of the B energy norm is given as*

$$
\frac{\mathrm{d}}{\mathrm{d}t} \|f(t)\|_B = \frac{\langle f'(t), f(t) \rangle_B + \langle f(t), f'(t) \rangle_B}{\|f(t)\|_B},
$$

and the first derivative of the B-scalar product is given as

$$
\frac{\mathrm{d}}{\mathrm{d}t} \|f(t)\|_B^2 = \frac{\mathrm{d}}{\mathrm{d}t} \langle f(t), f(t) \rangle_B = \langle f'(t), f(t) \rangle_B + \langle f(t), f'(t) \rangle_B.
$$

3.3 Monotonic Time Behavior of Vibrations by Algebraic Lyapunov Equations

In this section, we investigate the behavior of vibrations $x(t)$ as the solution to first order system of ODEs,

$$x'(t) = Ax(t), \tag{3.7}$$

where $A \in \mathbb{C}^{N\times N}$. The vibrational system can be rewritten as first order system of ODEs, see Section 2.5.4. In Section 3.4 we investigate a transformation of this kind in more detail. Here, we consider the vibrations $x(t)$ in certain norms $\|\cdot\|$ over time. Obviously, if A is stable, i.e., $\Lambda(A) \subset \mathbb{C}_-$, then the solution to the first order system of ODEs converges to zero, i.e., $\|x(t)\| \to 0$ as $t \to \infty$ in any norm $\|\cdot\|$. The vibrations $x(t)$ behave different in different norms and we derive with the help of the so-called algebraic Lyapunov equation (ALE) for a stable matrix A, a norm such that the norm convergence is monotonic. So let us start with defining the so-called algebraic Lyapunov equation (ALE), which will be heavily used in this thesis.

Definition 3.3.1. *Let $A, W, P \in \mathbb{C}^{N\times N}$. Then*

$$A^H P + PA = -W \tag{3.8}$$

is called an algebraic Lyapunov equation (ALE).

The algebraic Lyapunov equation (3.8) can be transformed with the help of the Kronecker product and the vectorization operation into

$$\left(I_n \otimes A^H + A^T \otimes I_n\right) \operatorname{vec}(P) = -\operatorname{vec}(W). \tag{3.9}$$

By the above transformation, we obtain the existence and uniqueness result in Theorem 3.3.2 of an algebraic Lyapunov equation (3.8), which can be found e.g. in [Ant05].

Theorem 3.3.2. *The algebraic Lyapunov equation has a unique solution P for all W if and only if A^H and $-A^T$ have no common eigenvalues.*

Proof. Based on the above transformation into Kronecker form (3.9). Details are given e.g. in [Ant05]. □

For a stable matrix A, the solution to an algebraic Lyapunov equation can be defined in the following lemma.

Lemma 3.3.3 (e.g. in [Ant05])**.** *Let $A, W \in \mathbb{C}^{N\times N}$ and A be stable, i.e., $\Lambda(A) \subseteq \mathbb{C}_-$. Then*

$$P = \int_0^\infty e^{A^H t} W e^{At}\,\mathrm{d}t$$

solves the algebraic Lyapunov equation (3.8)*.*

Theorem 3.3.4. *Given any $W > 0$, there exists a unique $P > 0$ satisfying the algebraic Lyapunov equation* (3.8) *if and only if the first order ODE system* (3.7) *is globally asymptotically stable.*

Proof. The proof is based on Lemma 3.3.3 and is given e.g. in [Ant05]. □

Now, we derive a matrix $P \in \mathbb{C}^{N\times N}$, which is Hermitian and positive definite, such that $x(t)$ behaves monotonic in the P-scalar product, i.e., $\|x(t)\|_P^2 = \langle x(t), x(t)\rangle_P \searrow 0$ as $t \to \infty$, where the vibrations $x(t)$ are solutions to the first order ODE system (3.7) for a stable matrix $A \in \mathbb{C}^{N\times N}$, i.e., $\Lambda(A) \subset \mathbb{C}_-$. Let us in the following assume that A is stable, i.e., $\Lambda(A) \subset \mathbb{C}_-$. The derivative of the P-scalar product, i.e., $\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|_P^2$, is given in Example 3.2.5, which can be rewritten as

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|_P^2 &= \langle x'(t), x(t)\rangle_P + \langle x(t), x'(t)\rangle_P\\ &= \langle Ax(t), x(t)\rangle_P + \langle x(t), Ax(t)\rangle_P\\ &= \langle Ax(t), Px(t)\rangle + \langle x(t), PAx(t)\rangle\\ &= \langle \left(A^HP + PA\right) x(t), x(t)\rangle,\end{aligned}$$

where $x(t)$ is the solution to the first order ODE system (3.7). Let $P \in \mathbb{C}^{N\times N}$ be the solution to the corresponding algebraic Lyapunov equation $A^HP + PA = -W$ for any $W \in \mathbb{C}^{N\times N}$, which is Hermitian and positive definite, i.e., $W > 0$, then P is Hermitian and positive definite by Theorem 3.3.4. Hence, $\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|_P^2$ is monotonic, since

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|_P^2 = \langle \left(A^HP + PA\right) x(t), x(t)\rangle = -\langle Wx(t), x(t)\rangle \leq 0, \tag{3.10}$$

for any $t \in \mathbb{R}$.
Moreover, any square matrix $A \in \mathbb{C}^{N\times N}$ can be decomposed in its symmetric and antisymmetric part, i.e.,

$$A = A_\mathrm{S} + A_\mathrm{A},$$

where A_S and A_A denote the symmetric and antisymmetric part of the matrix A, i.e.,

$$A_\mathrm{S} = \tfrac{1}{2}\left(A + A^H\right), \quad A_\mathrm{A} = \tfrac{1}{2}\left(A - A^H\right).$$

In the following we derive conditions for a matrix P such that $\|x(t)\|_P^2$ is monotonic. We suppose that the symmetric part of the matrix A is negative definite, i.e., $A_\mathrm{S} < 0$, which in many instances is not fulfilled. $A_\mathrm{S} < 0$ can be expressed as an algebraic Lyapunov equation

$$2A_\mathrm{S} = A + A^H = A^HP + PA = -W,$$

where $W > 0$ and $P = I$ is the solution to the algebraic Lyapunov equation. Then $\|x(t)\|^2 = \langle x(t), x(t)\rangle$ decreases monotonically along every nonzero trajectory, i.e., in equation (3.10) the matrix P is chosen as the identity matrix.
The stability of vibrations $x(t)$, where $x(t)$ are solutions to the first order ODE system (3.7), can be done by transient analysis. But in the following we analyze the stability of $x(t)$ by defining a matrix $R \in \mathbb{C}^{N\times N}$ such that $\|x(t)\|_R \searrow 0$ as $t \to \infty$. Now, let us derive the matrix R. We therefore consider the algebraic matrix eigenvalue problem,

$$YA^H + AY = \mu Y, \tag{3.11}$$

where all nontrivial $\mu \in \mathbb{C}$ and $Y \in \mathbb{C}^{N\times N}$ satisfying (3.11) are called eigenvalues and

eigenmatrices, respectively. The matrix eigenvalue problem can be transformed into a standard eigenvalue problem (SEP) 2.1.1 via (3.9). Hence, it has n^2 eigenvalues and eigenmatrices. Suppose $\mu = \nu + \lambda$, then (3.11) can be transformed into

$$Y\left(A^H - \nu I\right) + (A - \lambda I)\, Y = 0.$$

An obvious solution is obtained if both terms vanish. Here, we consider A to be non-defective, i.e., all its eigenvalues are simple or semi-simple. We choose λ to be an eigenvalue of A and ν as an eigenvalue of A^H (or correspondingly $\overline{\nu}$ is an eigenvalue of A), the columns of Y must then be (right) eigenvectors of A and the rows of Y must then be left eigenvectors of A^H. Therefore, μ and Y can be expressed by the eigenpairs of the standard eigenproblem (SEP) $Av_i = \lambda_i v_i$, i.e., $\mu_{ij} = \lambda_i + \overline{\lambda}_j$ and $R_{i,j} = v_i v_j^H$ for $i, j = 1, \ldots, n$. Since there exist n^2 eigenvalues and eigenmatrices of the matrix eigenvalue problem in (3.11), we have found all of them. The eigenpairs $(\mu_{ij}, R_{i,j})$ have been obtained e.g. in [Ant05; BM60].
We consider the solution $x(t)$ to (3.7) in the energy half norm $\|x(t)\|_{R_i}^2$, where $R_i = v_i v_i^H$ is Hermitian and $\mu_i = \lambda_i + \overline{\lambda}_i < 0$, since A is stable. In this norm, convergence is monotonic, since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t)\|_{R_i}^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\langle x(t), R_i x(t)\rangle = \langle Ax(t), R_i x(t)\rangle + \langle x(t), R_i Ax(t)\rangle \\ &= \langle\left(R_i A^H + AR_i\right) x(t), x(t)\rangle = \mu_i \langle R_i x(t), x(t)\rangle = \mu_i \|x(t)\|_{R_i}^2 \leq 0.\end{aligned}$$

The idea for deriving this norm is due to [Koh08], where the adjoint matrix eigenproblem has been considered. We now consider the more general case when A is defective and follow the presentation in [Koh08]. Let $v_k^{(i)}$ for $k = 1, \ldots, m_i$ be the chain of right principal vectors of A, i.e.,

$$Av_k^{(i)} = \lambda_i v_k^{(i)} + v_{k-1}^{(i)}$$

and $v_0^{(i)} = 0$ for $i = 1, \ldots, m$, corresponding to an eigenvalue λ_i of A. Let m be the number of Jordan blocks and m_i the algebraic multiplicity of the eigenvalue λ_i. Then we define the following matrices:

$$\begin{array}{lll} R_i^{(k,k)} & := & v_k^{(i)} v_k^{(i)^H} \quad \text{for } k = 1, \ldots, m_i \ , \ i = 1, \ldots, m, \\ R_i & := & \sum_{k=1}^{m_i} R_i^{(k,k)}, \\ R & := & \sum_{i=1}^{m} R_i. \end{array} \tag{3.12}$$

The matrices R_i are eigenmatrices of the matrix eigenvalue problem (3.11) corresponding to an eigenvalue $\mu = 2\,\mathrm{Re}(\lambda_i)$. We recall the following results given in Theorems 3.3.5, 3.3.6 and Lemma 3.3.7 from [Koh08] for a time-invariant system $x' = Ax$ and a possibly non-diagonalizable system matrix A.

Theorem 3.3.5 ([Koh08])**.**

1. *$R_i^{(k,k)} = v_k^{(i)} v_k^{(i)^H}$ are Hermitian and positive semidefinite for $k = 1, \ldots, m_i$ and $i = 1, \ldots, m$,*

2. *$R_i = \sum_{k=1}^{m_i} R_i^{(k,k)}$ are Hermitian and positive semidefinite for $i = 1, \ldots, m$ and*

3. $R = \sum_{i=1}^{m} R_i$ *is Hermitian and positive definite.*

Hence, $\|\cdot\|_R$ is a norm defined by $\|v\|_R^2 = \langle Rv, v\rangle$ for $v \in \mathbb{C}^n$ and $\|\cdot\|_{R_i}$ is a semi-norm defined by $\|v\|_{R_i}^2 = \langle R_i v, v\rangle$ for $v \in \mathbb{C}^n$. In general, $\|\cdot\|_{R_i}$ does not fulfill the definiteness property in (2.8). Furthermore, the square of the semi-norm $\|\cdot\|_{R_i}^2$ has a decoupling and filter effect shown by the next theorem [Koh08].

Theorem 3.3.6 ([Koh08])**.** *Let* $x(t)$ *be the solution to the IVP* (3.7)*,* $x' = Az\,,\ x(t_0) = x_0$*, and*

$$p_{x_0,k-1}^{(i)}(t) := \langle x_0, v_1^{(i)} \tfrac{t^{k-1}}{(k-1)!} + \ldots + v_{k-1}^{(i)} t + v_k^{(i)}\rangle, \tag{3.13}$$

for $k = 1, \ldots, m_i$*,* $i = 1, \ldots, m$*. Then*

$$\|x(t)\|_{R_i^{(k,k)}}^2 = \left|p_{x_0,k-1}^{(i)}(t)\right|^2 e^{2t\,\mathrm{Re}\,\lambda_i} \quad \textit{for } t \in \mathbb{R}, \tag{3.14}$$

and

$$\|x(t)\|_R^2 = \sum_{i=1}^{m}\sum_{k=1}^{m_i} \|z(t)\|_{R_i^{(k,k)}}^2 = \sum_{i=1}^{m}\sum_{k=1}^{m_i} \left|p_{x_0,k-1}^{(i)}(t)\right|^2 e^{2t\,\mathrm{Re}\,\lambda_i} \quad \textit{for } t \in \mathbb{R}.$$

The polynomials in $p_{x_0,k-1}^{(i)}(t)$ of equation (3.13) are due to the Jordan blocks, hence to the non-diagonalizability of the matrix A, i.e., if the matrix A is diagonalizable, then all polynomials in (3.13) are constant in time.

Lemma 3.3.7. *[Koh08] Let*

$$\psi_k^{(i)}(t) := p_{x_0,k-1}^{(i)}(t) e^{t\,\mathrm{Re}\,\lambda_i} \qquad \textit{for } t \in \mathbb{R}, \tag{3.15}$$

$\psi^{(i)}(t) = [\psi_1^{(i)}, \ldots, \psi_k^{(i)}, \ldots \psi_{m_i}^{(i)}]^T$ *for* $i = 1, \ldots, m$ *and* $k = 1, \ldots, m_i$ *and* $\psi(t) = [\psi^{(1)}(t)^T, \ldots, \psi^{(i)}(t)^T, \ldots, \psi^{(r)}(t)^T]^T$*. Then*

$$\|x(t)\|_R = \|\psi(t)\|_2 \qquad \textit{for } t \in \mathbb{R}. \tag{3.16}$$

Lemma 3.3.7 shows the connection to the Euclidean norm of the function ψ. By the equivalence of norms in finite-dimensional vector spaces, a two-sided bound $c\|\psi(t)\|_p \leq \|x(t)\|_R \leq C\|\psi(t)\|_p$ for $p \in [1, \infty]$ can be derived. For $p = 2$, the constants c and C can be chosen as unity by Lemma 3.3.7.

3.4 Vibration Reduction by Viscous Dampers

In this section we introduce various vibration reduction problems for viscous dampers that will be considered in this thesis. This section is structured as follows. First, we linearize the vibrational system (3.1) in Section 3.4.1. In Section 3.4.2 we investigate the damping matrix C for viscous dampers in more detail and obtain a structured linearization of the vibrational problem. In Section 3.4.3 we introduce the optimization criterion that will be used in this thesis and finally, in Section 3.4.4 we define various vibration reduction problems for viscous dampers that we will consider in Chapters 4 and 5.

3.4.1 Linearization

Here, we linearize the vibrational system (3.1), i.e., we transform the vibrational system to a first order system. There are infinitely many linearizations since any $N \in Gl_n(\mathbb{C})$ can be chosen in equation (2.48). In this section we want to focus on a linearization that exploits structural properties of the vibrational system (3.1). Since M and K in (3.1) are positive definite and real, there exists a Cholesky decomposition (see e.g. in [HJ85]) of $K = L_1 L_1^T$ and $M = L_2 L_2^T$ with $L_1, L_2 \in Gl_n(\mathbb{R})$ being lower triangular matrices. A linearization of the vibrational system (3.1) can be obtained as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} L_1^T x \\ L_2^T x' \end{bmatrix} = \begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1} L_1 & -L_2^{-1} C L_2^{-T} \end{bmatrix} \begin{bmatrix} L_1^T x \\ L_2^T x' \end{bmatrix}, \tag{3.17}$$

which we will investigate in the following. By a singular value decomposition (SVD) e.g. [HJ85, Theorem 7.3.5, p. 414], we obtain

$$L_2^{-1} L_1 = W_2 \Omega W_1^T, \tag{3.18}$$

where W_1, W_2 are real, orthogonal matrices and Ω is a diagonal matrix with its singular values on the diagonal in decreasing order, i.e.,

$$\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n), \tag{3.19}$$

where $\omega_1 \geq \ldots \geq \omega_n > 0$. The singular values ω_i for $i = 1, \ldots, n$ are the positive roots of $(L_2^{-1} L_1)(L_2^{-1} L_1)^T$, i.e., they are the eigenvalues of the generalized eigenproblem,

$$K\varphi = \omega^2 M\varphi. \tag{3.20}$$

The generalized eigenproblem (3.20) corresponds to a second order system of ordinary differential equations,

$$Mx'' + Kx = 0, \tag{3.21}$$

where the solution x is called free vibration and has the form (3.5), where the singular value ω_i is defined above and it is also known as the circular frequency of x_i in (3.3) for $i = 1, \ldots, 2n$. Let us then define Φ as

$$\Phi := L_2^{-T} W_2, \tag{3.22}$$

where φ_i is a column of $\Phi = [\varphi_1, \cdots, \varphi_n]$ for $i = 1, \ldots, n$, and φ_i is a (right) eigenvector for the generalized eigenproblem (3.20). φ_i is a left eigenvector of the GEP to the same eigenvalue ω_i^2 for $i = 1, \ldots, n$, since M and K are symmetric and positive definite. Hence, we have found a congruence transformation by Φ in view of Theorem 2.1.6, which diagonalizes M and K,

$$\begin{aligned} \Phi^T M \Phi &= W_2^T L_2^{-1} L_2 L_2^T L_2^{-T} W_2 = I, \\ \Phi^T K \Phi &= W_2^T L_2^{-1} L_1 L_1^T L_2^{-T} W_2 = W_2^T L_2^{-1} W_1 W_1^T L_1 L_1^T L_2^{-T} W_2 \\ &= \Omega \Omega^T = \Omega^2 = \mathrm{diag}(\omega_1^2, \ldots, \omega_n^2). \end{aligned} \tag{3.23}$$

Remark 3.4.1. *For vibrational modal analysis one considers the vibrational system*

(3.1), *where damping is generally ignored, i.e., one exactly considers* (3.21). *Due to the structure of* M *and* K, *it is often preferably to convert* (3.21) *by the congruence transformation* Φ *into a decoupled second order system of ordinary differential equations* $z'' + \Omega^2 z = 0$, *where* $x = \Phi z$.

We introduce the following lemma in order to rewrite the linearization (3.17).

Lemma 3.4.2. *The matrices* $\begin{bmatrix} 0 & \Omega \\ -\Omega & -\Phi^T C \Phi \end{bmatrix}$ *and* $\begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1} L_1 & -L_2^{-1} C L_2^{-T} \end{bmatrix}$ *are orthogonal equivalent.*

Proof. Let $W = \operatorname{blockdiag}(W_1, W_2)$, where W_1, W_2 are real, orthogonal matrices which are defined in (3.18). Hence,

$$\begin{aligned}
&\begin{bmatrix} W_1 & \\ & W_2 \end{bmatrix} \begin{bmatrix} 0 & \Omega \\ -\Omega & -\Phi^T C \Phi \end{bmatrix} \begin{bmatrix} W_1^T & \\ & W_2^T \end{bmatrix} \\
&= \begin{bmatrix} 0 & W_1 \Omega W_2^T \\ -W_2 \Omega W_1^T & -W_2 \Phi^T C \Phi W_2^T \end{bmatrix} \\
&= \begin{bmatrix} 0 & W_1 (W_2^T L_2^{-1} L_1 W_1)^T W_2^T \\ -W_2 (W_2^T L_2^{-1} L_1 W_1) W_1^T & -W_2 W_2^T L_2^{-1} C L_2^{-T} W_2 W_2^T \end{bmatrix} \\
&= \begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1} L_1 & -L_2^{-1} C L_2^{-T} \end{bmatrix}
\end{aligned}$$

□

Let $y_1 := W_1^T L_1^T x$ and $y_2 := W_2^T L_2^T x'$, then the linearization (3.17) can be rewritten by Lemma 3.4.2 as,

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = W^T \begin{bmatrix} 0 & L_1^T L_2^{-T} \\ -L_2^{-1} L_1 & -L_2^{-T} C L_2^{-1} \end{bmatrix} W \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & \Omega \\ -\Omega & -\Phi^T C \Phi \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{3.24}$$

Obviously, (3.24) is a linearization of the vibrational system (3.1) as well. The initial condition is given as $y_0 = y(0) = \begin{bmatrix} W_1^T L_1^T x_0 \\ W_2^T L_2^T x_0' \end{bmatrix}$.

3.4.2 Damping

Now, we investigate the damping matrix C of the vibrational system (3.1) in more detail. We assume that damping consist of internal damping C_{int} and external damping C_{ext}, which in our case is passive damping by viscous dampers, i.e.,

$$C = C_{\text{int}} + C_{\text{ext}}. \tag{3.25}$$

We say that the internal damping matrix C_{int} satisfies an orthogonality relation w.r.t. Φ, if

$$\varphi_i^T C_{\text{int}} \varphi_j = 0, \qquad \text{for } i \neq j, \tag{3.26}$$

i.e., the internal damping matrix C_{int} is diagonalized by the same congruence transformation Φ defined in (3.22). Classical damping models such as Rayleigh damping,

i.e.,

$$C_{\text{Rayleigh}} = \alpha M + \beta K,$$

for $\alpha, \beta \geq 0$ and modal damping

$$C_{\text{modal}} = 2\gamma M^{1/2}\sqrt{M^{-1/2}KM^{-1/2}}M^{1/2},$$

for $\gamma \geq 0$ fulfill the orthogonality relation (3.26). For notational simplicity we do not want to distinguish between diverse internal damping models. Hence, we choose modal damping as internal damping, i.e., $C_{\text{int}} = C_{\text{modal}}$ and therefore,

$$\Phi^T C_{\text{int}} \Phi = \Phi^T C_{\text{modal}} \Phi = 2\gamma\Omega.$$

We remark that the theory and algorithms that are derived in this thesis can be applied as long as the internal damping satisfies the orthogonality relation (3.26).
We distinguish in this thesis two diverse external damping matrices, which are needed for the definition of vibration reduction problems for viscous dampers in Section 3.4.4. Both external damping matrices are determined by passive viscous dampers, but the difference lies in the number of allowed external dampers. While the external damping matrix consists in the first case of $r \ll n$ external viscous dampers and in the second case of n external viscous dampers.

Case 1: The position of $r \ll n$ external viscous dampers, which are indexed as non-negative integers $j_i \in \{1, \ldots, n\}$, are fixed. Furthermore, the external damping matrix depends on the viscosities $\nu_i \geq 0$ for $i = 1, \ldots, r$,

$$C_{\text{ext}} = \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T. \tag{3.27}$$

Hence, $\operatorname{rank}(C_{\text{ext}}) = r$ and in general we are interested in only a few external dampers, i.e., $r \ll n$. $\Phi^T C_{\text{ext}} \Phi$ can be rewritten as

$$\Phi^T C_{\text{ext}} \Phi = V D(\nu) V^T, \tag{3.28}$$

where

$$V = \Phi^T \begin{bmatrix} e_{j_1} & e_{j_2} & \ldots & e_{j_r} \end{bmatrix} \tag{3.29}$$

and $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_r)$.

Case 2: Here, n external viscous dampers are allowed, which are indexed as non-negative integers $i \in \{1, \ldots, n\}$. Furthermore, the external damping matrix depends on the viscosities $\nu_i \geq 0$ for $i = 1, \ldots, n$,

$$C_{\text{ext}} = \sum_{i=1}^{n} \nu_i e_i e_i^T = D(\nu), \tag{3.30}$$

where $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_n)$. Even though C_{ext} in (3.30) exists of n external viscous dampers are realized, only $r \ll n$ external viscous dampers are realized. This is guaranteed by additional constraints in the vibration

reduction problems for viscous dampers in Section 3.4.4. $\Phi^T C_{\text{ext}} \Phi$ can be rewritten as

$$\Phi^T C_{\text{ext}} \Phi = V D(\nu) V^T, \tag{3.31}$$

where

$$V = \Phi^T \tag{3.32}$$

and $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_n)$.

In the following we derive a different state space representation such that the state space matrix can be represented by a block diagonal matrix and low-rank matrix. This representation favors our computation since the state space matrix can easily be inverted due to the Sherman-Morrison-Woodbury formula. We therefore introduce the perfect shuffle permutation P, which splits a set into two piles and interleaves them. More precisely, it is the permutation, which maps

$$k \mapsto \begin{cases} 2k-1, & k \leq n, \\ 2(k-n), & k > n, \end{cases} \tag{3.33}$$

for $k = 1, \ldots, 2n$, as in [BTT11]. We then define

$$z := Py = \begin{bmatrix} y_1 \\ y_{n+1} \\ y_2 \\ y_{n+2} \\ \vdots \\ y_n \\ y_{2n} \end{bmatrix}$$

and multiply (3.24) by P and obtain the following first order system of ordinary differential equations,

$$z' = Az, \tag{3.34}$$

where $A = P \begin{bmatrix} 0 & \Omega \\ -\Omega & -\Phi^T C \Phi \end{bmatrix} P^T$ depends on the damping matrix C. For simplicity we have assumed internal damping to be modal damping and hence, $\Phi^T C_{\text{int}} \Phi = 2\gamma\Omega$. Before, we have distinguished two cases for external damping, which in the following will be handled separately.

Case 1: If external damping C_{ext} is given in (3.27), i.e., $C_{\text{ext}} = \sum_i^r \nu_i e_{j_i} e_{j_i}^T$. We can rewrite $\Phi^T C \Phi$ as $2\gamma\Omega + V D(\nu) V^T$, where $V = \Phi^T \begin{bmatrix} e_{j_1} & \cdots e_{j_r} \end{bmatrix}$ and $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_r)$, see (3.28). Therefore, the matrix A can be decomposed into

$$\begin{aligned} A &= P \begin{bmatrix} 0 & \Omega \\ -\Omega & -2\gamma\Omega - V D(\nu) V^T \end{bmatrix} P^T \\ &= \underbrace{B^{(1)} \oplus B^{(2)} \oplus \ldots \oplus B^{(n)}}_{=:B} - \widehat{V} D(\nu) \widehat{V}^T, \end{aligned} \tag{3.35}$$

where

$$
\begin{aligned}
B &= \operatorname{blockdiag}(B^{(1)}, B^{(2)}, \ldots, B^{(n)}),\\
B^{(i)} &= \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -2\gamma\omega_i \end{bmatrix}, \qquad && \text{for } i = 1, \ldots, n,\\
\widehat{V} &= P \begin{bmatrix} 0 \\ V \end{bmatrix} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ v_{11} & v_{12} & \ldots & v_{1r} \\ 0 & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & v_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ v_{n1} & v_{n2} & \ldots & v_{nr} \end{bmatrix} \in \mathbb{R}^{2n\times r},\\
D(\nu) &= \operatorname{diag}(\nu_1, \ldots, \nu_r) \in \mathbb{R}^{r\times r}
\end{aligned}
\tag{3.36}
$$

and $(V)_{ij} = v_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, r$.

Case 2: If external damping C_{ext} is given in (3.30), i.e., $C_{\text{ext}} = \sum_i^n \nu_i e_i e_i^T = D(\nu)$, where $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_n)$, we can rewrite $\Phi^T C \Phi$ as $2\gamma\Omega + V D(\nu) V^T$, where $V = \Phi^T \begin{bmatrix} e_{j_1} & \cdots e_{j_r} \end{bmatrix}$ and $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_r)$, see (3.28). Therefore, the matrix A can be decomposed into

$$
\begin{aligned}
A &= P \begin{bmatrix} 0 & \Omega \\ -\Omega & -2\gamma\Omega - V D(\nu) V^T \end{bmatrix} P^T\\
&= \underbrace{B^{(1)} \oplus B^{(2)} \oplus \ldots \oplus B^{(n)}}_{=:B} - \widehat{V} D(\nu) \widehat{V}^T,
\end{aligned}
\tag{3.37}
$$

where

$$
\begin{aligned}
B &= \operatorname{blockdiag}(B^{(1)}, B^{(2)}, \ldots, B^{(n)}),\\
B^{(i)} &= \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -2\gamma\omega_i \end{bmatrix}, \qquad && \text{for } i = 1, \ldots, n,\\
\widehat{V} &= P \begin{bmatrix} 0 \\ V \end{bmatrix} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ v_{11} & v_{12} & \ldots & v_{1n} \\ 0 & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ v_{n1} & v_{n2} & \ldots & v_{nn} \end{bmatrix} \in \mathbb{R}^{2n\times n},\\
D(\nu) &= \operatorname{diag}(\nu_1, \ldots, \nu_n) \in \mathbb{R}^{n\times n}
\end{aligned}
\tag{3.38}
$$

and $(V)_{ij} = v_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

Remark 3.4.3. *We want to emphasize that $B \in \mathbb{R}^{2n\times 2n}$ is a block diagonal matrix with 2×2-blocks $B^{(i)}$ for $i = 1, \ldots, n$, i.e., $B = \operatorname{blockdiag}\left(B^{(1)}, \ldots, B^{(n)}\right)$. Hence, B is cheaply invertible. The matrix B is obtained from vibrational modal analysis (see Remark 3.4.1) and the corresponding congruence transformation by Φ defined in* (3.22).

Remark 3.4.4. *The matrix A in* (3.35) *and* (3.37) *is called asymptotically stable if all its eigenvalues are contained in the open left complex half-plane, i.e., $\Lambda(A) \subseteq \mathbb{C}_-$, compare Remark 2.5.17. This property is assumed for all parameters ν throughout this thesis.*

3.4.3 Optimization

In this thesis we investigate the vibrational systems (3.1) for a given set of masses and stiffnesses in order to determine a damping matrix such that optimal evanescence is ensured. Evanescence can be categorized via various optimization criteria that have been considered in the literature. The most well-known criteria to judge vibrations depend on the spectrum of the corresponding quadratic eigenproblem (3.2). In this context we mention in the following the spectral abscissa and the damping ratio. The spectral abscissa α is defined in (3.6) as the maximal real part of the eigenvalues of the quadratic eigenproblem (3.2), i.e., $\alpha = \max\{\operatorname{Re}\lambda : \lambda \in \Lambda\}$. We have seen the spectral abscissa before in the context of an upper bound on $\|x(t)\|$ as $f_u(t) = Ce^{\alpha t}$, see e.g. [Koh02]. The spectral abscissa gives the (asymptotic) behavior such that the energy of the system vanishes [Cox98], i.e.,

$$\alpha = \min\left\{w : \exists\kappa \text{ s.t. } E(t) \leq \kappa E(0)e^{2wt} \text{ for all } t > 0\right\},$$

where

$$E(t) = \frac{1}{2}\langle x'(t), Mx'(t)\rangle + \frac{1}{2}\langle x(t), Kx(t)\rangle \tag{3.39}$$

is the total energy of the system and $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. Here, the spectral abscissa is considered as a criterion in order to judge vibrations and their decay as e.g. in [Cox98; LF99; MS76]. As the inequality $E(t) \leq \kappa E(0)e^{2wt}$ indicates, the energy decays in general faster than at the asymptotic rate. In [LF99] conditions for constructing a damping matrix such that this inequality is tight for all times and all initial conditions are derived. Firstly, the QEP then has a single eigenvalue with multiplicity $2n$. In our setting this idea cannot be employed due to the structure of the damping matrix C which is given as $C = C_{\text{int}} + C_{\text{ext}}$. Secondly, only for the case $n = 2$ an exact construction of the damping matrix is given in [LF99].
In the following we consider the damping ratio for complex eigenvalues [MS76]. It is a measure how fast the system decays after displacement and it is defined as

$$\zeta = -\max\left\{\frac{\operatorname{Re}\lambda}{|\lambda|} : \lambda \in \Lambda\right\}. \tag{3.40}$$

A large damping ratio results not only in a fast decay of its mode in a few periods but it also guarantees small amplitudes for harmonically excited systems, i.e., the systems behave well w.r.t. perturbations [MS76].
The spectrum with $\operatorname{Re}(\lambda) \in [-1500, 0]$ and the respective spectral abscissa and damping ratio for the viscously damped beam shown in Figure 4.2 discretized by ten finite elements are shown in Figure 3.4. The viscously damped beam is further investigated in Chapters 4 and 5,
In the following we introduce the optimization criterion that we will consider through-

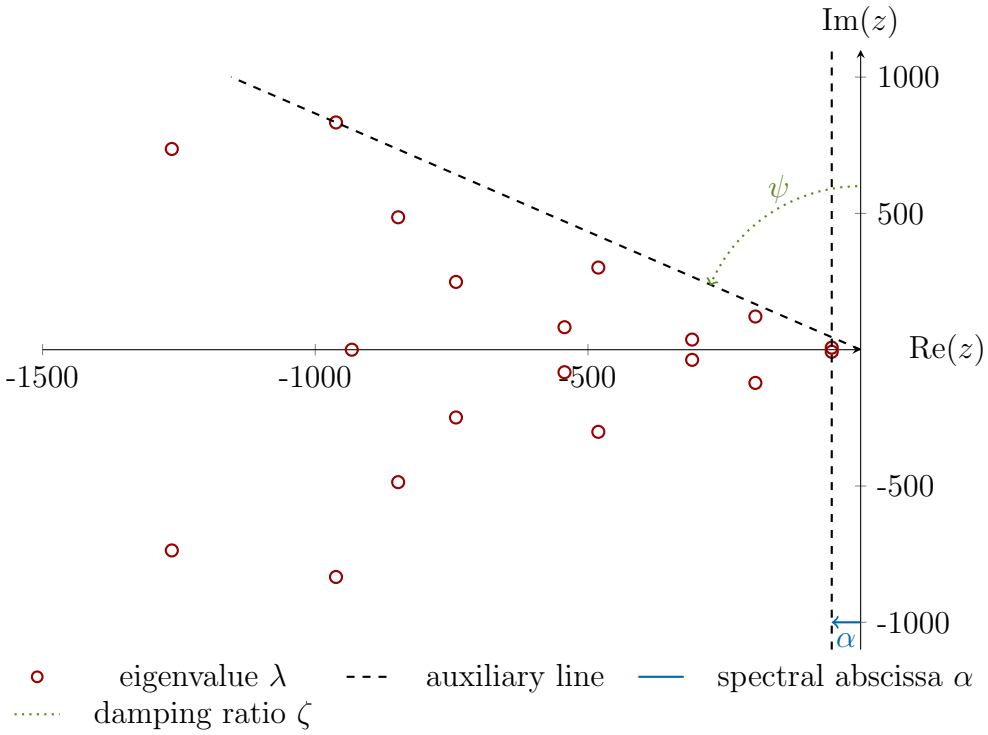


Figure 3.4: Spectrum with $\mathrm{Re}(\lambda) \in [-1500, 0]$ for the viscously damped beam discretized by 10 finite elements and respective spectral abscissa α and damping ratio $\zeta = \sin\psi$.

out this thesis. This criterion has several advantages w.r.t. the above criteria, namely the spectral abscissa and the damping ratio, which both depend solely on the eigenvalues of the system. As we will see in Chapter 4, the total energy of the system is twice continuously differentiable and hence, numerical methods such as steepest descent or Newton's method can be applied. Moreover, the solution of the Lyapunov equation provides rigorous bounds to the energy decay of a vibrating system [Ves97; Ves98; Ves03]. The criterion is based on the total energy $E(t)$ defined in (3.39) for a vibrational system (3.1). We rewrite the energy $E(t)$ w.r.t. the viscosity parameter ν as $E(t;\nu)$,

$$E(t;\nu) = \frac{1}{2}\langle x'(t;\nu), Mx'(t;\nu)\rangle + \frac{1}{2}\langle x(t;\nu), Kx(t;\nu)\rangle,$$

where $x(t;\nu)$ is the solution the vibrational system (3.1) w.r.t. the viscosity parameter ν. Since the vibrations $x(t;\nu)$ can be represented as (3.4), the energy can be represented by the spectrum of the QEP (3.2). But here we want to follow the presentation in [Ves90] by employing the minimization of the total energy w.r.t. viscosity,

$$\min_{\nu\geq 0} \int_0^\infty E(t;\nu)\,\mathrm{d}t. \tag{3.41}$$

We have chosen a linearization $z' = Az$ in (3.34) of the vibrational system such that

$$E(t;\nu) = \tfrac{1}{2}\langle z(t;\nu), z(t;\nu)\rangle.$$

Then we rewrite the optimization criterion by Lemma 3.3.3 as

$$\int_0^\infty E(t;\nu)\,\mathrm{d}t = \frac{1}{2}\int_0^\infty \langle z(t;\nu), z(t;\nu)\rangle\,\mathrm{d}t = \frac{1}{2}\int_0^\infty y_0^T e^{A^T t} e^{At} y_0\,\mathrm{d}t = \frac{1}{2} y_0^T X y_0,$$

where

$$X = \int_0^\infty e^{A^T t} e^{At}\mathrm{d}t. \tag{3.42}$$

By Lemma 3.3.3 and Theorem 3.3.4, X is symmetric and positive definite and X is the solution of the algebraic Lyapunov equation

$$A^T X + XA = -I. \tag{3.43}$$

We would like to minimize the total energy of the system. As mentioned above, it is dependent on the initial condition, and to fix this dependence, we normalize the criterion w.r.t. initial conditions with the same energy. Therefore, we consider the $2n$-dimensional unit ball $B = \{x \in \mathbb{R}^{2n} : \|x\|_2 \le 1\}$ and its surface, which is the $2n-1$-dimensional unit sphere $\partial B = \{x \in \mathbb{R}^{2n} : \|x\|_2 = 1\}$. We consider initial conditions on ∂B,

$$\min_{\nu\ge 0} \int_{\|y_0\|_2=1} \int_0^\infty E(t;\nu)\,\mathrm{d}t\,\mathrm{d}\sigma = \min_{\nu\ge 0} \frac{1}{2}\int_{\|y_0\|_2=1} y_0^T X y_0\,\mathrm{d}\sigma, \tag{3.44}$$

where σ is a measure on the sphere ∂B. In Example 3.4.5 we see how the integration on the unit sphere can be performed. We take the initial condition of the optimization criterion (3.44) into account and consider the map $X \to \frac{1}{2}\int_{\|y_0\|_2=1} y_0^T X y_0\,\mathrm{d}\sigma$, which is a linear functional on the space of symmetric matrices of size $2n \times 2n$. In this space a scalar product is defined by $\langle X, Y\rangle = \mathrm{tr}(Y^T X)$. By Riesz representation theorem, see e.g. [Kat95], a unique symmetric matrix Z exists such that

$$\frac{1}{2}\int_{\|y_0\|_2=1} y_0^T X y_0\,\mathrm{d}\sigma = \tfrac{1}{2}\langle Z, X\rangle = \tfrac{1}{2}\,\mathrm{tr}\,(X^T Z). \tag{3.45}$$

Let $x \in \mathbb{R}^{2n}$ be arbitrary and set $X = xx^T$. Then the matrix Z is positive semidefinite, i.e., $Z \ge 0$, since

$$0 \le \int_{\|y_0\|_2=1} \tfrac{1}{2} y_0^T X y_0\,\mathrm{d}\sigma = \tfrac{1}{2}\,\mathrm{tr}\,(X^T Z) = \tfrac{1}{2} x^T Z x.$$

In the following example we derive the matrix Z for the Lebesgue measure on ∂B.

Example 3.4.5. *Let $E_{ij} \in \mathbb{R}^{2n\times 2n}$ denote the matrix with all entries being zero except for the entry (i,j) that has value 1. Let $X = (X)_{ij} \in \mathbb{R}^{2n\times 2n}$ be an arbitrary symmetric matrix. Then we have*

$$\mathrm{tr}\,(X^T Z) = \mathrm{tr}\,(XZ) = \sum_{i,j} X_{ij}\,\mathrm{tr}\,(ZE_{ij}) = \sum_{i,j} X_{ij}\int_{\partial B} x^T E_{ij} x\,\mathrm{d}\sigma = \sum_{i,j} X_{ij}\int_{\partial B} x_i x_j\,\mathrm{d}\sigma,$$

hence,

$$Z_{ij} = \int_{\partial B} x_i x_j\,\mathrm{d}\sigma.$$

Let the vector field $F : U \to \mathbb{R}^{2n}$ be defined as

$$F(x_1, \ldots, x_{2n}) = (0, \ldots, 0, x_i, 0, \ldots, 0),$$

where x_i is at j-th position and $U \supseteq B$ is an open set in $\mathbb{R}^{2n}$. The divergence of F is given as

$$\operatorname{div} F(x) = \sum_{k=1}^{n} \frac{\partial F_k}{\partial x_k} = \delta_{ij}.$$

By the divergence theorem of Gauss, e.g. in [For12],

$$\int_{\partial B} \langle F(x), \nu(x) \rangle \, \mathrm{d}\sigma = \int_B \operatorname{div} F(x) \, \mathrm{d}^n x,$$

where $\nu(x)$ is an outer normal of B, we obtain

$$\int_{\partial B} x_i x_j \, \mathrm{d}\sigma = \int_B \delta_{ij} \, \mathrm{d}^n x = \operatorname{Vol}(B)\delta_{ij},$$

where $\operatorname{Vol}(B)$ denotes the volume of the $2n$-dimensional unit ball B. Hence, by $Z_{ij} = \int_{\partial B} x_i x_j \, \mathrm{d}\sigma = \operatorname{Vol}(B)\delta_{ij}$, we obtain $Z = \operatorname{Vol}(B) I$.

By Example 3.4.5 the optimization criterion based on the averaged total energy (3.44) is given as

$$\begin{aligned} \min\nolimits_{\nu \geq 0} \quad & \tfrac{\operatorname{Vol}(B)}{2} \operatorname{tr}(X) \\ \text{s.t.} \quad & A^T X + X A = -I, \end{aligned} \tag{3.46}$$

where $\frac{\operatorname{Vol}(B)}{2}$ is a scaling of the objective function that does not influence the optimal viscosity $\nu \geq 0$ and hence it is omitted in the following. We refer to [Nak02; Nak13] for further details on matrix Z. The reformulation of the total average energy as the trace of a solution of a Lyapunov equation has already been presented e.g. in [Ves90; Bra98; Nak02].

Definition 3.4.6. *Let $A, B, W, P \in \mathbb{C}^{N \times N}$ and $A = B + UCV^H$, where B is block diagonal, i.e., $B = \operatorname{blockdiag}(B^{(1)}, \ldots, B^{(\ell)})$ and $U, V \in \mathbb{C}^{N \times r}$, $C \in \mathbb{C}^{r \times r}$. Then we call*

$$A^H P + P A = -W \tag{3.47}$$

a structured algebraic Lyapunov equation (structured ALE). Moreover, we call UCV^H a low-rank perturbation if $r \ll N$.

Remark 3.4.7.

- *The algebraic Lyapunov equation $A^T X + X A = -I$ (3.43) is a structured algebraic Lyapunov equation due to the structure of $A = B - \widehat{V} D(\nu) \widehat{V}^T$, which is given in (3.35) and (3.37).*
- *The structured algebraic Lyapunov equation $A^T X + X A = -I$ in (3.43) is a parameter dependent structured algebraic Lyapunov equation, since the matrix $A = B - \widehat{V} D(\nu) \widehat{V}^T$, which is given in (3.35) and (3.37), depends on the viscosity*

ν, i.e., $A : \nu \mapsto A(\nu)$. Hence, the solution X to (3.43) *is parameter dependent as well, i.e., we can interpret X as a parameter dependent function $X : \nu \mapsto X(\nu)$.*

Hence, the solution $X(0)$ to the structured algebraic Lyapunov equation (3.47) *is simply given as*

$$X(0) = \widehat{X}_1 \oplus \widehat{X}_2 \oplus \ldots \oplus \widehat{X}_N,$$

where $\widehat{X}_i = \frac{1}{2\omega_i} \begin{bmatrix} \frac{2\gamma^2+1}{\gamma} & 1 \\ 1 & \frac{1}{\gamma} \end{bmatrix}$ *for* $i = 1, \ldots, N$.

- *The blocks $B^{(i)}$ are of size $N_i \times N_i$ for $i = 1, \ldots, \ell$, i.e., $\sum_{i=1}^{\ell} N_i = N$.*

- *The structured algebraic Lyapunov equation* (3.47) *has already been derived e.g. in [BTT11].*

3.4.4 Vibration Reduction Problems by Viscous Dampers

In this section we introduce three vibration reduction problems for viscous dampers that will be investigated in Chapters 4 and 5.

