Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

RESEARCH ARTICLE

Well-scaled, A-posteriori Error Estimation for Model Order Reduction of Large Second-order Mechanical Systems

Dennis Grunert[1] | Jörg Fehr*[1] | Bernard Haasdonk[2]

[1] Institute of Engineering and Computational Mechanics, University of Stuttgart, Germany
[2] Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, Germany

Correspondence
*Jörg Fehr, Institute of Engineering and Computational Mechanics, University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany. Email: joerg.fehr@itm.uni-stuttgart.de

Summary

Model Order Reduction is used to vastly speed up simulations but it also introduces an error to the simulation results, which needs to be controlled. The performance of the general to use, a-posteriori error estimator of Ruiner et al. for second-order systems is analyzed and a bottleneck is found in the offline stage making it unusable for larger models. We use the spectral theorem, power series expansions, monotonicity properties, and self-tailored algorithms to speed up the offline stage largely by one polynomial order both in terms of computation time as well as storage complexity. All properties are proven rigorously. This eliminates the aforementioned bottleneck. Hence, the error estimator of Ruiner et al. can finally be used for large, linear, second-order mechanical systems reduced by any model reduction method based on Petrov-Galerkin reduction. The examples show speedups of up to 28.000 and the ability to compute much larger systems with a fixed amount of memory.

KEYWORDS:
a-posteriori error estimation, model order reduction, mechanical system, spectral theorem

1 | INTRODUCTION

Mechanical systems modeled with the Finite Element Method become more and more detailed with increasing demands for precision. The finite element model of car nowadays has usually over 10 million degrees of freedom. Simulating such large systems in time, e.g., for car crashes, becomes a challenge even for large automobile manufacturers on high-performance computers when small modifications need to be checked quickly regarding crash safety or when optimizations need to be performed. This results in many simulations of large models to be performed in a small amount of time, usually over night.

Model Order Reduction (MOR) is a tool that is commonly used to reduce the computation time and storage complexity of the simulations significantly. Certain computer-aided engineering (CAE) task would not be possible without it. In model order reduction, the time-discretized ordinary differential equation is usually projected onto a space of much smaller dimension through Petrov-Galerkin reduction while retaining as much system properties as possible. We refer to, e.g., Schilders et al.[1] and Benner et al.[2] who give overviews of model order reduction. We will focus on linear mechanical systems given in second-order form. Model reduction techniques for this type of system can be found in Fehr[3].

Since information is typically lost when projecting to a lower-dimensional space, an error is introduced to the simulation of the reduced model compared to the full-order model. Error estimators are developed to quantify the size of the error introduced into the system, usually as an upper bound. Without any estimation of the error, model reduction would be of no use since it may produce arbitrarily bad results. This is especially important for safety critical systems like car crash simulations. Error estimators can also be used in an adaptive way to tune parameters of the used model reduction method.[4] A-priori error bounds

are able to estimate the error before the solution of the reduced system is available. The estimation cannot be better than the one of a-posteriori error estimators, which take also the individual result of the reduced simulation into account. For applications focussed on the frequency response of a system, error estimators in the frequency domain quantifying the error between the transfer functions of the reduced and the full-order system are also available.[5] This work will provide a vast speedup of an existing error estimator applicable independent of reduction techniques allowing it to scale for large models.

We focus on the a-posteriori error estimator in the time-domain described in Ruiner et al.[6], which is based on the work of Haasdonk and Ohlberger[7]. It defines the error as a weighted norm of the difference between reconstructed and full state. The error estimator is an upper bound on this error for each time instance after applying an arbitrary time integration scheme. Its main advantage is the general usability. While most error estimators are tailored for a specific model reduction method, the estimator by Ruiner et al. can be used for any Petrov-Galerkin projection independently on how the projection matrices were tailored. We will not discuss the quality of the error estimation as this was already done in Ruiner et al.[6] and Fehr et al.[8] Instead, we solely focus on the computational performance and scalability for large systems. For completeness, one example comparing the true error and error bound will be given in Section 6.

The error estimator of Ruiner et al.[6] involves taking the inverse of the mass matrix of the full-order system and computing the matrix exponential of a large matrix in each time step in order to compute specific constants in the offline stage. These computations are the bottleneck with respect to storage and computation time, respectively. We show that a straightforward computation will scale like $\mathcal{O}(N^3 n_T)$ in time and $\mathcal{O}(N^2)$ in storage with N being the system dimension and n_T the number of time steps in the discretized setting. This is infeasible for large-scale systems. The limits of a standard desktop computer are already reached for N around 4 000. It may be argued that arbitrary computation time is allowed in the offline stage but this is infeasible for real-world applications. Additionally, the storage bottleneck cannot be overcome with infinite time resources. Therefore, the bad scalability in the offline step is a major problem once the error estimator is applied for real-world, large-scale applications.

The novelty of this work is to speed up the offline step of the error estimator of Ruiner et al.[6] We improve the computational time to scale like $\mathcal{O}(N^{1.8}) + \mathcal{O}(n_T)$ with a storage complexity of $\mathcal{O}(N)$ in the case of a proportionally damped, second-order, linear mechanical system and a common choice for the error norm. The results of the undamped, less interesting case are similar. This way, the error estimator can be used on much larger systems. The examples show a speedup of up to 28 000 with the ability to solve 35 times larger systems before running out of memory on a standard desktop computer. For other examples and more powerful hardware, these numbers increase due to the better scalability of one order in time and storage complexity. This vast performance improvement is not only of practical nature but the theory developed may also serve as a template for performance improvements of other problems involving the calculation of a weighted spectral norm and the matrix exponential.

This work is structured as follows: In Section 2, the system of interest as well as the error estimator of Ruiner et al.[6] are described. The section ends with a short analysis of the performance to identify the bottlenecks, which are tackled in the next section: We prove a special case of the spectral theorem and apply it to the power series of the terms used to compute two constants needed in the online stage. These constants can now be described as the maximum of one-dimensional functions over the spectrum of a suitable system matrix. While this is still a considerable improvement since only one-dimensional operations are involved, the spectrum still needs to be computed fully. Hence, Section 4 focuses on analyzing these one-dimensional functions for monotonicity properties and finding bounds in certain areas. With these properties, algorithms are tailored which only need a handful of eigenvalues of the aforementioned system matrix. The examples presented in Section 6 show that the theoretical analysis of the performance of these algorithms in Section 5 is indeed in correspondence with numerical experiments. The paper ends with a conclusion and outlook in Section 7.

2 | A-POSTERIOR ERROR ESTIMATION

We first introduce the a-posteriori error estimator by Ruiner et al.[6] used in this work. After an analysis of the scalability, we argue that the error estimator in its original form is not scalable for large systems and identify the reason for that.

2.1 | System of Interest

The system of interest is any linear, second-order system which can be written as a multiple-input and multiple-output system (MIMO) of the form

$$\begin{aligned} \boldsymbol{M}\ddot{\boldsymbol{q}}(t) + \boldsymbol{D}\dot{\boldsymbol{q}}(t) + \boldsymbol{K}\boldsymbol{q}(t) &= \boldsymbol{B}\boldsymbol{u}(t)\,, \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{q}(t)\,, \\ \boldsymbol{q}(0) = \boldsymbol{q}_0,\ \dot{\boldsymbol{q}}(0) &= \dot{\boldsymbol{q}}_0\,, \end{aligned} \tag{1}$$

with system dimension N, where $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ is a symmetric, positive definite (SPD) mass matrix, $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ a symmetric, positive semi-definite (SPSD) damping matrix, $\boldsymbol{K} \in \mathbb{R}^{N\times N}$ an