First, we make a general assumption on the viscosities. In general a viscosity cannot attain a negative value and for practical application an upper bound exists, i.e., we assume that $0 \leq \nu_i \leq \nu_{\max}$ for $i = 1, \ldots, n$.

Now, let us introduce the first two vibration reduction problems. Here, the positions $j_i \in \{1, \ldots, n\}$ for $i = 1, \ldots, r$ of the external dampers are fixed, i.e., we consider the external damping matrix C_{ext} in (3.27), which is given as

$$C_{\text{ext}} = \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T.$$

The *first* problem that we will discuss is to compute the full spectrum for the vibrational system, i.e, we solve the quadratic eigenproblem (3.2) for fixed viscosities ν_i for $i = 1, \ldots, r$,

$$Q(\lambda)v = \left(M\lambda^2 + \left[C_{\text{modal}} + \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T \right] \lambda + K \right) v = 0. \qquad \text{(QEP)}$$

Solving (QEP) can be the basis of optimizing the spectrum of a vibrational system e.g. w.r.t. the spectral abscissa criterion (3.6) or the damping ratio (3.40).

The *second* problem is the minimization of the averaged total energy for a vibrational system (3.1) in (3.41), which has been rewritten in Section 3.4.3 as

$$
\begin{aligned}
\min_{\nu \geq 0} \quad & \operatorname{tr}(X) \\
\text{s.t.} \quad & A(\nu)^T X + X A(\nu) = -I, \\
& 0 \leq \nu_i \leq \nu_{\max}, \qquad\qquad i = 1, \ldots, r, \\
\text{where } & A(\nu) = B - \widehat{V} D(\nu) \widehat{V}^T, \\
& B = \operatorname{blockdiag}(B^{(1)}, B^{(2)}, \ldots, B^{(n)}), \\
& B^{(i)} = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -2\gamma\omega_i \end{bmatrix}, \qquad i = 1, \ldots, n, \\
& \widehat{V} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ v_{11} & v_{12} & \ldots & v_{1r} \\ 0 & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & v_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ v_{n1} & v_{n2} & \ldots & v_{nr} \end{bmatrix} \in \mathbb{R}^{2n \times r}, \\
& D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_r).
\end{aligned}
\tag{OPT 1}
$$

By minimizing the averaged total energy of a vibrational system w.r.t. viscosities in (OPT 1) we directly reduce the vibrations x_k in (3.3) for $k = 1, \ldots, 2n$. We remark that the spectrum of the matrix $A(\nu)$ in (OPT 1) with $\nu_i \geq 0$ for $i = 1, \ldots, r$ is in the open left complex half-plane, i.e., $\Lambda(A(\nu)) \subseteq \mathbb{C}_-$. The associated algebraic Lyapunov equation $A(\nu)^T X(\nu) + X(\nu) A(\nu) = -I$ is called stable algebraic Lyapunov equation and it has by Theorem 3.3.2 a unique solution. The algebraic Lyapunov equation is structured and parameter dependent, see Remark 3.4.7.

Finally, we come to the *third* vibration reduction problem, where the averaged total energy of a vibrational system is minimized. Now, not only the viscosities of the external dampers, but also the external damping positions have to be determined. If there are n damping positions, where an external damper can be placed and only $r \ll n$ external dampers should be placed (e.g. due to monetary constraints), we then call this the *placement problem for r viscous dampers*. Here, the idea is to first allow all external dampers to be present and then to restrict the number of damper realizations as described in Section 3.4.2. Hence, the external damping is given as

$$
C_{\text{ext}} = \sum_{i=1}^{n} \nu_i e_i e_i^T = D(\nu),
$$

where $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_n)$. Even though we have allowed n external dampers in C_{ext}, we only want to place r viscous dampers. In general we cannot expect the solution of the optimization problem (OPT 1), where the external damping matrix has been modified to $D(\nu)$, to be sparse. Hence, we would like to find a viscosity ν such that it has at most r elements that are non-zero. We can write this restriction as a cardinality function $|\nu|_0 \leq r$, where the cardinality function is defined as $|x|_0 := |\{i : x_i \neq 0\}|$. It is often called ℓ_0-"norm" but actually it is not a norm. It is widely used in the compressed

sensing community, which in its basic variant tries to find a sparse solution of an underdetermined linear system $Ax = b$, where $A \in \mathbb{R}^{m \times n}$ and $m < n$. Reconstruction of a sparse signal x is possible if the matrix A is nearly orthogonal, i.e., it fulfills the restricted isometry property (RIP) [CRT06b; CRT06a]. But here we do not consider an underdetermined linear system but a nonlinear function as we will see in Section 4.2. The reconstruction of a sparse signal in the setting of a Nonlinear Program is to the best of our knowledge not possible. Hence, we have to develop other methods to solve the problem.

In general there are $\binom{n}{r}$ combinations of external dampers where the r external dampers can be positioned. Hence, we have to face the combinatorial explosion of different damper combinations. For each of these combinations the viscosities have to be optimized as well, i.e., by solving the optimization problem (OPT 1). This results in an extremely expensive procedure, which has not been solved satisfactorily yet. The approach in the literature to solve this problem is essentially based on enumeration, i.e., that for each combination of external dampers, a corresponding Nonlinear Program (OPT 1) has to be solved. For reasonable large examples the damper space can only heuristically be searched due to the combinatorial explosion of damper configurations. The damper space is coarsely discretized and whenever a minimum may be obtained the discretization is refined. Of course this procedure does not guarantee to find a global minimum. For fixed damper positions the viscosities are optimized by solving the optimization problem (OPT 1). Local optima of (OPT 1) can be found by gradient or Newton-based methods as it will be described in Chapter 4. It is advantageous to compute the averaged total energy as the trace of an algebraic Lyapunov equation (ALE). Various methods have been considered in this context e.g. *Alternating Direction Implicit (ADI)* for an ALE with low-rank rhs [TV09], dimension reduction techniques with an error bound [BTT11; BTT13] or the structure exploiting sign function method [BD16], which we will describe in Chapter 4. Here, we want to introduce a different approach. To compensate the relaxation of allowing n external dampers in the external damping matrix C_{ext}, we introduce binary decision variables $b \in \{0,1\}^n$, which model the existence of external viscous dampers. $b_i = 1$ if and only if at position i an external damper is present for $i = 1, \ldots, n$. Furthermore, the constraint $\nu_i \leq b_i \nu_{\max}$ controls the viscosities such that they can only be non-zero if the respective external damper is present, i.e., $b_i = 1$ for $i = 1, \ldots, n$. Via the additional constraint $\sum_{i=1}^{n} b_i \leq r$ it is guaranteed that at most r external dampers are present. Hence, we end up with the following optimization problem:

$$
\begin{aligned}
\min_{\nu} \quad & \operatorname{tr}(X) \\
\text{s.t.} \quad & A(\nu)^T X + X A(\nu) = -I, \\
& 0 \le \nu_i \le \nu_{\max}, && i = 1, \dots, n, \\
& b_i \in \{0, 1\}, && i = 1, \dots, n, \\
& \nu_i \le b_i \nu_{\max}, && i = 1, \dots, n, \\
& \sum_{i=1}^{n} b_i \le r, \\
\text{where } & A(\nu) = B - \widehat{V} D(\nu) \widehat{V}^T, \\
& B = \operatorname{blockdiag}(B^{(1)}, B^{(2)}, \dots, B^{(n)}), \\
& B^{(i)} = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -2\gamma\omega_i \end{bmatrix}, && i = 1, \dots, n, \\
& \widehat{V} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ v_{11} & v_{12} & \dots & v_{1n} \\ 0 & 0 & \dots & 0 \\ v_{21} & v_{22} & \dots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \\ v_{n1} & v_{n2} & \dots & v_{nn} \end{bmatrix} \in \mathbb{R}^{2n \times n}, \\
& D(\nu) = \operatorname{diag}(\nu_1, \dots, \nu_n).
\end{aligned}
\tag{OPT 2}
$$

The difference between the second and the third optimization problem, namely (OPT 1) and (OPT 2), is the external damping matrix C_{ext}. While in the second optimization problem (OPT 1) the positions for r external dampers are fixed, i.e., $C_{\text{ext}} = \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T$ and hence $\operatorname{rank}(C_{\text{ext}}) = \operatorname{rank}(D(\nu)) = r$, the positions for the external dampers in the third optimization problem (OPT 2) can be varied and the external damping matrix is relaxed such that all positions are allowed, i.e., $C_{\text{ext}} = \sum_{i=1}^{n} \nu_i e_i e_i^T = \operatorname{diag}(\nu)$. We note that in (OPT 2) the number of external dampers is restricted by the constraint $\sum_{i=1}^{n} b_i \le r$ and hence, $\operatorname{rank}(C_{\text{ext}}) = \operatorname{rank}(D(\nu)) \le r$.

4

Vibration Reduction by Viscous Dampers

In this chapter we reduce the vibrations

$$x_k(t) = e^{\lambda_k t} = e^{d_k t}\left(\cos(t\omega_k)\,\mathrm{Re}(v_k) - \sin(t\omega_k)\,\mathrm{Im}(v_k)\right)$$

given in (3.3) for $k = 1, \ldots, 2n$ of the vibrational system (3.1) at once. We consider r external dampers whose positions are fixed but their viscosities may be varied. As discussed in Chapter 3, the vibrations are dependent on the viscosities ν, hence, they can be expressed more explicitly as

$$x_k(t;\nu) = e^{\lambda_k(\nu) t} = e^{d_k(\nu) t}\left(\cos(t\omega_k(\nu))\,\mathrm{Re}(v_k(\nu)) - \sin(t\omega_k(\nu))\,\mathrm{Im}(v_k(\nu))\right)$$

We investigate the two problems that were introduced in Section 3.4.4, namely solving the quadratic eigenproblem (QEP),

$$Q(\lambda)v = \left(\lambda^2 M + \lambda\left[C_{\mathrm{modal}} + \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T\right] + K\right) v = 0$$

and the minimization of the total energy (OPT 1),

$$
\begin{aligned}
& A(\nu)^T X(\nu) + X(\nu) A(\nu) = -I, \\
& 0 \leq \nu_i \leq \nu_{\max}, && i = 1, \ldots, r, \\
\text{where } & A(\nu) = B - \widehat{V} D(\nu) \widehat{V}^T, \\
& B = \operatorname{blockdiag}(B^{(1)}, B^{(2)}, \ldots, B^{(n)}), \\
& B^{(i)} = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -2\gamma\omega_i \end{bmatrix}, && i = 1, \ldots, n, \\
& \widehat{V} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ v_{11} & v_{12} & \ldots & v_{1r} \\ 0 & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & v_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ v_{n1} & v_{n2} & \ldots & v_{nr} \end{bmatrix} \in \mathbb{R}^{2n \times r}, \\
& D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_r).
\end{aligned}
$$

In Section 4.1 we introduce a structure-exploiting variant of the Ehrlich-Aberth iteration in order to compute the spectrum of the vibrational system by solving the above quadratic eigenproblem. In Section 4.2 we turn to the second problem (OPT 1) and classify it as a Nonlinear Program (NLP). Furthermore, we show differentiability of the objective function and the constraints functions of the NLP. In Section 4.3 we introduce the so-called structure-exploiting sign function method in order to solve the optimization problem (OPT 1). Finally, in Section 4.4 we show numerical results for both numerical algorithms – the structure-exploiting Ehrlich-Aberth iteration and the structure-exploiting sign function method.

4.1 Eigenvalue Algorithm - Ehrlich-Aberth Iteration

We start the investigation of the vibrational problem (3.1) by deriving a method that computes the eigenvalues of the corresponding quadratic eigenproblem (3.2). Hence, we solve the first optimization problem (QEP) in Section 3.4.4, namely,

$$
Q(\lambda)v = \left(\lambda^2 M + \lambda \left[C_{\text{modal}} + \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T \right] + K \right) v = 0
$$

The algorithm has been presented in [BD15] for any internal damping that fulfills (3.26), but for notational simplicity we restrict ourselves in the following to modal damping as internal damping. In absence of the external damping $C_{\text{ext}} = \sum_{i=1}^{r} \nu_i e_{j_i} e_{j_i}^T$, the vibrational problem (3.1) and the corresponding QEP (3.2) have a lot of nice properties. Especially, their ability to be simultaneously diagonalized by a congruence transformation by Φ as in (3.23),

$$
Iz'' + \left(2\gamma\Omega + V D(\nu) V^T \right) z' + \Omega^2 z = 0, \tag{4.1}
$$

where $z = \Phi x$ and $\Phi^T C_{\text{ext}}(\nu)\Phi^T = VD(\nu)V^T$ with $D(\nu) = \operatorname{diag}(\nu_1, \ldots, \nu_r)$, compare (3.28) and (3.29). The corresponding QEP is

$$\left(\lambda^2 I + \lambda\left(2\gamma\Omega + VD(\nu)V^T\right) + \Omega^2\right) w = 0, \tag{4.2}$$

where λ are the eigenvalues and w are the corresponding eigenvectors. In the following we assume that the viscosities ν_i for $i = 1, \ldots, r$ are fixed and for brevity we omit the viscosity dependence of D, i.e.,

$$D := \operatorname{diag}(\nu_1, \ldots, \nu_r).$$

The eigenvalues of the QEPs (3.2) and (4.2) coincide if the corresponding viscosities coincide. Especially, if the external damping $C_{\text{ext}}(\nu)$ is absent, i.e., $\nu = 0$, the eigenvalues of the QEPs are given as $\lambda_i = \left(-\gamma \pm \sqrt{\gamma^2 - 1}\right)\omega_i$, where ω_i is defined in (3.23) for $i = 1, \ldots, n$. The eigenvectors of the QEPs (3.2) and (4.2) have been transformed by Φ in (3.23), i.e., $v = \Phi w$. Our goal is to design an algorithm that computes the eigenvalues of the QEP by taking the external damping into account, but not to increase the complexity significantly.

Let $q(x)$ be a polynomial of degree $2n$. The Ehrlich-Aberth iteration [Abe73; Ehr67] determines the roots $\lambda_1, \lambda_2, \ldots, \lambda_{2n}$ of q, i.e., $q(\lambda_k) = 0$ for $k = 1, \ldots, 2n$, simultaneously and iteratively. Let $\lambda_1^{(i)}, \ldots, \lambda_{2n}^{(i)} \in \mathbb{C}$ be the current approximations of the zeros of $q(x)$. The Ehrlich-Aberth iteration takes the form

$$\lambda_k^{(i+1)} = \lambda_k^{(i)} - \frac{\frac{q(\lambda_k^{(i)})}{q'(\lambda_k^{(i)})}}{1 - \frac{q(\lambda_k^{(i)})}{q'(\lambda_k^{(i)})}\left(\sum_{j<k} \frac{1}{\lambda_k^{(i)} - \lambda_j^{(i+1)}} + \sum_{j>k} \frac{1}{\lambda_k^{(i)} - \lambda_j^{(i)}}\right)}, \tag{4.3}$$

for $k = 1, \ldots, 2n$, where each new approximation $\lambda_j^{(i+1)}$ is used as soon as it is computed, i.e., in a Gauß-Seidel fashion. In [BN13] the Ehrlich-Aberth iteration was used for finding eigenvalues of regular matrix polynomials. In [Tas15] it was used for QEPs with external low-rank damping. In [BD15] this method has been generalized to QEPs with external low-rank damping and internal damping. As for any iterative algorithm, a stopping criteria is needed. In [BN13] it was suggested to stop updating when the condition number of $Q(\lambda_i)$ is sufficiently large or when the associated backward error is sufficiently small. The first criterion can only be used if the approximated eigenvalues are semi-simple. Let us first derive the steps that are needed for external low-rank damping. Let $Q(\lambda)$ be a second order matrix polynomial, the Ehrlich-Aberth iteration is then applied to the equation $\det Q(\lambda) = q(\lambda) = 0$. The evaluation of $q(\lambda)/q'(\lambda)$ is the crucial part of the update, by using Jacobi's formula e.g. in [MN99],

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \det Q(\lambda) = \operatorname{tr}\left(Q(\lambda)^{-1} Q'(\lambda)\right) \det Q(\lambda),$$

and hence,

$$\frac{q'(\lambda)}{q(\lambda)} = \operatorname{tr}\left(Q(\lambda)^{-1} Q'(\lambda)\right).$$

If no structure is exploited, each update costs $\mathcal{O}(n^3)$ flops. We derive in the following a formula for $\operatorname{tr}\left(Q(\lambda)^{-1}Q'(\lambda)\right)$ such that the essential step in the Ehrlich-Aberth iteration, i.e., the computation of $\frac{q(\lambda)}{q'(\lambda)}$ can be computed in $\mathcal{O}(r^2n)$, where $r \ll n$. In the following deduction the low-rank of the external damping is exploited and matrices are derived and denoted by $\widetilde{A}, \widetilde{B}, \ldots, \widetilde{G}$, which are then summarized in Algorithm 1. Let us first start with the diagonalization of the QEP (4.2), i.e., after the congruence transformation with Φ, then $Q(\lambda)$ and $Q'(\lambda)$ are given as

$$\begin{aligned} Q(\lambda) &= \lambda^2 I_n + 2\lambda\gamma\Omega + \lambda V D V^T + \Omega^2, \\ Q'(\lambda) &= 2\lambda I_n + 2\gamma\Omega + V D V^T, \end{aligned}$$

where $I_n, \Omega, \Omega^2 \in \mathbb{R}^{n\times n}$ are diagonal matrices. To compute the trace, first we use the Sherman-Morrison-Woodbury formula in order to compute the inverse $Q(\lambda)^{-1}$ cheaply:

$$Q(\lambda)^{-1} = \left(\widetilde{A} + \lambda V D V^T\right)^{-1} = \widetilde{A}^{-1} - \lambda\widetilde{A}^{-1}V\left(D^{-1} + \lambda V^T\widetilde{A}^{-1}V\right)^{-1}V^T\widetilde{A}^{-1}, \quad (4.4)$$

where $\widetilde{A} = \lambda^2 I_n + 2\lambda\gamma\Omega + \Omega^2$. Since $\widetilde{A}$ and D are diagonal matrices, their inverses are given as the reciprocal of the diagonal matrix, i.e., $\widetilde{A}^{-1} = \operatorname{diag}\left(1/\widetilde{d}_{11}, \ldots, 1/\widetilde{d}_{nn}\right)$ and $D^{-1} = \operatorname{diag}\left(1/d_{11}, \ldots, 1/d_{nn}\right)$. Hence,

$$\begin{aligned} Q(\lambda)^{-1}Q'(\lambda) &= 2\lambda\widetilde{A}^{-1} + \widetilde{A}^{-1}(2\gamma\Omega + V D V^T) \\ &\quad -2\lambda^2\widetilde{A}^{-1}V\left(D^{-1} + \lambda V^T\widetilde{A}^{-1}V\right)^{-1}V^T\widetilde{A}^{-1} \\ &\quad -\lambda\widetilde{A}^{-1}V\left(D^{-1} + \lambda V^T\widetilde{A}^{-1}V\right)^{-1}V^T\widetilde{A}^{-1}(2\gamma\Omega + V D V^T) \end{aligned}$$

and with $\widetilde{B} = \widetilde{A}^{-1}V$, $\widetilde{B}^T = V^T\widetilde{A}^{-1}$ and $\widetilde{C} = V^T\widetilde{B}$

$$\begin{aligned} Q(\lambda)^{-1}Q'(\lambda) &= 2\lambda\widetilde{A}^{-1} + \widetilde{A}^{-1}(2\gamma\Omega + V D V^T) - 2\lambda^2\widetilde{B}\left(D^{-1} + \lambda\widetilde{C}\right)^{-1}\widetilde{B}^T \\ &\quad -\lambda\widetilde{B}\left(D^{-1} + \lambda\widetilde{C}\right)^{-1}\widetilde{B}^T(2\gamma\Omega + V D V^T). \end{aligned}$$

We apply the properties of the trace, especially its linearity and invariance under cyclic permutations, i.e., $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$, to obtain

$$\begin{aligned} \operatorname{tr}\left(Q(\lambda)^{-1}Q'(\lambda)\right) &= 2\lambda\operatorname{tr}\left(\widetilde{A}^{-1}\right) + 2\gamma\operatorname{tr}\left(\widetilde{A}^{-1}\Omega\right) + \operatorname{tr}\left(D\widetilde{C}\right) \\ &\quad -2\lambda^2\operatorname{tr}\left(\left(D^{-1} + \lambda\widetilde{C}\right)^{-1}\widetilde{B}^T\widetilde{B}\right) \\ &\quad -2\gamma\lambda\operatorname{tr}\left(\left(D^{-1} + \lambda\widetilde{C}\right)^{-1}\widetilde{B}^T\Omega\widetilde{B}\right) \\ &\quad -\lambda\operatorname{tr}\left(\left(D^{-1} + \lambda\widetilde{C}\right)^{-1}\widetilde{C}^T D\widetilde{C}\right). \end{aligned}$$

Therefore, we end up with Algorithm 1 for the computation of $\operatorname{tr}(Q(\lambda)^{-1}Q'(\lambda))$. The total flop count is dominated by line 3 and the linear system solves in lines 5 to

7 of Algorithm 1. The computation of the matrix $\widetilde{C}$ can be done in $\mathcal{O}(r^2n)$ flops since $\widetilde{B}$ and V are of size $n \times r$. The computational complexity of forming matrices $\widetilde{B}^T\widetilde{B}, \widetilde{B}^T\Omega\widetilde{B}, \widetilde{C}^T D\widetilde{C}$ in a naive way is $\mathcal{O}(r^2n)$ and solving an $r \times r$ linear system can be done in $\mathcal{O}(r^3)$ flops. Hence, the total flop count of Algorithm 1 is of order $\mathcal{O}(r^2n)$, where $r \ll n$, instead of order $\mathcal{O}(n^3)$ without exploiting its structure.

Algorithm 1 Computes $\operatorname{tr}(Q(\lambda)^{-1}Q'(\lambda))$

Require: $\Omega = \operatorname{diag}(\omega_1, \ldots, \omega_n) \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{n\times r}$ and $D \in \mathbb{R}^{r\times r}$ and $\gamma \in \mathbb{R}$, $\lambda \in \mathbb{C}$
Ensure: $\operatorname{tr}(Q(\lambda)^{-1}Q'(\lambda))$
1: $\widetilde{A} \leftarrow \lambda^2 I_n + 2\gamma\lambda\Omega + \Omega^2$
2: $\widetilde{B} \leftarrow \widetilde{A}^{-1}V$
3: $\widetilde{C} \leftarrow V^T\widetilde{B}$
4: $\widetilde{D} \leftarrow D^{-1} + \lambda\widetilde{C}$
5: $\widetilde{E} \leftarrow \widetilde{D}^{-1}(\widetilde{B}^T\widetilde{B})$
6: $\widetilde{F} \leftarrow \widetilde{D}^{-1}(\widetilde{B}^T\Omega\widetilde{B})$
7: $\widetilde{G} \leftarrow \widetilde{D}^{-1}(\widetilde{C}^T D\widetilde{C})$
8: **return** $2\lambda \operatorname{tr}(\widetilde{A}^{-1}) + 2\gamma \operatorname{tr}(\widetilde{A}^{-1}\Omega) + \operatorname{tr}(D\widetilde{C}) - 2\lambda^2 \operatorname{tr}(\widetilde{E}) - 2\gamma\lambda \operatorname{tr}(\widetilde{F}) - \lambda \operatorname{tr}(\widetilde{G})$.

In line 3 in Algorithm 2 the starting points $\lambda^{(0)}$ are locked if they are eigenvalues of the QEP (4.2).

Algorithm 2 Computes all eigenpairs of (4.2) with modal internal damping

Require: $\Omega, \Phi \in \mathbb{R}^{n\times n}$, $\gamma \in \mathbb{R}$, starting points $\lambda^{(0)}$
Ensure: (λ_i, w_i) for $i = 1, \ldots, 2n$.
1: Lock starting points $\lambda^{(0)}$ that are eigenvalues of (4.2):

$$\left(\lambda^2 I_n + \lambda\left[2\gamma\Omega + VD(\nu)V^T\right] + \Omega^2\right) w = 0.$$

2: Compute the eigenvalues λ_i for $i = 1, \ldots, 2n$ of (4.2) by the Ehrlich-Aberth iteration (4.3).
3: Compute the eigenvectors w_i for $i = 1, \ldots, 2n$ of (4.2) by inverse iteration.
4: **return** (λ_i, w_i) for $i = 1, \ldots, 2n$.

4.2 Preliminaries

Now, we come to the second optimization problem (OPT 1). Let $f : \mathbb{R}^r \to \mathbb{R}$ be the objective function of (OPT 1), i.e., it is defined as $f(\nu) = \operatorname{tr}(X(\nu))$ s. t. $A(\nu)^T X(\nu) + X(\nu)A(\nu) = -I$. By Lemma 3.3.3 the function f can be written as

$$f : \nu \mapsto \operatorname{tr}\left(\int_0^\infty e^{A(\nu)^T t} e^{A(\nu)t} \mathrm{d}t\right), \tag{4.5}$$

where $A(\nu) = B - \widehat{V}D(\nu)\widehat{V}^T$ is given in (3.35) and $\widehat{V}$ and $D(\nu)$ are given in (3.36). f is well-defined since $A(\nu)$ is stable. First, we note that f is invariant under any permu-

tation due to the cyclic property of the trace. Hence, it is invariant under the perfect shuffle permutation P defined in (3.33). In the following section we investigate the smoothness of the function f and the optimization problem (OPT 1) in more detail. We start by computing the gradient and Hessian of the objective function f defined in (4.5) by deducing the first and second derivative of the trace of the structured algebraic Lyapunov equation (3.47). The partial derivative of the trace for the solution of an algebraic Lyapunov equation (3.43) has been derived for an algebraic Lyapunov equation $AY + YA^T = -Z$ without the perfect shuffle permutation in [Bra98]. We obtain an equivalent result for the structured algebraic Lyapunov equation 3.4.6. We differentiate the parameter dependent algebraic Lyapunov equation $A(\nu)^T X(\nu) + X(\nu)A(\nu) = -I$ given in (3.43) w.r.t. the viscosity ν_j and obtain

$$A(\nu)^T \frac{\partial}{\partial \nu_j} X(\nu) + \frac{\partial}{\partial \nu_j} X(\nu) A(\nu) = -\frac{\partial}{\partial \nu_j} A(\nu)^T X(\nu) - X(\nu) \frac{\partial}{\partial \nu_j} A(\nu),$$

with

$$\frac{\partial}{\partial \nu_j} A(\nu)^T = \frac{\partial}{\partial \nu_j} A(\nu) = -\widehat{V}_j \widehat{V}_j^T = - \begin{bmatrix} 0 \\ v_{1j} \\ 0 \\ v_{2j} \\ \vdots \\ 0 \\ v_{nj} \end{bmatrix} \begin{bmatrix} 0 \\ v_{1j} \\ 0 \\ v_{2j} \\ \vdots \\ 0 \\ v_{nj} \end{bmatrix}^T$$

where $\widehat{V}_j$ is the j-th column of $\widehat{V}$. We obtain the following structured algebraic Lyapunov equation

$$A(\nu)^T \frac{\partial}{\partial \nu_j} X(\nu) + \frac{\partial}{\partial \nu_j} X(\nu) A(\nu) = \widehat{V}_j \widehat{V}_j^T X(\nu) + X(\nu) \widehat{V}_j \widehat{V}_j^T. \tag{4.6}$$

Due to the linearity and cyclic property of the trace, the j-th component of the gradient fulfills

$$\begin{aligned} \frac{\partial f(\nu)}{\partial \nu_j} &= \frac{\partial \operatorname{tr}(X(\nu))}{\partial \nu_j} = \operatorname{tr}\left(\frac{\partial X}{\partial \nu_j}(\nu) \right) \\ &= -\operatorname{tr}\left(\int_0^\infty e^{A(\nu)^T t} \left(\widehat{V}_j \widehat{V}_j^T X(\nu) + X(\nu) \widehat{V}_j \widehat{V}_j^T \right) e^{A(\nu) t} \mathrm{d}t \right) \\ &= -\operatorname{tr}\left(\widehat{V}_j^T X(\nu) \int_0^\infty e^{A(\nu) t} e^{A(\nu)^T t} \mathrm{d}t \widehat{V}_j \right) \\ &\quad -\operatorname{tr}\left(\widehat{V}_j^T \int_0^\infty e^{A(\nu) t} e^{A(\nu)^T t} \mathrm{d}t X(\nu) \widehat{V}_j \right) \\ &= -\widehat{V}_j^T X(\nu) Y(\nu) \widehat{V}_j - \widehat{V}_j^T Y(\nu) X(\nu) \widehat{V}_j, \end{aligned} \tag{4.7}$$

where $Y(\nu)$ is the solution of the algebraic Lyapunov equation $A(\nu)Y(\nu) + Y(\nu)A(\nu)^T = -I$. Since $JA(\nu)^T J = A(\nu)$, where

$$J = \operatorname{diag}(1, -1, 1, -1, \ldots, 1, -1),$$

it follows that $Y(\nu)$ is given as $Y(\nu) = JX(\nu)J$ and hence, (4.7) simplifies to

$$\begin{aligned}\frac{\partial \operatorname{tr}(X(\nu))}{\partial \nu_j} &= -\widehat{V}_j^T \left[X(\nu)JX(\nu)J + JX(\nu)JX(\nu)\right] \widehat{V}_j \\ &= 2\widehat{V}_j^T X(\nu)JX(\nu)\widehat{V}_j \end{aligned} \tag{4.8}$$

since $J\widehat{V}_j = -\widehat{V}_j$ and $J = J^T$.
We use the same idea in order to derive the Hessian of f. We differentiate (4.6) w.r.t. ν_i and obtain the same structured algebraic Lyapunov equation, but with a different right-hand side:

$$\begin{aligned}A(\nu)^T \frac{\partial^2 X(\nu)}{\partial \nu_i \partial \nu_j} + \frac{\partial^2 X(\nu)}{\partial \nu_i \partial \nu_j} A(\nu) &= \widehat{V}_i \widehat{V}_i^T \frac{\partial X(\nu)}{\partial \nu_j} + \frac{\partial X(\nu)}{\partial \nu_j} \widehat{V}_i \widehat{V}_i^T \\ &\quad + \widehat{V}_j \widehat{V}_j^T \frac{\partial X(\nu)}{\partial \nu_i} + \frac{\partial X(\nu)}{\partial \nu_i} \widehat{V}_j \widehat{V}_j^T. \end{aligned} \tag{4.9}$$

Due to the linearity and cyclic property of the trace, the Hessian $H_{ij} = \frac{\partial^2 f}{\partial \nu_i \partial \nu_j}$ fulfills

$$\begin{aligned}\frac{\partial^2 f(\nu)}{\partial \nu_i \partial \nu_j} &= \frac{\partial^2 \operatorname{tr}(X(\nu))}{\partial \nu_i \partial \nu_j} = \operatorname{tr}\left(\frac{\partial^2 X}{\partial \nu_i \partial \nu_j}(\nu)\right) \\ &= -\operatorname{tr}\left(\int_0^\infty e^{A(\nu)^T t} \left(\widehat{V}_i \widehat{V}_i^T \frac{\partial X(\nu)}{\partial \nu_j} + \frac{\partial X(\nu)}{\partial \nu_j} \widehat{V}_i \widehat{V}_i^T\right) e^{A(\nu)t} \mathrm{d}t\right) \\ &\quad -\operatorname{tr}\left(\int_0^\infty e^{A(\nu)^T t} \left(\widehat{V}_j \widehat{V}_j^T \frac{\partial X(\nu)}{\partial \nu_i} + \frac{\partial X(\nu)}{\partial \nu_i} \widehat{V}_j \widehat{V}_j^T\right) e^{A(\nu)t} \mathrm{d}t\right) \\ &= -\widehat{V}_i^T \int_0^\infty \frac{\partial X(\nu)}{\partial \nu_j} e^{A(\nu)t} e^{A(\nu)^T t} + e^{A(\nu)t} e^{A(\nu)^T t} \frac{\partial X(\nu)}{\partial \nu_j} \mathrm{d}t\, \widehat{V}_i \\ &\quad -\widehat{V}_j^T \int_0^\infty \frac{\partial X(\nu)}{\partial \nu_i} e^{A(\nu)t} e^{A(\nu)^T t} + e^{A(\nu)t} e^{A(\nu)^T t} \frac{\partial X(\nu)}{\partial \nu_i} \mathrm{d}t\, \widehat{V}_j \\ &= -\widehat{V}_i^T \left(\frac{\partial X(\nu)}{\partial \nu_j} Y(\nu) + Y(\nu) \frac{\partial X(\nu)}{\partial \nu_j}\right) \widehat{V}_i \\ &\quad -\widehat{V}_j^T \left(\frac{\partial X(\nu)}{\partial \nu_i} Y(\nu) + Y(\nu) \frac{\partial X(\nu)}{\partial \nu_i}\right) \widehat{V}_j, \end{aligned} \tag{4.10}$$

where $Y(\nu)$ is defined as above, i.e., it is the solution of the structured algebraic Lyapunov equation $A(\nu)Y(\nu)+Y(\nu)A^T(\nu) = -I$, and it is given as $Y(\nu) = JX(\nu)J$. With the above calculations we can state that the objective function $f : \nu \mapsto \operatorname{tr}(X(\nu))$ of the optimization problem (OPT 1) is twice continuously differentiable, see e.g. [Ves90; Bra98].

Lemma 4.2.1. *The mapping $f : \mathbb{R}^r \to \mathbb{R}$ with*

$$f : \nu \mapsto \operatorname{tr}\left(\int_0^\infty e^{A(\nu)^T t} e^{A(\nu)t} \mathrm{d}t\right)$$

is twice continuously differentiable, i.e., $f \in \mathcal{C}^2$.

Proof. Since the spectrum of the matrix $A(\nu)$ for any ν is in the open left complex half-plane, the mapping is well-defined. The first two derivatives of f are given in (4.8) and (4.10). Since $\frac{\partial X(\nu)}{\partial \nu_i}$, $\frac{\partial X(\nu)}{\partial \nu_j}$ and $Y(\nu)$ of (4.10) are continuous, it follows that $\frac{\partial^2 f(\nu)}{\partial \nu_i \partial \nu_j}$ is continuous. Hence, f is twice continuously differentiable. □

In [Ves90] the mapping f in Lemma 4.2.1 has been analytically derived for a single damper. It has been shown that this mapping is convex for a single damper. Convexity cannot be generalized to more than one damper. A counter example for two external viscous dampers is given in [TV09].

Lemma 4.2.2. *The optimization problem* (OPT 1) *is a Nonlinear Program (NLP).*

Proof. For theoretical investigation we use the integral formulation of the solution of the algebraic Lyapunov equation (3.42) by Lemma 3.3.3, i.e., $X(\nu) = \int_0^\infty e^{A(\nu)^T t} e^{A(\nu)t} \mathrm{d}t$, which is obviously nonlinear. Then the optimization problem (OPT 1) can be rewritten as the following NLP:

$$\begin{aligned} &\min f(\nu) \\ &\text{s.t. } g_i(\nu), \qquad\qquad \text{for} \qquad i = 1, \ldots, 2r, \end{aligned}$$

where

$$\begin{aligned} f(\nu) &:= \int_0^\infty \operatorname{tr}\left(e^{A(\nu)^T t} e^{A(\nu)t}\right) \mathrm{d}t \\ g_i(\nu) &:= -\nu_i \le 0, && \text{for } i = 1, \ldots, r, && (4.11) \\ g_{r+i}(\nu) &:= \nu_i - \nu_{\max} \le 0, && \text{for } i = 1, \ldots, r. && (4.12) \end{aligned}$$

□

Theorem 4.2.3. *The objective function f defined in* (4.5) *and the constraint functions g_i for $i = 1, \ldots, n+1$ defined in* (4.11) *and* (4.12) *of the Nonlinear Program* (OPT 1) *are twice continuously differentiable.*

Proof. Since the constraint functions g_i for $i = 1, \ldots, n+1$ are linear, it is obvious that they are twice continuously differentiable. The objective function f is smooth by Lemma 4.2.1 and their first and second partial derivatives are given in (4.8) and (4.10). □

We have shown that the optimization problem (OPT 1) is a Nonlinear Program (NLP) in Lemma 4.2.2 and that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, 2n$, are twice continuously differentiable in Theorem 4.2.3. Hence, smooth optimization methods based on gradient and/or Hessian can be applied in order to find a local minimum of f. The function f has a local minimum at viscosity ν^* if ν^* fulfills the first order necessary conditions, the so-called Karush-Kuhn-Tucker

(KKT) conditions [Kar39; KT51] given in (4.13)-(4.16), i.e., there exists λ^* such that

$$\nabla f(\nu^*) + \sum_{i=1}^{2n} \lambda_i^* \nabla g_i(\nu^*) = 0, \tag{4.13}$$

$$g_i(\nu^*) \leq 0, \quad \text{for } i = 1, \ldots, 2r, \tag{4.14}$$

$$\lambda_i^* \geq 0, \quad \text{for } i = 1, \ldots, 2r, \tag{4.15}$$

$$\lambda_i^* g_i(\nu^*) = 0, \quad \text{for } i = 1, \ldots, 2r. \tag{4.16}$$

The first condition (4.13) implies stationarity of the solution. While the inequality (4.14) guarantees feasibility for the primal, the inequality (4.15) guarantees feasibility for the dual problem. Equation (4.16) is a complementary slackness condition, i.e., that the KKT multiplier λ_i^* or the respective constraint $g_i(\nu^*)$ is tight. In general the conditions (4.13)-(4.16) are only necessary but not sufficient conditions for a minimizer.

4.3 Sign Function Method

To efficiently solve the Nonlinear Program (OPT 1), we derive in this section a method that solves the structured algebraic Lyapunov equation. The method exploits the structure of the algebraic Lyapunov equation and by the considerations in Section 4.2, we then can cheaply compute the Jacobian and Hessian of the objective function. Here, we assume that for each iteration the viscosity ν is fixed. Hence, for brevity we omit the dependence on ν in this section. This situation occurs e.g. when the NLP (OPT 1) is solved by a numerical method such as steepest descent or Newton's method, where the objective function and its derivatives have to be evaluated in each iteration in order to update the viscosities. We derive in this section an efficient way of evaluating the objective function via the so-called structure-exploiting sign function method. The sign function method has been introduced in [Rob80] for solving algebraic Riccati equations of the form $-XGX + A^TX + XA = W$, where $A, G, Q, W \in \mathbb{R}^{N\times N}$, $G = G^T$, $W = W^T$ and $X = X^T$ is the unknown solution matrix. In [Rob80] the matrix sign function was used to solve stable algebraic Lyapunov equations (3.8) as well,

$$A^TX + XA = W,$$

where $A, X, W \in \mathbb{R}^{N\times N}$ and A is stable, i.e., $\Lambda(A) \subseteq \mathbb{C}_-$. First, let us define the matrix sign function of $Z \in \mathbb{R}^{N\times N}$. Let Z have no eigenvalues on the imaginary axis, i.e., $\Lambda(Z) \cap \imath\mathbb{R} = \emptyset$. There are several ways to define the matrix sign function, see e.g. the survey paper [KL95]. Let $Z = UJU^{-1}$ be the Jordan decomposition of Z as in Theorem 2.1.4, where

$$J = \begin{bmatrix} J^- & 0 \\ 0 & J^+ \end{bmatrix}$$

and $J^- \in \mathbb{C}^{M\times M}$ and $J^+ \in \mathbb{C}^{(N-M)\times(N-M)}$ consist of the Jordan blocks corresponding to the eigenvalues in the open left and open right half-planes, respectively. Then the

matrix sign function of Z is defined as

$$\operatorname{sign}(Z) := U \begin{bmatrix} -I_M & 0 \\ 0 & I_{N-M} \end{bmatrix} U^{-1}.$$

Note that $\operatorname{sign}(Z)$ is unique and independent of the order of the eigenvalues in the Jordan decomposition of Z.

The sign function can be computed via the Newton iteration for the equation $Z^2 = I$, where the starting point is chosen as Z, i.e.,

$$Z_0 := Z, \qquad Z_{k+1} := \frac{Z_k + Z_k^{-1}}{2}, \qquad k = 0, 1, 2, \ldots. \tag{4.17}$$

In [Rob80] it has been shown that this procedure converges to the matrix sign function of Z, i.e., $\operatorname{sign}(Z) = \lim_{k\to\infty} Z_k$. The Newton iteration converges quadratically when the iterates are being sufficiently close to a root, but the initial convergence may be slow. Scaling can be introduced in order to accelerate the convergence, i.e., $Z_{k+1} := \frac{1}{2}c_k Z_k + \frac{1}{2}c_k^{-1} Z_k^{-1}$. Various scalings have been proposed in the literature, for a summary see [BD93]. Here, we use the scaling

$$c_k = \sqrt{\frac{\|Z_k^{-1}\|_F}{\|Z_k\|_F}},$$

which is known as Frobenius scaling, see e.g. [KL92]. For a summary of other schemes to compute the sign function of a matrix see [KL95]. When the sign function method (SFM) (4.17) is applied to

$$Z = \begin{bmatrix} A & 0 \\ W & -A^T \end{bmatrix},$$

where the matrices A, W, A^T are from the real and stable algebraic Lyapunov equation $A^T X + XA = -W$, we obtain the sign function iteration Z_k for $k = 0, 1, \ldots$ with

$$\lim_{k\to\infty} Z_k = \begin{bmatrix} -I_N & 0 \\ 2X & I_N \end{bmatrix}, \tag{4.18}$$

see e.g. [Ant05; Rob80]. Hence, the solution X of the real and stable algebraic Lyapunov equation $A^T X + XA = -W$ can be read off from (4.18) directly. Applying the generalized Newton iteration (4.17) to the matrix Z and exploiting the block-triangular structure of all matrices involved, it is easy to see that (4.17) boils down to

$$\begin{aligned} &A_0 := A, \quad W_0 := W, \\ &A_{k+1} := \tfrac{1}{2}\left(A_k + A_k^{-1}\right), \\ &W_{k+1} := \tfrac{1}{2}\left(W_k + A_k^{-T} W_k A_k^{-1}\right), \end{aligned} \qquad \text{for k=0,1,...,} \tag{4.19}$$

which is summarized in Algorithm 3. The solution of the algebraic Lyapunov equation is then $X = \frac{1}{2}\lim_{k\to\infty} W_k$.

Algorithm 3 Sign function method for ALE

Require: Matrices $A, W \in \mathbb{R}^{N\times N}, tol$.
Ensure: $Y \in \mathbb{R}^{N\times N}$ such that $Y \approx X$.

$A_0 \leftarrow A$
$W_0 \leftarrow W$
$k = 0$
while $\|A + I\| > tol$ **do**
  $c_k \leftarrow \sqrt{\frac{\|A_k^{-1}\|_F}{\|A_k\|_F}}$
  $A_{k+1} \leftarrow \frac{1}{2}\left(c_k A_k + c_k^{-1} A_k^{-1}\right)$
  $W_{k+1} \leftarrow \frac{1}{2}\left(c_k W_k + c_k^{-1} A_k^{-T} W_k A_k^{-1}\right)$
  $k = k + 1$
end while
return $Y = \frac{1}{2} W_k$

4.3.1 Structure Exploiting Sign Function Method

In this section we consider a special case of the stable algebraic Lyapunov equation (3.8),

$$A^T X + X A = W,$$

where $A, X, W \in \mathbb{R}^{N\times N}$ and A is stable. Namely, we consider the structured algebraic Lyapunov equation (3.47), i.e., the matrix A can be expressed as $A = B + UCV^T$, where $B \in \mathbb{R}^{N\times N}$, $C \in \mathbb{R}^{r\times r}$, $U, V \in \mathbb{R}^{N\times r}$ and B is a block diagonal matrix, i.e., $B = \text{blockdiag}\left(B^{(1)}, \ldots, B^{(\ell)}\right)$, which is cheaply invertible and UCV^T is a low-rank perturbation on B with rank $r \ll N$.

We later show in Theorem 4.3.1 that the above structure can be kept throughout the sign function iteration in Algorithm 3. For now, let us assume A_k can be expressed by a block diagonal matrix and a low-rank factor, i.e., $A_k = B_k + U_k C_k V_k^T$, where $B_k \in \mathbb{R}^{N\times N}$, $U_k, V_k \in \mathbb{R}^{N\times r}$, $C_k \in \mathbb{R}^{r\times r}$ and $r \ll N$. Then the inverse of A_k is given as

$$A_k^{-1} = \left(B_k + U_k C_k V_k^T\right)^{-1} = B_k^{-1} - B_k^{-1} U_k \left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1} V_k^T B_k^{-1} \tag{4.20}$$

by the Sherman-Morrison-Woodbury formula. Let us denote with N_B the maximal block size of the matrix B, i.e., $N_B = \max_{i=1}^{\ell} N_i$. Hence, computing the inverse A_k^{-1} can be done by inverting the block diagonal matrix B_k in $\mathcal{O}(\ell N_B^3)$ and an $r \times r$ system in $\mathcal{O}(r^3)$. Forming matrix products such as $B_k^{-1} U_k$ is in $\mathcal{O}(\ell N_B^3)$ due to the block diagonal structure of B_k^{-1}. The complexity to compute the inverse of A_k is $\mathcal{O}(N^3)$ without exploiting the structure, compared to $\mathcal{O}(r^3 + \ell N_B^3)$ by using the Sherman-Morrison-Woodbury formula (4.20). We now initialize $U_0 := U$, $V_0 := V$ and $C_0 := C$ and rewrite the sign function method for A_{k+1} as

$$A_{k+1} = B_{k+1} + U_{k+1} C_{k+1} V_{k+1}^T,$$

where

$$B_{k+1} = \text{blockdiag}\left(B_{k+1}^{(1)}, \ldots, B_{k+1}^{(\ell)}\right) \in \mathbb{R}^{N\times N}, \tag{4.21}$$

$$B_{k+1}^{(i)} = \tfrac{1}{2}\left(c_k B_k^{(i)} + c_k^{-1} {B_k^{(i)}}^{-1}\right) \in \mathbb{R}^{N_i\times N_i} \text{ for } i = 1, \ldots, \ell, \tag{4.22}$$

$$U_{k+1} = \begin{bmatrix} U_k & B_k^{-1}U_k \end{bmatrix}, \tag{4.23}$$

$$V_{k+1} = \begin{bmatrix} V_k & B_k^{-T}V_k \end{bmatrix}, \tag{4.24}$$

$$C_{k+1} = \tfrac{1}{2}\text{blockdiag}\left(c_k C_k, -c_k^{-1}\left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1}\right), \tag{4.25}$$

for $k = 0, 1, \ldots$. Hence, we can formulate the following theorem.

Theorem 4.3.1. *Let $A_k = B_k + U_k C_k V_k^T$, where $B_k = \text{blockdiag}\left(B_k^{(1)}, \ldots, B_k^{(\ell)}\right) \in \mathbb{R}^{N\times N}$, $U_k, V_k \in \mathbb{R}^{N\times r}$, $C_k \in \mathbb{R}^{r\times r}$ and $r \ll n$. Then the next sign function iterate A_{k+1} can be expressed as*

$$A_{k+1} = B_{k+1} + U_{k+1} C_{k+1} V_{k+1}^T, \tag{4.26}$$

where $B_{k+1} = \text{blockdiag}\left(B_{k+1}^{(1)}, \ldots, B_{k+1}^{(\ell)}\right) \in \mathbb{R}^{N\times N}$, $U_{k+1}, V_{k+1} \in \mathbb{R}^{N\times 2r}$, $C_{k+1} \in \mathbb{R}^{2r\times 2r}$ are defined in (4.21)-(4.25).

Proof. $A_{k+1} = B_{k+1} + U_{k+1} C_{k+1} V_{k+1}^T$ can be reformulated as

$$A_{k+1} = \frac{1}{2}\left(c_k B_k + c_k^{-1} B_k^{-1} + c_k U_k C_k V_k^T - c_k^{-1} B_k^{-1} U_k \left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1} V_k^T B_k^{-1}\right)$$

with the definitions in (4.21)-(4.25). □

Note that the rank of the factors $U_{k+1}, C_{k+1}, V_{k+1}$ doubles per sign function iteration. Let r_0 denote the initial rank of $U_0 C_0 V_0^T$. Hence, after $k+1$ sign function iterations, the rank of $U_{k+1} C_{k+1} V_{k+1}^T$ is $r_{k+1} = 2^{k+1} r_0$, but the numerical rank, see e.g. [GV96], may be smaller.
In order to avoid a large workspace and to reduce the computational costs, we therefore propose to compute a Rank-Revealing QR factorization of $U_{k+1} \in \mathbb{R}^{N\times r_{k+1}}$ and $V_{k+1} \in \mathbb{R}^{N\times r_{k+1}}$ such that $U_{k+1} = Q_U R_U$ and $V_{k+1} = Q_V R_V$, where $Q_U, Q_V \in \mathbb{R}^{N\times r_{k+1}}$ contain r_{k+1} orthonormal columns and $R_U, R_V \in \mathbb{R}^{r_{k+1}\times r_{k+1}}$ are upper triangular matrices. Hence, $U_{k+1} C_{k+1} V_{k+1}^T$ can be rewritten as

$$U_{k+1} C_{k+1} V_{k+1}^T = Q_U R_U C_{k+1} R_V^T Q_V^T,$$

where $R_U C_{k+1} R_V^T$ has dimension $r_{k+1} \times r_{k+1}$. We compute its singular value decomposition (SVD),

$$R_U C_{k+1} R_V^T = U \Sigma V^T,$$

where U, V are orthonormal matrices and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{r_{k+1}})$ is a diagonal matrix. Now, we truncate the singular values if they are below some predefined tolerance ε and cut the respective singular vectors as well, i.e.,

$$R_U C_{k+1} R_V^T \approx \widehat{U}\widehat{\Sigma}\widehat{V}^T, \tag{4.27}$$

where $\widehat{U}, \widehat{V} \in \mathbb{R}^{r_{k+1} \times \widehat{r}_{k+1}}$, $\widehat{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\widehat{r}_{k+1}}) \in \mathbb{R}^{\widehat{r}_{k+1} \times \widehat{r}_{k+1}}$ such that the singular values $\sigma_{\widehat{r}_{k+2}}, \ldots, \sigma_{r_{k+1}}$ are below the given tolerance ε, i.e., $\sigma_{\widehat{r}_{k+2}}, \ldots, \sigma_{r_{k+1}} < \varepsilon$. Hence, the numerical rank $\widehat{r}_{k+1}$ is at most r_{k+1}, i.e., $\widehat{r}_{k+1} \leq r_{k+1}$, and the new iterates are given as

$$\begin{aligned} U_{k+1} &:= Q_U \widehat{U}, \\ V_{k+1} &:= Q_V \widehat{V}, \\ C_{k+1} &:= \widehat{\Sigma}, \end{aligned}$$

where $U_{k+1}, V_{k+1} \in \mathbb{R}^{N \times \widehat{r}_{k+1}}$ and $C_{k+1} \in \mathbb{R}^{\widehat{r}_{k+1} \times \widehat{r}_{k+1}}$. The rank of $U_{k+1} C_{k+1} V_{k+1}^T$ has been reduced to $\widehat{r}_{k+1}$. By the above considerations we end up with a structure preserving sign function method defined in Algorithm 4 for the structured algebraic Lyapunov equation 3.4.6. If $C_{k+1} = \widehat{\Sigma}$, then the computation of the inverse C_{k+1}^{-1} for the next iterate comes for free since it is a diagonal matrix. But even though, reducing the rank takes additional computational time, therefore, the reduction of the rank is only worth it if the rank r_{k+1} is sufficiently large. Hence, the compression in line 16 to 20 of Algorithm 4 is omitted for the first m sign function iterations until the rank r_{k+1} is reasonable large.

Algorithm 4 Structure exploiting sign function method for structured ALE

Require: Matrices $B, U, C, V, W, tol, \tau, m$.

Ensure: $Y \in \mathbb{R}^{N\times N}$ such that Y approx. solves structured ALE.

$B_0 \leftarrow B$
$U_0 \leftarrow U$
$V_0 \leftarrow V$
$C_0 \leftarrow C$
$W_0 \leftarrow W$
$k = 0$
while $\|B_k + U_k C_k V_k^T + I\| > tol$ **do**
  $A_k^{-1} \leftarrow B_k^{-1} - B_k^{-1}U_k \left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1} V_k^T B_k^{-1}$
  $c_k \leftarrow \sqrt{\frac{\|A_k^{-1}\|_F}{\|A_k\|_F}}$
  $W_{k+1} \leftarrow \frac{1}{2}\left(c_k W_k + c_k^{-1} A_k^{-T} W_k A_k^{-1}\right)$
  $B_{k+1} \leftarrow \frac{1}{2}\left(c_k B_k + c_k^{-1} B_k^{-1}\right)$
  $U_{k+1} \leftarrow \begin{bmatrix} U_k & B_k^{-1}U_k \end{bmatrix}$
  $V_{k+1} \leftarrow \begin{bmatrix} V_k & B_k^{-T}V_k \end{bmatrix}$
  $C_{k+1} \leftarrow \frac{1}{2}\text{blockdiag}\left(c_k C_k, -c_k^{-1}\left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1}\right)$
  if $k > m$ **then**
    Compute RR-QR: $Q_U R_U = U_{k+1}$, $Q_V R_V = V_{k+1}$.
    Compute truncated SVD: $\widehat{U}\widehat{\Sigma}\widehat{V}^T \approx R_U C_{k+1} R_V^T$ w.r.t. threshold τ.
    $U_{k+1} \leftarrow Q_U \widehat{U}$
    $V_{k+1} \leftarrow Q_V \widehat{V}$
    $C_{k+1} \leftarrow \widehat{\Sigma}$
  end if
  $k = k + 1$
end while
return $Y = \frac{1}{2} W_k$

4.3.2 Structure Exploiting Sign Function Method with low-rank rhs

We consider a structured algebraic Lyapunov equation 3.4.6, where the "right-hand side" $W = FF^T$ has low-rank, i.e.,

$$A^T X + XA = FF^T, \tag{4.28}$$

where $F \in \mathbb{R}^{N\times r_2}$ with $r_2 \ll n$, see [BQO99]. Furthermore, we assume as above that A can be expressed by a block diagonal matrix $B = \text{blockdiag}\left(B^{(1)}, \ldots, B^{(\ell)}\right)$ plus a low-rank perturbation UCV^T of B, i.e.,$A = B + UCV^T$, where $A, B \in \mathbb{R}^{N\times N}$, $C \in \mathbb{R}^{r_1\times r_1}$, $U, V \in \mathbb{R}^{N\times r_1}$ with $r_1 \ll n$. Since the matrix B is block diagonal, it is cheaply invertible.

By the stability assumption on A and Theorem 3.3.2, the Lyapunov equation (4.28) has a unique solution X. It can be factored as $X = YY^T$, where Y is a full-rank factors of X, i.e., $Y \in \mathbb{R}^{N \times \operatorname{rank}(X)}$ is a rectangular matrix. Even though the solution X of the structured algebraic Lyapunov equation (4.28) may be nonsingular, often its numerical rank or ε-rank [GV96] is very small [Gra04; Pen00b]. Hence, it can be approximated by $\widehat{Y} \in \mathbb{R}^{N \times N_y}$, where N_y is the numerical rank and, so that

$$\frac{\|X - \widehat{Y}\widehat{Y}^T\|_2}{\|X\|_2} \leq \varepsilon,$$

whereas the tolerance threshold determines the numerical rank. This observation has led to various methods to solve Lyapunov equations based on low-rank factorization of the solution [BQO99; LW02; Pen00a]. We use this methodology and rewrite the sign function iteration for W_{k+1} in (4.19) as full-rank factors,

$$F_{k+1}F_{k+1}^T = W_{k+1} = \frac{1}{2}\left(c_k W_k + c_k^{-1} A_k^{-T} W_k A_k^{-1}\right) = \frac{1}{2}\left(F_k F_k^T + \left(A_k^{-T} F_k\right)\left(A_k^{-T} F_k\right)^T\right),$$

where

$$F_{k+1} \;:=\; \tfrac{1}{\sqrt{2}}\begin{bmatrix} F_k & A_k^{-T} F_k \end{bmatrix}, \tag{4.29}$$

with $F_0 := F$.

Remark 4.3.2. *The workspace of F_{k+1} doubles per sign function iteration in* (4.29)*, since $F_k \in \mathbb{R}^{N \times p_k}$ and $F_{k+1} \in \mathbb{R}^{N \times 2p_k}$.*

In order to limit workspace, we compute a full-rank factorization in each iteration step based on a QR decomposition. This idea with scaling is summarized in Algorithm 5.

Algorithm 5 Full-rank factor Y of sign function method with low-rank rhs

Require: Matrices $B, U, C, V, F, tol, \tau_1, \tau_2$.

Ensure: Approximation to full-rank factor Y of the solution X.

$B_0 \leftarrow B$
$U_0 \leftarrow U$
$V_0 \leftarrow V$
$C_0 \leftarrow C$
$F_0 \leftarrow F$
$k = 0$
while $\|B_k + U_k C_k V_k^T + I\| > tol$ **do**
  $A_k^{-1} \leftarrow B_k^{-1} - B_k^{-1} U_k \left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1} V_k^T B_k^{-1}$
  $c_k \leftarrow \sqrt{\frac{\|A_k^{-1}\|_F}{\|A_k\|_F}}$
  $F_{k+1} \leftarrow \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{c_k} F_k & \frac{1}{\sqrt{c_k}} A_k^{-T} F_k \end{bmatrix}$
  Compress columns of F_{k+1} using a RR-QR with threshold τ_1.
  $B_{k+1} \leftarrow \frac{1}{2}\left(c_k B_k + c_k^{-1} B_k^{-1}\right)$
  $U_{k+1} \leftarrow \begin{bmatrix} U_k & B_k^{-1} U_k \end{bmatrix}$
  $V_{k+1} \leftarrow \begin{bmatrix} V_k & B_k^{-T} V_k \end{bmatrix}$
  $C_{k+1} \leftarrow \frac{1}{2}\text{blockdiag}\left(c_k C_k, -c_k^{-1}\left(C_k^{-1} + V_k^T B_k^{-1} U_k\right)^{-1}\right)$
  Compute RR-QR: $Q_U R_U = U_{k+1}$, $Q_V R_V = V_{k+1}$.
  Compute truncated SVD: $\widehat{U}\widehat{\Sigma}\widehat{V}^T \approx R_U C_{k+1} R_V^T$ w.r.t. threshold τ_2.
  $U_{k+1} \leftarrow Q_U \widehat{U}$
  $V_{k+1} \leftarrow Q_V \widehat{V}$
  $C_{k+1} \leftarrow \widehat{\Sigma}$
  $k = k + 1$
end while
return $Y = F_k$

4.3.3 Error Analysis of the structure-exploiting sign function method

In this section, we derive error bounds for the Algorithms 4 and 5. In the analysis, scaling is not taken into account and we neglect errors introduced by the truncated SVD. We consider the following basic definitions and general assumptions in analogy to the sign function method using hierarchical matrices, see e.g. [Gra01; GH07; BB06]. Any numerical computation underlies numerical errors caused by finite precision of computations involving floating-point or integer values. Hence, let $\widetilde{A}_k, \widetilde{W}_k$ and $\widetilde{F}_k$ denote the perturbed iterates arising during Algorithm 4 or 5, respectively. Furthermore, we denote with $\text{Inv}_{SMW}(A_k)$ the inversion by the Sherman-Morrison-Woodbury formula in (4.20).

Definition 4.3.3. *By δ we denote the maximal numerical error for using the Sherman-Morrison-Woodbury formula* (4.20)*,*

$$\delta = \max_{k=0,\dots,k_{max}} \|\mathrm{Inv}_{SMW}(\widetilde{A}_k) - \widetilde{A}_k^{-1}\|_2, \tag{4.30}$$

and for the distance of the exact and perturbed iterates we define

$$\eta_k := \|\widetilde{A}_k - A_k\|_2 \quad \text{for } k = 0, 1, \dots, k_{max}. \tag{4.31}$$

Assumption 4.3.4. *We assume with*

$$\Delta := \max_{k=0,\dots,k_{max}} \|A_k^{-1}\|_2 \tag{4.32}$$

that

$$\eta_k \Delta < 1 \quad \text{for } k = 0, 1, \dots, k_{max}.$$

Lemma 4.3.5. *With Definition 4.3.3 and Assumption 4.3.4, we obtain the following bound on the forward error of the perturbed iterates in line 8 of Algorithm 4 and in line 8 of Algorithm 5, respectively,*

$$\eta_{k+1} = \|\widetilde{A}_{k+1} - A_{k+1}\|_2 \leq \frac{1}{2}\left(\delta + \eta_k + \frac{\eta_k \Delta^2}{1 - \eta_k \Delta}\right).$$

Proof. By Assumption 4.3.4, we know $\left(I - A_k^{-1}\widetilde{A}_k\right)^k \to 0$ as $k \to \infty$. Hence, $A_k^{-1}\widetilde{A}_k$ can be expressed as a von Neumann series, since $\eta_k \Delta < 1$ for $k = 0, 1, \dots, k_{max}$, see e.g. [GV96] and for the inverse $\widetilde{A}_k^{-1}$ it holds,

$$\widetilde{A}_k^{-1} = \sum_{i=0}^{\infty} \left(A_k^{-1}\left(A_k - \widetilde{A}_k\right)\right)^i A_k^{-1}.$$

Hence,

$$\widetilde{A}_k^{-1} - A_k^{-1} = \sum_{i=1}^{\infty} \left(A_k^{-1}\left(A_k - \widetilde{A}_k\right)\right)^i A_k^{-1},$$

which can be bounded by using Assumption 4.3.4 on the geometric series,

$$\left\|\widetilde{A}_k^{-1} - A_k^{-1}\right\|_2 \leq \Delta \sum_{i=1}^{\infty} \left\|A_k^{-1}\left(A_k - \widetilde{A}_k\right)\right\|^i \leq \frac{\eta_k \Delta^2}{1 - \eta_k \Delta}.$$

Using Definition 4.3.3 yields

$$\begin{aligned}
\|\widetilde{A}_{k+1} - A_{k+1}\|_2 &\leq \tfrac{1}{2}\left\|\widetilde{A}_k - A_k\right\|_2 + \tfrac{1}{2}\left\|\mathrm{Inv}_{SMW}\left(\widetilde{A}_k\right) - A_k^{-1}\right\|_2 \\
&\leq \tfrac{1}{2}\left\|\widetilde{A}_k - A_k\right\|_2 + \tfrac{1}{2}\left\|\mathrm{Inv}_{SMW}\left(\widetilde{A}_k\right) - \widetilde{A}_k^{-1}\right\|_2 + \frac{1}{2}\left\|\widetilde{A}_k^{-1} - A_k^{-1}\right\|_2 \\
&\leq \tfrac{1}{2}\left(\delta + \eta_k + \frac{\eta_k \Delta^2}{1 - \eta_k \Delta}\right).
\end{aligned}$$

Hence, we have obtained the upper bound of the iterates. □

Theorem 4.3.6. *With the Assumption 4.3.4, the forward error for computing the approximate solution $\widetilde{W}_k$ in line 10 of Algorithm 4 can be bounded by*

$$\|\widetilde{W}_{k+1} - W_{k+1}\|_2 \leq \left(\tfrac{1+\Delta^2}{2} + \Theta_k^W\right) \|\widetilde{W}_k - W_k\|_2 + \Theta_k^W \|W_k\|_2, \tag{4.33}$$

where

$$\Theta_k^W = \Delta\delta + \frac{\delta^2}{2} + \frac{(\Delta+\delta)\,\eta_k\Delta^2}{1-\eta_k\Delta} + \frac{\eta_k^2\Delta^4}{2(1-\eta_k\Delta)^2}.$$

Proof. For notational simplicity $\mathrm{Inv}(A)$ denotes the inverse of A w.r.t. the Sherman-Morrison-Woodbury formula in (4.20), i.e., $\mathrm{Inv}(A) = \mathrm{Inv}_{SMW}(A)$. We then reformulate,

$$\begin{aligned}
&\mathrm{Inv}(\widetilde{A}_k^T)\widetilde{W}_k\mathrm{Inv}(\widetilde{A}_k) - A^{-T}W_kA_k^{-1} \\
&= \left(\mathrm{Inv}(\widetilde{A}_k^T) - \widetilde{A}_k^{-T}\right)\widetilde{W}_k\left(\mathrm{Inv}(\widetilde{A}_k) - \widetilde{A}_k^{-1}\right) + \widetilde{A}_k^{-T}\widetilde{W}_k\left(\mathrm{Inv}(\widetilde{A}_k) - \widetilde{A}_k^{-1}\right) \\
&\quad + \left(\mathrm{Inv}(\widetilde{A}_k^T) - \widetilde{A}_k^{-T}\right)\widetilde{W}_k\widetilde{A}_k^{-1} + \widetilde{A}_k^{-T}\widetilde{W}_k\widetilde{A}_k^{-1} - A^{-T}W_kA_k^{-1} \\
&= \left(\mathrm{Inv}(\widetilde{A}_k^T) - \widetilde{A}_k^{-T}\right)\widetilde{W}_k\left(\mathrm{Inv}(\widetilde{A}_k) - \widetilde{A}_k^{-1}\right) + \left(\widetilde{A}_k^{-T} - A_k^{-T}\right)\widetilde{W}_k\left(\mathrm{Inv}(\widetilde{A}_k) - \widetilde{A}_k^{-1}\right) \\
&\quad + A_k^{-T}\widetilde{W}_k\left(\mathrm{Inv}(\widetilde{A}_k) - \widetilde{A}_k^{-1}\right) + \left(\mathrm{Inv}(\widetilde{A}_k^T) - \widetilde{A}_k^{-T}\right)\widetilde{W}_k\left(\widetilde{A}_k^{-1} - A_k^{-1}\right) \\
&\quad + \left(\mathrm{Inv}(\widetilde{A}_k^T) - \widetilde{A}_k^{-T}\right)\widetilde{W}_kA_k^{-1} + \left(\widetilde{A}_k^{-T} - A_k^{-T}\right)\widetilde{W}_k\left(\widetilde{A}_k^{-1} - A_k^{-1}\right) \\
&\quad + \left(\widetilde{A}_k^{-T} - A_k^{-T}\right)\widetilde{W}_kA_k^{-1} + A_k^{-T}\widetilde{W}_k\left(\widetilde{A}_k^{-1} - A_k^{-1}\right) + A^{-T}\left(\widetilde{W}_k - W_k\right)A_k^{-1}.
\end{aligned}$$

Hence, we obtain with Assumption 4.3.4,

$$\begin{aligned}
&\left\|\mathrm{Inv}(\widetilde{A}_k^T)\widetilde{W}_k\mathrm{Inv}(\widetilde{A}_k) - A^{-T}W_kA_k^{-1}\right\|_2 \\
&\leq \left(\delta^2 + \tfrac{\delta\eta_k\Delta^2}{1-\eta_k\Delta^2} + \Delta\delta + \tfrac{\delta\eta_k\Delta^2}{1-\eta_k\Delta^2} + \Delta\delta + \tfrac{\eta_k^2\Delta^4}{(1-\eta_k\Delta^2)^2} + \tfrac{2\delta\eta_k\Delta^3}{1-\eta_k\Delta^2}\right)\left\|\widetilde{W}_k\right\|_2 + \Delta^2\left\|\widetilde{W}_k - W_k\right\|_2.
\end{aligned}$$

Therefore, we obtain,

$$\begin{aligned}
\left\|\widetilde{W}_{k+1} - W_{k+1}\right\|_2 &\leq \tfrac{1}{2}\left\|\widetilde{W}_k - W_k\right\|_2 + \tfrac{1}{2}\left\|\mathrm{Inv}(\widetilde{A}_k^T)\widetilde{W}_k\mathrm{Inv}(\widetilde{A}_k) - A_k^{-T}W_kA_k^{-1}\right\|_2 \\
&\leq \tfrac{1}{2}\left(1+\Delta^2\right)\left\|\widetilde{W}_k - W_k\right\|_2 + \Theta_k^W\left\|\widetilde{W}_k\right\|_2,
\end{aligned}$$

from which the bound (4.33) follows. □

Corollary 4.3.7. *With the Assumption 4.3.4, we obtain the following bound on the forward error for computing $\widetilde{W}_k$ in line 10 of Algorithm 4*

$$\|\widetilde{W}_{k+1} - W_{k+1}\|_2 \leq \sum_{j=0}^{k}\Theta_j^W\|W_j\|_2\prod_{i=j+1}^{k}\left(\frac{1+\Delta^2}{2} + \Theta_i^W\right),$$

and for the relative error we obtain

$$\frac{\|\widetilde{W}_{k+1} - W_{k+1}\|_2}{\|W_{k+1}\|_2} \leq \sum_{j=0}^{k} \Theta_j^W \prod_{i=j+1}^{k} \left(\frac{1+\Delta^2}{2} + \Theta_i^W \right),$$

where

$$\Theta_k^W = \Delta\delta + \frac{\delta^2}{2} + \frac{(\Delta+\delta)\,\eta_k\Delta^2}{1-\eta_k\Delta} + \frac{\eta_k^2\Delta^4}{2(1-\eta_k\Delta)^2},$$

for $k = 0, \ldots, k_{max}$.

Proof. We use that the iterates are bounded,

$$\|W_{k-1}\|_2 \leq \|W_k\|_2 \leq 2\sigma_1, \quad \text{for } k = 1, \ldots, k_{max},$$

where σ_1 is the largest singular value of the solution to the structured algebraic Lyapunov equation, i.e., $X = \frac{1}{2}\lim_{k\to} W_k$, [GV96]. □

Theorem 4.3.8. *With the Assumption 4.3.4, the forward error for computing the approximate full-rank factor* $\widetilde{F}_k$ *in line 10 of Algorithm 5 can be bounded by*

$$\|\widetilde{F}_{k+1} - F_{k+1}\|_2 \leq \left(\tfrac{1+\Delta}{\sqrt{2}} + \Theta_k^F \right) \|\widetilde{F}_k - F_k\|_2 + \Theta_k^F \|F_k\|_2, \tag{4.34}$$

where $\Theta_k^F = \frac{1}{\sqrt{2}} \left(\delta + \frac{\eta_k\Delta^2}{1-\eta_k\Delta} \right)$.

Proof. We prove the bound of the forward error of the approximate full-rank factor $\widetilde{F}_k$. As in the proof of Lemma 4.3.5, $A_k^{-1}\widetilde{A}_k$ can be expressed as a von Neumann series, see e.g. [GV96].

$$\begin{aligned}
\sqrt{2}\|\widetilde{F}_{k+1} - F_{k+1}\|_2 \leq & \|\widetilde{F}_k - F_k\|_2 + \|\mathrm{Inv}_{SMW}(\widetilde{A}_k^T)\widetilde{F}_k - A_k^{-T}F_k\|_2 \\
\leq & \|\widetilde{F}_k - F_k\|_2 + \|\mathrm{Inv}_{SMW}(\widetilde{A}_k^T)\widetilde{F}_k - \widetilde{A}_k^{-T}\widetilde{F}_k\|_2 + \|\widetilde{A}_k^{-T}\widetilde{F}_k - A_k^{-T}F_k\|_2 \\
\leq & \|\widetilde{F}_k - F_k\|_2 + \delta\|\widetilde{F}_k\|_2 + \|\widetilde{A}_k^{-T}\widetilde{F}_k - A_k^{-T}\widetilde{F}_k\|_2 + \|A_k^{-T}\widetilde{F}_k - A_k^{-T}F_k\|_2 \\
\leq & (1+\Delta)\,\|\widetilde{F}_k - F_k\|_2 + \delta\|\widetilde{F}_k\|_2 + \Delta\sum_{i=1}^{\infty} \left\| A_k^{-1}\left(A_k - \widetilde{A}_k\right) \right\|_2^i \|\widetilde{F}_k\|_2 \\
\leq & (1+\Delta)\,\|\widetilde{F}_k - F_k\|_2 + \left(\delta + \frac{\eta_k\Delta^2}{1-\eta_k\Delta} \right) \|\widetilde{F}_k\|_2,
\end{aligned}$$

from which (4.34) follows. □

Corollary 4.3.9. *With the Assumption 4.3.4, we obtain the following bound on the forward error for computing the approximate full-rank factor* $\widetilde{F}_k$ *in line 10 of Algorithm 5*

$$\|\widetilde{F}_{k+1} - F_{k+1}\|_2 \leq \sum_{j=0}^{k} \Theta_j^F \|F_j\|_2 \prod_{i=j+1}^{k} \left(\frac{1+\Delta}{\sqrt{2}} + \Theta_i^F \right),$$

and for the relative error we obtain

$$\frac{\|\widetilde{F}_{k+1} - F_{k+1}\|_2}{\|F_{k+1}\|_2} \leq \sum_{j=0}^{k} \Theta_j^F \prod_{i=j+1}^{k} \left(\frac{1+\Delta}{\sqrt{2}} + \Theta_i^F \right),$$

where $\Theta_k^F = \frac{1}{\sqrt{2}} \left(\delta + \frac{\eta_k \Delta^2}{1-\eta_k \Delta} \right)$ *for* $k = 1, \ldots, k_{max}$.

Proof. We use that the iterates are bounded,

$$\|F_{k-1}\|_2 \leq \|F_k\|_2 \leq 2\sigma_1, \qquad \text{for } k = 1, \ldots, k_{max},$$

where σ_1 is the largest singular value of the solution to the structured algebraic Lyapunov equation, i.e., $X = \frac{1}{2} \lim_{k\to} F_k F_k^T$, [GV96]. □

By the above analysis, we expect an increase of the errors in the W_k's and F_k's in each iteration proportional to δ and μ_k, where μ_k is again proportional to δ. Hence, it is sufficient to choose the threshold τ_1 for the RR-QR in line 11 of Algorithm 5 of the same order as δ. The bound for the Sherman-Morrison-Woodbury inversion error of Algorithm 4 can be controlled by an adaptive rank choice for the SVD w.r.t. threshold τ. The bound for the Sherman-Morrison-Woodbury inversion error of Algorithm 5 can be controlled by an adaptive rank choice for the RR-QR and the SVD w.r.t. thresholds τ_1 and τ_2, respectively.

4.4 Numerical Results

In this section we show numerical results concerning the eigenpairs of a QEP and the viscosity optimization w.r.t. the averaged total energy, namely we solve the problems (QEP) and (OPT 1). The eigenpairs of (QEP) are determined by the Ehrlich-Aberth iteration in Algorithm 2. The structure exploiting sign function method in Algorithm 4 is used as the basic ingredient to compute the averaged total energy by the structured algebraic Lyapunov equation of the NLP (OPT 1).

We consider two vibrational systems — a triple chain oscillator and a FEM model of a viscously damped beam. The first model is a triple chain oscillator from [BTT11], which is shown in Figure 4.1. The governing equations of motion for the triple-chain oscillator are given by a quadratic ordinary differential equation defined in Section 2.5.4,

$$Mx'' + Cx' + Kx = 0,$$

where mass matrix $M \in \mathbb{R}^{(3d+1)\times(3d+1)}$ and stiffness matrix $K \in \mathbb{R}^{(3d+1)\times(3d+1)}$ are

defined as

$$
\begin{aligned}
M &= \operatorname{diag}(m_1, \ldots, m_{3d+1}),\\
K &= \begin{bmatrix} K_{11} & & & -k_1 e_d \\ & K_{22} & & -k_2 e_d \\ & & K_{33} & -k_3 e_d \\ -k_1 e_d^T & -k_2 e_d^T & -k_3 e_d^T & k_1 + k_2 + k_3 + k_4 \end{bmatrix},\\
K_{ii} &= k_i \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix},
\end{aligned}
\tag{4.35}
$$

where $e_d \in \mathbb{R}^d$ is the d-th unit vector and $k_1, \ldots, k_4$ are given stiffnesses. We consider two configurations with 1000 and 1500 masses in each chain, i.e, in total 3001 and 4501 masses. Each chain has $d+1$ springs, which have the same stiffness. A damper with viscosity ν is attached in the middle of the first chain, i.e., at mass $m_{d/2}$. After transforming the quadratic ODE into first order form (see Section 4.2 and linearization of a QEP in Section 2.1.2), we obtain a first order system dimension of 6002 and 9002, respectively. In both configurations the masses are defined as $m_i = i$ for $i = 1, \ldots, 3d+1$, and the stiffnesses are defined as $k_1 = 1$, $k_2 = 50$, $k_3 = 100$ and $k_4 = 200$. The damping matrix is given as the sum of internal and external damping, i.e.,

$$C = C_{\text{int}} + \nu e_{d/2} e_{d/2}^T,$$

where $e_{d/2} \in \mathbb{R}^{3d+1}$ is the $\frac{d}{2}$-th unit vector and internal damping is given as modal damping $C_{\text{int}} = 2\gamma M^{1/2} \sqrt{M^{-1/2} K M^{-1/2}} M^{1/2}$ with $\gamma = \frac{1}{200}$.

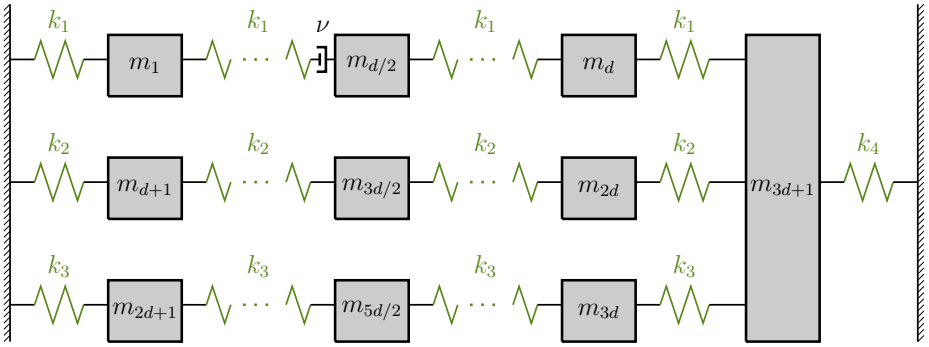


Figure 4.1: Oscillator with $3d+1$ masses, $d+4$ springs and a single viscous damper at mass $m_{d/2}$.

The second example is a slender beam, which is simply supported at both ends and viscously damped in the middle. It is shown in Figure 4.2. The example is taken from [Hig+08], but internal damping is added. The original example is part of the collection of nonlinear eigenvalue problems [Bet+]. The governing equation of motion for the transverse displacement $u(x,t)$ are given by the Euler-Bernoulli equation of the beam

$$\rho A \frac{\partial^2 u}{\partial t^2} + c(\nu) \frac{\partial u}{\partial t} + EI \frac{\partial^4 u}{\partial x^4} = 0,$$

where ρA is the mass unit per length, E is Young's modulus, I is the area moment of inertia of the cross-section, $c(\nu) = c_{\text{int}} + c_{\text{ext}}(\nu)$ represents damping which is given as internal damping c_{int} and external viscous damping $c_{\text{ext}}(\nu)$, where the viscosity fulfills $\nu \geq 0$. The boundary conditions are $u(0,t) = u_{xx}(0,t) = u(L,t) = u_{xx}(L,t) = 0$, where L is the length of the beam. Making the separation ansatz $u(x,t) = X(x)e^{\lambda t}$, we obtain the boundary value problem

$$\lambda^2 \rho A X(x) + \lambda c(\nu) X(x) + EI \cdot X^{(4)}(x) = 0, \tag{4.36}$$

with $X(0) = X''(0) = X(L) = X''(L) = 0$. We discretize the boundary value problem by finite elements using cubic Hermite polynomials as interpolation shape functions,

$$\left(\lambda^2 M + \lambda C(\nu) + K\right) v = 0,$$

which is a quadratic eigenproblem and which is exactly the first optimization problem (QEP). As we have discussed in Section 2.5.4, the quadratic eigenproblem can be transformed with $x(t) = Ve^{Jt}c$ to a corresponding quadratic ordinary differential

$$Mx'' + C(\nu)x' + K = 0,$$

where (V, J) is a Jordan pair and $c \in \mathbb{C}^{2n}$ is a vector of arbitrary constants. The damping matrix C is the sum of internal and external damping,

$$C = C_{\text{int}} + C_{\text{ext}},$$

where internal damping is given as modal damping $C_{\text{int}} = 2\gamma M^{1/2}\sqrt{M^{-1/2}KM^{-1/2}}M^{1/2}$ with $\gamma = \frac{1}{200}$ and external damping is given as viscous damping $C_{\text{ext}} = \nu e_{n/2} e_{n/2}^T$. We compute the averaged total energy for the geometric and material properties $E = 7 \times 10^{10} \frac{N}{m^2}$, $I = \frac{0.05 \times 0.005^3}{12} m^4$, $L = 1m$ and $\rho A = 0.674kg$ as in [Hig+08]. We used 3000 and 4000 finite elements for the beam, which results in a first order system $x' = Ax$ such that the matrix A is of size 6000×6000 and 8000×8000, respectively.

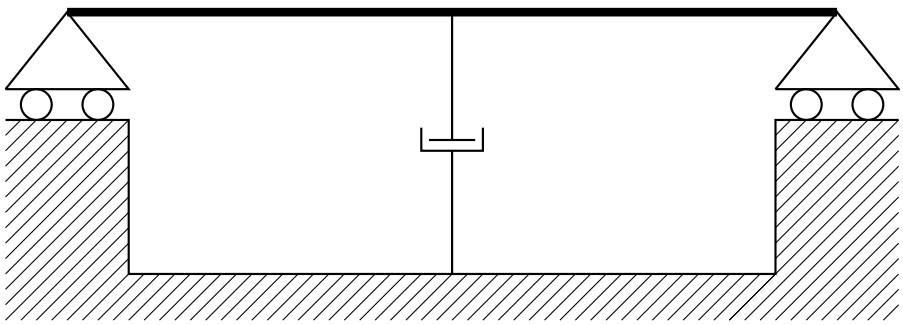


Figure 4.2: Viscously damped beam.

4.4.1 Numerical results for (QEP)

We solve the first problem (QEP), i.e., we solve the quadratic eigenproblem of the vibrational system for fixed viscosities ν_i for $i = 1, \ldots, r$. Hence, the full spectrum of

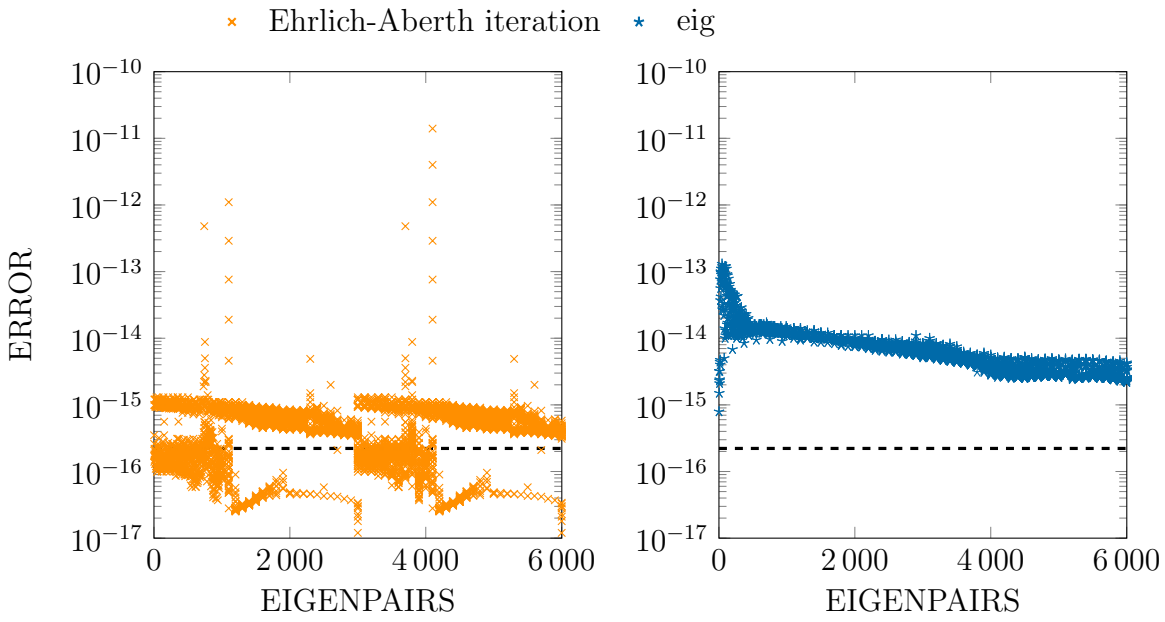


Figure 4.3: Backward error for eigenpairs computed by EAI (left) and *eig* from MATLAB (right) for triple-chain oscillator with 3001 masses.

the vibrational system is computed. In order to judge the quality of the eigenvalues we compute the eigenvectors and the backward errors of the eigenpairs. The eigenpairs are once computed on the linearization of the QEP as described in Section 2.5.4 by the *eig* function from MATLAB, which implements the QZ algorithm. Here, we used scaled matrices in the linearization, which have been introduced in [FLVD04]. The importance of scaling is shown w.r.t. accuracy and stability in [Hig+08].
We compute the eigenvectors in Algorithm 2 by a single inverse iteration for each eigenvalue since this is often sufficient as the backward errors in Figures 4.3, 4.4, 4.5 and 4.6 for the triple-chain oscillator and the viscously damped beam. The numbering of the eigenpairs in Figures 4.3, 4.4, 4.5 and 4.6 is artificial, i.e., the numbering of the backward errors of Algorithm 2 and the *eig* function from MATLAB do in general not correspond to the same eigenpair. Hence, in these figures the backward errors cannot be compared individually. But overall, for Algorithm 2 the backward errors are sufficiently good and close to machine precision. In most instances they are much smaller than the backward error for the *eig* function and in some rare cases they are worse. The backward errors can still be improved by further employing the inverse iteration on the eigenvectors. But the most promising feature of Algorithm 2 can be observed in Table 4.1, the Ehrlich-Aberth iteration with inverse iteration is much faster then the *eig* function from MATLAB.

4.4.2 Numerical results for NLP (OPT 1)

We minimize the averaged total energy for a vibrational system by transforming the above quadratic ODE into first order form (see Section 4.2). Hence, we solve the NLP (OPT 1), where the rank of the initial perturbation UCV^T for the structure-exploiting

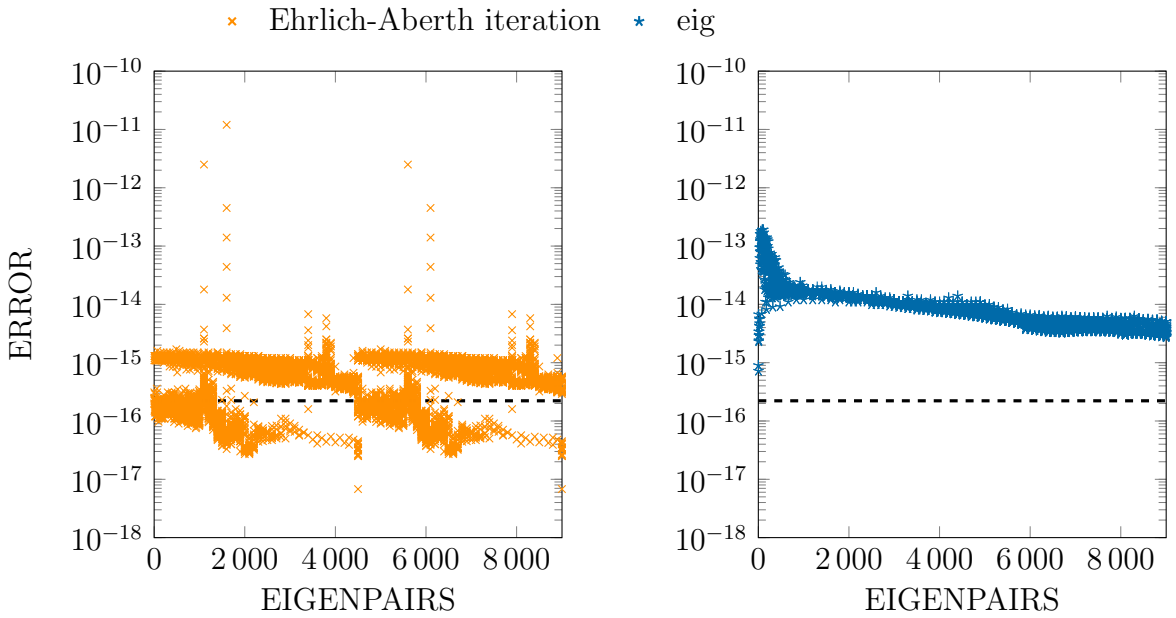


Figure 4.4: Backward error for eigenpairs computed by EAI (left) and *eig* from MATLAB (right) for triple-chain oscillator with 4501 masses.

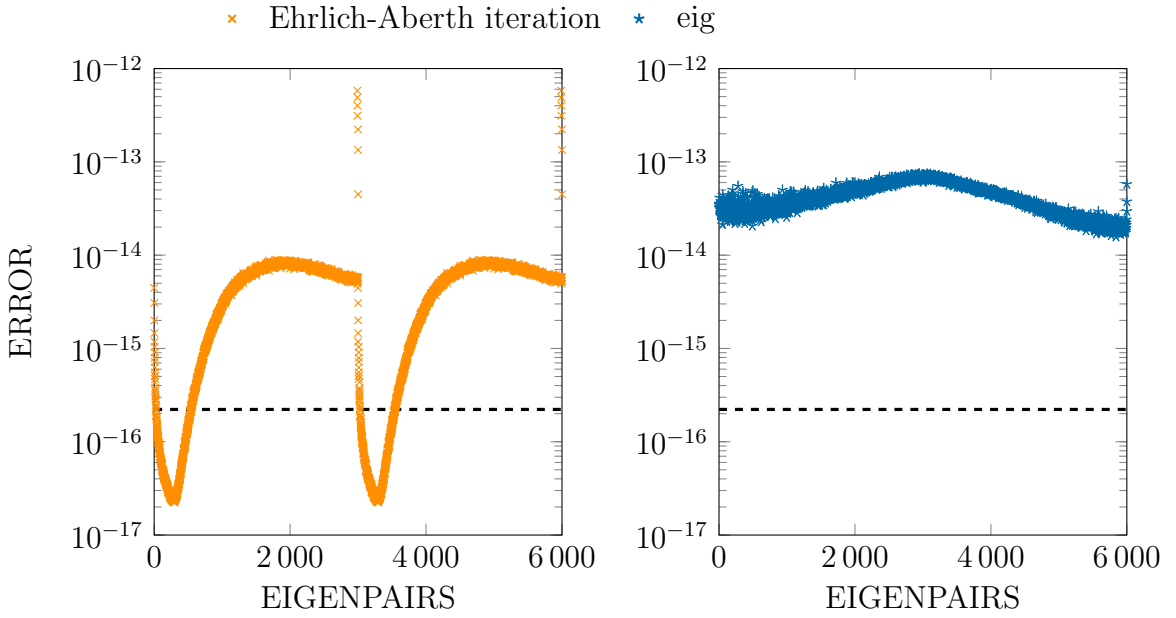


Figure 4.5: Backward error for eigenpairs computed by EAI (left) and *eig* from MATLAB (right) for beam with 3000 finite elements.

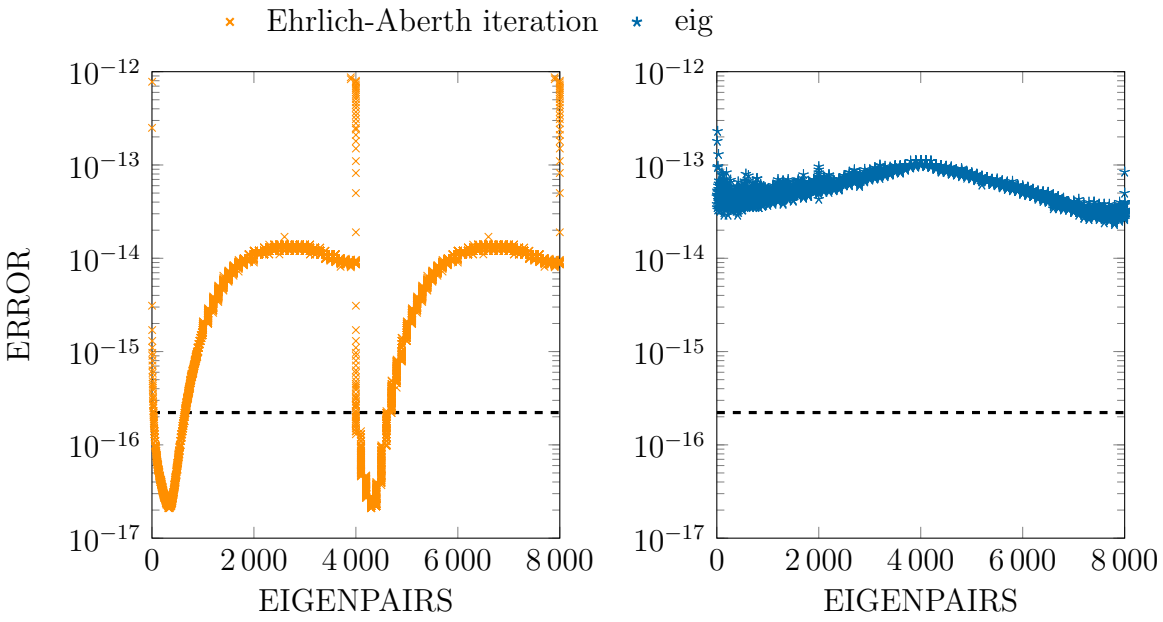


Figure 4.6: Backward error for eigenpairs computed by EAI (left) and *eig* from MATLAB (right) for beam with 4000 finite elements.

Example	**EAI**	***eig* from MATLAB**
oscillator, $n = 6002$	30.1779	5542.4930
oscillator, $n = 9002$	65.9200	29809.5515
beam, $n = 6000$	31.8541	5120.4564
beam, $n = 8000$	60.4226	17197.5731

Table 4.1: Running times for computation of eigenpairs by EAI incl. single inverse iteration and *eig* function from MATLAB for various examples.

sign function method defined in Algorithm 4 is exactly the number of viscous dampers. Due to the result from [Ves90] the global optimization of the viscosities in (OPT 1) can be performed by the NLP solver *fmincon* from MATLAB, which implements an interior point algorithm. During the iteration of *fmincon* the solution of the structured algebraic Lyapunov equation and its gradient $\frac{\partial \operatorname{tr} X(\nu)}{\partial \nu_j}$ given in (4.8) for $j = 1, \ldots, r$ was provided. We computed the solution of the structured algebraic Lyapunov equation by the structure-exploiting sign function method defined in Algorithm 4. We chose as a stopping criterion $\|A_k + I\|_F \leq 10^{-4}$ and accelerated the convergence with Frobenius scaling, i.e., $c_k^2 = \|A_k^{-1}\|_F / \|A_k\|_F$ for the first 12 sign function iterations. We omitted the rank compression in line 16 to 20 of Algorithm 4 for all numerical examples for the first seven iterations of Algorithm 4, i.e., as a parameter $m = 7$ was chosen. The results for the structure-exploiting sign function method defined in Algorithm 4 with $\varepsilon = 10^{-4}$ are shown in Table 4.2.

Obviously, the standard and the structure-exploiting sign function method defined in Algorithms 3 and 4 do not compute exactly the same results. Hence, the NLP solver *fmincon* vary for both methods. Therefore, the corresponding running times

Example	**structure-exploiting SFM with $\varepsilon = 10^{-4}$**		
	tot. energy	**#*fmincon*-iteration**	**time**
oscillator $n = 6002$	$9.99 \cdot 10^{6}$	10	$1.54 \cdot 10^{5}$
oscillator $n = 9002$	$1.98 \cdot 10^{7}$	15	$1.49 \cdot 10^{6}$
beam $n = 6000$	$3.10 \cdot 10^{-1}$	13	$1.52 \cdot 10^{5}$
beam $n = 8000$	$3.11 \cdot 10^{-1}$	20	$8.83 \cdot 10^{5}$

Table 4.2: Global optimal solutions to NLP (5.24) obtained by NLP solver *fmincon* from MATLAB with structure-exploiting sign function method defined in Algorithm 4 with $\varepsilon = 10^{-4}$.

highly depend on the in particular chosen example, especially the number of *fmincon*-iterations. Therefore, we investigate in the following the running times for the standard and the structure-exploiting sign function method defined in Algorithms 3 and 4 for fixed viscosities.

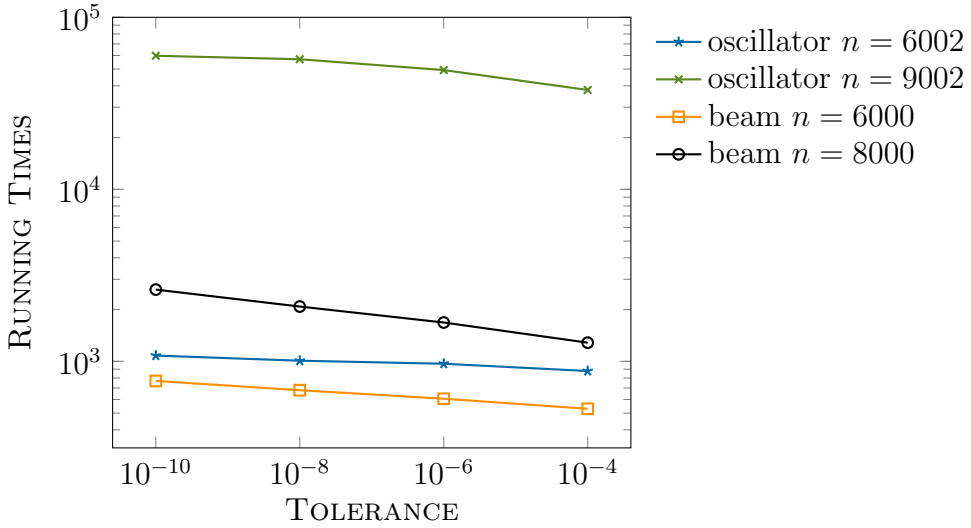


Figure 4.7: Running times of structure-exploiting sign function method (Algorithm 4) w.r.t. truncation tolerance ε.

In Figure 4.7 the running times of the structure-exploiting sign function method w.r.t. truncation tolerance ε for the SVD are shown. As expected, the running time of the structure-exploiting SFM increases when the truncation tolerance ε for the SVD is decreased.
In Figure 4.8 the speedup of the structure-exploiting sign function method (Algorithm 4) vs. the standard sign function method (Algorithm 3) is shown. Based on the truncation tolerance between 7-56% computational time is saved by applying the structure-exploiting SFM described in Algorithm 4 instead of the standard SFM defined in Algorithm 3.

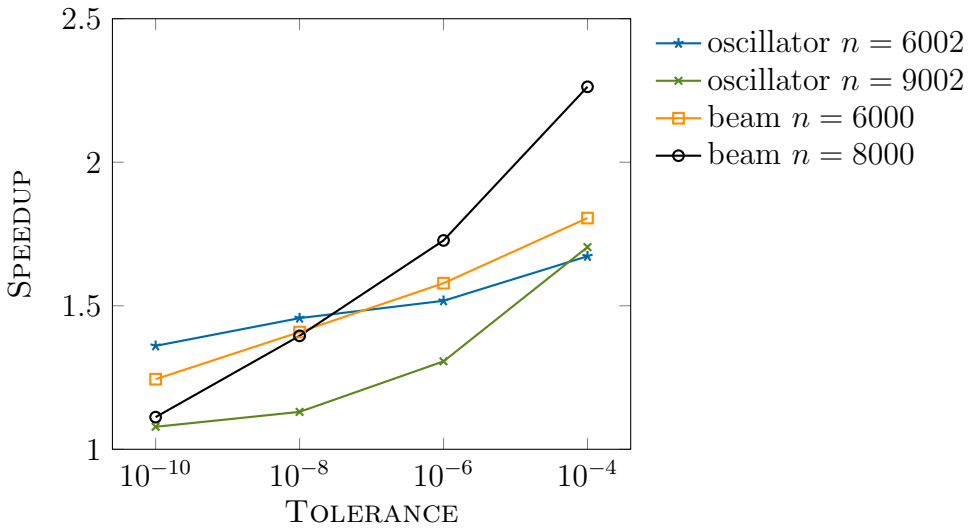


Figure 4.8: Speedup of structure-exploiting sign function method (Algorithm 4) w.r.t. truncation tolerance ε.

5

Placement of Viscous Dampers

In this chapter we aim at reducing the vibrations

$$x_k = e^{\lambda_k t} = e^{d_k t}\left(\cos(t\omega_k)\,\mathrm{Re}(v_k) - \sin(t\omega)\,\mathrm{Im}(v_k)\right)$$

given in (3.3) for $k = 1, \ldots, 2n$ of the vibrational system (3.1) by r viscous dampers at once. Here, the r external dampers have to be placed and their viscosities have to be optimized. Hence, we investigate the placement problem for r viscous dampers that was introduced in Section 3.4.4, namely the minimization of the averaged total energy w.r.t. the placement of r viscous dampers (OPT 2),

$$\begin{aligned}
\min_{\nu} \quad & \mathrm{tr}(X) \\
\text{s.t.} \quad & A(\nu)^T X + X A(\nu) = -I, \\
& 0 \le \nu_i \le \nu_{\max}, && i = 1, \ldots, n, \\
& b_i \in \{0, 1\}, && i = 1, \ldots, n, \\
& \nu_i \le b_i \nu_{\max}, && i = 1, \ldots, n, \\
& \sum_{i=1}^{n} b_i \le r, \\
\text{where } & A(\nu) = B - \widehat{V} D(\nu) \widehat{V}^T, \\
& B = \mathrm{blockdiag}(B^{(1)}, B^{(2)}, \ldots, B^{(n)}), \\
& B^{(i)} = \begin{bmatrix} 0 & \omega_i \\ -\omega_i & -2\gamma\omega_i \end{bmatrix}, && i = 1, \ldots, n, \\
& \widehat{V} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ v_{11} & v_{12} & \ldots & v_{1n} \\ 0 & 0 & \ldots & 0 \\ v_{21} & v_{22} & \ldots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \\ v_{n1} & v_{n2} & \ldots & v_{nn} \end{bmatrix} \in \mathbb{R}^{2n \times n}, \\
& D(\nu) = \mathrm{diag}(\nu_1, \ldots, \nu_n).
\end{aligned}$$

In Chapter 4 we have established the theory to determine the optimal viscosity w.r.t. the averaged total energy. In this chapter we do not only want to pose the question of how to damp a vibrational system, but also where a damper should be placed. Therefore, we consider a set of external damper positions, but the number of realizations is restricted. Hence, this chapter is an extension to Chapter 4.
This chapter is structured as follows. In Section 5.1 we briefly introduce the basic concepts for solving Mixed Integer Nonlinear Programs (MINLPs). In Section 5.2 we introduce and derive properties of the viscous damper placement problem and encode it as a Mixed Integer Nonlinear Program. Linearization techniques based on McCormick envelopes and piecewise linear functions are introduced in Section 5.3 and Section 5.4. A novel ℓ_1-penalization heuristic for the viscous damper placement problem is presented in Section 5.5, which is applicable to medium-scale problems. In Section 5.6 we derive an efficient algorithm that determines the optimal damper positions in $\mathcal{O}(n^2)$ for sufficiently small viscosities, i.e., in this case we do not need the MINLP formulation and its linearizations for the viscous damper placement problem. Finally, we show numerical results in Section 5.7.

5.1 Algorithmic Treatment of Solving MINLPs

We briefly discuss the algorithmic treatment of solving Mixed Integer Nonlinear Programs (MINLPs) of the form (2.51). We therefore introduce the basic principle for solving MINLPs: relaxation and constraint enforcement.

Relaxation

Enlarging the feasible set of the MINLP is called a relaxation. We are interested in relaxations that are substantially easier to solve than the MINLP itself e.g. by integrality relaxations. Via a relaxation a lower bound on the optimal solution of (2.51) is computed. Together with an upper bound that can be obtained from any feasible point, relaxations allow us to terminate the search for a solution whenever the lower bound is larger than the current upper bound.

Constraint Enforcement

After the feasible set is enlarged by a relaxation, we have to exclude solutions that are feasible to the relaxation, but not to the original MINLP. This procedure is called constraint enforcement which can be accomplished by refining or tightening the relaxation, often by adding valid inequalities (cuts), or by branching, where the relaxation is divided into two or more separate problems. (x, y) is called a feasible solution of the MINLP (2.51) if $g_i(x, y) \leq 0$ for $i \in I$. If $y \in \mathbb{Z}^{n_y}$, then y is called integral, i.e., $y_i \in \mathbb{Z}$ for $i = 1, \ldots, n_y$. Via branching we obtain some form of search tree, i.e., if the solution (x, y) is feasible but y is not integral, i.e., there exists a non-integral variable e.g. $y_i \notin \mathbb{Z}$ and then we branch on it. Branching introduces two new child nodes in the search tree. The first child has a new upper bound on the branching variable $u_i = \lfloor y_i \rfloor$ and the second child has a new lower bound $l_i = \lceil y_i \rceil$.

Remark 5.1.1. *Matrix constraints such as the algebraic Lyapunov equation cannot be solved by general MINLP solvers and the aforementioned methods cannot be applied. Hence, matrix constraints have to be encoded in terms of the Kronecker product if possible in order to solve them by a linear system solver, which is a subroutine of any MINLP/MILP solver.*

5.2 Mixed Integer Nonlinear Programming Formulation

In this chapter we discuss the problem to find the optimal set of indices $\{j_1, \ldots, j_r\}$ and viscosities $\{\nu_{j_1}, \ldots, \nu_{j_r}\}$ such that optimal evanescence w.r.t. the averaged total energy is ensured. While in the optimization problem (OPT 1) the positions of the r external dampers were fixed, we now consider n viscous dampers in the parameter dependent structured algebraic Lyapunov equation $A(\nu)^T X + XA(\nu) = -I$ in (OPT 2). Hence, the objective functions in (OPT 1) and (OPT 2) are different. The objective function of the optimization problem (OPT 2) can be formulated as $f : \mathbb{R}^n \to \mathbb{R}$,

$$f : \nu \mapsto \operatorname{tr}\left(\int_0^\infty e^{A(\nu)^T t} e^{A(\nu)t} \mathrm{d}t\right), \tag{5.1}$$

where $A(\nu)$ is defined in (OPT 2).

Lemma 5.2.1. *The mapping $f : \mathbb{R}^n \to \mathbb{R}$ as in* (5.1) *is twice continuously differentiable, i.e., $f \in \mathcal{C}^2$.*

Proof. The proof of Lemma 5.2.1 can be carried out as in Lemma 4.2.1. The first and second partial derivatives of f are given in (4.8) and (4.10), where $A(\nu) = B - \widehat{V} D(\nu) \widehat{V}^T$ is now different, since $\widehat{V} \in \mathbb{R}^{2n \times n}$ and $D(\nu) \in \mathbb{R}^{n \times n}$ are now defined in (OPT 2). □

The state space matrix $A(\nu)$ and therefore, the objective function f has changed w.r.t. Chapter 4, but their qualitative behavior as a nonlinear function has not changed. Since we have introduced binary variables $b_i \in \{0, 1\}$ for $i = 1, \ldots, n$ in (OPT 2), the optimization problem (OPT 2) is a MINLP.

Corollary 5.2.2. *The optimization problem* (OPT 2) *is a MINLP.*

Proof. For theoretical investigation we use the integral formulation of the solution of the algebraic Lyapunov equation (3.42) by Lemma 3.3.3, i.e., $X(\nu) = \int_0^\infty e^{A(\nu)^T t} e^{A(\nu)t} \mathrm{d}t$, which is obviously nonlinear. Since we have binary variables $b_i \in \{0, 1\}$ for $i = 1, \ldots, n$, the optimization problem (OPT 2) is obviously a MINLP:

$$\begin{aligned}
\min \quad & \int_0^\infty \operatorname{tr}\left(e^{A(\nu)^T t} e^{A(\nu)t}\right) \mathrm{d}t \\
\text{s.t.} \quad & -\nu_i \leq 0, && \text{for } i = 1, \ldots, n, \\
& \nu_i - \nu_{\max} \leq 0 && \text{for } i = 1, \ldots, n, \\
& b_i \in \{0, 1\}, && i = 1, \ldots, n, \\
& \nu_i \leq b_i \nu_{\max}, && i = 1, \ldots, n, \\
& \sum_{i=1}^n b_i \leq r.
\end{aligned}$$

□

In Section 4.2, the objective function was in general non-convex for more than one external damper. Now in this chapter we enlarge the number of external dampers, but the objective function f defined in (5.1) remains non-convex as shown by an example in [Bra98]. Global optimization e.g. in [Han92; HT96] of f is challenging since f is non-convex and so many local minima may exist.

By Corollary 5.2.2 the optimization problem (OPT 2) is a MINLP. For theoretical investigation we have considered the objective function $f : \mathbb{R}^n \to \mathbb{R}$ defined in (5.1), which was nonlinear. But this representation is not beneficial for numerical computations and hence, we use its equivalent formulation as the solution of an algebraic Lyapunov equation as in (OPT 2). Therefore, the type of nonlinearity for the optimization problem (OPT 2) is not obvious, since it is due to the Lyapunov equation via the product of its solution X and the viscosities $\nu_1, \ldots, \nu_n$ of the system matrix $A(\nu)$. So we have a closer look at the nonlinearity and investigate $A(\nu)$, which can be rewritten as

$$A(\nu) = B - \widehat{V} D(\nu) \widehat{V}^T = B - \sum_{i=1}^{n} \nu_i \widehat{V}_i \widehat{V}_i^T,$$

where $\widehat{V}_i$ is the i-th column of $\widehat{V}$ and $\widehat{V}_i^T$ is its transpose. Due to Remark 5.1.1 matrix constraints cannot be solved by a general MINLP solver. A way out is the reformulation by the Kronecker product. We then transform the structured algebraic Lyapunov equation $A(\nu)^T X + X A(\nu) = -I$ via the Kronecker product (3.9) into

$$\begin{aligned} \left(I_{2n} \otimes A(\nu)^T + A(\nu) \otimes I_{2n}\right) \operatorname{vec}(X) &= \operatorname{vec}(I), \\ \left(I_{2n} \otimes \left(B^T - \sum_{i=1}^{n} \nu_i \widehat{V}_i \widehat{V}_i^T\right) + \left(B - \sum_{i=1}^{n} \nu_i \widehat{V}_i \widehat{V}_i^T\right) \otimes I_{2n}\right) \operatorname{vec}(X) &= \operatorname{vec}(I), \end{aligned}$$

which can be rewritten as

$$\left(\mathcal{A}^{(0)} - \sum_{i=1}^{n} \nu_i \mathcal{A}^{(i)}\right) \operatorname{vec}(X) = \operatorname{vec}(I), \tag{5.2}$$

where $\mathcal{A}^{(0)} = I_{2n} \otimes B^T + B \otimes I_{2n}$ and $\mathcal{A}^{(i)} = I_{2n} \otimes \widehat{V}_i \widehat{V}_i^T + \widehat{V}_i \widehat{V}_i^T \otimes I_{2n}$ for $i = 1, \ldots, n$ and $\widehat{V}_i \widehat{V}_i^T \in \mathbb{R}^{2n \times 2n}$ is given as

$$\widehat{V}_i \widehat{V}_i^T = \begin{bmatrix} 0 \\ v_{1i} \\ 0 \\ v_{2i} \\ \vdots \\ 0 \\ v_{ni} \end{bmatrix} \begin{bmatrix} 0 \\ v_{1i} \\ 0 \\ v_{2i} \\ \vdots \\ 0 \\ v_{ni} \end{bmatrix}^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & v_{1i}^2 & 0 & v_{1i}v_{2i} & 0 & \cdots & 0 & v_{1i}v_{ni} \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & v_{2i}v_{1i} & 0 & v_{2i}^2 & 0 & \cdots & 0 & v_{2i}v_{ni} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & v_{ni}v_{1i} & 0 & v_{ni}v_{2i} & 0 & \cdots & 0 & v_{ni}^2 \end{bmatrix}.$$

Hence, $\mathcal{A}^{(i)} \in \mathbb{R}^{4n^2 \times 4n^2}$ is given as

$$\mathcal{A}^{(i)} = \begin{bmatrix} \widehat{V}_i\widehat{V}_i^T & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & v_{1i}^2 I_{2n} + \widehat{V}_i\widehat{V}_i^T & 0 & v_{1i}v_{2i}I_{2n} & 0 & \cdots & 0 & v_{1i}v_{ni}I_{2n} \\ 0 & 0 & \widehat{V}_i\widehat{V}_i^T & 0 & 0 & \cdots & 0 & 0 \\ 0 & v_{2i}v_{1i}I_{2n} & 0 & v_{2i}^2 I_{2n} + \widehat{V}_i\widehat{V}_i^T & 0 & \cdots & 0 & v_{2i}v_{ni}I_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \widehat{V}_i\widehat{V}_i^T & 0 \\ 0 & v_{ni}v_{1i}I_{2n} & 0 & v_{ni}v_{2i}I_{2n} & 0 & \cdots & 0 & v_{ni}^2 I_{2n} + \widehat{V}_i\widehat{V}_i^T \end{bmatrix},$$

for $i = 1, \ldots, n$. We considered the structured algebraic Lyapunov equation in (5.2). An equivalent formulation for an algebraic Lyapunov equation without the perfect shuffle permutation (3.33) has been derived in [Tru04].
We can rewrite (5.2) as a bilinear map $b : \mathbb{R}^n \times \mathbb{R}^{4n^2} \to \mathbb{R}^{4n^2}$,

$$b(\nu, x) := \mathcal{A}^{(0)}x - \sum_{i=1}^{n} \nu_i \mathcal{A}^{(i)}x - \operatorname{vec}(I), \tag{5.3}$$

where $x = \operatorname{vec}(X) \in \mathbb{R}^{4n^2}$ such that $b(\nu, x) = 0$. The bilinearity is given in $\nu_i\mathcal{A}^{(i)}x$ for $i = 1, \ldots, n$. Hence, we obtain the following corollary.

Corollary 5.2.3. *$A(\nu)^T X + XA(\nu) = -I$ in* (OPT 2) *can be rewritten as $b(\nu, x) = 0$, where $b(\cdot, \cdot)$ is defined in* (5.3)*. Hence, the optimization problem* (OPT 2) *is nonlinear due to the bilinear constraint $b(\nu, x) = 0$.*

By Remark 5.1.1, the algebraic Lyapunov equation and the respective bilinear constraint $b(\nu, x) = 0$, where $b(\cdot, \cdot)$ is defined in (5.3), have to be encoded as bilinear constraints. We then obtain the following $4n^2$ bilinear constraints which include $4n^3$ bilinear products, namely $\nu_i x_j$,

$$\sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{k=1}^{4n^2} a_{kj}^{(i)} \nu_i x_j = \delta_{k \bmod (2n+1),1}, \tag{5.4}$$

where $a_{kj}^{(i)} = \left(\mathcal{A}^{(i)}\right)_{kj}$ for $i = 0, \ldots, n$, $j, k = 1, \ldots, 4n^2$ and $\delta_{k \bmod (2n+1),1}$ is the Kronecker delta.

Remark 5.2.4. *Let $a_p^{(i)}$ denote the p-th column of $\mathcal{A}^{(i)}$, where $(i, p) \in N \times M$. Due to the structure of $\widehat{V}_i\widehat{V}_i^T$ and $\mathcal{A}^{(i)}$ for $i = 1, \ldots, n$, it follows that n columns of $\mathcal{A}^{(i)}$ are zero columns, i.e., $a_{4n(j-1)+2k-1}^{(i)} = 0$, where $(j, k) \in N \times N$. Hence, the bilinear constraint $b(\nu, x) = 0$, where the bilinear map $b(\cdot, \cdot)$ is defined in* (5.3)*, can be encoded by $4n^2$ bilinear constraints given in* (5.4)*, which contain overall $3n^3$ bilinear products $\nu_i \cdot x_{4n(j-1)+2k}$ and $\nu_i \cdot x_{4nj+\ell}$, where $i, j, k = 1, \ldots, n$ and $\ell = 1, \ldots, 2n$.*

For notational simplicity we omit the observation of Remark 5.2.4 in the MINLP (5.5), which we derive in the following. In Section 5.1 we have seen that one needs bounds on the variables for solving MINLPs by a general MINLP solver. Hence, we assume that

bounds on the solution of the algebraic Lyapunov equations are given, $x_j^\ell \leq x_j \leq x_j^u$, where $j \in M = \{1, \ldots, 4n^2\}$. [AMK96] surveys bounds for the solution of an algebraic Lyapunov equation, but most bounds are not applicable, since a major assumption is the stability for the symmetric part of the system matrix, which does not hold for the structured algebraic Lyapunov equation (3.43). A bound which does not need this assumption can be found e.g. in [PS08].
We mention that the trace of the matrix X in terms of the vectorization operation can be rewritten as

$$\operatorname{tr}(X) = \sum_{i=1}^{2n} x_{2n(i-1)+i},$$

where $x = \operatorname{vec}(X)$. In order to apply general MINLP solvers to the viscous damper placement problem (OPT 2), we rewrite it as the following MINLP:

$$\begin{aligned}
\min_{\nu} \quad & \sum_{i=1}^{2n} x_{2n(i-1)+i} \\
\text{s.t.} \quad & \sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} \nu_i x_j = \delta_{k \bmod (2n+1),1}, && k \in M, \\
& 0 \leq \nu_i \leq \nu_{\max}, && i \in N, \\
& b_i \in \{0,1\}, && i \in N, \\
& \nu_i \leq b_i \nu_{\max}, && i \in N, \\
& x_j^\ell \leq x_j \leq x_j^u, && j \in M, \\
& \sum_{i=1}^{n} b_i \leq r, \\
& \text{where } N = \{1, \ldots, n\} \text{ and } M = \left\{1, \ldots, 4n^2\right\}.
\end{aligned} \tag{5.5}$$

We state the main drawback of the MINLP formulation (5.5) in the following remark.

Remark 5.2.5. *Let the viscosities ν be fixed. Then a direct solve of equation* (5.2) *or* (5.4)*, respectively, by e.g. Gaussian elimination or LU decomposition, has a complexity of $\mathcal{O}(n^6)$, since the matrices $\mathcal{A}^{(i)}$ are of size $4n^2 \times 4n^2$ for $i = 0, \ldots, n$. Therefore, solving the MINLP* (5.5) *is very demanding, which is the main drawback of the encoding in terms of the bilinear map in* (5.2)*. Hence, computation of the solution to the MINLP* (5.5) *by a general MINLP solver is limited to small-scale problems.*

We compare the computational complexity for computing solutions to the algebraic Lyapunov equation by various methods. The main drawback of the MINLP formulation (5.5) is its computational complexity given in Remark 5.2.5. For solving algebraic Lyapunov equations, direct methods such as the Bartels-Stewart method [BS72] and Hammarling's method [Ham82] are based on transforming the coefficient matrix into Schur form. Both methods have cubic computational complexity since the computational complexity is dominated by transforming the coefficient matrix into Schur form. Moreover, iterative algorithms for the solution of algebraic Lyapunov equations exist.

In Chapter 4 various iterative methods for algebraic Lyapunov equation were introduced, namely the sign function method in Algorithm 3, the structure exploiting sign function method in Algorithm 4 and the structure exploiting sign function method with low-rank rhs in Algorithm 5. A further iterative method is the Alternating Direction Implicit (ADI) method, which has drawn a lot of attention for solving algebraic Lyapunov equation and has been investigated in the context of damping optimization in [TV09].

We conclude this section with an outlook on the following topics which we will discuss in this chapter. The reformulated objective function f defined in (5.1) is non-convex as mentioned above and so is the optimization problem (OPT 2) and its reformulation as a MINLP (5.5). Even when the integer decision variables are relaxed to be continuous, the feasible region may be non-convex. Non-convex Mixed Integer Nonlinear Programming is much more challenging since then the continuous relaxation of integer decision variables is still a global optimization problem [Han92; HT96]. Therefore, an efficiently solvable convex relaxation for the branch-and-bound framework is needed. An outer approximation of the feasible set of the MINLP is obtained by McCormick envelopes in Section 5.3. The outer approximation (relaxation) then turns out to be convex and polyhedral, i.e., it is a Mixed Integer Linear Program (MILP), which can be solved efficiently. We test in Section 5.7 if this outer approximation is tight enough such that it yields further information for positioning external dampers from solving the resulting MILP.

In Section 5.4 another idea is employed, namely the approximation of the nonlinearity (5.3). We add artificial functions in order to model the bilinear product $\nu_i \cdot x_j$ for $i, j, k = 1, \ldots, n$. These functions are then separable and hence, they can be approximated by piecewise linear functions such that the resulting approximation is a MILP as well.

5.3 McCormick Envelopes

We describe the basic idea of McCormick envelopes [McC76] of a bilinear function. Here, the bilinear function is the product of two variables, namely $\nu_i \cdot x_j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, 4n^2$. First, we introduce a new and artificial variable, which is defined as

$$y_{ij} := \nu_i x_j, \tag{5.6}$$

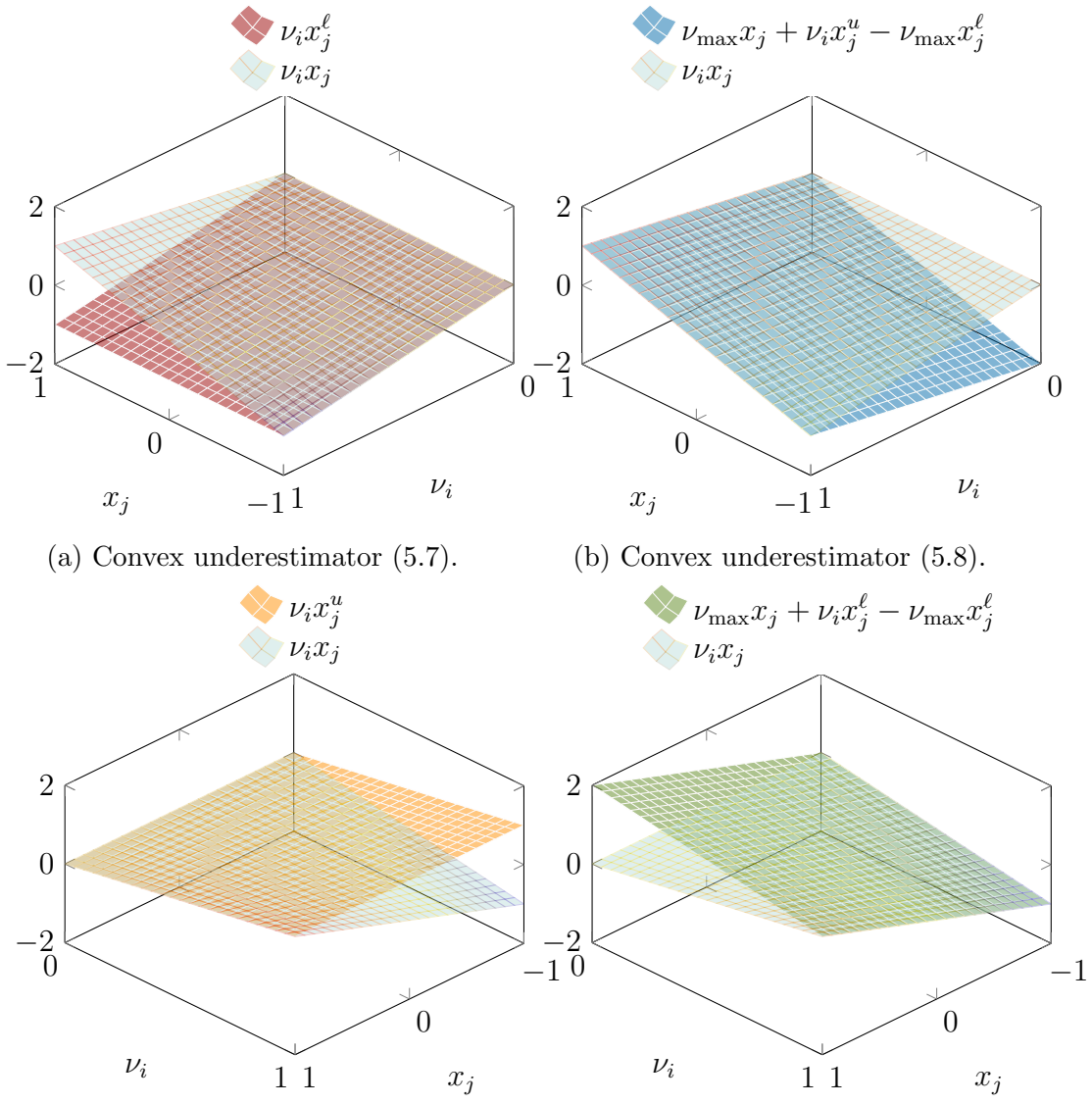
where the variables ν_i and x_j have the same bounds as in the previous section, i.e., $\nu_{\min} \leq \nu_i \leq \nu_{\max}$ and $x_j^\ell \leq x_j \leq x_j^u$ for $i = 1, \ldots, n$ and $j = 1, \ldots, 4n^2$. The variable y_{ij} is relaxed by the following four inequalities which are known to be McCormick envelopes [McC76],

$$\begin{aligned}
y_{ij} &\geq \nu_{\min} x_j + \nu_i x_j^\ell - \nu_{\min} x_j^\ell, \\
y_{ij} &\geq \nu_{\max} x_j + \nu_i x_j^u - \nu_{\max} x_j^u, \\
y_{ij} &\leq \nu_{\min} x_j + \nu_i x_j^u - \nu_{\min} x_j^u, \\
y_{ij} &\leq \nu_{\max} x_j + \nu_i x_j^\ell - \nu_{\max} x_j^\ell.
\end{aligned}$$

Since the lower bound on the viscosities in the MINLP (5.5) is zero, i.e., $\nu_{\min} = 0$, the McCormick envelopes in simplify to

$$
\begin{aligned}
y_{ij} &\geq \nu_i x_j^{\ell}, && (5.7)\\
y_{ij} &\geq \nu_{\max} x_j + \nu_i x_j^{u} - \nu_{\max} x_j^{u}, && (5.8)\\
y_{ij} &\leq \nu_i x_j^{u}, && (5.9)\\
y_{ij} &\leq \nu_{\max} x_j + \nu_i x_j^{\ell} - \nu_{\max} x_j^{\ell}. && (5.10)
\end{aligned}
$$

The inequalities (5.7) and (5.8) are known as convex underestimators and they are shown in Figure 5.1a and 5.1b. The inequalities (5.9) and (5.10) are known as convex overestimators and they are shown in Figure 5.1c and 5.1d.



(a) Convex underestimator (5.7).

(b) Convex underestimator (5.8).

(c) Convex overestimator (5.9).

(d) Convex overestimator (5.10).

Figure 5.1: McCormick envelopes

The MINLP (5.5) is linearized by McCormick envelopes (5.7)-(5.10) and hence, we obtain the following MILP:

$$
\begin{aligned}
\min_{\nu} \quad & \sum_{i=1}^{2n} x_{2n(i-1)+i} \\
\text{s.t.} \quad & \sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} y_{ij} = \delta_{k \bmod (2n+1),1}, && k \in M, \\
& 0 \le \nu_i \le \nu_{\max}, && i \in N, \\
& b_i \in \{0,1\}, && i \in N, \\
& \nu_i \le b_i \nu_{\max}, && i \in N, \\
& x_j^{\ell} \le x_j \le x_j^{u}, && j \in M, \\
& \nu_i x_j^{\ell} \le y_{ij} \le \nu_{\max} x_j + \nu_i x_j^{\ell} - \nu_{\max} x_j^{\ell}, && (i,j) \in N \times M, \\
& \nu_{\max} x_j + \nu_i x_j^{u} - \nu_{\max} x_j^{u} \le y_{ij} \le \nu_i x_j^{u}, && (i,j) \in N \times M, \\
& \sum_{i=1}^{n} b_i \le r, \\
\text{where } & N = \{1, \ldots, n\} \text{ and } M = \left\{1, \ldots, 4n^2\right\}.
\end{aligned}
\tag{5.11}
$$

Remark 5.3.1. *Linearizing the MINLP* (5.5) *via McCormick envelopes by* (5.6) *and* (5.7)-(5.10)*, yields the Mixed Integer Linear Program* (5.11)*, which has* $3n^3$ *additional variables* y_{ij}*, where* $(i,j) \in N \times M$*, compared to the MINLP* (5.5)*, see Remark 5.2.4. Moreover, the MILP* (5.11) *has* $12n^3$ *additional linear constraints compared to the MINLP* (5.5)*.*

5.4 Piecewise Linear Approximation

Piecewise linear approximation of a separable nonlinear function has been investigated e.g. in [Fou92; FM92]. In this section we describe the basic idea of adding auxiliary separable functions and then piecewise linearizing of the bilinear product $\nu_i \cdot x_j$, where $i \in N = \{1, \ldots, n\}$ and $j \in M = \{1, \ldots, 4n^2\}$. First, we define separability of a function f. f is called separable if it can be defined as the sum of functions of scalar variables. Hence, we mean additive separability of a function in the following definition.

Definition 5.4.1. *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *is said to be separable if it can be represented as*

$$f(x) = \sum_{j=1}^{n} g_j(x_j),$$

where $x \in \mathbb{R}^n$ *and* $g_j : \mathbb{R} \to \mathbb{R}$ *for* $j = 1, \ldots, n$*.*

We introduce two auxiliary variables $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$, which are defined as

$$y_{ij}^{(1)} := \tfrac{\nu_i + x_j}{2} \text{ and } y_{ij}^{(2)} := \tfrac{\nu_i - x_j}{2}, \tag{5.12}$$

such that

$$\nu_i x_j = f_1^{ij}\left(y_{ij}^{(1)}\right) - f_2^{ij}\left(y_{ij}^{(2)}\right), \tag{5.13}$$

where $(i,j) \in N \times M$ and $f_1^{ij}(y_{ij}^{(1)}) = \left(y_{ij}^{(1)}\right)^2$ and $f_2^{ij}(y_{ij}^{(2)}) = \left(y_{ij}^{(2)}\right)^2$. Hence, we have now obtained quadratic separable functions f_1^{ij} and f_2^{ij} for $(i,j) \in N \times M$. We use their separability later on, but first we derive bounds on the variables $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$. Let $0 \le \nu_i \le \nu_{\max}$ and $x_j^\ell \le x_j \le x_j^u$ denote some bounds on the variables ν_i and x_j for $(i,j) \in N \times M$. Then lower and upper bounds on the variables $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$ can be derived with (5.12),

$$\begin{aligned} \tfrac{1}{2}x_j^\ell \le y_{ij}^{(1)} &\le \tfrac{1}{2}(\nu_{\max} + x_j^u), \\ -\tfrac{1}{2}x_j^u \le y_{ij}^{(2)} &\le \tfrac{1}{2}(\nu_{\max} - x_j^\ell), \end{aligned}$$

for $(i,j) \in N \times M$.

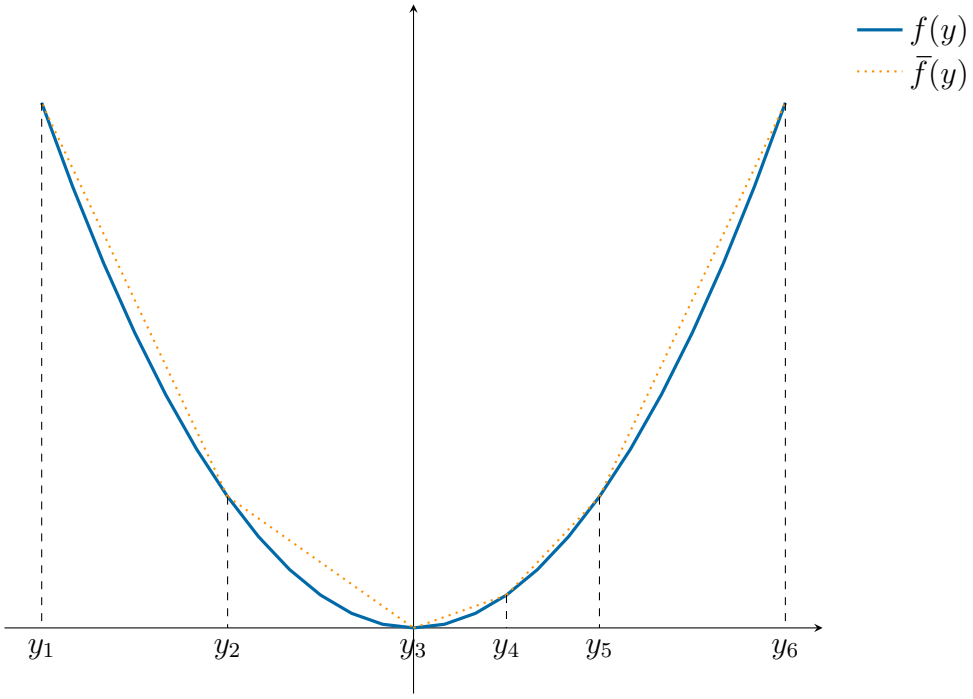


Figure 5.2: Piecewise linear approximation $\overline{f}(y)$ of the quadratic function $f(y) = \frac{1}{2}y^2$.

We approximate the nonlinear, but separable function $f(y) = \frac{1}{2}y^2$ by $\overline{f}(y)$, where $\overline{f}$ is piecewise linear. The quadratic function f is plotted by a solid graph and its piecewise linearizations $\overline{f}$ by dotted lines in Figure 5.2. Here, the piecewise linear approximation is shown for $s = 6$ approximation nodes. Many approaches exist in order to model a piecewise linear approximation of a nonlinear but separable function, e.g. the *multiple choice model* [JL84], the *disaggregated convex combination model* [Mey76], the *incremental model* [MM57] and the *convex combination model* [Dan60]. The convex combination model uses binary variables to select the correct piece of the piecewise linear approximation. Here, we use a variant of the convex combination model, which uses special ordered sets of variables of type 2 instead of binary variables, but for

completeness the convex combination model with binary variables is given below. A special ordered set of type 2 (SOS2) is an ordered set of non-negative variables, of which at most two can be non-zero, and if two are non-zero these must be consecutive in their ordering. Special ordered sets of type 2 were introduced in [BT70].

Now, we come to the convex combination model for the piecewise linear approximation of a function $f : \mathbb{R} \to \mathbb{R}$, where its argument y has a lower and upper bound, i.e., $y^\ell \leq y \leq y^u$. First, predefined and fixed approximation nodes $y_1, \ldots, y_s$ have to be chosen such that the first and last node coincide with the bounds of y, i.e., $y_1 = y^\ell$ and $y_s = y^u$. Then non-negative weights $\lambda_\ell \geq 0$ for each approximation node y_ℓ are introduced, i.e., $\lambda_\ell \geq 0$ for $\ell = 1, \ldots, s$. Any $y \in [y^\ell, y^u]$ can be given by a convex combination of approximation nodes w.r.t. the weights λ_ℓ for $\ell = 1, \ldots, s$, see (5.15)-(5.17). Every convex combination of two points lies on the line segment between the points. Hence, for any $y \in [y^\ell, y^u]$ the convex combination in (5.15) can be chosen such that only the two weights of the adjacent nodes can be positive. Assume $y \in [y_k, y_{k+1}]$, then choose $\lambda_\ell = 0$ for $\ell \neq k, k+1$. Hence, we obtain the piecewise linear approximation $\overline{f}$ at y as the convex combination of $f(y_k)$ and $f(y_{k+1})$ w.r.t. the two adjacent weights λ_k and λ_{k+1},

$$\overline{f}(y) = \lambda_k f(y_k) + \lambda_{k+1} f(y_{k+1}),$$

where $y \in [y_k, y_{k+1}]$. Since $\lambda_\ell = 0$ for $\ell \neq k, k+1$, we can rewrite $\overline{f}(y)$ globally as a convex combination $\overline{f}(y) = \sum_{\ell=1}^{s} \lambda_\ell f(y_\ell)$, see (5.14) for any $y \in [y^\ell, y^u]$. But we have to be assure that $\{\lambda_1, \ldots, \lambda_s\}$ is a special ordered set of variables of type 2 (SOS2). Then the convex combination model for the piecewise linear approximation $\overline{f}$ of f can be summarized as

$$\sum_{\ell=1}^{s} \lambda_\ell f(y_\ell) = \overline{f}(y), \tag{5.14}$$

where its domain is defined by

$$\sum_{\ell=1}^{s} \lambda_\ell y_\ell = y, \tag{5.15}$$

$$0 \leq \lambda_\ell \leq 1, \qquad \text{for } \ell = 1, \ldots, s, \tag{5.16}$$

$$\sum_{\ell=1}^{s} \lambda_\ell = 1. \tag{5.17}$$

Many MILP solver support SOS2 and it is preferably to declare $\{\lambda_1, \ldots, \lambda_s\}$ as a SOS2. If $\{\lambda_1, \ldots, \lambda_s\}$ is declared as a SOS2 constraint a different branching strategy is selected. In an LP relaxed solution where the SOS2 condition is violated by the relaxed solution, the set $\{\lambda_1, \ldots, \lambda_s\}$ is divided into two disjoint subsets S_1 and S_2 such that $S_1 \cup S_2 = \{\lambda_1, \ldots, \lambda_s\}$ and the relaxed solution has a non-zero entry in each of the subsets. Then branching is done such that each branch corresponds to a subset S_i for $i = 1, 2$. For example $\lambda = 0$ is enforced for all $\lambda \in S_1$ on the first branch. Obviously, this branching strategy does not exclude any feasible solutions. In addition, the current relaxed solution, that violates the SOS2 condition, is eliminated from both branches, enabling the relaxation bound to improve and ensuring that the SOS2 condition will be satisfied after a finite number of branches. Valid inequalities based on the SOS2

condition, analogous to the use of valid inequalities for mixed integer programming, can be derived [KFN06].

If the MILP solver does not support SOS2, we instead use the convex combination model [Dan60], which is given below for completeness. It fulfills the model (5.14)-(5.17) but $\{\lambda_1, \ldots, \lambda_s\}$ cannot be declared as a SOS2 and hence, it has to be modeled differently. First, binary variables d_i for $i = 1, \ldots, s-1$ are introduced. $d_i = 1$ if and only if $x \in [x_i, x_{i+1})$. These binary variables with the following constraints model then a SOS2

$$
\begin{aligned}
d_i &\in \{0,1\}, && \text{for } i = 1, \ldots, s-1, && (5.18)\\
\sum_{i=1}^{s-1} d_i &= 1, && && (5.19)\\
\lambda_i + \lambda_{i+1} &\le d_i, && \text{for } i = 1, \ldots, s-1, && (5.20)\\
x &< y_{i+1} + M(1-d_i), && \text{for } i = 1, \ldots, s-1, && (5.21)\\
x &\ge y_i - M(1-d_i), && \text{for } i = 1, \ldots, s-1, && (5.22)
\end{aligned}
$$

for a suitably large constant M. By (5.19)-(5.22) at most two adjacent weights λ_i and λ_{i+1} are greater than zero. If $d_j = 1$, then the constraints (5.21) and (5.22) force $x_j \le x < x_{j+1}$ and are vacuous for all $i \neq j$ (assuming M is correctly chosen).

In the following we linearize the MINLP (5.5) by piecewise linear functions, i.e., by (5.14)-(5.17). We further assume that special ordered sets of type 2 can be declared in the MILP solver. Let us assume that each bilinear product $\nu_i x_j$ is given by two quadratic separable functions $f_1^{(i,j)}(y_{ij}^{(1)})$ and $f_2^{(i,j)}(y_{ij}^{(2)})$ as in (5.13), where $y_{ij}^{(1)} = \frac{\nu_i + x_j}{2}$ and $y_{ij}^{(2)} = \frac{\nu_i - x_j}{2}$ as in (5.12) for $(i,j) \in N \times M$. The functions $f_1^{(i,j)}$ and $f_2^{(i,j)}$ are piecewise linear approximated by $\overline{f}_1^{(i,j)}$ and $\overline{f}_2^{(i,j)}$ with $s-1$ pieces for $(i,j) \in N \times M$. Hence, we obtain the following constraints for the piecewise linear approximation of $f_1^{(i,j)}$ by $\overline{f}_1^{(i,j)}$,

$$
\begin{aligned}
&\sum_{\ell=1}^{s} \lambda_\ell^{(ij)} f(t_\ell^{(ij)}) = \overline{f}_1^{(ij)}(y_{ij}^{(1)}),\\
&\sum_{\ell=1}^{s} \lambda_\ell^{(ij)} t_\ell^{(ij)} = y_{ij}^{(1)},\\
&0 \le \lambda_\ell^{(ij)} \le 1, && \ell \in S,\\
&\sum_{\ell=1}^{s} \lambda_\ell^{(ij)} = 1,\\
&\left\{\lambda_1^{(ij)}, \ldots, \lambda_s^{(ij)}\right\} \text{ is a SOS2},
\end{aligned}
$$

where $(i,j) \in N \times M$. The following constraints for the piecewise linear approximation

of $f_2^{(i,j)}$ by $\overline{f}_2^{(i,j)}$ are obtained

$$\begin{aligned}
&\sum_{\ell=1}^{s} \kappa_\ell^{(ij)} f(t_\ell^{(ij)}) = \overline{f}_2^{(ij)}(y_{ij}^{(2)}), \\
&\sum_{\ell=1}^{s} \kappa_\ell^{(ij)} t_\ell^{(ij)} = y_{ij}^{(2)}, \\
&0 \leq \kappa_\ell^{(ij)} \leq 1, && \ell \in S, \\
&\sum_{\ell=1}^{s} \kappa_\ell^{(ij)} = 1, \\
&\left\{\kappa_1^{(ij)}, \ldots, \kappa_s^{(ij)}\right\} \text{is a SOS2}, ,
\end{aligned}$$

where $(i,j) \in N \times M$. With the piecewise linear approximation we rewrite the constraint $\sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} \nu_i x_j = \delta_{k \bmod (2n+1),1}$ of the MINLP (5.5) as

$$\sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} (\overline{f}_1^{(ij)}(y_{ij}^{(1)}) - \overline{f}_2^{(ij)}(y_{ij}^{(2)})) = \delta_{k \bmod (2n+1),1},$$

where $k \in M$. Therefore, we obtain the MILP (5.23), which is obtained by piecewise linear approximation of the MINLP (5.5).

$$
\begin{aligned}
\min_{\nu} \quad & \sum_{i=1}^{2n} x_{2n(i-1)+i} \\
\text{s.t.} \quad & \sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n}\sum_{j=1}^{4n^2} a_{kj}^{(i)} (\overline{f}_1^{(ij)}(y_{ij}^{(1)}) - \overline{f}_2^{(ij)}(y_{ij}^{(2)})) = \delta, && k \in M, \\
& 0 \le \nu_i \le \nu_{\max}, && i \in N, \\
& b_i \in \{0,1\}, && i \in N, \\
& \nu_i \le b_i \nu_{\max}, && i \in N, \\
& x_j^{\ell} \le x_j \le x_j^{u}, && j \in M, \\
& y_{ij}^{(1)} = \tfrac{\nu_i + x_j}{2}, && j \in M, \\
& y_{ij}^{(2)} = \tfrac{\nu_i - x_j}{2}, && j \in M, \\
& \sum_{\ell=1}^{s} \lambda_{\ell}^{(ij)} f(t_{\ell}^{(ij)}) = \overline{f}_1^{(ij)}(y_{ij}^{(1)}), && (i,j) \in N \times M, \\
& \sum_{\ell=1}^{s} \lambda_{\ell}^{(ij)} t_{\ell}^{(ij)} = y_{ij}^{(1)}, && (i,j) \in N \times M, \\
& 0 \le \lambda_{\ell}^{(ij)} \le 1, && (i,j,\ell) \in N \times M \times S, \\
& \sum_{\ell=1}^{s} \lambda_{\ell}^{(ij)} = 1, && (i,j) \in N \times M, \\
& \left\{ \lambda_1^{(ij)}, \ldots, \lambda_s^{(ij)} \right\} \text{is a SOS2}, && (i,j) \in N \times M, \\
& \sum_{\ell=1}^{s} \kappa_{\ell}^{(ij)} f(t_{\ell}^{(ij)}) = \overline{f}_2^{(ij)}(y_{ij}^{(2)}), && (i,j) \in N \times M, \\
& \sum_{\ell=1}^{s} \kappa_{\ell}^{(ij)} t_{\ell}^{(ij)} = y_{ij}^{(2)}, && (i,j) \in N \times M, \\
& 0 \le \kappa_{\ell}^{(ij)} \le 1, && (i,j,\ell) \in N \times M \times S, \\
& \sum_{\ell=1}^{s} \kappa_{\ell}^{(ij)} = 1, && (i,j) \in N \times M, \\
& \left\{ \kappa_1^{(ij)}, \ldots, \kappa_s^{(ij)} \right\} \text{is a SOS2}, && (i,j) \in N \times M, \\
& \tfrac{1}{2} x_j^{\ell} \le y_{ij}^{(1)} \le \tfrac{1}{2}(\nu_{\max} + x_j^{u}), && (i,j) \in N \times M, \\
& -\tfrac{1}{2} x_j^{u} \le y_{ij}^{(2)} \le \tfrac{1}{2}(\nu_{\max} - x_j^{\ell}), && (i,j) \in N \times M, \\
& \sum_{i=1}^{n} b_i \le r, \\
\text{where } & \delta = \delta_{k \bmod (2n+1),1},\ N = \{1, \ldots, n\}, \\
& M = \left\{1, \ldots, 4n^2\right\},\ S = \{1, \ldots, s\}.
\end{aligned}
\tag{5.23}
$$

Remark 5.4.2. *Linearizing the MINLP* (5.5) *via the piecewise linear approximation* (5.14)-(5.17)*, yields a Mixed Integer Linear Program (MILP) with* $6n^3(s+1)$ *additional variables and* $18n^3$ *additional linear constraints, see Remark 5.2.4.* (5.16) *must not necessarily be encoded as a linear constraint, since it can be viewed as a bound on the variables and hence, it is omitted in the count of additional linear constraints.*
Either $6n^3$ *special ordered sets of variables of type 2 (SOS2) have to be communicated to the MILP solver or* $6n^3(s-1)$ *additional binary variables by* (5.18) *and* $6n^3+18n^3(s-1)$ *linear constraints by* (5.19)-(5.22) *have to be added to the MILP.*

5.5 Heuristic Determination of Damper Position by ℓ_1-Penalization

Some real-world applications cannot be solved to global optimality, because the problems are too large, generate a huge search tree, or must be solved in real time. Encoding the viscous damper problem as a MINLP (5.5) generates a huge problem due to Remark 5.2.5, which cannot be solved to global optimality using the linearization techniques described in Sections 5.3 and 5.4. Then it is often more desirable to obtain a good solution quickly than to wait for an optimal solution. In this section, we propose a heuristic framework to find good but not necessarily optimal damping positions. Hence, we reduce the number of damper positions such that we obtain a subset D_μ of good damper positions w.r.t. a parameter $\mu \geq 0$, i.e., $D_\mu \subseteq D_0$, where D_0 are the initial damper positions. In our case we assume that to each mass an external damper can be attached, i.e., $D_0 = \{1, \ldots, n\}$. Once, the good damper positions D_μ are identified, we solve a series of NLPs (OPT 1) in order to determine the best r damper positions in this subset D_μ with respective viscosities.
First, we return to the mixed integer nonlinear programming formulation in (OPT 2). If we omit all binary variables of (OPT 2) and the corresponding inequalities where they occur, we obtain a Nonlinear Program (NLP),

$$
\begin{aligned}
\min_{\nu} \quad & \sum_{i=1}^{2n} x_{2n(i-1)+i} \\
\text{s.t.} \quad & \sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} \nu_i x_j = \delta_{k \bmod (2n+1),1}, && k \in M, \\
& 0 \leq \nu_i \leq \nu_{\max}, && i \in N, \\
& x_j^\ell \leq x_j \leq x_j^u, && j \in M, \\
& \text{where } N = \{1, \ldots, n\} \text{ and } M = \left\{1, \ldots, 4n^2\right\}.
\end{aligned}
\tag{5.24}
$$

We cannot expect the solution of the NLP (5.24) to be sparse. If it is sparse it would not be necessary to consider the MINLP (5.5) in the first place. Our idea is to modify the NLP (5.24) such that the solution is sparse. There are several ways to induce sparsity to a solution, e.g., by adding an ℓ_p penalty to the objective function. The

function $\ell_p : \mathbb{R}^n \to \mathbb{R}$ for $0 < p \leq 1$ is defined as

$$\|x\|_p := \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{\frac{1}{p}}.$$

By (2.10), $\|x\|_p$ is a regular norm for $p \in [1, \infty]$. $\|x\|_p$ for $p \in (0, 1)$ is often called "norm", even though it is not subadditive. Furthermore, $\|x\|_p$ is not convex for $p \in (0, 1)$, which is a disadvantage w.r.t. optimization, since convex functions on an open set have no more than a single minimum. Penalization with the Manhattan norm $\|\cdot\|_1$ sparsifies the solution of an optimization problem (see Figure 3.3). Therefore, we add an ℓ_1-term to the objective function which is regulated by a parameter $\mu > 0$, namely,

$$\mu\|\nu\|_1 = \mu \sum_{i=1}^{n} |\nu_i|.$$

The viscosities are non-negative, i.e., $\nu_i \geq 0$ for $i = 1, \ldots, n$. Hence, we can reformulate the ℓ_1-term as $\mu \sum_{i=1}^{n} \nu_i$ and we obtain the Nonlinear Program (5.25).

$$\begin{aligned}
\min_{\nu} \quad & \sum_{i=1}^{2n} x_{2n(i-1)+i} + \mu \sum_{i=1}^{n} \nu_i \\
\text{s.t.} \quad & \sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} \nu_i x_j = \delta_{k \bmod (2n+1),1}, \quad && k \in M, \\
& 0 \leq \nu_i \leq \nu_{\max}, && i \in N, \\
& x_j^{\ell} \leq x_j \leq x_j^{u}, && j \in M, \\
& \text{where } N = \{1, \ldots, n\} \text{ and } M = \left\{1, \ldots, 4n^2\right\}.
\end{aligned} \tag{5.25}$$

Theorem 5.5.1. *The optimization problem* (5.25) *is a Nonlinear Program (NLP), where the objective and constraint functions are twice continuously differentiable.*

Proof. The only qualitatively newly added function to the relaxed MINLP (OPT 2) is the ℓ_1-term $\mu \sum_{i=1}^{n} \nu_i$ in the objective function, which in fact is linear and hence, it is twice continuously differentiable. The remaining proof can be carried over from Lemma 4.2.1 and 5.2.1. □

Let $x \in \mathbb{R}^n$ be a real vector. $x \in \mathbb{R}^n$ is called k-sparse if it has k non-zero elements, i.e., $|x|_0 = |\{i : x_i \neq 0\}| = k$. Due to numerical computation, a vector is often not exactly but almost k-sparse. We therefore relax the notion of sparsity w.r.t. a threshold ϑ. $x \in \mathbb{R}^n$ is called k-sparse w.r.t. a threshold $\vartheta > 0$ if it has k elements which are absolutely larger than ϑ, i.e., $|\{i : x_i \notin [-\vartheta, \vartheta]\}| = k$. In the following sparsity is meant in the sense of a small threshold $\vartheta > 0$, where the threshold $\vartheta > 0$ is in general problem specific (see Section 5.7).

Remark 5.5.2. *Let ν_μ^* be the optimal solution of the NLP* (5.25)*. The tradeoff between sparsity and optimality of ν_μ^* can be varied by parameter μ. Optimality of the solution is meant in the sense of the NLP* (OPT 2)*, i.e., without penalization $\mu = 0$.*

By Remark 5.5.2 we can regulate sparsity and optimality of the solution ν_μ^* by varying the parameter μ, i.e., $\left|\nu_\mu^*\right|_0 = 0$ as $\mu \to \infty$. The non-zero indices of the optimal viscosities ν_μ^* represent the positions of external dampers, i.e., the set of "good" damper positions D_μ w.r.t. the penalization parameter μ, which is then given as

$$D_\mu = \left\{ i : \left(\nu_\mu^*\right)_i > \vartheta \right\}. \tag{5.26}$$

We solve a series of Nonlinear Programs (NLPs) (5.25), where μ is varied until the optimal solution ν_μ^* is sufficiently sparse. Here, sufficient sparsity is meant in the sense that the computation of the r optimal damper positions within the set of "good" damper positions is feasible. Let ν_μ^* be k-sparse, then it follows that $\frac{k!}{(k-r)!}$ is sufficiently small, where r is the predefined number of external dampers that should be placed. Here, the computation of optimal damper positions can be performed by the methods described in Chapter 4 or by the MINLP formulation in Section 5.2.

Remark 5.5.3. *The set of "good" damper positions D_μ, which is defined in* (5.26) *and is computed by solving the NLP* (5.25) *w.r.t. parameter μ, does not necessarily include the optimal r damping positions of* (OPT 2).

In fact, in Section 5.7 a numerical example is presented, where the set of "good" damper positions D_μ does not include the optimal damper position. Hence, we obtain the following corollary.

Corollary 5.5.4. *The framework of identifying "good" damper positions D_μ by solving the NLP* (5.25) *w.r.t. parameter μ and then computing the optimal damper position within the set D_μ is a heuristic.*

5.6 Optimal Damper Positions for sufficiently small Viscosities

In this section we derive an algorithm, which determines the optimal damper positions w.r.t. the averaged total energy for sufficiently small viscosities, i.e., for sufficiently small viscosities it is not necessary to consider the MINLP formulation (5.5). The algorithm is based on a linear approximation of the function $f : \mathbb{R}^n \to \mathbb{R}$ defined in (5.1), where $\nu \mapsto \operatorname{tr}(X(\nu))$. Here f is the reformulated objective function of the MINLP (5.5) as in Lemma 5.2.1. f is smooth by Lemma 5.2.1 and hence, it can be expressed by a multivariate Taylor series (its single variable equivalent is given in Theorem 2.3.9). Here, we use a linear approximation of the averaged total energy

$$\operatorname{tr}(X(\nu)) = f(\nu) = f(0) + \nu^T \nabla f(0) + \ldots, \tag{5.27}$$

where ∇f is the gradient of f. The partial derivative $\frac{\partial f}{\partial \nu_j}(\nu)$ for the structured algebraic Lyapunov equation is defined in (4.8). Hence,

$$\frac{\partial f}{\partial \nu_j}(0) = 2\widehat{V}_j^T X(0) J X(0) \widehat{V}_j.$$

$X(0)$ is given in Remark 3.4.7 as $X(0) = \widehat{X}_1 \oplus \widehat{X}_2 \oplus \ldots \oplus \widehat{X}_n$, where $\widehat{X}_i = \frac{1}{2\omega_i} \begin{bmatrix} \frac{2\gamma^2+1}{\gamma} & 1 \\ 1 & \frac{1}{\gamma} \end{bmatrix}$ for $i = 1, \ldots, n$ and hence,

$$\begin{aligned} \widehat{X}_i \cdot \operatorname{diag}(1,-1) \cdot \widehat{X}_i &= \frac{1}{4\omega_i^2} \begin{bmatrix} \frac{2\gamma^2+1}{\gamma} & 1 \\ 1 & \frac{1}{\gamma} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{2\gamma^2+1}{\gamma} & 1 \\ 1 & \frac{1}{\gamma} \end{bmatrix} \\ &= \frac{1}{4\omega_i^2} \begin{bmatrix} \frac{(2\gamma^2+1)^2}{\gamma^2} - 1 & \frac{2\gamma^2+1}{\gamma} - \frac{1}{\gamma} \\ \frac{2\gamma^2+1}{\gamma} - \frac{1}{\gamma} & 1 - \frac{1}{\gamma^2} \end{bmatrix}. \end{aligned}$$

Since $\widehat{V} = P \begin{bmatrix} 0 \\ V \end{bmatrix}$ in (3.38), where P is the perfect shuffle permutation (3.33), we obtain

$$\frac{\partial f}{\partial \nu_j}(0) = \tfrac{1}{2}\left(1 - \tfrac{1}{\gamma^2}\right) V_j^T \Omega^{-2} V_j, \tag{5.28}$$

where Ω is defined in (3.19) and V_j is the j-th column of V, which is defined in (3.32). The formula (5.28) is the partial derivative of the averaged total energy w.r.t. the viscosity ν_j. Since $\Omega^{-2} = \operatorname{diag}(\omega_1^{-2}, \ldots, \omega_n^{-2})$ is a diagonal matrix, (5.28) can be rewritten as

$$\frac{\partial f}{\partial \nu_j}(0) = \tfrac{1}{2}\left(1 - \tfrac{1}{\gamma^2}\right) \sum_{i=1}^{n} \tfrac{1}{\omega_i^2} v_{ij}^2, \tag{5.29}$$

where $v_{ij} = (V)_{ij}$. Since $0 < \gamma \ll 1$ and $\omega_i > 0$ for $i = 1, \ldots, n$, it follows that $\frac{\partial f}{\partial \nu_j}(0) < 0$. Hence, (5.29) shows the descent of the averaged total energy w.r.t. viscosity ν_j. The steepest descent gives us the optimal damping position for sufficiently small viscosities. Therefore, we denote with d_j the j-th descent which is given as $\frac{\partial f}{\partial \nu_j}(0)$ for $j = 1, \ldots, n$. We sort them in ascending order and the smallest r descents are then the best positions to place r external dampers for sufficiently small viscosities. This procedure is summarized in Algorithm 6. For larger viscosities higher order terms in the approximation (5.27) may dominate and hence, the restriction on the viscosities is due to the linear approximation in (5.27).

Algorithm 6 Optimal damper positions for sufficiently small viscosities

Require: Matrix $\Omega \in \mathbb{R}^{n \times n}$, $V_i \in \mathbb{R}^n$ for $i = 1, \ldots, n$ and $0 < \gamma \ll 1$.
Ensure: Ordered damper positions $j_1, \ldots, j_n \in \{1, \ldots, n\}$.

1: **for** $j = 1, \ldots, n$ **do**
2: $\quad d_j \leftarrow \frac{1}{2}\left(1 - \frac{1}{\gamma^2}\right) \sum_{i=1}^{n} \frac{1}{\omega_i^2} v_{ij}^2$
3: **end for**
4: Sort d_j for $j = 1, \ldots, n$ in ascending order, i.e., $d_{j_1} \leq d_{j_2} \leq \ldots \leq d_{j_n}$.
5: **return** $j_1, \ldots, j_n$

Theorem 5.6.1. *The optimal damper positions for sufficiently small viscosities can be computed in $\mathcal{O}(n^2)$.*

Proof. A single execution of line 2 of Algorithm 6 can be done in $\mathcal{O}(n)$. Hence, the

computational complexity is $\mathcal{O}(n^2)$ for n times executing line 2. Sorting in line 4 of Algorithm 6 can be done in $\mathcal{O}(n\log(n))$ e.g. by quick sort. □

5.7 Numerical Results

In this section we show numerical results for the optimization problem (OPT 2). As an example we consider the viscously damped beam shown in Figure 4.2, which is discretized by finite elements as in Section 4.4. In Chapter 4 the position of the viscous damper is fixed in the middle of the beam at $\frac{L}{2}$ and hence, we considered in Chapter 4 the corresponding NLP (OPT 1). But now the position of the viscous damper is not necessarily in the middle of the beam anymore, but rather we want to determine the best viscous damper positions w.r.t. the averaged total energy of the beam. The dampers have to be attached to the beam, i.e., their positions have to be in $[0, L]$. The boundary value problem (4.36) is discretized by finite elements and the same space for external dampers is used, i.e., a viscous damper can be attached to any finite element. We used ten cubic Hermite polynomials as interpolation shape functions and assume that a viscous damper can be attached to any finite element. The internal damping is given as modal damping, i.e., $C_{\text{int}} = 2\gamma M^{1/2}\sqrt{M^{-1/2}KM^{-1/2}}M^{1/2}$, where $\gamma = \frac{1}{100}$.
We encoded the viscous damper placement problem for the viscously damped beam as the non-convex MINLP (5.5) in AMPL. We considered two variants of the non-convex MINLP (5.5), namely determination of the damper positions and viscosities for one and two viscous dampers. The lower and upper bounds on x and ν are given by the results of Chapter 4, i.e., $x^\ell \le x_j \le x^{\text{u}}$ for $j = 1, \ldots, 4n^2$ and $0 \le \nu_i \le \nu_{\max}$ for $i = 1, \ldots, n$. Obviously, other bounds can be used as well. A survey on bounds for the solution of an algebraic Lyapunov equation is given in [AMK96]. A bound on the solution for an algebraic Lyapunov equation is given in [PS08], where the assumption of stability for the symmetric part of the system matrix can be dropped, which would be necessary for the structured algebraic Lyapunov equation (3.43) given in the MINLP (5.5).

#dampers	**1st damper position**	**2nd damper position**	**tot. energy**	**time in s**
$r = 1$	$i_1 = 1$	—	0.1418	197665.75
$r = 1$	$i_1 = 10$	—	0.1418	197665.75
$r = 2$	$i_1 = 3$	$i_2 = 5$	0.0951	5543529.21
$r = 2$	$i_1 = 7$	$i_2 = 9$	0.0951	5543529.21

Table 5.1: Solving MINLP (5.5) with BARON to determine positions of $r = 1, 2$ external dampers, averaged total energy and running time.

The running tests for solving the MINLP (5.5) were performed with the non-convex MINLP solver BARON. In Table 5.1 the results are summarized, where we denote with i_j the optimal position for j-th external viscous damper for $j = 1, \ldots, r$. We found two optimal sets for the damper positions for one and two external dampers (i.e. $r = 1$ and $r = 2$), namely the sets of positions $\{1\}$ and $\{10\}$ for a single external damper and the sets of positions $\{3, 5\}$ and $\{7, 9\}$ for two external dampers. Solving the non-convex MINLP to global optimality is very demanding as it can bee seen at

the running times — even for the small-scale example that we used (compare Remark 5.2.5). In order to apply the optimization problem (OPT 2) to examples with a larger scale, we test in the following methods that do not guarantee global optimality, but have a reduced complexity. These methods can be classified into the following three categories: linearization of the nonlinearity, successive selection of damper positions and heuristics. We start the discussion with the linearization of the non-convex MINLP (5.5) by McCormick envelopes.

5.7.1 Linearization

Linearization by McCormick Envelopes

We first want to analyze the sensitivity of the linearization of the non-convex MINLP (5.5) by McCormick envelopes. Hence, we analyze the MILP (5.11). In order to analyze the structure of the linearization we repeatedly solve the optimization problems (5.5) and (5.11) for each damper combination, i.e., we fix b_i for $i = 1, \dots, n$ before we solve the optimization problem (5.5) and (5.11), which then turn out to be an NLP and LP, respectively, since no integer variables occur. We encoded the resulting NLPs and LPs in AMPL and solved these by the build-in NLP and LP solver from CPLEX. For clarification we denote with f_{MINLP} and $f_{\text{McCormick}}$ the objective values of the optimization problems (5.5) and (5.11) even though the optimization problems are NLPs and LPs as we have fixed the integer variables in advance. We compute the relative linearization error by McCormick envelopes as

$$err_{\text{McCormick}} = \frac{|f_{\text{MINLP}} - f_{\text{McCormick}}|}{|f_{\text{MINLP}}|}. \tag{5.30}$$

In Figures 5.3 and 5.4 we show the relative linearization error by McCormick envelopes for one and two external viscous dampers, respectively. The relative linearization error by McCormick envelopes is very sensitive w.r.t. the position of the dampers, since it varies between 0.2 and 2.2 as shown in Figures 5.3 and 5.4.
Furthermore, we encoded the MILP (5.11) in AMPL and solved it with CPLEX. The results are shown in Table 5.2. Here, we show the approximated total energy of the MILP (5.11) the positions of the external dampers and the computation time for one and two external viscous dampers. The averaged total energy of the MINLP (5.5) is not approximated well with a negative value of -0.2137199306 by the MILP (5.11) and the approximated value does not decrease for an increased number of external dampers (see Table 5.2).

#dampers	approx. tot. energy	Position	time
$r = 1$	-0.2137199306	9	1.35608
$r = 2$	-0.2137199306	9,10	0.992061

Table 5.2: Results for MILP (5.11) for viscously damped beam discretized by ten finite elements.

Due to the sensitivity of the McCormick linearization error w.r.t. the positions of the

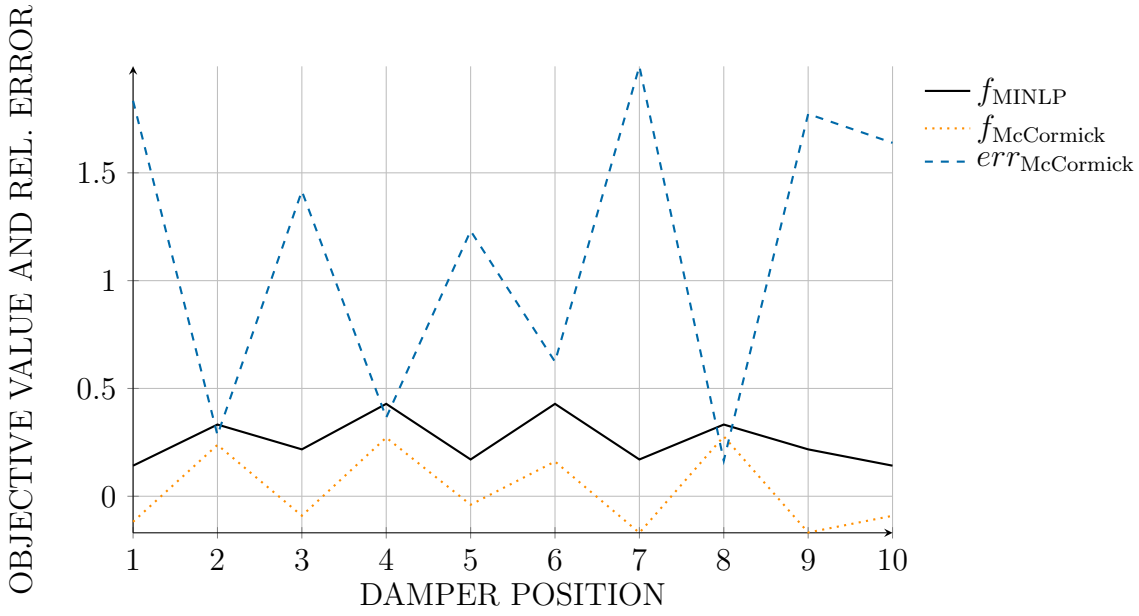


Figure 5.3: Optimum and relative linearization error by McCormick envelopes defined in (5.30) for the viscously damped beam w.r.t. the position of a single damper.

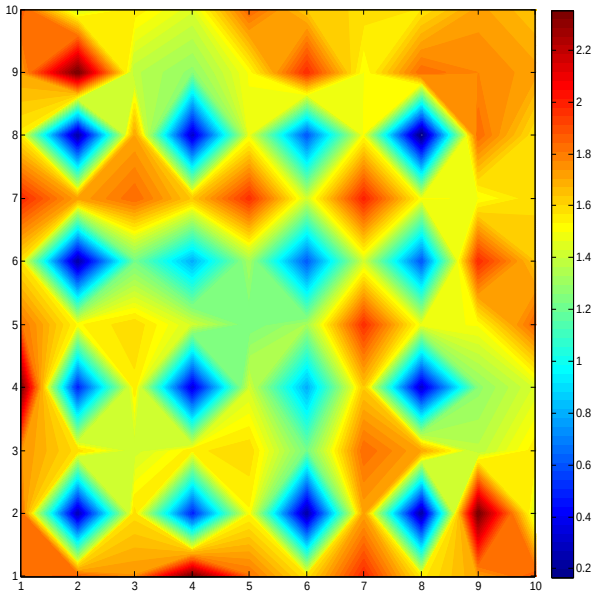


Figure 5.4: Contour plot of the relative linearization error by McCormick envelopes defined in (5.30) of the viscously damped beam for two external dampers.

external dampers and since increasing the number of external dampers, does not necessarily decrease the approximated total energy, we conclude that the MILP (5.11) does not yield a sufficiently good approximation of the MINLP (5.5) in order to determine good damper positions for this particular example.

Linearization by piecewise linear function

In this section we show results for the linearization by piecewise linear function for the nonlinearity of the MINLP (5.5). Hence, we consider the MILP (5.23), which was encoded in AMPL and solved by CPLEX. Unfortunately, solving the MILPs (5.23) is very demanding and for most test cases neither the lower bound could be improved nor an integer solution was found during the branch-&-bound algorithm. Hence, the MILP (5.23) obtained by an approximation of the quadratic functions by $s = 3, 4, 8, 16$ piecewise linear functions and $r = 1, 2$ external dampers could not be solved since CPLEX ran out of memory.
In order to examine the MILP (5.23) further, we considered a smaller example, namely, the viscously damped beam discretized by two finite elements. We denote with i_1 the position of the first external viscous damper. Since the viscously damped beam is discretized by two finite elements the position of the first external viscous damper is $i_1 \in \{1, 2\}$. Furthermore, we denote with f_{pwl} the objective value of the MILP (5.23). We compute the relative linearization error by piecewise linear functions,

$$err_{\text{pwl}} = \frac{|f_{\text{MINLP}} - f_{\text{pwl}}|}{|f_{\text{MINLP}}|}. \tag{5.31}$$

In Table 5.3 we show the approximated averaged total energy f_{pwl}, the relative piecewise linearization error err_{pwl} and the CPLEX running time for one external damper at the positions $i_1 = 1$ and $i_1 = 2$, where the nonlinearity of the MINLP (5.5) is approximated by $s = 3, 4, 8, 16$ piecewise linear functions.

#pw. linear functions	**damper at $i_1 = 1$**			**damper at $i_1 = 2$**		
	f_{pwl}	err_{pwl}	time	f_{pwl}	err_{pwl}	time
$s = 3$	infeasible	—	1135.1	0.05690374	0.1973	17.6
$s = 4$	infeasible	—	10978.8	0.05740511	0.2079	386.2
$s = 8$	0.06049407	0.2729	2114.6	0.05824296	0.2255	1325.0
$s = 16$	0.05947558	0.2514	25982.9	0.05846105	0.2301	24777.9

Table 5.3: Results for piecewise linear approximation of the viscously damped beam discretized by two finite elements.

First, we want to analyze the CPLEX method and its computation time. CPLEX could not obtain an integer feasible solution for the damper at position $i_1 = 1$ and $s = 3, 4$ piecewise linear functions. In the remaining cases CPLEX computed an optimal solution. Increasing the number of pieces for the approximation, yields in most cases a larger solving time (excluding if the problem has been integer infeasible before), since for a minimization problem an integer solution solution serves as an upper bound and nodes of the branch-&-bound tree are pruned if its lower bound is larger than an existing upper bound. If a problem is integer infeasible, no upper bound exists and all nodes of the branch-&-bound tree have to be visited. Then the branch-&-bound method amounts to brute-force enumeration. This results in a very large computation time as for a damper at position $i_1 = 1$ and $s = 3, 4$ piecewise linear functions in Table 5.3. If the number of external dampers was increased to two, CPLEX ran out of

memory as well.

Secondly, we analyze the relative piecewise linearization errors err_{pwl}. The relative errors err_{pwl} increase as the number of piecewise linear function increases for the damper at position $i_1 = 2$ as shown in Table 5.3. Hence, improving the piecewise linear approximation of the MINLP (5.5), does not necessarily approximate the averaged total energy f_{MINLP} better. The relative error of the piecewise linearization is much less than the relative McCormick error, which is given for this example by $err_{\text{McCormick}} = 0.6486$ for $i_1 = 1$ and $err_{\text{McCormick}} = 1.1504$ for $i_1 = 2$. We conclude that the approximation of the nonlinearity in (5.5) by piecewise linear function is favorable compared to McCormick envelopes if a solution to the MILP (5.23) can be computed.

In summary, the MILP (5.23) could not be solved efficiently for this particular example. Solving it is very memory demanding even for this particular small-scale example and a small number of nodes for approximation with piecewise linear functions.

5.7.2 Successive selection of damper positions

In this section we describe a method that successively selects positions for external dampers, i.e., we do not consider the MINLP (5.5), but we replace it with a series optimization problems. We will later specify the nature of the optimization problems. As before, we assume that r is the number of external dampers. In this context r is then the number of positions for external viscous dampers. The general idea is that each of these optimization problems determines a single position for an external damper. Hence, the number of damper combinations is reduced from $\frac{n!}{(n-r)!}$ for the MINLP (5.5) to $\frac{(2n-r)r}{2}$ for the series of optimization problems, since the selection process is sequential. We outline the idea of successive selection of damper positions in Algorithm 7.

Algorithm 7 Successive selection of damper positions

Require: Set of possible damper positions $P_{\text{possible}} = \{1, \ldots, n\}$.
Ensure: Set of selected damper positions $P_{\text{selected}} = \{i_1, \ldots, i_r\}$.
1: $P_{\text{selected}} \leftarrow \emptyset$
2: **for** $j = 1, \ldots, r$ **do**
3: Select damper position $i_j \in P_{\text{possible}}$
4: $P_{\text{possible}} \leftarrow P_{\text{possible}} \setminus \{i_j\}$
5: $P_{\text{selected}} \leftarrow P_{\text{selected}} \cup \{i_j\}$
6: **end for**
7: **return** P_{selected}

The selection of the position for the external viscous damper in line 3 of Algorithm 7 has not been specified yet and in the following we will discuss two ideas. But every selection strategy cannot guarantee to find the optimal damper positions of the MINLP (5.5).

Successive selection of local optimal damper positions

We denote with $f_{P_{\text{selected}}}$ the averaged total energy w.r.t. the set of selected positions for external viscous dampers P_{selected}. The strategy that we want to introduce here selects an optimal damper positions i_j in line 3 of Algorithm 7, i.e.,

$$f_{P_{\text{selected}} \cup \{i_j\}} \leq f_{P_{\text{selected}} \cup \{i\}} \tag{5.32}$$

for all $i \in P_{\text{possible}}$. The selection of a damper positions is locally optimal due to (5.32) and it can be described by the following MINLP:

$$\begin{aligned}
\min_{\nu} \quad & \sum_{i=1}^{2n} x_{2n(i-1)+i} \\
\text{s.t.} \quad & \sum_{j=1}^{4n^2} a_{kj}^{(0)} x_j - \sum_{i=1}^{n} \sum_{j=1}^{4n^2} a_{kj}^{(i)} \nu_i x_j = \delta_{k \bmod (2n+1),1}, && k \in M, \\
& 0 \leq \nu_i \leq \nu_{\max}, && i \in N, \\
& b_i \in \{0,1\}, && i \in P_{\text{possible}}, \\
& \nu_i \leq b_i \nu_{\max}, && i \in N, \\
& x_j^{\ell} \leq x_j \leq x_j^{u}, && j \in M, \\
\text{where } & N = \{1, \ldots, n\} \text{ and } M = \left\{1, \ldots, 4n^2\right\}.
\end{aligned} \tag{5.33}$$

We conclude this section with the observation that the optimal position for a single external damper is not included in the set of optimal positions for two external dampers, see Table 5.1 and Figure 5.5. We conclude that for this particular example it is computationally too expensive to employ the strategy to select local optimal damper positions (compare Table 5.1), since the local optimal choice of damper positions i_j for $j = 1, \ldots, r$ does in general not yield global optimality of P_{selected} and the series of MINLPs (5.33) cannot be solved efficiently.

Successive selection of damper positions w.r.t. steepest descent

The strategy that we want to introduce here selects a damper positions i_j in line 3 of Algorithm 7 that has the steepest descent. We therefore have to compute the gradient, which can be done by Algorithm 6. We have shown in Section 5.6 that the damping positions determined by Algorithm 6 are optimal if the perturbation is sufficiently small. Hence, for sufficiently small viscosities, the strategies to select damper positions w.r.t. local optimality and steepest descent coincide. Moreover, both strategies yield global optimality of P_{selected} if the corresponding viscosities are sufficiently small.
We computed the gradient (5.29) for the viscously damped beam by Algorithm 6 implemented in MATLAB. The results are shown in Figure 5.6. The running time for computing the gradient by Algorithm 6 is $0.00282s$. The best damper positions for a sufficiently small viscosity were determined, which in fact were the positions $i_1 = 1$ and $i_1 = 10$, where the steepest descent of the gradient is attained with 45.70971. These

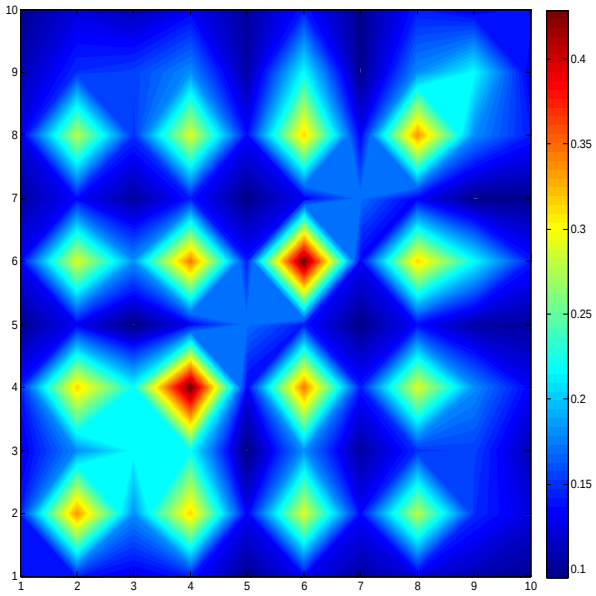


Figure 5.5: Contour plot of the averaged total energy for viscously damped beam with two viscous dampers.

positions are the optimal positions for placing external viscous dampers (see Table 5.1 and Figure 5.5), but finding the optimal damper's positions is not a general property of Algorithm 6, since it does not guarantee to find the optimal positions for arbitrary viscosities as mentioned in Section 5.6.

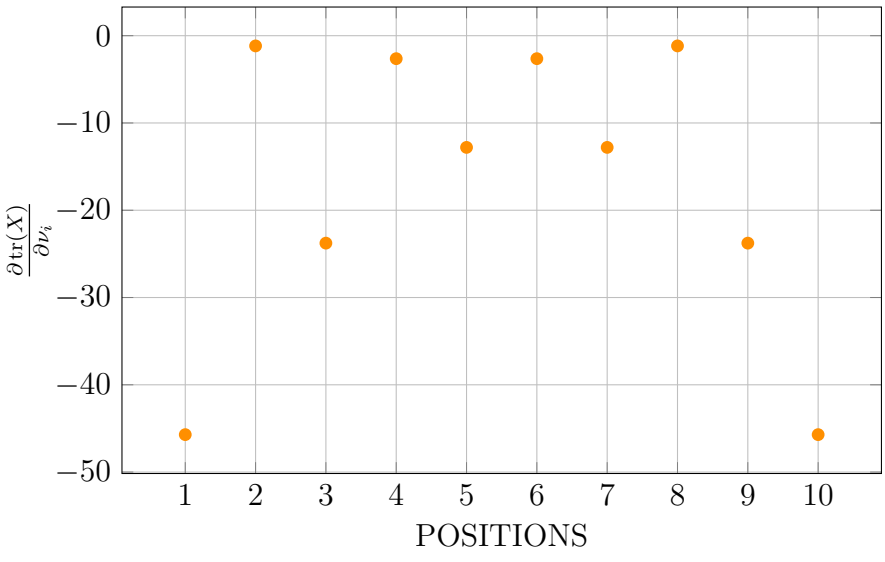


Figure 5.6: Gradient of the damped beam with ten finite elements.

We conclude that the selected damper positions were good for this particular example and it is computationally cheap to employ the strategy to select damper positions w.r.t. the steepest descent of the gradient by Algorithm 6.

5.7.3 Heuristics

Damper positions by ℓ_1-Penalization

In this section we show results for the viscously damped beam discretized by ten finite elements described by the NLP (5.25). We used the *global search strategy* with the NLP solver *fmincon* from MATLAB to solve the NLP (5.25). We varied the ℓ_1-penalization parameter μ of the NLP (5.25), i.e., $\mu \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and for each penalization parameter, the corresponding optimal viscosities are shown in Figure 5.7. The optimal viscosities ν^* of the NLP (5.25) without a penalization, i.e., $\mu = 0$, are in general non-sparse, see Figure 5.7. But as we increased the penalization parameter μ, we increased the sparsity of the optimal solution ν_μ^* of the NLP (5.25). "Good" damper positions in terms of (5.26) can be deduced for $\mu \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, since the optimal viscosities ν_μ^* at some positions are five times larger in magnitude than the remaining ones, as shown in Figure 5.7. Hence, for $\vartheta = 10^{-3}$ we obtain the "good" damper positions $D_\mu = \{1, 10\}$ which are actually the optimal damper positions for placing a single external damper (see Table 5.1).

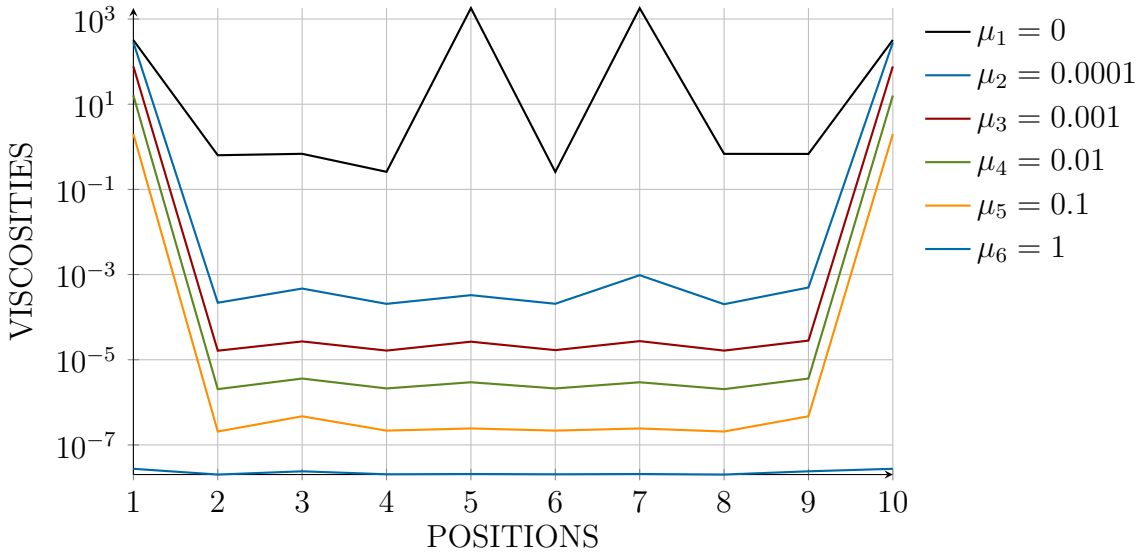


Figure 5.7: Optimal viscosities of NLP (5.25) for the viscously damped beam for various ℓ_1-penalizations μ.

For $\vartheta = 10^{-3}$ and placing two dampers, we obtain the same set of "good" damper positions $D_\mu = \{1, 10\}$ for $\mu \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, while the optimal positions are at $(3, 5)$ and $(7, 9)$ (see Table 5.1 and Figure 5.5). The NLP (5.25) can efficiently been solved and the position of the external damper w.r.t. the penalization parameter μ are good solutions.

Non-convex MINLP solver

We encoded the viscous damper placement problem for the damped beam as the non-convex MINLP (5.5) in AMPL. In this section we used the convex MINLP solver MINOTAUR to solve the non-convex MINLP (5.5). Obviously, global optimality of the solution to the non-convex MINLP (5.5) cannot be guaranteed, but the running

time of the convex solver is much less since for the relaxation convexity is assumed and hence, no outer-approximation of the feasible set is needed, compare the results in Tables 5.1 and 5.4.

#dampers	1st damper position	2nd damper position	tot. energy	time in s
$r = 1$	$i_1 = 10$	—	0.1418	14.95
$r = 2$	$i_1 = 1$	$i_2 = 10$	0.0986	11.84

Table 5.4: Solving MINLP (5.5) with non-convex MINLP solver MINOTAUR to determine positions of $r = 1, 2$ external dampers, averaged total energy and running time.

5.7.4 Summary of numerical results

We have reduced the complexity of the non-convex MINLP (5.5) by various ideas, namely, linearization of the nonlinearity by McCormick envelopes and piecewise linear functions, strategies of successive selection of damper positions and heuristics to determine good damper positions by ℓ_1-Penalization in the NLP (5.25) and using a non-convex MINLP solver. We have tested the above methods on various examples and presented here results for the viscously damped beam.
Unfortunately, the results on linearization by McCormick envelopes and by piecewise linear functions described in Section 5.7.1 were not promising, since due to the linearization, the approximation of the objective function was too weak or even the linearization was too complex and no efficiency increase was gained by solving the approximated problem. The strategy of successive selection of local optimal damper positions described in Section 5.7.2 was too expensive, since the computational costs were too high for obtaining a non-optimal global set of damper positions.
We give an outlook on the prospective methods that can be applied to examples of a larger scale. From our experiments the strategy of successive selection of damper positions by steepest descent described in Section 5.7.2 is promising due to the very efficient calculation of the gradient and the optimality of the resulting set of damper positions for sufficiently small viscosities. Furthermore, the heuristic computation of damper positions by ℓ_1-Penalization and a non-convex MINLP solver described in Section 5.7.3 are from our point of view also promising methods due to the quality of the resulting damper positions and the running times of the corresponding solvers.

6

Two-Sided Bounds on the Solution of Time-Periodic Systems

Linear time-periodic systems arise in many fields of application, e.g. in parametrically excited systems and anisotropic rotor-bearing systems. Often they are obtained by linearizing a nonlinear system about a periodic trajectory. Knowledge of the systems components is necessary to understand its transient behavior completely, which may not be applicable for very complex and large-scale systems. Understanding system characteristics such as its transient behavior in a certain norm, stability or robustness may often be sufficient. A general time-periodic system is given in (2.43) as

$$\begin{array}{rcll} \dot{x} & = & A(t)x\,, & \text{for } t \in I, \\ A(t) & = & A(t+t_p)\,, & \text{for } t \geq 0, \\ x(0) & = & x_0, & \end{array} \tag{6.1}$$

where for notational simplicity the initial condition is at $t_0 = 0$, i.e., $x(0) = x_0$. In this chapter we investigate the transient behavior of its solution $\|x(t)\|$ as $t \to \infty$ by two-sided bounds, i.e., $f_\ell(t) \leq \|x(t)\| \leq f_u(t)$ for all $t \geq 0$. The structure of a solution for a linear time-periodic system is known (Floquet's Theorem 2.5.18). But nevertheless, it has to be approximated since in general it cannot be given in closed form. Important physical properties such as stability and robustness can be lost due to (numerical) approximations. In order to guarantee such properties for the original solution and not only for the approximation, one can derive analytic results on the solution or the approximation error has to be incorporated in the analysis. This is the key idea of this chapter: bounds that solely depend on the solution structure or bounds that incorporate the approximation error. Firstly, we were able to generalize results from the linear time-invariant to time-periodic setting. The time-invariant setting was discussed in Section 3.3. For the time-periodic setting we were able to derive a time-varying norm that captures important properties such as *decoupling*, *filtering* and *monotonicity*. Secondly, we used two different methodologies where the approximation error is incorporated in the two-sided bound. In the first one, an approximated solution is obtained due to time discretization and a trigonometric spline approximation. The upper bound depends on the discretization grid of the trigonometric spline solution and converges to the original solution. The derived two-sided bound is an extension

to work on the solution of ODEs by trigonometric splines [Nik93; Nik04; NS05]. In the second case we used a general framework — the linear time-periodic system is approximated by Chebyshev projections e.g. in [Tre13]. We generalized results from [SW91; SB96] w.r.t. convergence and convergence rates and most importantly we could incorporate the two approximation errors of the Chebyshev projections into the rigorous two-sided bound. While the first approximation error is due to the polynomial approximation of the linear time-periodic system, the second error is due to solving the approximated system. The polynomial approximation of the linear time-periodic system yields properties of the solution such that its solution can be represented by an infinite series. Truncation of this series yields the second error. A series representation of the solution is not necessarily possible for the original system.
In summary, the bounds converge to the original solution of the linear time-periodic system as the number of splines or the degree of the Chebyshev projections is increased. For a smooth time-periodic system, the spectral bound in general superiors the trigonometric spline bound due to its faster convergence. In all cases the bounds converge to the norm of the solution if and only if the approximation converges to the solution. The computational complexity and convergence rate for the trigonometric spline bound and the spectral bound are stated. The applicability of all bounds and stability analysis of linear time-periodic systems is demonstrated by means of various examples which include a Jeffcott rotor and a parametrically excited Cantilever beam. While in [BDK13; BDK17] rigorous upper bounds on quadratic trigonometric spline interpolation and in [BD14; BDK17] rigorous upper bounds on Chebyshev projections were presented, we now generalize these bounds twofold. Firstly, the order of the spline interpolant is increased, i.e., now cubic trigonometric splines can be considered and convergence results are derived. Secondly, the rigorous bounds have been extended such that they are now two-sided, i.e., an additional lower bound on the solution norm has been derived which e.g. can be used as a certificate to prove instability for the transient behavior of the solution of a time-periodic system.

6.1 Time-Periodic Bounds

This section is based on the Floquet-Lyapunov transformation in Remark 2.5.19, which transforms the time-periodic system (6.1) into a constant linear system

$$\begin{aligned} \dot{z} &= Lz\,, \qquad \text{for } t \in I, \\ z(0) &= x_0. \end{aligned} \tag{6.2}$$

The solution to the time-periodic system (6.1) is given by Theorem 2.5.18 as $x(t) = Z(t)e^{L(t-t_0)}x_0$ and the solution to (6.2) has been given in Section 2.5.2 as $z(t) = e^{Lt}x_0$. We remark that for notational simplicity the initial condition is at $t_0 = 0$ and in this section we denote with $\Phi(t)$ a fundamental matrix with initial condition at $t_0 = 0$, i.e., $\Phi(t) := \Phi(t, 0)$.
Let us start this investigation by defining a time-dependent matrix function $\widetilde{R} : \mathbb{R} \to \mathbb{R}^{n\times n}$ as

$$\widetilde{R}(t) = Z^{-H}(t)RZ^{-1}(t),$$

for $t \in \mathbb{R}$, where $R \in \mathbb{R}^{n\times n}$ and Z is the time-periodic matrix function $Z : \mathbb{R} \to \mathbb{R}^{n\times n}$

from the general solution $x(t) = Z(t)e^{L(t-t_0)}x_0$ in Theorem 2.5.18. By Remark 2.5.19 $Z(t)$ has full rank for any $t \in \mathbb{R}$, i.e., $Z(t) \in Gl_n(\mathbb{R})$ for any $t \in \mathbb{R}$. First, we show that the matrix $\widetilde{R}(t)$ is Hermitian, positive definite and bounded for any $t \in \mathbb{R}$ under the proper assumptions on the matrices R and Z.

Lemma 6.1.1. *Let R be Hermitian and positive definite and $\widetilde{R}(t) = Z^{-H}(t)RZ^{-1}(t)$, where $Z(t)$ is defined by Floquet's normal form in (2.46). Then*

1. *$\widetilde{R}(t)$ is positive definite for all $t \in \mathbb{R}$,*
2. *$\widetilde{R}(t)$ is Hermitian for all $t \in \mathbb{R}$,*
3. *$\widetilde{R}(t)$ is t_p-periodic, i.e., $\widetilde{R}(t) = \widetilde{R}(t + t_p)$, for all $t \in \mathbb{R}$ and*
4. *$\widetilde{R}(t)$ is bounded, i.e. there exist $c, C > 0 : c \leq \|\widetilde{R}(t)\| \leq C$ for all $t \in \mathbb{R}$.*

Proof.

1. Choose u and t arbitrarily but fixed and let $\widetilde{u} = Z^{-1}(t)u$, then

$$u^H \widetilde{R}(t)u = u^H Z^{-H}(t)RZ^{-1}(t)u = \widetilde{u}^H R\widetilde{u} \geq 0,$$

   for all $\widetilde{u} \in \mathbb{C}^n$ since R is positive definite. Now,

$$\widetilde{u}^H R\widetilde{u} = 0 \Leftrightarrow \widetilde{u} = 0 \Leftrightarrow \widetilde{u} = Z^{-1}(t)u = 0 \Leftrightarrow u = 0,$$

   since $Z(t)$ has full rank and is invertible for all t.

2. $\widetilde{R}(t)$ is Hermitian, since R is Hermitian.

3. $\widetilde{R}(t)$ is t_p-periodic, since $Z(t)$ is t_p-periodic.

4. $Z^{-1}(t) = e^{Lt}\Phi^{-1}(t)$ and $Z^{-H}(t) = \Phi^{-H}(t)e^{L^H t}$ are continuous and periodic with periodicity t_p. Note, that $\Phi(t)$ is a fundamental matrix (see Definition 2.5.11), $\Phi^{-1}(t) = \Phi(-t)$ holds, see e.g. [MS76]. $\widetilde{R}(t)$ is continuous and periodic, hence, the function $p : t \mapsto \|\widetilde{R}(t)\|$ is continuous and periodic as well. Due to the extreme value theorem e.g. in [For11], p attains its minimum c and maximum C in $t_c \in [0, t_p]$ and $t_C \in [0, t_p]$, respectively. Since p is periodic, it can be bounded globally: $c \leq \|\widetilde{R}(t)\| \leq C$. Since $\widetilde{R}(t)$ has full rank for all $t \in \mathbb{R}$, $\widetilde{R}(t_c)$ has full rank and hence, $\widetilde{R}(t_c) \neq 0$ and therefore $c > 0$, i.e.,

$$\exists c, C > 0 : c \leq \|\widetilde{R}(t)\| \leq C \quad \forall t \in \mathbb{R}.$$

□

Let R be be Hermitian and positive definite, then $\|\cdot\|_R$ is a norm, see (2.14). We call $\|\cdot\|_R$ a global norm and then we define a local (time-dependent) norm $\|\cdot\|_{\widetilde{R}(t)}$, see e.g. [SM85], which is defined as

$$\|u\|_{\widetilde{R}(t)} := \langle Z^{-H}(t)RZ^{-1}(t)u, u\rangle^{\frac{1}{2}}.$$

By Lemma 6.1.1, $\|\cdot\|_{\widetilde{R}(t)}$ is well-defined and fulfills the axioms of a norm. Furthermore,

$$\begin{aligned}\|x(t)\|_{\widetilde{R}(t)} &= \langle Z^{-H}(t)RZ^{-1}(t)x(t), x(t)\rangle^{\frac{1}{2}} \\ &= \langle RZ^{-1}(t)x(t), Z^{-1}(t)x(t)\rangle^{\frac{1}{2}} \\ &= \|Z^{-1}(t)x(t)\|_R = \|z(t)\|_R = \|e^{Lt}x_0\|_R, \end{aligned} \tag{6.3}$$

for any $t \in \mathbb{R}$ holds. Now, we choose R to be given with the help of the matrix eigenvalue problem $YL^H + LY = \mu Y$,

$$\begin{aligned} R_i^{(k,k)} &:= v_k^{(i)} v_k^{(i)^H} \qquad \text{for } k = 1, \ldots, m_i\ ,\ i = 1, \ldots, m, \\ R_i &:= \sum_{k=1}^{m_i} R_i^{(k,k)}, \\ R &:= \sum_{i=1}^{m} R_i, \end{aligned}$$

where $v_k^{(i)}$ is the chain of right principal vectors of L, i.e., $Lv_k^{(i)} = \lambda_i v_k^{(i)} + v_{k-1}^{(i)}$ for $k = 1, \ldots, m_i$ and R_i is an eigenmatrix of the matrix eigenvalue problem corresponding to an eigenvalue $\mu = 2\,\mathrm{Re}(\lambda_i)$, see (3.11). In the following we generalize results from Section 3.3 to the time-periodic setting with the help of the norm $\|\cdot\|_{\widetilde{R}(t)}$.

Theorem 6.1.2 (Decoupling and filter effect of the norm $\|\cdot\|_{\widetilde{R}(t)}$)**.** *Let L be a complex matrix such that it fulfills* (2.45) *and z be the solution to the IVP* (6.2)*, $\dot{z} = Lz\,,\ z(0) = x_0$. Then*

$$\|x(t)\|_{\widetilde{R}(t)}^2 = \|z(t)\|_R^2 = \sum_{i=1}^{m}\sum_{k=1}^{m_i} \left|p_{x_0,k-1}^{(i)}(t)\right|^2 e^{2t\,\mathrm{Re}\,\lambda_i} \quad \textit{for } t \in \mathbb{R}, \tag{6.4}$$

where $p_{x_0,k-1}^{(i)}(t)$ for $k = 1, \ldots, m_i$ and $i = 1, \ldots, m$ are defined in (3.13)*.*

Proof. The relation $\|x(t)\|_{\widetilde{R}(t)}^2 = \|z(t)\|_R^2$ for all $t \in \mathbb{R}$ is given in (6.3) and

$$\|z(t)\|_R^2 = \sum_{i=1}^{m}\sum_{k=1}^{m_i} \left|p_{x_0,k-1}^{(i)}(t)\right|^2 e^{2t\,\mathrm{Re}\,\lambda_i}$$

for $t \in \mathbb{R}$ is given by Theorem 3.3.6. □

Theorem 3.3.6 has shown a *decoupling* and *filter effect* on the semi-norms $\|\cdot\|_{R_i^{(k,k)}}^2$ for $k = 1, \ldots, m_i$ and $i = 1, \ldots, m$, which carries over to the norm $\|\cdot\|_R^2$ by Theorem 3.3.6 and to $\|\cdot\|_{\widetilde{R}(t)}^2$ by Theorem 6.1.2. *Decoupling* and *filtering* are meant in the sense that we obtain a system of decoupled differential equations, where only the real part of the eigenvalues is passed and the imaginary parts are suppressed. By the following corollary the semi-norms suppress vibrations in the sense of decoupling and filtering given by Theorem 6.1.2.

Corollary 6.1.3 (Vibration-suppression property of $\|x(t)\|_{\widetilde{R}(t)}$)**.**

- *If L is diagonalizable, then*

$$\|x(t)\|^2_{\widetilde{R}(t)} = \sum_{i=1}^{n} \|x_0\|^2_{R_i} \, e^{2t \operatorname{Re} \lambda_i} \quad \textit{for } t \in \mathbb{R}.$$

- *If L is defective, i.e., it is non-diagonalizable, then*

$$\|x(t)\|^2_{\widetilde{R}(t)} = \sum_{i=1}^{m} \sum_{k=1}^{m_i} \left| p^{(i)}_{x_0,k-1}(t) \right|^2 e^{2t \operatorname{Re} \lambda_i} \quad \textit{for } t \in \mathbb{R}.$$

The monotonic behavior for linear systems was discussed in [Koh13], which we extend to the time-periodic case. If the spectral abscissa $\nu[L] = \max_{i=1,\dots,m} \operatorname{Re} \lambda_i$ is negative, i.e. $\nu[L] < 0$, and $d = \max_{i=1,\dots,m} \max_{k=1,\dots,m_i} \operatorname{degree}(p^{(i)}_{x_0,k-1}(t))$, then $\|x(t)\|_{\widetilde{R}(t)}$ behaves essentially in a way similar to $t^d e^{-t}$, i.e., there exist $t_1 > 0$ such that $\|x(t)\|_{\widetilde{R}(t)} \searrow 0$ (monotonic decrease) for $t \geq t_1$ as $t \to \infty$. If the matrix L is diagonalizable and the spectral abscissa is nonzero, then we can conclude a monotonic behavior in $\|\cdot\|_{\widetilde{R}(t)}$ since no Jordan block occurs.
Corollary 6.1.3 does not state, that in the linear time-periodic system (6.1) the vibrations are suppressed, but in the $\widetilde{R}(t)$-norm of its solution due to the decoupling and filtering effect of the norm. We would like to mention the following two cases of monotonic behavior:

1. If the spectral abscissa $\nu[L] = \max_{i=1}^{n} \operatorname{Re} \lambda_i < 0$ for a diagonalizable matrix L, then $\|x(t)\|_{\widetilde{R}(t)}$ tends monotonically to zero, i.e., $\|x(t)\|_{\widetilde{R}(t)} \searrow 0$ as $t \to \infty$.

2. If all eigenvalue have positive real part, i.e., $\operatorname{Re} \lambda_i > 0$ for $i = 1, \dots, r$, then $\|x(t)\|_{\widetilde{R}(t)}$ tends monotonically to infinity, i.e., $\|x(t)\|_{\widetilde{R}(t)} \nearrow \infty$ as $t \to \infty$. If $\|x(t)\|_{\widetilde{R}(t)} \nearrow \infty$ as $t \to \infty$, then the physical system is vibrating with an increasing amplitude and it will eventually collapse.

The monotonic behavior of $\|x(t)\|_{\widetilde{R}(t)}$ can be used to derive upper bounds on the amplitude of $\|x(t)\|_\infty$.

6.2 Trigonometric Spline Bounds

In [LT67] a method of spline approximation is introduced in order to solve ODEs numerically. This idea was further developed by e.g. [Nik93; Nik04; NS05]. Here, trigonometric B-splines of second and third order are used to solve a nonlinear ODE. We use a modified approach in order to apply it to a linear system of ODEs and further equip the computation with rigorous bounds. The unknown quantities are the coefficients of the trigonometric splines. While in the nonlinear approach one has to solve a series of nonlinear systems, this simplifies to a series of structured linear systems. Hence, a decrease of computational complexity and an effective speed-up is achieved. For further details on trigonometric splines we refer the interested reader to [Sch64; Sch81].

First, we need some mathematical basics. $(\mathbb{R}^n, \|\cdot\|_\infty)$ is a normed vector space and let $L_\infty([0,t_p],\mathbb{R}^n)$ be the space of measurable and essentially bounded functions from $[0,t_p]$ to $\mathbb{R}^n$. For a function $f \in L_\infty([0,t_p],\mathbb{R}^n)$, its essential supremum serves as an appropriate norm, which we denote in this section by $\|f\|_\infty = \|f\|_{L_\infty[0,t_p]}$ as in Definition 2.3.7.

The key idea of this section is to approximate the solution $x(t)$ to the time-periodic system (6.1) by splines. Due to the periodicity of the time-periodic system (6.1), trigonometric splines are chosen which mimic the behavior of the time-periodic matrix function $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ with $A(t) = A(t+t_p)$ for all $t \in \mathbb{R}$. Every quadratic trigonometric spline can be expressed as $\sum_{i=-1}^{r} \alpha_i S_i(t)$ and every cubic trigonometric spline as $\sum_{i=-2}^{r} \alpha_i S_i(t)$. For representing a quadratic or cubic spline, the summation index i is from -1 or -2, respectively to r which does not represent the number of nodes, but the number of intervals $[t_i, t_{i+1}]$ for $i = -1,\dots,r$ or $i = -2,\dots,r$, see Section 2.4.2. The coefficients α_i of the spline are unknown and have to be determined. In the following we describe a method how to compute the coefficients α_i. First, let us generalize the 1-dimensional quadratic trigonometric B-splines $s_2(t)$ defined in (2.26) and the cubic trigonometric B-splines $s_3(t)$ defined in (2.27) to a vector of n dimensions:

$$\begin{aligned} s_2(t) &= \sum_{i=-2}^{r} \alpha^{(i)} S_i^2(t), \\ s_3(t) &= \sum_{i=-3}^{r} \alpha^{(i)} S_i^3(t), \end{aligned}$$

where $t \in [0,t_p]$, $\alpha^{(i)} \in \mathbb{R}^n$ for $i = -2,\dots,r$ and $i = -3,\dots,r$, respectively. The subscript of $s(t)$ denotes the order of the trigonometric spline. The idea is to approximate $x(t)$ by $s(t)$, i.e., $x(t) \approx s(t)$ for $t \in [0,t_p]$, where the unknown coefficients of the trigonometric B-splines are given by the coefficient vectors $\alpha^{(i)} \in \mathbb{R}^n$. The approximation is performed by demanding that the spline s fulfills the ODE of the time-periodic system (6.1) at the node t_i, i.e., $\dot{s}(t_i) = A(t_i)s(t_i)$ for $i = 0,\dots,r$. Depending on the spline-order, i.e., quadratic or cubic, we obtain two different schemes, which are investigated in the following.

6.2.1 Quadratic Trigonometric Splines

By demanding

$$\begin{aligned} s_2(t_0) &= x_0, \\ \dot{s}_2(t_i) &= A(t_i)s_2(t_i), \end{aligned} \tag{6.5}$$

for $i = -2,\dots,r$, we obtain a sequence of $r+2$ linear systems

$$A^{(i)}\alpha^{(i)} = b^{(i)}, \tag{6.6}$$

for the coefficient vector $\alpha^{(i)}$. It is a sequence since the coefficient matrix $A^{(i)}$ and the right-hand side $b^{(i)}$ change w.r.t. node t_i,

$$\begin{aligned} A^{(i)} &= I_n - \tan\left(\tfrac{h}{2}\right) A(t_i), && \text{for } i = -1, \dots, r, \\ b^{(i)} &= \left(I_n + \tan\left(\tfrac{h}{2}\right) A(t_i)\right) \alpha^{(i-1)}, && \text{for } i = -1, \dots, r, \end{aligned}$$

where I_n is the n-dimensional identity matrix and the initial condition $s(t_0) = x_0$ yields $\alpha^{(-2)} = \cos\left(\frac{h}{2}\right) x_0 - \sin\left(\frac{h}{2}\right) A(t_0) x_0$. In [Nik93] this procedure has been investigated for nonlinear systems, where one does not solve a sequence of linear systems but a sequence of nonlinear systems by an iterative method such as the Newton method. Trigonometric splines are L-splines by Lemma 2.5.6, where the L corresponds to a certain linear differential operator, which in the case of quadratic splines is $L_3 x := x''' + x'$, where x is the solution to the time-periodic system (6.1). The convergence result from nonlinear systems carries over to the linear case and is stated in Theorem 6.2.1.

Theorem 6.2.1 ([Nik93])**.** *For $A \in \mathcal{C}^2([0, t_p], \mathbb{R}^{n\times n})$, the quadratic trigonometric spline converges quadratically to the solution, more precisely $\|x - s_2\|_\infty = \mathcal{O}(\|L_3 x\|_\infty r^{-2})$.*

Theorem 6.2.1 is proven in [Nik93]. The Theorem 6.2.1 was extended by an upper bound on $\|x(t)\|$ for the maximum norm, i.e., $\|x(t)\|_\infty$, in [BDK17]. The upper bounds on the errors at the node t_i in (6.16), (6.17) and (6.18) and the general upper bound given in (6.21) for any $t \in [0, t_p]$ are given in [BDK17] and yield a new result in terms of the following theorem, which generalizes the upper bound from [BDK17] twofold. Firstly, a two sided bound is derived, i.e., a rigorous lower bound on $\|x(t)\|$ is given as well. Secondly, the rigorous bound is valid for any norm $\|\cdot\|$ and not only the maximum norm $\|\cdot\|_\infty$.

Theorem 6.2.2. *Let $A \in \mathcal{C}^2([0, t_p], \mathbb{R}^{n\times n})$. Then, $L_3 x \in L_\infty([0, t_p], \mathbb{R}^n)$ and*

$$\|s_2(t)\| - \|L_3 x\|_\infty \Theta^{(2)}(t) \le \|x(t)\| \le \|s_2(t)\| + \|L_3 x\|_\infty \Theta^{(2)}(t), \tag{6.7}$$

where

$$\Theta^{(2)}(t) = \Theta_1^{(2)} \Theta_2^{(2)}(t) + \Theta_3^{(2)}(t) + \Theta_4^{(2)}(t), \tag{6.8}$$

$$\Theta_1^{(2)} = \frac{\left|2\tan\left(\frac{h}{2}\right) - h\right|}{L|\sin(h)| + L\left|\tan\left(\frac{h}{2}\right)\right|} \left[\left(\frac{1 + L\,|\sin(h)|}{1 - L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^i - 1\right], \tag{6.9}$$

$$\Theta_2^{(2)}(t) = 1 + L|\sin(t - t_i)| + L\left|\frac{1 - \cos(t - t_i)}{\sin(h)}\right| \left(|\cos(h)| + \frac{\left|2\tan\left(\frac{h}{2}\right) - h\right|}{1 - L\left|\tan\left(\frac{h}{2}\right)\right|}\right), \tag{6.10}$$

$$\Theta_3^{(2)}(t) = L\frac{\left|2\tan\left(\frac{h}{2}\right) - h\right|}{1 - L\left|\tan\left(\frac{h}{2}\right)\right|} \left|\frac{1 - \cos(t - t_i)}{\sin(h)}\right|, \tag{6.11}$$

$$\Theta_4^{(2)}(t) = \left|\tan\left(\tfrac{h}{2}\right)(1 - \cos(t - t_i))\right| + |t - t_i - \sin(t - t_i)|, \tag{6.12}$$

for $t \in (t_i, t_{i+1}]$ and h is sufficiently small, i.e., $L|\tan\left(\frac{h}{2}\right)| < 1$ for L being the Lipschitz constant for the ODE of the time-periodic system (6.1)*, and $L_3 x = x''' + x'$.*

Proof. Since $A \in \mathcal{C}^2([0, t_p], \mathbb{R}^{n\times n})$ and $x \in \mathcal{C}^3([0, t_p], \mathbb{R}^n)$, $L_3 x \in \mathcal{L}_\infty([0, t_p], \mathbb{R}^n)$ is obvious. We split the remaining proof in two parts. We define the error of the quadratic

spline as

$$e_2(t) := x(t) - s_2(t),$$

where $t \in [0, t_p]$. First, we prove an upper bound on the error $e_2(t) = x(t) - s_2(t) \in \mathbb{R}^n$ at the node $t = t_i$ between the solution $x(t)$ and its spline approximation $s_2(t)$. Secondly, we derive an upper bound on the error for any $t \in [0, t_p]$. The null space (see Definition 2.5.3) for the linear differential operator $L_3 = \frac{\mathrm{d}}{\mathrm{d}t} + \frac{\mathrm{d}^3}{\mathrm{d}t^3}$ is given as

$$N_{L_3} = \left\{ x \in L_1^3[0, t_p] : L_3 x(t) = 0,\, t \in [0, t_p] \right\},$$

where the Sobolev space $L_1^3[0, t_p]$ is defined in (2.15). Any set of three functions spanning the null space N_{L_3} form a fundamental solution of L_3. By Example 2.5.8 a fundamental solution is given by $N_{L_3} = \{1, \cos(t), \sin(t)\}$. The associated Green's function for L_3 is

$$G_{L_3}(t, \xi) = \begin{cases} 0 & \text{for } t \leq \xi, \\ 2\sin^2\left(\frac{t-\xi}{2}\right) & \text{for } t > \xi. \end{cases}$$

Green's function is given in (2.16) and can be found e.g. in [Sch81]. L-splines fulfill an extended Taylor formula by Remark 2.5.7, which in the case of L_3 for $t \in [t_i, t_{i+1}]$ is given as

$$\begin{aligned} x(t) &= u_x(t) + \int_{t_i}^{t} G_{L_3}(t, \xi) L_3 x(\xi) \mathrm{d}\xi\,, \text{ with} \\ u_x(t) &= x(t_i) + \dot{x}(t_i)\sin(t - t_i) + \ddot{x}(t_i)(1 - \cos(t - t_i)). \end{aligned} \tag{6.13}$$

$u_x(t)$ is the unique element in N_{L_3} such that $u_x(t_i) = x(t_i)$, $\dot{u}_x(t_i) = \dot{x}(t_i)$ and $\ddot{u}_x(t_i) = \ddot{x}(t_i)$, see e.g. [Sch81]. The derivative of the extended Taylor formula for $t \in [t_i, t_{i+1}]$ is

$$\begin{aligned} \dot{x}(t) &= \dot{u}_x(t) + \int_{t_i}^{t} \sin(t - \xi)\, L_3 x(\xi) \mathrm{d}\xi\,, \text{ with} \\ \dot{u}_x(t) &= \dot{x}(t_i)\cos(t - t_i) + \ddot{x}(t_i)\sin(t - t_i). \end{aligned} \tag{6.14}$$

1. First, we want to bound the error $\|e_2(t_i)\| = \|x(t_i) - s_2(t_i)\|$. Therefore, we bound the error for $t = t_1$ first and then derive a recursive formula for the i-th error. We can use the extended Taylor formula (6.13) since trigonometric splines are L-splines,

$$x(t_1) = x(t_0) + \ddot{x}(t_0) + \dot{x}(t_0)\sin(h) - \ddot{x}(t_0)\cos(h) + \int_{t_0}^{t_1} G_{L_3}(t_1, \xi) L_3 x(\xi) \mathrm{d}\xi.$$

   The spline s_2 fulfills the extended Taylor formula as well, but since $L_3 s_2(t) = 0$, it holds

$$s_2(t_1) = s_2(t_0) + \ddot{s}_2(t_0) + \dot{s}_2(t_0)\sin(h) - \ddot{s}_2(t_0)\cos(h).$$

Hence,

$$\begin{aligned} e_2(t_1) &= (\ddot{x}(t_0) - \ddot{s}_2(t_0)) - (\ddot{x}(t_0) - \ddot{s}_2(t_0))\cos(h) - \int_{t_0}^{t_1} G_{L_3}(t_1,\xi) L_3 x(\xi) \mathrm{d}\xi \\ &= 2(\ddot{x}(t_0) - \ddot{s}_2(t_0))\sin^2\left(\tfrac{h}{2}\right) - \int_{t_0}^{t_1} G_{L_3}(t_1,\xi) L_3 x(\xi) \mathrm{d}\xi. \end{aligned}$$

We can apply (6.14) for the derivatives $\dot{x}$ and $\dot{s}_2$,

$$\begin{aligned} \dot{x}(t_1) &= \dot{x}(t_0)\cos(h) + \ddot{x}(t_0)\sin(h) + \int_{t_0}^{t_1} \sin(t_1 - \xi) L_3 x(\xi) \mathrm{d}\xi, \\ \dot{s}_2(t_1) &= \dot{s}_2(t_0)\cos(h) + \ddot{s}_2(t_0)\sin(h), \end{aligned}$$

and subtraction yields

$$\ddot{x}(t_0) - \ddot{s}_2(t_0) = \frac{\dot{x}(t_1) - \dot{s}_2(t_1)}{\sin(h)} + \int_{t_0}^{t_1} \frac{\sin(t_1 - \xi)}{\sin(h)} L_3 x(\xi) \mathrm{d}\xi. \tag{6.15}$$

Hence,

$$\|e_2(t_1)\| = \left\| 2(\ddot{x}(t_0) - \ddot{s}_2(t_0))\sin^2\left(\tfrac{h}{2}\right) - \int_{t_0}^{t_1} G_{L_3}(t_1,\xi) L_3 x(\xi) \mathrm{d}\xi \right\|$$

and substituting (6.15) yields

$$\begin{aligned} \|e_2(t_1)\| &\leq \left\| (\dot{x}(t_1) - \dot{s}_2(t_1)) \tan\left(\tfrac{h}{2}\right) \right\| \\ &\quad + \left\| \int_{t_0}^{t_1} \left[\tan\left(\tfrac{h}{2}\right) \sin(t_1 - \xi) - G_{L_3}(t_1,\xi) \right] L_3 x(\xi) \mathrm{d}\xi \right\| \\ &\leq L \|e_2(t_1)\| \left| \tan\left(\tfrac{h}{2}\right) \right| + \|L_3 x\|_\infty \left| 2\tan\left(\tfrac{h}{2}\right) - h \right|, \end{aligned}$$

where L is the Lipschitz constant of the ODE of the time-periodic system (6.1), i.e., it fulfills the Lipschitz condition $\|\dot{x}(t) - \dot{s}_2(t)\| = \|A(t)(x(t) - s_2(t))\| \leq L\|x(t) - s_2(t)\|$, since $A \in \mathcal{C}([0, t_p], \mathbb{R}^{n\times n})$ and by periodicity of the matrix function $A : \mathbb{R} \to \mathbb{R}^{n\times n}$ by $A(t) = A(t + t_p)$ for any $t \in R$, the matrix function is bounded by $\|A\|_\infty \leq L$. For $L\left|\tan\left(\tfrac{h}{2}\right)\right| < 1$, it follows

$$\|e_2(t_1)\| \leq \|L_3 x\|_\infty \frac{\left|2\tan\left(\tfrac{h}{2}\right) - h\right|}{1 - L\left|\tan\left(\tfrac{h}{2}\right)\right|}. \tag{6.16}$$

The right-hand side of (6.16) tends to zero, especially $\frac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} \to 0$ as $h \to 0$. With the same analysis, the i-th discrete error can be bounded by

$$\|e_2(t_i)\| = \|x(t_i) - s_2(t_i)\| \leq \|e_2(t_{i-1})\| \frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} + \|L_3 x\|_\infty \frac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}. \tag{6.17}$$

The bound of the error at the i-th node consists of the error at the previous node $\|e_2(t_{i-1})\|$ with the factor $\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}$ and a cubic order term $\mathcal{O}(\|L_3 x\|_\infty h^3)$.

Additionally, we obtain an explicit upper bound for the i-th discrete error by recursively expanding the series:

$$\begin{aligned}
\|e_2(t_i)\| &\leq \|e_2(t_{i-1})\| \tfrac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} + \|L_3 x\|_\infty \tfrac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} \\
&\leq \|e_2(t_{i-2})\| \left(\tfrac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^2 + \|L_3 x\|_\infty \tfrac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} \left[1 + \tfrac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right] \\
&\leq \|e_2(t_1)\| \left(\tfrac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^{i-1} + \|L_3 x\|_\infty \tfrac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} \sum_{j=0}^{i-2} \left(\tfrac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^j
\end{aligned}$$

and with (6.16), it follows

$$\|e_2(t_i)\| \leq \|L_3 x\|_\infty \frac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} \sum_{j=0}^{i-1} \left(\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^j.$$

Since $\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} \neq 1$, the $(i-1)$-st partial sum of the (finite) geometric series can be simplified to

$$\sum_{j=0}^{i-1} \left(\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^j = \frac{\left(\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^i - 1}{\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} - 1} = \frac{\left(\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^i - 1}{L\frac{|\sin(h)|+\left|\tan\left(\frac{h}{2}\right)\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}}$$

and hence,

$$\|e_2(t_i)\| \leq \|L_3 x\|_\infty \Theta_1^{(2)}, \tag{6.18}$$

where $\Theta_1^{(2)} = \frac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{L|\sin(h)|+L\left|\tan\left(\frac{h}{2}\right)\right|} \left[\left(\frac{1+L|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|}\right)^i - 1\right]$. The right-hand side of (6.18) tends to zero as the number of nodes r tends to infinity, i.e., the error $|e_2(t_i)|$ for any $i = 0, \ldots, r$ tends to zero as well for $r \to \infty$ (Theorem 6.2.1).

2. Now we want to bound the error $e_2(t) = x(t) - s_2(t)$ for any $t \in [0, t_p]$. Therefore, let $t \in [0, t_p]$ be fixed and choose i such that $t \in (t_i, t_{i+1}]$ and apply the extended Taylor formula (6.13) to the solution and the spline:

$$\begin{aligned}
x(t) &= x(t_i) + \dot{x}(t_i)\sin(t-t_i) + \ddot{x}(t_i)(1-\cos(t-t_i)) + \int_{t_i}^{t} G_{L_3}(t,\xi) L_3 x(\xi) \mathrm{d}\xi, \\
s_2(t) &= s_2(t_i) + \dot{s}_2(t_i)\sin(t-t_i) + \ddot{s}_2(t_i)(1-\cos(t-t_i)).
\end{aligned}$$

The mean value theorem for integrals e.g. in [For11] yields, that there exists $\gamma_i \in (t_i, t)$ such that

$$x(t) = x(t_i) + \dot{x}(t_i)\sin(t-t_i) + \ddot{x}(t_i)(1-\cos(t-t_i)) + L_3 x(\gamma_i)\left(t - t_i - \sin(t-t_i)\right).$$

Then, for the error, it follows

$$e_2(t) = e_2(t_i)+\dot{e}_2(t_i)\sin(t-t_i)+\ddot{e}_2(t_i)(1-\cos(t-t_i))+L_3x(\gamma_i)\,(t-t_i-\sin(t-t_i))\,. \tag{6.19}$$

Differentiation leads to

$$\dot{e}_2(t) = \dot{x}(t)-\dot{s}_2(t) = \dot{e}_2(t_i)\cos(t-t_i)+\ddot{e}_2(t_i)\sin(t-t_i)+L_3x(\gamma_i)\,(1-\cos(t-t_i))$$

and evaluation at $t = t_{i+1}$

$$\begin{aligned} & \dot{e}_2(t_{i+1}) = \dot{e}_2(t_i)\cos(h)+\ddot{e}_2(t_i)\sin(h)+L_3x(\gamma_i)\,(1-\cos(h)) \\ \Leftrightarrow\quad & \ddot{e}_2(t_i) = -\dot{e}_2(t_i)\frac{\cos(h)}{\sin(h)}+\frac{\dot{e}_2(t_{i+1})}{\sin(h)}-L_3x(\gamma_i)\tan\left(\tfrac{h}{2}\right). \end{aligned}$$

The spline s_2 and the solution x fulfill the ODE (6.1) at the time-points t_i for $i = 0, 1, \ldots, r$, and as mentioned above, both are Lipschitz-continuous, hence

$$\|\dot{e}_2(t_i)\| = \|\dot{x}(t_i)-\dot{s}_2(t_i)\| \le \|A\|_\infty \|x(t_i)-s_2(t_i)\| = L\,\|e_2(t_i)\|\,. \tag{6.20}$$

Hence,

$$\|\ddot{e}_2(t_i)\| \le L\,\|e_2(t_i)\|\,|\cot(h)| + L\,\|e_2(t_{i+1})\|\,\tfrac{1}{|\sin(h)|} + \|L_3x\|_\infty \left|\tan\left(\tfrac{h}{2}\right)\right|$$

and by equation (6.19), (6.20) and the triangle inequality, we obtain the following upper bound on the error

$$\begin{aligned} \|e_2(t)\| \le & \|e_2(t_i)\| + \|\dot{e}_2(t_i)\|\,|\sin(t-t_i)| + \|\ddot{e}_2(t_i)\|\,|1-\cos(t-t_i)| \\ & + \|L_3x\|_\infty\,|t-t_i-\sin(t-t_i)| \\ \le & \|e_2(t_i)\| + L\,\|e_2(t_i)\|\,|\sin(t-t_i)| + |1-\cos(t-t_i)| \\ & \cdot\left(L\,\|e_2(t_i)\|\,|\cot(h)| + L\,\|e_2(t_{i+1})\|\,\tfrac{1}{|\sin(h)|} + \|L_3x\|_\infty \left|\tan\left(\tfrac{h}{2}\right)\right|\right) \\ & + \|L_3x\|_\infty\,|t-t_i-\sin(t-t_i)| \end{aligned}$$

Using the recursive bound on the error at t_{i+1}, i.e., inequality (6.17),

$$\|e_2(t_{i+1})\| \le \|e_2(t_i)\|\frac{1+L\,|\sin(h)|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|} + \|L_3x\|_\infty \frac{\left|2\tan\left(\frac{h}{2}\right)-h\right|}{1-L\left|\tan\left(\frac{h}{2}\right)\right|},$$

yields

$$\|e_2(t)\| \le \|e_2(t_i)\|\,\Theta_2^{(2)} + \|L_3x\|_\infty\,\Theta_3^{(2)} + \|L_3x\|_\infty\,\Theta_4^{(2)}.$$

By the upper bound on the i-th discrete error in (6.18), we obtain a general upper bound on the error

$$\|e_2(t)\| \le \|L_3x\|_\infty\,\Theta^{(2)}(t)., \tag{6.21}$$

where $\Theta^{(2)}, \Theta_1^{(2)}, \Theta_2^{(2)}, \Theta_3^{(2)}$ and $\Theta_4^{(2)}$ are defined in (6.8)-(6.12). Applying the triangle inequality to $\|e_2(t)\|$, we obtain $\|x(t)\| - \|s_2(t)\| \le \|x(t)-s_2(t)\|$ and

$\|s_2(t)\| - \|x(t)\| \leq \|x(t) - s_2(t)\|$. Therefore,

$$\|x(t)\| - \|s_2(t)\| \leq \|x(t) - s_2(t)\| \leq \|L_3 x\|_\infty \Theta^{(2)}(t),$$
$$\|s_2(t)\| - \|x(t)\| \leq \|x(t) - s_2(t)\| \leq \|L_3 x\|_\infty \Theta^{(2)}(t),$$

which concludes the proof.

□

The spline and the upper bound converge to the solution resp., to the norm of the solution by Theorem 6.2.1 resp. Theorem 6.2.2 as $h \to 0$.

6.2.2 Cubic Trigonometric Splines

In this section we derive two-sided bounds with cubic trigonometric splines. Compared to quadratic splines, an additional degree of freedom can be used for cubic splines, which is used for smoothness of the solution. We demand further regularity of the spline solution at the nodes. Hence, in addition to (6.5), we demand $\ddot{s}_3(t_i) = \dot{A}(t_i)s_3(t_i) + A(t_i)\dot{s}_3(t_i)$ for $i = 0, \ldots, r$. Overall, we obtain

$$\begin{aligned} s_3(t_0) &= x_0, \\ \dot{s}_3(t_i) &= A(t_i)s_3(t_i), \\ \ddot{s}_3(t_i) &= \dot{A}(t_i)s_3(t_i) + A(t_i)\dot{s}_3(t_i), \end{aligned} \tag{6.22}$$

for $i = 0, \ldots, r$. We obtain a sequence of $r+1$ linear systems of dimension $2n \times 2n$,

$$A^{(i)} \begin{bmatrix} \alpha^{(i-2)} \\ \alpha^{(i-1)} \end{bmatrix} = b^{(i)} \tag{6.23}$$

for the coefficient vector $\begin{bmatrix} \alpha^{(i-2)} \\ \alpha^{(i-1)} \end{bmatrix}$, where $i = 0, \ldots, r$. It is a sequence since the coefficient matrix $A^{(i)}$ and the right-hand side $b^{(i)}$ change w.r.t. the i-th node t_i,

$$\begin{aligned} A^{(i)} &= \begin{bmatrix} 0_n & \frac{3}{2}\cos(\frac{h}{2})\sin(\frac{h}{2})I_n \\ -3\cos(\frac{h}{2})I_n & \frac{3}{8}(3\cos(h)+1)I_n \end{bmatrix} - \begin{bmatrix} 2\sin(\frac{h}{2})\sin(h)A(t_i) & \sin^2(\frac{h}{2})A(t_i) \\ 0_n & \frac{3}{4}\sin(h)A(t_i) \end{bmatrix} \\ &\quad - \begin{bmatrix} 0_n & 0_n \\ 2\sin(\frac{h}{2})\sin(h) & \sin^2(\frac{h}{2})A'(t_i) \end{bmatrix}, \\ b^{(i)} &= \begin{bmatrix} \frac{3}{2}\sin(\frac{h}{2})\cos(\frac{h}{2})I_n + \sin^2(\frac{h}{2})A(t_i) \\ -\frac{3}{8}(3\cos(h)+1)I_n - \frac{3}{2}\sin(\frac{h}{2})\cos(\frac{h}{2})A(t_i) + \sin^2(\frac{h}{2})A'(t_i) \end{bmatrix} \alpha^{(i-3)}, \end{aligned}$$

for $i = 0, \ldots, r$, where I_n is the n-dimensional identity matrix and 0_n is the n-dimensional zero matrix. The coefficient vectors $\alpha^{(-3)}, \alpha^{(-2)}, \alpha^{(-1)}$ can be determined by the initial condition $s_3(t_0) = x_0$ and the additional linear system (6.23) for $i = 0$, i.e., at $t = t_0$. Hence in order to determine the coefficients $\alpha^{(-3)}, \alpha^{(-2)}, \alpha^{(-1)}$, we end

up solving the following $3n \times 3n$ linear system $\mathcal{A}x = b$, where

$$\mathcal{A} = \begin{bmatrix} \sin^2(\frac{h}{2})I_n & 2\sin(\frac{h}{2})\sin(h)I_n & \sin^2(\frac{h}{2})I_n \\ -\frac{3}{4}\sin(h)I_n & 0_n & \frac{3}{4}\sin(h)I_n \\ \frac{3}{8}(3\cos(h)+1)I_n & -3\cos_3(\frac{h}{2})I_n & \frac{3}{8}(3\cos(h)+1)I_n \end{bmatrix},$$

$$x = \begin{bmatrix} \alpha^{(-3)} \\ \alpha^{(-2)} \\ \alpha^{(-1)} \end{bmatrix},$$

$$b = \sin(h)\sin(\tfrac{3h}{2}) \begin{bmatrix} x_0 \\ A(t_0)x_0 \\ A'(t_0)x_0 + A(t_0)^2 x_0 \end{bmatrix}.$$

Trigonometric splines are L-splines by Lemma 2.5.6, where the L corresponds to a certain linear differential operator, which in the case of cubic trigonometric splines is

$$L_4 x := x^{(4)} + \frac{5}{2}x'' + \frac{9}{16}x.$$

The following rigorous two-sided bound is an extension to Theorem 6.2.2 w.r.t. the order of the trigonometric spline.

Theorem 6.2.3. *Let $A \in \mathcal{C}^3([0,t_p],\mathbb{R}^{n\times n})$. Then, $L_4 x \in L_\infty([0,t_p],\mathbb{R}^n)$ and*

$$\|s_3(t)\| - \Theta^{(3)}(t) \leq \|x(t)\| \leq \|s_3(t)\| + \Theta^{(3)}(t), \tag{6.24}$$

where

$$\Theta^{(3)}(t) = \tfrac{4\overline{\overline{L}}}{3}\left|\sin^3\left(\tfrac{h}{2}\right)\right|\left(\Theta_1^{(3)}\right)^{i-1}\Theta_2^{(3)}(t) + \|L_4 x\|_\infty\left(\Theta_2^{(3)}(t)\Theta_3^{(3)}\sum_{k=0}^{i-1}\left(\Theta_1^{(3)}\right)^k + \Theta_4^{(3)}(t)\right), \tag{6.25}$$

$$\begin{aligned}\Theta_1^{(3)} =& \tfrac{1}{4}\left(5-\cos(h)\right)\left|\cos\left(\tfrac{h}{2}\right)\right| + \tfrac{L}{6}\left(13-\cos(h)\right)\left|\sin\left(\tfrac{h}{2}\right)\right| + \overline{L}\left|\sin\left(\tfrac{h}{2}\right)\sin(h)\right| \\ &+ \tfrac{4\overline{\overline{L}}}{3}\left|\sin^3\left(\tfrac{h}{2}\right)\right|,\end{aligned} \tag{6.26}$$

$$\begin{aligned}\Theta_2^{(3)}(t) =& \tfrac{1}{4}\left(5-\cos(t-t_i)\right)\left|\cos\left(\tfrac{t-t_i}{2}\right)\right| + \tfrac{L}{6}\left(13-\cos(t-t_i)\right)\left|\sin\left(\tfrac{t-t_i}{2}\right)\right| \\ &+ \overline{L}\left|\sin\left(\tfrac{t-t_i}{2}\right)\sin(t-t_i)\right| + \tfrac{4\overline{\overline{L}}}{3}\left|\sin^3\left(\tfrac{t-t_i}{2}\right)\right|,\end{aligned} \tag{6.27}$$

$$\Theta_3^{(3)} = \tfrac{64+32\cos\left(\frac{h}{2}\right)}{9}\left|\sin^4\left(\tfrac{h}{4}\right)\right|, \tag{6.28}$$

$$\Theta_4^{(3)}(t) = \tfrac{64+32\cos\left(\frac{t-t_i}{2}\right)}{9}\left|\sin^4\left(\tfrac{t-t_i}{4}\right)\right|, \tag{6.29}$$

for $t \in (t_i, t_{i+1}]$, $i = -1,\ldots,r-1$, $L, \overline{L}, \overline{\overline{L}}$ being the Lipschitz-type constants for the ODE of the time-periodic system (6.1) *defined in* (6.32)-(6.34)*, and $L_4 x = x^{(4)} + \frac{5}{2}x'' + \frac{9}{16}x$.*

Proof. Since $A \in \mathcal{C}^3([0,t_p],\mathbb{R}^{n\times n})$ and $x \in \mathcal{C}^4([0,t_p],\mathbb{R}^n)$, $L_4 x \in L_\infty([0,t_p],\mathbb{R}^n)$ is obvious. We proceed as in the proof of Theorem 6.2.2, hence we split the remaining proof in two parts. Let the error between the solution of the time-periodic system and

the its cubic spline approximation be defined as

$$e_3(t) := x(t) - s_3(t),$$

where $t \in [0, t_p]$. Obviously, $e_3(t) \in \mathbb{R}^n$. First, we proof an upper bound on $e_3(t)$ at the node $t = t_i$, i.e., $e_3(t_i)$. Secondly, we derive a two-sided bound on the error for any $t \in [0, t_p]$. By Example 2.5.9 the fundamental system of L_4 is spanned by $\sin\left(\frac{t}{2}\right)$, $\cos\left(\frac{t}{2}\right)$, $\sin\left(\frac{3t}{2}\right)$ and $\cos\left(\frac{3t}{2}\right)$. The associated Green's function for L_4 is

$$G_{L_4}(t,\xi) = \begin{cases} 0 & \text{for } t \le \xi, \\ \frac{4}{3}\sin^3\left(\frac{t-\xi}{2}\right) & \text{for } t > \xi, \end{cases}$$

as given in e.g. [Sch81]. By Lemma 2.5.6 trigonometric splines are L-splines and they fulfill an extended Taylor formula by Remark 2.5.7, which in case for L_4 and $x \in \mathcal{C}^4([0, t_p], \mathbb{R}^n)$ is:

$$x(t) = u_x(t) + \int_{t_i}^{t} G_{L_4}(t,\xi) L_4 x(\xi)\,\mathrm{d}\xi, \tag{6.30}$$

for $t \in (t_i, t_{i+1}]$ with

$$\begin{aligned} u_x(t) =& x(t_i)\frac{9\cos\left(\frac{t-t_i}{2}\right)-\cos\left(\frac{3(t-t_i)}{2}\right)}{8} + \dot{x}(t_i)\frac{27\sin\left(\frac{t-t_i}{2}\right)-\sin\left(\frac{3(t-t_i)}{2}\right)}{12} \\ &+ \ddot{x}(t_i)\frac{\cos\left(\frac{t-t_i}{2}\right)-\cos\left(\frac{3(t-t_i)}{2}\right)}{2} + \dddot{x}(t_i)\frac{3\sin\left(\frac{t-t_i}{2}\right)-\sin\left(\frac{3(t-t_i)}{2}\right)}{3} \end{aligned}$$

1. First, we prove a bound on the error $e_3(t) = x(t) - s_3(t) \in \mathbb{R}^n$ at the node $t = t_{i+1}$ between the solution $x(t)$ and the cubic trigonometric spline $s_3(t)$.

   Applying the extended Taylor formula (6.30) to the solution $x(t)$ and the spline $s_3(t)$, we obtain the $i+1$-st discretization error at $t = t_{i+1}$:

   $$\begin{aligned} e_3(t_{i+1}) =& e_3(t_i)\tfrac{5-\cos(h)}{4}\cos\left(\tfrac{h}{2}\right) + \dot{e}_3(t_i)\tfrac{13-\cos(h)}{6}\sin\left(\tfrac{h}{2}\right) + \ddot{e}_3(t_i)\sin\left(\tfrac{h}{2}\right)\sin(h) \\ &+ \dddot{e}_3(t_i)\tfrac{4}{3}\sin^3\left(\tfrac{h}{2}\right) - \tfrac{64+32\cos\left(\frac{h}{2}\right)}{9}\int_{t_i}^{t} G_{L_4}(t,\xi) L_4 x(\xi)\,\mathrm{d}\xi. \end{aligned} \tag{6.31}$$

   Since $A \in \mathcal{C}^3[0, t_p]$, we differentiate the linear time-periodic system and obtain

   $$\begin{aligned} \dot{x} &= A(t)x, \\ \ddot{x} &= \left[\dot{A}(t) + (A(t))^2\right] x, \\ \dddot{x} &= \left[\ddot{A}(t) + 2\dot{A}(t)A(t) + A(t)\dot{A}(t) + (A(t))^3\right] x, \end{aligned}$$

   for $t \in [0, t_p]$. The right hand-sides of the derivatives are Lipschitz continuous in

x since they are linear in x and can be bounded by

$$\|A(t)\| \leq L, \tag{6.32}$$

$$\|\dot{A}(t) + (A(t))^2\| \leq \overline{L}, \tag{6.33}$$

$$\|\ddot{A}(t) + 2\dot{A}(t)A(t) + A(t)\dot{A}(t)\,(A(t))^3\| \leq \overline{\overline{L}}, \tag{6.34}$$

for any $t \in [0, t_p]$, since a continuous function attains its maximum in a compact set $[0, t_p]$ and $A \in \mathcal{C}^3[0, t_p]$. Hence, $\|\dot{x}\| \leq L\|x\|$, $\|\ddot{x}\| \leq \overline{L}\|x\|$, $\|\dddot{x}\| \leq \overline{\overline{L}}\|x\|$ and $\|\dot{e}_3(t_i)\| \leq L\|e_3(t_i)\|$, $\|\ddot{e}_3(t_i)\| \leq \overline{L}\|e_3(t_i)\|$, $\|\dddot{e}_3(t_i)\| \leq \overline{\overline{L}}\|e_3(t_i)\|$. Using these upper bounds on the errors $\|\dot{e}_3(t_i)\|$, $\|\ddot{e}_3(t_i)\|$ and $\|\dddot{e}_3(t_i)\|$ in (6.31), we obtain

$$\|e_3(t_{i+1})\| \leq \|e_3(t_i)\|\,\Theta_1^{(3)} + \|L_4 x\|_\infty \Theta_3^{(3)}, \tag{6.35}$$

for $i = 0, \ldots, r-1$. By Example 2.3.12 on the Taylor series, it is obvious that $\Theta_1^{(3)} \in \mathcal{O}(1)$ and $\Theta_2^{(3)} \in \mathcal{O}(h^4)$. The error between the solution and the spline and their first two derivatives at node t_0 is by construction zero, i.e., $e_3(t_0) = \dot{e}_3(t_0) = \ddot{e}_3(t_0) = 0$. We then obtain the following bound on the error at the node t_1:

$$\|e_3(t_1)\| \leq \tfrac{4\overline{\overline{L}}}{3}\left|\sin^3\left(\tfrac{h}{2}\right)\right| + \|L_4 x\|_\infty \Theta_3^{(3)}. \tag{6.36}$$

Via recursively applying (6.35) with (6.36), we obtain the following bound for the $i+1$-discrete error

$$\|e_3(t_{i+1})\| \leq \tfrac{4\overline{\overline{L}}}{3}\left|\sin^3\left(\tfrac{h}{2}\right)\right| \left(\Theta_1^{(3)}\right)^i + \|L_4 x\|_\infty \Theta_3^{(3)} \sum_{k=0}^{i} \left(\Theta_1^{(3)}\right)^k \tag{6.37}$$

Obviously, $e_3(t_{i+1}) \in \mathcal{O}(h^3)$.

2. Secondly, we want to bound the error $e_3(t) = x(t) - s_3(t)$ for any $t \in [0, t_p]$. Therefore, let $t \in [0, t_p]$ be fixed and choose i such that $t \in (t_i, t_{i+1}]$ and apply the extended Taylor formula (6.30) to the solution $x(t)$ and the spline $s_3(t)$, then we obtain a bound on the error. Here, we distinguish two cases:

$$\|e_3(t)\| \leq \begin{cases} \frac{4}{3}\overline{\overline{L}}\sin^3(\frac{t-t_0}{2}) + \|L_4 x\|_\infty \Theta_4^{(3)}(t), & \text{for } i = 0, \\ \|e_3(t_i)\|\Theta_2^{(3)}(t) + \|L_4 x\|_\infty \Theta_4^{(3)}(t), & \text{for } i > 0. \end{cases}$$

Using the recursive bound on the i-th discrete error (6.37) for $i > 0$, we obtain

$$\begin{aligned} \|e_3(t)\| &\leq \tfrac{4\overline{\overline{L}}}{3}\left|\sin^3\left(\tfrac{h}{2}\right)\right| \left(\Theta_1^{(3)}\right)^{i-1} \Theta_2^{(3)}(t) \\ &\quad + \|L_4 x\|_\infty \left(\Theta_2^{(3)}(t)\Theta_3^{(3)} \sum_{k=0}^{i-1} \left(\Theta_1^{(3)}\right)^k + \Theta_4^{(3)}(t)\right) \\ &= \Theta^{(3)}(t). \end{aligned}$$

By applying the triangle inequality to $\|e_3(t)\|$, we obtain $\|x(t)\| - \|s_3(t)\| \leq$

$\|x(t) - s_3(t)\|$ and $\|s_3(t)\| - \|x(t)\| \leq \|x(t) - s_3(t)\|$. Hence,

$$\|x(t)\| - \|s_3(t)\| \leq \|x(t) - s_3(t)\| \leq \Theta^{(3)}(t),$$
$$\|s_3(t)\| - \|x(t)\| \leq \|x(t) - s_3(t)\| \leq \Theta^{(3)}(t),$$

which concludes the proof.

□

The two-sided bound in Theorem 6.2.3 shows cubic convergence of the spline s_3 to the solution x, which we summarize in Corollary 6.2.4.

Corollary 6.2.4. *For $A \in \mathcal{C}^3([0, t_p], \mathbb{R}^{n\times n})$, the cubic trigonometric spline converges cubically to the solution, more precisely $\|x - s_3\|_\infty = \mathcal{O}(\|L_4 x\|_\infty r^{-3})$.*

Actually, a false proof of cubic convergence for cubic trigonometric splines to nonlinear ODEs is given in [NS05].

6.3 Spectral Bounds

The key idea is to replace the system (6.1) by an approximation. We use the spectral method [GO77; Tre00] in the setting of polynomial approximation of linear ordinary differential equations [BD14; Fun92]. The solution of the approximated system is entire and hence, the truncation error of the approximated solution can be given. Here, we approximate the system matrix by Chebyshev polynomials [Che54] and use results from approximation theory [Tre13] in order to derive rigorous bounds on the original solution $x(t)$.
We need some results from approximation theory, here we focus on Chebyshev polynomials introduced in [Che54] and Chebyshev projections in Section 2.4.1. In the following we explain the general idea of the spectral method and how we use the results from approximation theory in order to derive bounds. The resulting bound depends heavily on how well the original system is approximated.
We now return to our original problem of a linear time-periodic system (6.1) but instead of solving it directly, we first approximate it by the following system

$$\begin{aligned} \dot{y}(t) &= (P_m A)(t) y(t) \qquad \forall t \in [0, t_p], \\ y(0) &= x_0, \end{aligned} \tag{6.38}$$

where $(P_m A)$ denotes the component-wise Chebyshev projection of A, see (2.22). If $(P_m A)(t_1)$ commutes with $(P_m A)(t_2)$ for all times t_1 and t_2, then the solution to the approximated system (6.38) is given by $y(t) = \exp\left(\int_0^t (P_m A)(\tau)\mathrm{d}\tau\right) x_0$ as shown in Section 2.5.2. $y(t) = \exp\left(\int_0^t (P_m A)(\tau)\mathrm{d}\tau\right) x_0$ is an entire function (see Definition 2.3.13), since polynomials and their exponentials are entire (see Remark 2.3.14). But in general the commutativity of $(P_m A)(t)$ is a rather strong assumption. By Theorem 2.5.13, it follows that the solution $y(t)$ is entire since the function $(P_m A)(t)$ is a polynomial which by definition is entire. If the approximation is exact, i.e. $a_{ij}(t)$ is a polynomial of degree at most m for $i, j = 1, \ldots, n$, then $x(t)$ and $y(t)$ coincide. In

order to prove rigorous upper bounds on $x(t)$, we use Theorem 2.4.3 and 2.4.5 to bound the difference between the original function A and its Chebyshev projection. These bounds depend on the smoothness of the system matrix A.
We assume in this section that $\|\cdot\|$ is given as the Manhattan norm, the Euclidean norm or the maximum norm, i.e.,

$$\|\cdot\| = \|\cdot\|_p, \qquad \text{for } p \in \{1, 2, \infty\} \tag{6.39}$$

see (2.11), (2.12) and (2.13).

Theorem 6.3.1. *If $a_{ij} \in \mathcal{AC}^{k-1}[0, t_p]$ and the k-th derivative $a_{ij}^{(k)}$ is of bounded variation V for all $i, j = 1, \ldots, n$, then for any $m > k > 0$:*

$$\|y(t)\| - \Psi_{BV}(t) \leq \|x(t)\| \leq \|y(t)\| + \Psi_{BV}(t), \tag{6.40}$$

where $\Psi_{BV}(t) = \frac{2nVe^{Lt}}{\pi k(m-k)^k} \int_0^t \|y(s)\| \, \mathrm{d}s$ and L is the Lipschitz constant for the ODE of the time-periodic system (6.1).

Theorem 6.3.2. *If a_{ij} is analytic in $[0, t_p]$ and analytically continuable to the open Bernstein ellipse $\mathcal{E}_\rho$, where it satisfies $|a_{ij}(t)| \leq M$ for all $i, j = 1, \ldots, n$ for some M, then for any $m \geq 0$:*

$$\|y(t)\| - \Psi_{analytic}(t) \leq \|x(t)\| \leq \|y(t)\| + \Psi_{analytic}(t), \tag{6.41}$$

where $\Psi_{analytic}(t) = \frac{2Mn\rho^{-m}e^{Lt}}{\rho - 1} \int_0^t \|y(s)\| \, \mathrm{d}s$ and L is the Lipschitz constant for the ODE of the time-periodic system (6.1).

Now, we prove Theorem 6.3.1 and 6.3.2.

Proof. $x(t)$ and $y(t)$ fulfill the integral formulation of the ODE

$$\begin{aligned} x(t) - y(t) &= \int_0^t A(s)x(s) - (P_m A)(s)y(s)\mathrm{d}s \\ &= \int_0^t A(s)x(s) - A(s)y(s) + A(s)y(s) - (P_m A)(s)y(s)\mathrm{d}s \\ &= \int_0^t A(s)\left[x(s) - y(s)\right] + \left[A(s) - (P_m A)(s)\right] y(s)\mathrm{d}s \end{aligned}$$

Taking the norm $\|\cdot\|$ on both sides (see (6.39)) and using the triangle inequality yields

$$\|x(t) - y(t)\| \leq \int_0^t \left(\|A(s)\|\|x(s) - y(s)\| + \|A(s) - (P_m A)(s)\|\|y(s)\|\right) \mathrm{d}s$$

We remark that $\|A(s)\|$ denotes the induced matrix norm w.r.t. $\|\cdot\|$ (see Definition 2.2.6), which is a compatible matrix norm by Remark 2.2.7. The case of $A \equiv 0$ and $x = const$, is trivial. Otherwise, we define β in Gronwall's lemma 2.5.2 as the Lipschitz constant for the ODE of the time-periodic system (6.1), i.e., $\|A\|_\infty \leq L =: \beta$.

1. If the assumptions of Theorem 6.3.1 are fulfilled, then we conclude in the following $\|A(s) - (P_m A)(s)\|_p \leq \frac{2nV}{\pi k(m-k)^k}$, where $p \in \{1, 2, \infty\}$.

For the Manhattan norm, i.e., $\|\cdot\| = \|\cdot\|_1$, it follows

$$\|A(s) - (P_m A)(s)\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} \underbrace{|a_{ij}(s) - (P_m a_{ij})(s)|}_{\le \frac{2V}{\pi k(m-k)^k}} \le \frac{2nV}{\pi k(m-k)^k}.$$

For the maximum norm, i.e., $\|\cdot\| = \|\cdot\|_\infty$, it follows

$$\|A(s) - (P_m A)(s)\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} \underbrace{|a_{ij}(s) - (P_m a_{ij})(s)|}_{\le \frac{2V}{\pi k(m-k)^k}} \le \frac{2nV}{\pi k(m-k)^k}.$$

By $\|A\|_2^2 = \lambda_{\max}(A^H A) \le \|A^H A\|_1 \le A^H\|_1 \|A\|_1 = \|A\|_\infty \|A\|_1$, we obtain

$$\begin{aligned} \|A(s) - (P_m A)(s)\|_2 &\le \sqrt{\|A(s) - (P_m A)(s)\|_1 \|A(s) - (P_m A)(s)\|_\infty} \\ &\le \frac{2nV}{\pi k(m-k)^k}. \end{aligned}$$

Overall, for any norm $\|\cdot\| = \|\cdot\|_p$, where $p \in \{1, 2, \infty\}$ (see (6.39)), it follows

$$\|A(s) - (P_m A)(s)\| \le \frac{2nV}{\pi k(m-k)^k}.$$

Therefore,

$$\|x(t) - y(t)\| \le \beta \int_0^t \|x(s) - y(s)\| \,\mathrm{d}s + \frac{2nV}{\pi k(m-k)^k} \int_0^t \|y(s)\| \,\mathrm{d}s$$

and applying Gronwall's lemma 2.5.2 with

$$\begin{aligned} g(t) &= \|x(t) - y(t)\|, \\ \alpha(t) &= \frac{2nV}{\pi k(m-k)^k} \int_0^t \|y(s)\| \,\mathrm{d}s \end{aligned}$$

and $\beta = \text{const} > 0$ yields

$$\|x(t) - y(t)\| \le \frac{2nVe^{Lt}}{\pi k(m-k)^k} \int_0^t \|y(s)\| \,\mathrm{d}s. \tag{6.42}$$

With the reverse triangle inequality the theorem follows.

2. If the assumptions of Theorem 6.3.2 are fulfilled, then

$$\|A(s) - (P_m A)(s)\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} \underbrace{|a_{ij}(s) - (P_m a_{ij})(s)|}_{\le \frac{2M\rho^{-m}}{\rho-1}} \le \frac{2nM\rho^{-m}}{\rho - 1}.$$

The remaining proof is analogous to the previous case.

□

Remark 6.3.3. *In order to apply Theorem 6.3.1 and 6.3.2, the ODE system* (6.38) *has to be solved nevertheless. But the solution y to the IVP* (6.38) *is entire due to Theorem 2.5.13.*

Corollary 6.3.4. *Since y is entire by Remark 6.3.3 and Theorem 2.4.5, it follows that*

$$\|y - P_m y\|_{L_\infty[0,t_p]} \leq \frac{2M\rho^{-m}}{\rho - 1}$$

in the Bernstein ellipse $\mathcal{E}_\rho$, where $|y_i(t)| \leq M$ is satisfied for $i = 1, \ldots, n$ and some M.

The Chebyshev projections of A and y do not necessarily have the same degree, hence in the following we distinguish them by their subscripts. The index A refers to the matrix function A and an index y to the solution of the IVP (6.38).

Corollary 6.3.5. *If $a_{ij} \in \mathcal{AC}^{k-1}[0, t_p]$ and the k-th derivative $a_{ij}^{(k)}$ is of bounded variation V for all $i, j = 1, \ldots, n$, then $P_{m_y} y$ converges to x at rate k, i.e., $\|x - P_{m_y} y\|_{L_\infty[0,t_p]} = \mathcal{O}(V m_A^{-k})$.*

Proof. Due to equation (6.42) in the proof of Theorem 6.3.1 and Theorem 2.4.5, we obtain

$$\begin{aligned} \|x - P_{m_y} y\|_{L_\infty[0,t_p]} &\leq \|x - y\|_{L_\infty[0,t_p]} + \|y - P_{m_y} y\|_{L_\infty[0,t_p]} \\ &\leq \frac{2nVe^{Lt_p}t_p}{\pi k(m_A - k)^k}\|y\|_{L_\infty[0,t_p]} + \frac{2M_y\rho_y^{-m_y}}{\rho_y - 1}. \end{aligned} \tag{6.43}$$

We choose the same approximation degree, i.e., $m_A = m_y$. Since $\|y\|_{L_\infty[0,t_p]}$ is bounded, the right-hand sides of (6.43) tend to zero as $m_A, m_y \to \infty$. Hence, $P_{m_y} y$ converges to the original solution x at a rate of order k. □

Corollary 6.3.6. *If a_{ij} is analytic in $[0, t_p]$ and analytically continuable to the open Bernstein ellipse $\mathcal{E}_\rho$, where it satisfies $|a_{ij}(t)| \leq M_A$ for all $i, j = 1, \ldots, n$ for some M, then $P_{m_y} y$ converges to x geometrically, i.e., $\|x - P_{m_y} y\|_\infty = \mathcal{O}(M\rho_A^{-m_A})$.*

Proof. As previously, due to equation (6.42) in the proof of Theorem 6.3.1 and Theorem 2.4.5, we obtain

$$\|x - P_{m_y} y\|_{L_\infty[0,t_p]} \leq \frac{2M_A n\rho_A^{-m_A}e^{Lt_p}t_p}{\rho_A - 1}\|y\|_{L_\infty[0,t_p]} + \frac{2M_y\rho_y^{-m_y}}{\rho_y - 1}. \tag{6.44}$$

We choose the same approximation degree, i.e., $m_A = m_y$ and $M = \max\{M_A, M_y\}$. Since $\|y\|_{L_\infty[0,t_p]}$ is bounded, the right-hand sides of (6.44) tend to zero as $m_A, m_y \to \infty$. Hence, $P_{m_y} y$ converges to the original solution x at a geometric rate. □

For a better Chebyshev approximation, i.e., for larger approximation levels m_A and m_y, one hopes to have sharper bounds indicated by Corollary 6.3.5 and 6.3.6. This does not necessarily mean that the sharper bound is closer to the norm of the solution for any $t \in [0, t_p]$. Sharper is meant w.r.t. the convergence rate given by Corollary 6.3.5 and 6.3.6.

Theorem 6.3.7. *If $a_{ij} \in \mathcal{AC}^{k-1}[0,t_p]$ and the k-th derivative $a_{ij}^{(k)}$ is of bounded variation V for all $i,j=1,\dots,n$, then for any $m_A > k > 0$:*

$$\left\|(P_{m_y}y)(t)\right\| - \varepsilon_{BV}(t) \le \|x(t)\| \le \left\|(P_{m_y}y)(t)\right\| + \varepsilon_{BV}(t), \tag{6.45}$$

where $\varepsilon_{BV}(t) = \frac{2nVe^{Lt}}{\pi k(m_A-k)^k}\int_0^t \left\|(P_{m_y}y)(s)\right\| \mathrm{d}s + \frac{2M_y\rho_y^{-m_y}}{\rho_y-1}\left(1+\frac{2nVe^{Lt}t}{\pi k(m_A-k)^k}\right)$ and L is the Lipschitz constant for the ODE of the time-periodic system (6.1).

Proof. Due to equation (6.42) in the proof of Theorem 6.3.1 and Theorem 2.4.5, we obtain

$$\begin{aligned}\|x(t)-(P_{m_y}y)(t)\| &\le \|x(t)-y(t)\| + \|y(t)-(P_{m_y}y)(t)\| \\ &\le \frac{2nVe^{Lt}}{\pi k(m_A-k)^k}\int_0^t \|y(s)\|\mathrm{d}s + \frac{2M_y\rho_y^{-m_y}}{\rho_y-1}.\end{aligned} \tag{6.46}$$

By Corollary 6.3.4, we obtain

$$\|y(t)\| \le \|(P_m y)(t)\| + \frac{2M_y\rho_y^{-m}}{\rho_y-1}, \qquad \text{for } t\in[0,t_p],$$

which inserted in (6.46) yields

$$\begin{aligned}\|x(t)-(P_{m_y}y)(t)\| &\le \frac{2nVe^{Lt}}{\pi k(m_A-k)^k}\int_0^t \left\|(P_{m_y}y)(s)\right\| \mathrm{d}s + \frac{2M_y\rho_y^{-m_y}}{\rho_y-1}\left(1+\frac{2nVe^{Lt}t}{\pi k(m_A-k)^k}\right) \\ &= \varepsilon_{BV}(t).\end{aligned}$$

Applying the triangle and reverse triangle inequality concludes the proof. □

Remark 6.3.8. *We note that $\varepsilon_{BV}(t) \to 0$ as $m_y, m_A \to \infty$ for any $t \in [0,t_p]$.*

Theorem 6.3.9. *If a_{ij} is analytic in $[0,t_p]$ and analytically continuable to the open Bernstein ellipse $\mathcal{E}_\rho$, where it satisfies $|a_{ij}(t)| \le M$ for all $i,j=1,\dots,n$ for some M, then for any $m \ge 0$:*

$$\left\|(P_{m_y}y)(t)\right\| - \varepsilon_{analytic}(t) \le \|x(t)\| \le \left\|(P_{m_y}y)(t)\right\| + \varepsilon_{analytic}(t), \tag{6.47}$$

where $\varepsilon_{analytic}(t) = \frac{2M_An\rho_A^{-m_A}e^{Lt}}{\rho_A-1}\int_0^t \left\|(P_{m_y}y)(t)\right\| \mathrm{d}s + \frac{2M_y\rho_y^{-m_y}}{\rho_y-1}\left(1+\frac{2M_An\rho_A^{-m_A}e^{Lt}t}{\rho_A-1}\right)$ and L is the Lipschitz constant for the ODE of the time-periodic system (6.1).

Proof. Due to equation (6.42) in the proof of Theorem 6.3.1 and Theorem 2.4.5, we obtain

$$\|x(t)-(P_{m_y}y)(t)\| \le \frac{2M_An\rho_A^{-m_A}e^{Lt}}{\rho_A-1}\int_0^t \|y(s)\|\,\mathrm{d}s + \frac{2M_y\rho_y^{-m_y}}{\rho_y-1}. \tag{6.48}$$

By Corollary 6.3.4, we obtain

$$\|y(t)\| \le \|(P_m y)(t)\| + \frac{2M_y\rho_y^{-m}}{\rho_y-1}, \qquad \text{for } t\in[0,t_p],$$

which inserted in (6.48) yields

$$\begin{aligned}\|x(t) - (P_{m_y}y)(t)\| &\leq \frac{2M_A n\rho_A^{-m_A}e^{Lt}}{\rho_A - 1}\int_0^t \left\|(P_{m_y}y)(t)\right\| \,\mathrm{d}s + \frac{2M_y\rho_y^{-m_y}}{\rho_y - 1}\left(1 + \frac{2M_A n\rho_A^{-m_A}e^{Lt}t}{\rho_A - 1}\right)\\ &= \varepsilon_{analytic}(t).\end{aligned}$$

Applying the triangle and reverse triangle inequality concludes the proof. □

Remark 6.3.10. *We note that* $\varepsilon_{analytic}(t) \to 0$ *as* $m_y, m_A \to \infty$ *for any* $t \in [0, t_p]$.

The rigorous two sided bounds in Theorem 6.3.7 and 6.3.9 tend to the norm of the solution $\|x(t)\|$ as $m_A, m_y \to \infty$ by Remark 6.3.8 and 6.3.10.
If the matrix function A is analytic, one does not need to replace the original system by (6.38) since even for the original system the solution is analytic by Theorem 2.5.13. But for the sake of completeness we also derived bounds in this case and the bounds are very tight for moderate m_A as shown in Section 6.4.
Similar results can be obtained for interpolation instead of Chebyshev projection. In this context, the main question concerns the interpolation points. If Chebyshev points are chosen, then the Chebyshev interpolant satisfies Theorem 2.4.3 and 2.4.5 with an additional factor 2, see e.g. [Tre13]. Hence, one can obtain results in view of Theorem 6.3.1 and 6.3.2 with the same additional factor.

6.4 Numerical Results

While the two-sided bounds on the solution of a time-periodic system in Theorem 6.2.2 and 6.2.3 were established for any norm $\|\cdot\|$, the two-sided spectral bounds defined in equations (6.45) and (6.47) were established for a norm $\|\cdot\|_p$, where $p \in \{1, 2, \infty\}$. Hence, the maximum norm is chosen in order to compare the two-sided bounds for the transient behavior of the solution to a time-periodic system, i.e., $f_\ell(t) \leq \|x(t)\|_\infty \leq f_u(t)$. The rigorous two-sided bounds $f_\ell(t)$ and $f_u(t)$ are illustrated by two examples that can be described by a time-periodic system of the form (6.1). The first example is chosen such that the solution to the time-periodic system can be given analytically and the second example is a Jeffcott rotor. Both examples are discussed below in more detail. The two-sided bounds $f_\ell(t)$ and $f_u(t)$ are given by Theorem 6.2.2 for quadratic trigonometric splines, by Theorem 6.2.3 for cubic trigonometric splines and equations (6.45) and (6.47) by Chebyshev projections. The approximation degree r and m_A for each example is given in Table 6.1. In the following, the parameters r and m_A of the trigonometric spline or the spectral bound, respectively, are chosen such that firstly, a visible difference between the solution and its respective two-sided bounds can be seen and secondly, an effect of the parameters can be noticed. In the Figures 6.1 and 6.3 the spectral bound cannot be distinguished from the original solution if the order of the Chebyshev projection m_A is increased slightly This observation does not hold for the quadratic trigonometric spline bound and cubic trigonometric spline bound since its convergence is slower, see Table 6.3 and Figure 6.5 compared to Figure 6.6. But for a larger number of nodes r, the quadratic trigonometric spline bound and the cubic trigonometric spline bound tend to the solution quadratically and cubically by Theorem 6.2.1 and 6.2.3, respectively, compare Figures 6.1 and 6.3.

Example	**Dim.**	**Trigonometric spline** quadratic	cubic	**Spectral**
$\dot{x} = \lvert\sin(2\pi t)\rvert^3 x$	$n = 1$	$r = 15, 20$	$r = 15, 20$	$m_A = 10, 13$
Jeffcott rotor	$n = 4$	$r = 5000$	$r = 1000, 3000$	$m_A = 33, 34, 35$

Table 6.1: Setting for trigonometric spline bound and spectral bound

Computation of global extrema is not an easy task due to the possibly large number of local minima and maxima of the objective function, see e.g. [HT96]. We used MATLAB's *Global Search* strategy with *fminsearch*, which is an NLP solver that uses the simplex search method [Lag+98], in order to determine $L, \|L_3 x\|_{L_\infty}$ and $\|L_4 x\|_{L_\infty}$. The computed values for $L, \|L_3 x\|_\infty$ and $\|L_4 x\|_{L_\infty}$ are given in Table 6.2. They are used in the figures mentioned above and also appear in the convergence rates of the methods in Table 6.3. Note, that the parameters ρ_A and M_A with respect to the spectral bound are not unique, especially any Bernstein ellipse can be chosen since the function is entire. Here, we chose ρ_A with respect to the decay of the Chebyshev coefficients $|c_k|$ given by (2.21) but for the sake of simplicity the derivation is omitted and for the appropriate examples ρ_A is given in Table 6.2. M_A is determined by the strategy mentioned above, i.e., by a combination of *fminsearch* and *Global Search*.

Example	**Trigonometric Spline Bound**			**Spectral Bound** (6.47)		(6.45)
	L	$\|L_3 x\|_\infty$	$\|L_4 x\|_\infty$	M_A	ρ_A	V
$\dot{x} = \lvert\sin(2\pi t)\rvert^3 x$	1	$3.3 \cdot 10^2$	$4.0 \cdot 10^3$	—	—	$4\pi^3$
Jeffcott rotor	1.15	$3.7 \cdot 10^3$	$1.3 \cdot 10^4$	1.12	2.57	—

Table 6.2: Constants used for trigonometric spline bound and spectral bound

The first example is a one-dimensional IVP $\dot{x}(t) = |\sin(2\pi t)|^3 x(t)$ with initial condition $x(0) = 1$. The function of the right hand-side $A(t) = |\sin(2\pi t)|^3$ is thrice differentiable and $A^{(k)}$ is absolutely continuous, i.e., $A \in \mathcal{AC}^3[0, t_p]$. We use this example as here, we are able to compare our results to the analytical solution, which is

$$x(t) = \begin{cases} \exp\left(\frac{\cos_3(2\pi t)}{6\pi} - \frac{\cos(2\pi t)}{2\pi}\right) & \text{if } t \in [0, 0.5), \\ \exp\left(-\frac{\cos_3(2\pi t)}{6\pi} + \frac{\cos(2\pi t)}{2\pi} + \frac{2}{3\pi}\right) & \text{if } t \in [0.5, 1]. \end{cases}$$

The results of the trigonometric spline bound and the spectral bound are shown in Figure 6.1. For better approximation levels, the trigonometric spline and spectral bound are closer to the original solution $\|x(t)\|_\infty$ as indicated by the convergence results. The convergence rates are quadratic and cubic as shown in Table 6.3.
Figure 6.2 shows the solution of the first example in the Euclidean norm and the weighted time-dependent norm $\|\cdot\|_{\widetilde{R}(t)}$. For the one-dimensional example, the Euclidean norm and the maximum norm coincide with the absolute value, i.e. $|\cdot| = \|\cdot\|_2 = \|\cdot\|_\infty$. Furthermore, the weighted R-norm is a scaling, but since the single eigenvector is normalized, $|\cdot| = \|\cdot\|_2 = \|\cdot\|_\infty = \|\cdot\|_R$ holds. The weighted time-dependent

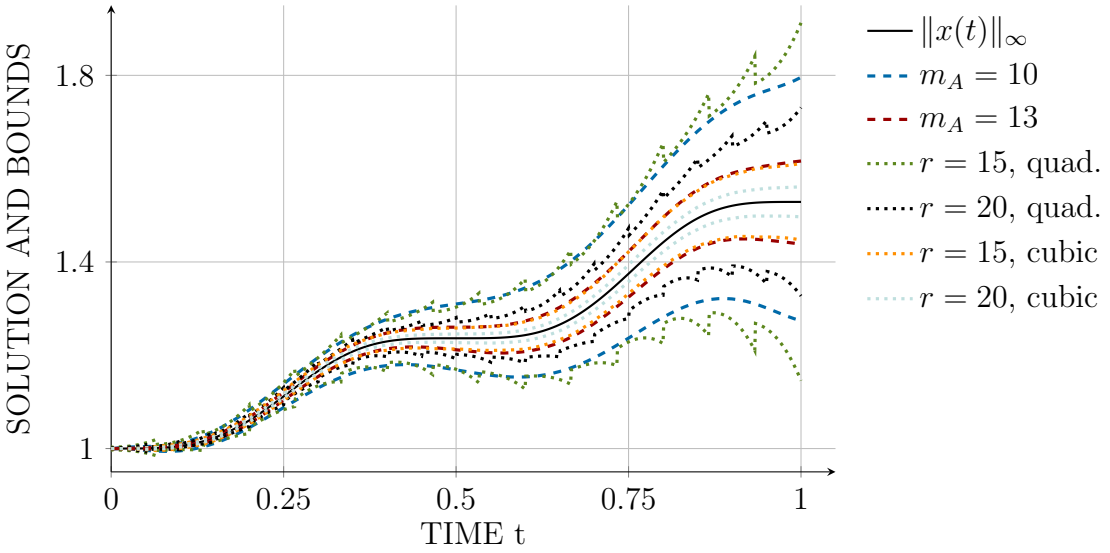


Figure 6.1: Solution for $A(t) = |\sin(2\pi t)|^3$ for $t \in [0, 1]$.

norm $\|\cdot\|_{\widetilde{R}(t)}$ suppresses the oscillations and since the spectral abscissa is positive, $\nu[L] = 0.424413181578411 > 0$, a monotonic increase can be observed, see Corollary 6.1.3.

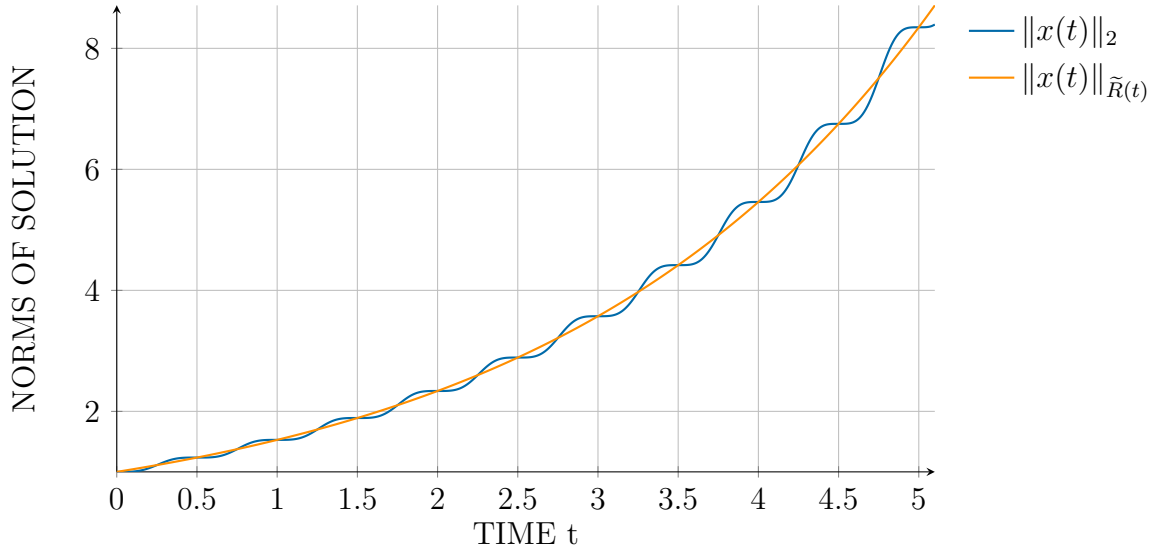


Figure 6.2: Solution for $A(t) = |\sin(2\pi t)|^3$ for $t \in [0, 5]$.

The second example is a Jeffcott rotor on an anisotropic shaft supported by anisotropic bearings [All09]. It can be modeled as a linear time-periodic system (6.1) where $A(t)$ is entire with system dimension $n = 4$. The same parameter values are chosen as in [All09]. This is an asymptotically stable system since the maximal Lyapunov exponent is $\nu[L] = -0.002000131812440 < 0$. The results are illustrated in Figure 6.3. The quadratic trigonometric spline bound for $r = 50,000$ is highly oscillatory such that some components of its graph in Figure 6.3 cannot be distinguished anymore. But nevertheless, the upper bound is valid. In order to sharpen this bound we can increase

the number of nodes r or we can increase the order of the trigonometric spline, e.g., the cubic trigonometric spline bound is been visualized for $r = 1000$ and $r = 3000$.

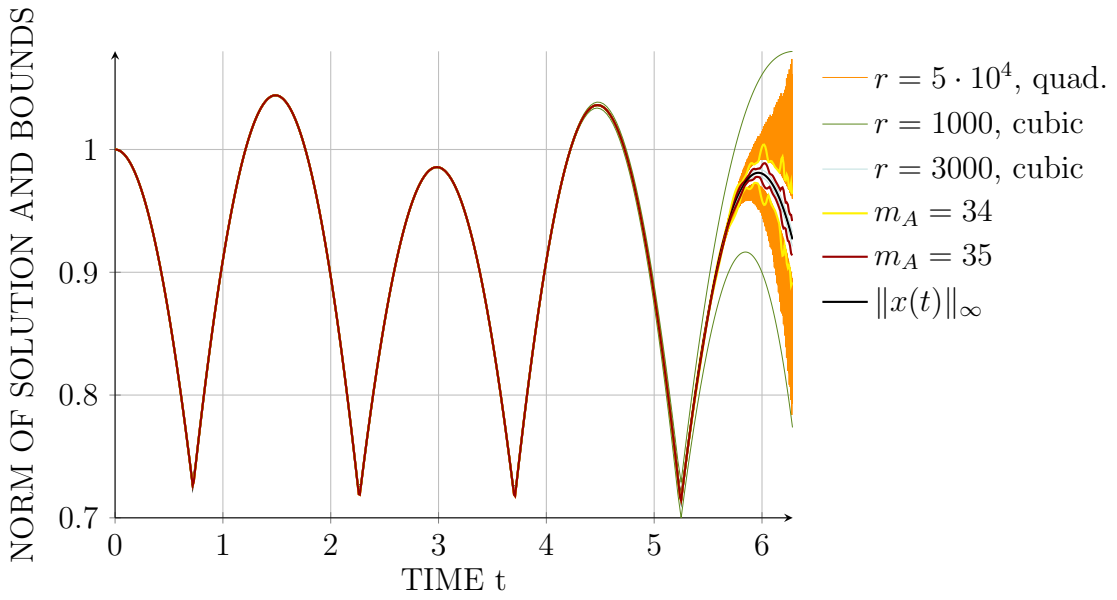


Figure 6.3: Jeffcott rotor on an anisotropic shaft for $t \in [0, 2\pi]$.

Figure 6.4 shows the solution of the Jeffcott rotor over time in the interval $[0, 10\pi]$ in various norms, the Euclidean norm, the maximum norm, the weighted time-invariant R-norm and the weighted time-dependent $\widetilde{R}(t)$-norm. The weighted time-dependent norm $\|\cdot\|_{\widetilde{R}(t)}$ suppresses the oscillations and since the matrix L is diagonalizable and the spectral abscissa is negative, $\nu[L] < 0$, a monotonic decrease can be observed, see Corollary 6.1.3.

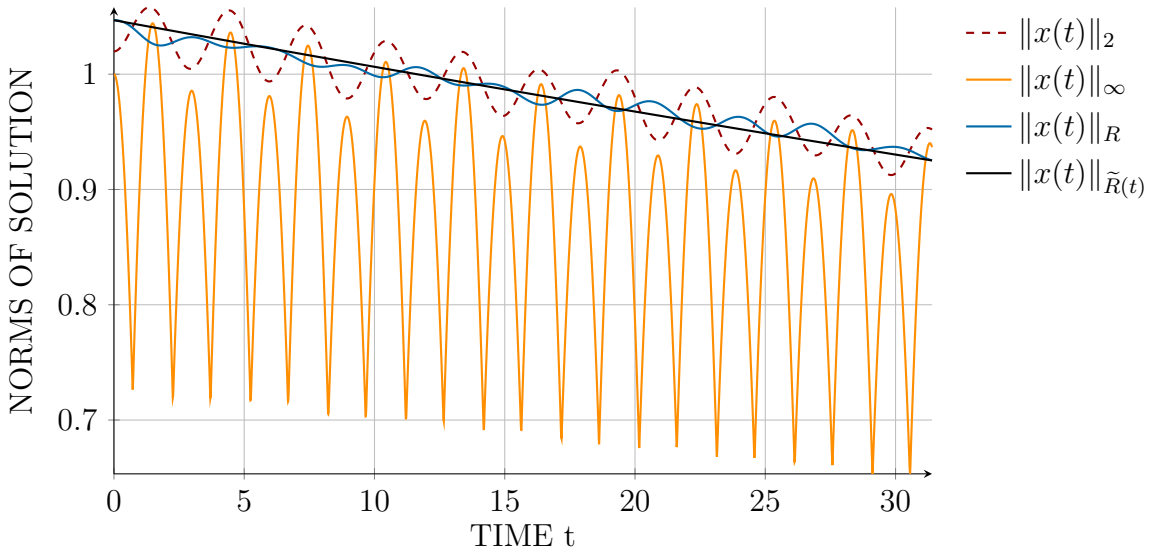


Figure 6.4: Jeffcott rotor on an anisotropic shaft for $t \in [0, 10\pi]$.

Finally, we discuss the convergence of trigonometric splines and of the spectral method depending on the smoothness of A indicated by Theorem 6.2.2, 6.2.3, 2.4.3, and 2.4.5.

Smoothness of A	**Trigonometric** quadratic	**Spline Bound** cubic	**Spectral Bound**
$\mathcal{C}^2$	$\mathcal{O}(\|L_3x\|_\infty r^{-2})$	—	$\mathcal{O}(Vm_A^{-1})$
$\mathcal{C}^3$	$\mathcal{O}(\|L_3x\|_\infty r^{-2})$	$\mathcal{O}(\|L_4x\|_\infty r^{-3})$	$\mathcal{O}(Vm_A^{-2})$
$\mathcal{AC}^{k-1}$, $1 \leq k < 3$	—	—	$\mathcal{O}(Vm_A^{-k})$
$\mathcal{AC}^2$	$\mathcal{O}(\|L_3x\|_\infty r^{-2})$	—	$\mathcal{O}(Vm_A^{-3})$
$\mathcal{AC}^{k-1}$, $k \geq 4$	$\mathcal{O}(\|L_3x\|_\infty r^{-2})$	$\mathcal{O}(\|L_4x\|_\infty r^{-3})$	$\mathcal{O}(Vm_A^{-k})$
analytic	$\mathcal{O}(\|L_3x\|_\infty r^{-2})$	$\mathcal{O}(\|L_4x\|_\infty r^{-3})$	$\mathcal{O}(M_A\rho_A^{-m_A})$

Table 6.3: Convergence for trigonometric spline and spectral bound

In Table 6.3 the convergence rates for the trigonometric spline bound defined in Theorem 6.2.2 and 6.2.3 and the spectral bounds defined in equations (6.45) and (6.47) are given for various function classes and they are visualized in Figures 6.5 and 6.6. The computational complexity for the trigonometric spline bound is dominated by computing the spline solution. Trigonometric splines with compact support, i.e. trigonometric B-splines, are chosen due to the local influence of each spline. For general splines, a linear system of dimension $n(r+1) \times n(r+1)$ has to be solved while for B-splines, $r+1$ systems of dimension $n \times n$ have to be solved. Hence, the computational complexity for trigonometric B-splines is $\mathcal{O}(n^3(r+1))$. For the spectral bound, each element of the system matrix A has to be approximated, which can be done by Fast Fourier Transformations (FFT) in $\mathcal{O}((m+1)\log(m+1))$. The convergence of the trigonometric spline bound is local, i.e., a trigonometric spline S_i converges on its support to the solution of the time-periodic system. The support of the trigonometric splines is given by $\text{supp}(S_i^2) = \{t \in [0, t_p] : S_i^2(t) \neq 0\} = [t_i, t_{i+3}]$ for quadratic splines visualized in Figure 2.2 and $\text{supp}(S_i^3) = \{t \in [0, t_p] : S_i^3(t) \neq 0\} = [t_i, t_{i+4}]$ for cubic splines visualized in Figure 2.3. The spectral bound converges globally, i.e., on the whole interval $[0, t_p]$, to the solution of the time-periodic system.

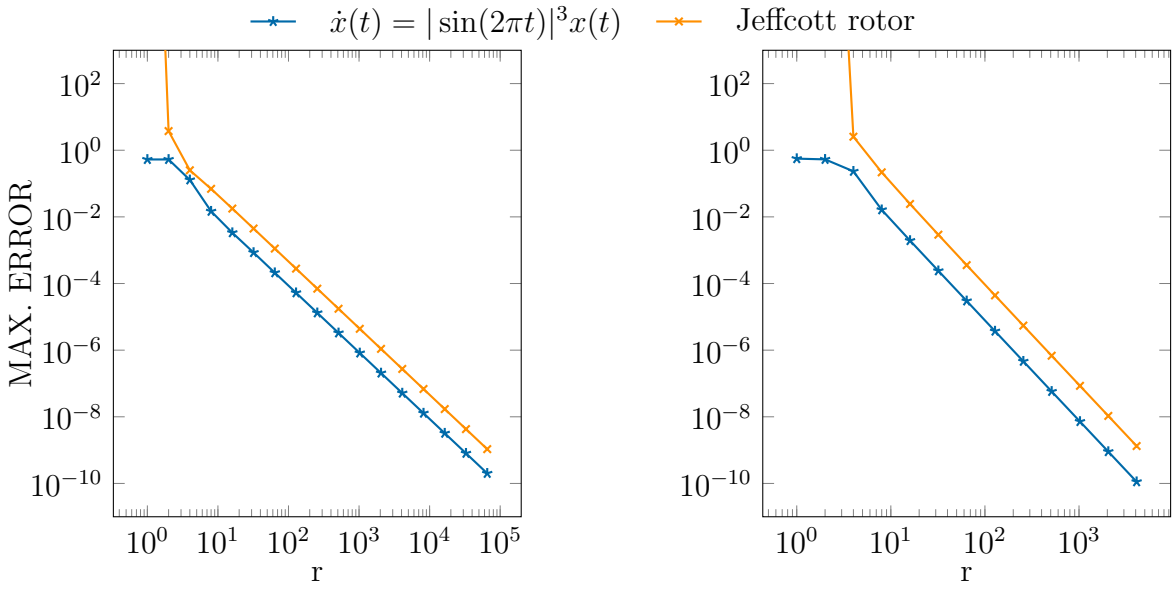


Figure 6.5: Convergence rates for quadratic (left) and cubic (right) trigonometric splines.

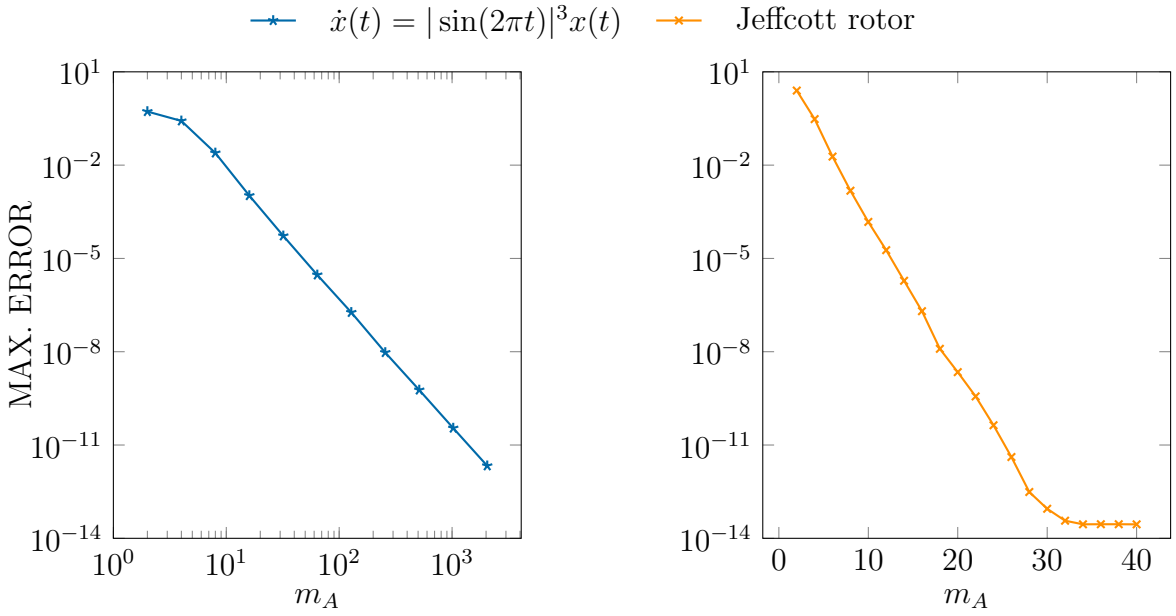


Figure 6.6: Convergence rates for Chebyshev projection method.

7

Summary and Outlook

In this thesis we have advanced the theory for damped linear systems as well as the time behavior of its solution. The contributions of this thesis are of both theoretical and numerical nature.

In Chapter 4 we have related the time behavior of a damped linear system to the systems energy and to the solution of a structured algebraic Lyapunov equation. Vibration reduction by optimizing passive damping was classified as a Nonlinear Program (NLP) with a nonlinear objective function, namely the trace of a solution to a algebraic Lyapunov equation. We therefore introduced the structure exploiting sign function method, which can efficiently solve the structured algebraic Lyapunov equation. We were able to show that the structure of the algebraic Lyapunov equation is kept throughout the sign function iteration. Moreover, the structure of the algebraic Lyapunov equation can efficiently been used to compute the gradient and Hessian of the objective function in order to improve the convergence of the NLP solver. In addition, the system's energy can be characterized by its eigenvalues and eigenvectors. We have derived a structure exploiting variant of the Ehrlich-Aberth iteration, which computes all eigenvalues simultaneous and iteratively. By an inverse iteration the corresponding eigenvectors can be determined.

In Chapter 5 the global optimization of passive damping w.r.t. external dampers positions was firstly considered and it could be encoded as a Mixed Integer Nonlinear Program. We have shown that the optimal positions can be computed in $\mathcal{O}(n^2)$ for sufficiently small viscosities. Linearization strategies based on McCormick envelopes and piecewise linear functions are given and a heuristic approach to find good damping positions are given.

Finally, in Chapter 6 we have analyzed vibrations and the time behavior of a solution to a time-periodic linear system. We could relate time-periodic linear systems to linear systems by the Floquet-Lyapunov transformation and therefore, we have obtained results on a certain norm of the solution, which guarantees two-sided rigorous bounds, decoupling, filtering and monotonicity. Moreover, its time behavior can be characterized by two-sided bounds for the Manhattan norm, the Euclidean norm and the maximum norm. Here, we have used two different ideas in order to derive two sided-bounds. While in the first method we have approximated the solution of the time-periodic linear system by trigonometric splines and then two-sided bounds have

been established on the quality of the approximation, the second method has approximated the time-periodic linear system, which then has turned out to be entire. Hence, its solution can be represented as an infinite series. Depending on the smoothness of the time-periodic system, we have formulated two-sided bounds which have incorporated the approximation error of the linear time-periodic system and the truncation error of the series representation. We have shown the order of convergence of the two-sided bounds to the solution of a linear time-periodic system depending on the smoothness of the linear time-periodic system.

Future possible research topics are manifold. For instance, in this thesis we have only considered linear systems. However, many of the concepts considered here, are also extendable to matrix pencils. There are infinitely many linearizations of a vibrational system in the sense of Definition 2.1.9, especially if the mass matrix is ill-conditioned or even singular, a matrix pencils (E, A) is obtained and the corresponding algebraic Lyapunov equation is then called generalized algebraic Lyapunov equation, which can be solved by the sign function method [BQO99]. We strongly believe that this holds for the generalized structured algebraic Lyapunov equation as well.

The global optimization of passive damping w.r.t. the external damper's positions in Chapter 5 has not been solved satisfactorily yet. A promising linearization idea of MINLP formulation is to discretize the viscosity space for each external damper, i.e., the viscosities are not continuous anymore but attain discrete values. It turns out that this problem can be reformulated as a Mixed Integer Linear Program.

The two-sided bounds on the solution of a time-periodic system in Chapter 6 can be derived for other norms as well since in a finite-dimensional vector space all norms are equivalent. A possible future research topic is the extension of two-sided bounds to nonlinear time-periodic systems. Some advances in this direction have been obtained for nonlinear systems [Koh07].

Bibliography

[Abe73] O. Aberth. Iteration methods for finding all zeros of a polynomial simultaneously. *Math. Comp.* 27 (1973), pp. 339–344.

[AMK96] S. C. Ahn, Y. S. Moon, and W. H. Kwon. Bounds in algebraic Riccati and Lyapunov equations: a survey and some new results. *Internat. J. Control* 64.3 (1996), pp. 377–389.

[All09] M. S. Allen. Frequency-Domain Identification of Linear Time-Periodic Systems Using LTI Techniques. *J. Comput. Nonlinear Dynam.* 4 (2009), pp. 041004:1–041004:6.

[Ant05] A. C. Antoulas. Approximation of Large-Scale Dynamical Systems. Philadelphia, PA: SIAM Publications, 2005.

[BD93] Z. Bai and J. Demmel. Design of a Parallel Nonsymmetric Eigenroutine Toolbox, Part I. *Proc. of the 6th SIAM Conf. on Parallel Processing for Scientific Computing* (1993).

[BS72] R. H. Bartels and G. W. Stewart. Solution of the Matrix Equation $AX + XB = C$: Algorithm 432. *Comm. ACM* 15 (1972), pp. 820–826.

[BB06] U. Baur and P. Benner. Factorized solution of Lyapunov equations based on hierarchical matrix arithmetic. *Computing* 78.3 (2006).

[BT70] E. M. L. Beale and J. A. Tomlin. Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *Proceedings of the Fifth International Conference on Operational Research*. Tavistock Publications, 1970, pp. 447–454.

[Bel43] R. Bellman. The stability of solutions of linear differential equations. *Duke Math. J.* 10.4 (Dec. 1943), pp. 643–647.

[Bel+13] P. Belotti et al. Mixed-integer nonlinear optimization. *Acta Numer.* 22 (2013), pp. 1–131.

[BD15] P. Benner and J. Denißen. Ehrlich-Aberth iteration for vibrational systems. *Proceedings of ICoEV 2015* (2015), pp. 1540–1548.

[BD16] P. Benner and J. Denißen. Numerical solution to low rank perturbed Lyapunov equations by the sign function method. *Proc. Appl. Math. Mech.* 16.1 (2016), pp. 723–724.

[BD14] P. Benner and J. Denißen. Spectral bounds on the solution of linear time-periodic systems. *Proc. Appl. Math. Mech.* 14.1 (2014), pp. 863–864.

[BDK13] P. Benner, J. Denißen, and L. Kohaupt. Bounds on the Solution of Linear Time-Periodic Systems. *Proc. Appl. Math. Mech.* 13.1 (2013), pp. 447–448.

[BDK17] P. Benner, J. Denißen, and L. Kohaupt. Trigonometric spline and spectral bounds for the solution of linear time-periodic systems. *J. Appl. Math. Comput.* 54.1 (2017), pp. 127–157.

[BQO99] P. Benner and E. S. Quintana-Ortí. Solving Stable Generalized Lyapunov Equations with the Matrix Sign Function. *Numer. Algorithms* 20.1 (1999), pp. 75–100.

[BTT11] P. Benner, Z. Tomljanović, and N. Truhar. Dimension reduction for damping optimization in linear vibrating systems. *Z. Angew. Math. Mech.* 91.3 (2011), pp. 179–191.

[BTT13] P. Benner, Z. Tomljanović, and N. Truhar. Optimal damping of selected eigenfrequencies using dimension reduction. *Numer. Lin. Alg. Appl.* 20.1 (2013), pp. 1–17.

[BT97] D. Bertsimas and J. N. Tsitsiklis. Introduction to linear optimization. Athena scientific series in optimization and neural computation. Athena Scientific, 1997.

[Bet+] T. Betcke et al. NLEVP: A Collection of Nonlinear Eigenvalue Problems. `http://www.mims.manchester.ac.uk/research/numerical-analysis/nlevp.html`.

[BM60] W. G. Bickley and J. McNamee. Matrix and other Direct Methods for the Solution of Systems of Linear Difference Equations. *Philos. Trans. Roy. Soc. A* 252.1005 (1960), pp. 69–131.

[BN13] D. A. Bini and V. Noferini. Solving polynomial eigenvalue problems by means of the Ehrlich-Aberth method. *Linear Algebra Appl.* 439.4 (2013), pp. 1130–1149.

[BD08] F. Bornemann and P. Deuflhard. Numerische Mathematik. II. Vol. 2. Walter de Gruyter GmbH & Co. KG, 2008.

[Bra98] K. Brabender. Optimale Dämpfung von linearen Schwingungssystemen. PhD thesis. Fernuniversität Hagen, 1998.

[BV10] M. R. Bussieck and S. Vigerske. MINLP Solver Software. *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Ltd, 2010.

[CRT06a] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* 52.2 (Feb. 2006), pp. 489–509.

[CRT06b] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* 59.8 (2006), pp. 1207–1223.

[Che54] P. L. Chebyshev. Théorie des mécanismes connus sous le nom de parallélogrammes. *Mémoires des Savants étrangers présentés à l'Académie de Saint-Pétersbourg* 7 (1854), pp. 539–568.

[CC97] A. Coddington and R. Carlson. Linear Ordinary Differential Equations. SIAM, 1997.

[CL55] E. A. Coddington and N. Levinson. Theory of Ordinary Differential Equations. English. McGraw-Hill, 1955.

[Col90] L. Collatz. Differentialgleichungen: eine Einführung unter besonderer Berücksichtigung der Anwendungen. seventh. Vieweg+Teubner Verlag, 1990.

[Coo71] S. A. Cook. The Complexity of Theorem-proving Procedures. *Proceedings of the Third Annual ACM Symposium on Theory of Computing*. STOC '71. New York, NY, USA: ACM, 1971, pp. 151–158.

[Cox98] S. J. Cox. Designing for Optimal Energy Absorption, 1: Lumped Parameter Systems. *J. Vib. Acoust.* 120 (2 1998), pp. 339–345.

[Dan60] G. B. Dantzig. On the Significance of Solving Linear Programming Problems with Some Integer Variables. *Econometrica* 28.1 (1960), pp. 30–44.

[DB01] C. De Boor. A practical guide to splines. Applied mathematical sciences. Berlin: Springer-Verlag, 2001.

[DH08] P. Deuflhard and A. Hohmann. Numerische Mathematik. I. Vol. 1. Walter de Gruyter GmbH & Co. KG, 2008.

[Duf18] G. Duffing. Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung (1918).

[Ehr67] L. W. Ehrlich. A Modified Newton Method for Polynomials. *Commun. ACM* 10.2 (Feb. 1967), pp. 107–108.

[Fab14] G. Faber. Über die interpolatorische Darstellung stetiger Funktionen. *Jahresber. Dtsch. Math.-Ver.* 23 (1914), pp. 192–210.

[FLVD04] H.-Y. Fan, W.-W. Lin, and P. Van Dooren. Normwise Scaling of Second Order Polynomial Matrices. *SIAM J. Matrix Anal. Appl.* 26.1 (2004), pp. 252–256.

[Flo83] G. Floquet. Sur les équations différentielles linéaires à coefficients périodiques. *Annales Scientifiques de l'École Normale Supérieure* 12.2 (1883), pp. 47–88.

[For11] O. Forster. Analysis 1. German. Vieweg+Teubner Verlag, 2011.

[For10] O. Forster. Analysis 2. German. Vieweg+Teubner Verlag, 2010.

[For12] O. Forster. Analysis 3. German. Vieweg+Teubner Verlag, 2012.

[Fou92] R. Fourer. A simplex algorithm for piecewise-linear programming III: Computational analysis and applications. *Math. Program.* 53.1 (1992), pp. 213–235.

[FGK89] R. Fourer, D. M. Gay, and B. W. Kernighan. AMPL: A Mathematical Programing Language. *Algorithms and Model Formulations in Mathematical Programming*. Berlin, Heidelberg: Springer-Verlag, 1989, pp. 150–151.

[FM92] R. Fourer and R. E. Marsten. Solving Piecewise-Linear Programs: Experiments with a Simplex Approach. *ORSA Journal on Computing* 4.1 (1992), pp. 16–31.

[Fun92] D. Funaro. Polynomial approximation of differential equations. Lecture Notes in Physics. Berlin, Heidelberg: Springer-Verlag, 1992.

[GLR09] I. Gohberg, P. Lancaster, and L. Rodman. Matrix Polynomials. Classics in Applied Mathematics. SIAM, 2009.

[GV96] G. H. Golub and C. F. Van Loan. Matrix Computations. third. Baltimore: Johns Hopkins University Press, 1996.

[GO77] D. Gottlieb and S. A. Orszag. Numerical Analysis of Spectral Methods: Theory and Applications. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1977.

[Gra04] L. Grasedyck. Existence of a low rank or H-matrix approximant to the solution of a Sylvester equation. *Numer. Lin. Alg. Appl.* 11 (2004), pp. 371–389.

[Gra01] L. Grasedyck. Theorie und Anwendungen hierarchischer Matrizen. PhD thesis. Christian-Albrechts-Universität zu Kiel, 2001.

[GH07] L. Grasedyck and W. Hackbusch. A Multigrid Method to Solve Large Scale Sylvester Equations. *SIAM J. Matrix Anal. Appl.* 29.3 (2007), pp. 870–894.

[Gro19] T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. English. *Ann. Math. (2)* 20 (1919), pp. 292–296.

[Ham82] S. J. Hammarling. Numerical Solution of the Stable, Non-negative Definite Lyapunov Equation. *IMA J. Numer. Anal.* 2 (1982), pp. 303–323.

[HH91] G. Hämmerlin and K.-H. Hoffmann. Numerische Mathematik. second. Grundwissen Mathematik. Berlin, Heidelberg: Springer-Verlag, 1991.

[Han92] E. Hansen. Global optimization using interval analysis. Monographs and textbooks in pure and applied mathematics. New York: M. Dekker, 1992.

[Hig+08] N. J. Higham et al. Scaling, sensitivity and stability in the numerical solution of quadratic eigenvalue problems. *Internat. J. Numer. Methods Engrg.* 73.3 (2008), pp. 344–360.

[HJ85] R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, 1985.

[HT96] R. Horst and H. Tuy. Global Optimization. Berlin, Heidelberg: Springer-Verlag, 1996.

[Ibm] IBM ILOG CPLEX Optimization Studio CPLEX User's Manual. English. Version Version 12. 2016.

[Jer73] R. G. Jeroslow. There Cannot be any Algorithm for Integer Programming with Quadratic Constraints. *Oper. Res.* 21.1 (1973), pp. 221–224.

[JL84] R. G. Jeroslow and J. K. Lowe. Modelling with integer variables. *Mathematical Programming at Oberwolfach II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1984, pp. 167–184.

[KM78] R. Kannan and C. L. Monma. On the Computational Complexity of Integer Programming Problems. *Optimization and Operations Research: Proceedings of a Workshop Held at the University of Bonn, October 2–8, 1977*. Berlin, Heidelberg: Springer-Verlag, 1978, pp. 161–172.

[Kar72] R. M. Karp. Reducibility among Combinatorial Problems. *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations*. Boston, MA: Springer-Verlag, 1972, pp. 85–103.

[Kar39] W. Karush. Minima of Functions of Several Variables with Inequalities as Side Constraints. MA thesis. University of Chicago, 1939.

[Kat95] T. Kato. Perturbation Theory for Linear Operators. Classics in Mathematics. Berlin, Heidelberg: Springer-Verlag, 1995.

[KFN06] A. B. Keha, I. R. de Farias, and G. L. Nemhauser. A Branch-and-Cut Algorithm Without Binary Variables for Nonconvex Piecewise Linear Optimization. *Oper. Res.* 54.5 (2006), pp. 847–858.

[KL92] C. Kenney and A. J. Laub. On Scaling Newton's Method for Polar Decomposition and the Matrix Sign Function. *SIAM J. Matrix Anal. Appl.* 13.3 (1992), pp. 688–706.

[KL95] C. Kenney and A. J. Laub. The Matrix Sign Function. *IEEE Trans. Automat. Control* 40.8 (1995), pp. 1330–1348.

[Koh02] L. Kohaupt. Differential calculus for p-norms of complex-valued vector functions with applications. *J. Comput. Appl. Math.* 145.2 (2002), pp. 425–457.

[Koh07] L. Kohaupt. New upper bounds for excited vibration systems with applications of the differential calculus of norms. *Int. J. Comput. Math.* 84.7 (2007), pp. 1035–1053.

[Koh13] L. Kohaupt. On the vibration-suppression property and monotonicity behavior of a special weighted norm for dynamical systems $\dot{x} = Ax$, $x(t_0) = x_0$. *Appl. Math. Comput.* 222.0 (2013), pp. 307–330.

[Koh08] L. Kohaupt. Solution of the matrix eigenvalue problem $VA^* + AV = \mu V$ with applications to the study of free linear dynamical systems. *J. Comput. Appl. Math.* 213.1 (2008), pp. 142–165.

[KT51] H. W. Kuhn and A. W. Tucker. Nonlinear Programming. *Proc. Second Berkeley Symp. on Math. Statist. and Prob.* (1951), pp. 481–492.

[Lag+98] J. C. Lagarias et al. Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions. *SIAM J. Optim.* 9.1 (1998), pp. 112–147.

[Lan02] P. Lancaster. Lambda-matrices and vibrating systems. Dover Publications, 2002.

[LF99] P. Lancaster and P. Freitas. On the Optimal Value of the Spectral Abscissa for a System of Linear Oscillators. *SIAM J. Matrix Anal. Appl.* 21.1 (1999), pp. 195–208.

[LT85] P. Lancaster and M. Tismenetsky. The Theory of Matrices. second. Orlando: Academic Press, 1985.

[LW02] J.-R. Li and J. White. Low Rank Solution of Lyapunov Equations. *SIAM J. Matrix Anal. Appl.* 24.1 (2002), pp. 260–280.

[LT67] F. R. Loscalzo and T. D. Talbot. Spline function approximations for solutions of ordinary differential equations. English. *Bull. Am. Math. Soc.* 73 (1967), pp. 438–442.

[LY15] D. G. Luenberger and Y. Ye. Linear and Nonlinear Programming. Springer-Verlag, 2015.

[LW79] T. Lyche and R. Winther. A stable recurrence relation for trigonometric B-splines. *J. Approx. Theory* 25.3 (1979), pp. 266 –279.

[MN99] J. R. Magnus and H. Neudecker. Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, 1999.

[Mah+11] A. Mahajan et al. MINOTAUR: A toolkit for solving mixed-integer nonlinear optimization. 2011. URL: `http://wiki.mcs.anl.gov/minotaur` (visited on 03/23/2017).

[MM57] H. M. Markowitz and A. S. Manne. On the Solution of Discrete Programming Problems. *Econometrica* 25.1 (1957), pp. 84–110.

[McC76] G. P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I — Convex underestimating problems. English. *Math. Program.* 10.1 (1976), pp. 147–175.

[Mey76] R. R. Meyer. Mixed integer minimization models for piecewise-linear functions of a single variable. *Discrete Math.* 16.2 (1976), pp. 163–171.

[MS76] P. C. Müller and W. Schiehlen. Lineare Schwingungen. Akademische Verlagsgesellschaft Wiesbaden, 1976.

[Nak13] I. Nakić. Integration of positive linear functionals on a sphere in $\mathbb{R}^{2n}$ with respect to Gaussian surface measures. *Math. Commun.* 18.2 (2013), pp. 349–158.

[Nak02] I. Nakić. Optimal damping of vibrational systems. PhD thesis. Fernuniversität Hagen, 2002.

[Nik04] A. Nikolis. Numerical solutions of ordinary differential equations with quadratic trigonometric splines. eng. *Appl. Math. E-Notes* 4 (2004), pp. 142–149.

[Nik93] A. Nikolis. Trigonometrische Splines und ihre Anwendung zur numerischen Behandlung von Integralgleichungen. PhD thesis. Ludwig-Maximilians-Universität München, 1993.

[NS05] A. Nikolis and I. Seimenis. Solving dynamical systems with cubic trigonometric splines. eng. *Appl. Math. E-Notes* 5 (2005), pp. 116–123.

[Pen00a] T. Penzl. A cyclic low rank Smith method for large sparse Lyapunov equations. *SIAM J. Sci. Comput.* 21.4 (2000), pp. 1401–1418.

[Pen00b] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Systems Control Lett.* 40 (2 2000), pp. 139–144.

[Pol20] B. Van der Pol. A theory of the amplitude of free and forced triode vibration. *Radio Review* 1 (1920), pp. 701–720.

[PS08] I. P. Popchev and S. G. Savov. New Upper Bounds for the CALE. *International Journal of Control, Automation and Systems* 6.2 (2008), pp. 288–294.

[Rao07] S. S. Rao. Vibration of Continuous Systems. John Wiley & Sons, Ltd, 2007.

[Rob80] J. D. Roberts. Linear Model Reduction and Solution of the Algebraic Riccati Equation by Use of the Sign Function. *Internat. J. Control* 32 (1980). (Reprint of Technical Report No. TR-13, CUED/B-Control, Cambridge University, Engineering Department, 1971), pp. 677–687.

[Roy88] H. L. Royden. Real analysis. third. New York: Macmillan, 1988.

[Sch64] I. Schoenberg. On Trigonometric Spline Interpolation. *Indiana Univ. Math. J.* 13 (5 1964), pp. 795–825.

[Sch71] A. Schönhage. Approximationstheorie. De Gruyter Lehrbuch. Walter de Gruyter GmbH & Co. KG, 1971.

[Sch86] A. Schrijver. Theory of Linear and Integer Programming. New York, NY, USA: John Wiley & Sons, Ltd, 1986.

[Sch81] L. L. Schumaker. Spline Functions: Basic Theory. John Wiley & Sons, Ltd, 1981.

[SB96] S. C. Sinha and E. A. Butcher. Solution and stability of a set of p-th order linear differential equations with periodic coefficients via Chebyshev polynomials. *Math. Probl. Eng.* 2 (2 1996), pp. 165–190.

[SW91] S. C. Sinha and D.-H. Wu. An efficient computational scheme for the analysis of periodic systems. *J. Sound Vib.* 151 (1991), pp. 91–117.

[SM85] G. Söderlind and R. M. M. Mattheij. Stability and Asymptotic Estimates in Nonautonomous Linear Differential Systems. *SIAM J. Math Anal.* 16.1 (1985), pp. 69–92.

[Tas15] L. Taslaman. An Algorithm for Quadratic Eigenproblems with Low Rank Damping. *SIAM J. Matrix Anal. Appl.* 36.1 (2015), pp. 251–272.

[TS05] M. Tawarmalani and N. V. Sahinidis. A polyhedral branch-and-cut approach to global optimization. *Math. Program.* 103 (2 2005), pp. 225–249.

[TM01] F. Tisseur and K. Meerbergen. The Quadratic Eigenvalue Problem. *SIAM Rev.* 43.2 (2001), pp. 235–286.

[Tre13] L. N. Trefethen. Approximation Theory and Approximation Practice. SIAM, 2013, pp. viii + 305.

[Tre08] L. N. Trefethen. Is Gauss Quadrature Better than Clenshaw-Curtis? *SIAM Rev.* 50.1 (Feb. 2008), pp. 67–87.

[Tre00] L. N. Trefethen. Spectral Methods in MatLab. Philadelphia, PA, USA: SIAM, 2000.

[Tru04] N. Truhar. An efficient algorithm for damper optimization for linear vibrating systems using Lyapunov equation. *J. Comput. Appl. Math.* 172.1 (2004), pp. 169–182.

[TV09] N. Truhar and K. Veselić. An Efficient Method for Estimating the Optimal Dampers' Viscosity for Linear Vibrating Systems Using Lyapunov Equation. *SIAM J. Matrix Anal. Appl.* 31.1 (2009), pp. 18–39.

[Ves03] K. Veselić. Bounds for exponentially stable semigroups. *Linear Algebra Appl.* 358.1-3 (2003), pp. 309–333.

[Ves98] K. Veselić. Estimating the operator exponential. *Linear Algebra Appl.* 280.2-3 (1998), pp. 241–244.

[Ves97] K. Veselić. Exponential Decay of Semigroups in Hilbert Space. English. *Semigroup Forum* 55.3 (1997), pp. 325–331.

[Ves90] K. Veselić. On linear vibrational systems with one dimensional damping II. English. *Integral Equations Operator Theory* 13.6 (1990), pp. 883–897.

[Wal70] W. Walter. Differential- und Integral-Ungleichungen. second. Springer Tracts in Natural Philosophy. Berlin, Heidelberg: Springer-Verlag, 1970.

[Wal90] W. Walter. Gewöhnliche Differentialgleichungen. Berlin, Heidelberg: Springer-Verlag, 1990.

Declaration of Honor

I hereby declare that I produced this thesis without prohibited assistance and that all sources of information that were used in producing this thesis, including my own publications, have been clearly marked and referenced.

In particular I have not willfully:

- Fabricated data or ignored or removed undesired results.
- Misused statistical methods with the aim of drawing other conclusions than those warranted by the available data.
- Plagiarized data or publications or presented them in a distorted way.

I know that violations of copyright may lead to injunction and damage claims from the author or prosecution by the law enforcement authorities.

This work has not previously been submitted as a doctoral thesis in the same or a similar form in Germany or in any other country. It hast not previously been published as a whole.

Magdeburg, December 18th, 2018

Jonas Denißen

Ehrenerklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; verwendete fremde und eigene Quellen sind als solche kenntlich gemacht.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert oder verzerrt wiedergegeben.

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadenersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen kann.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 18.12.2018

_____

Jonas Denißen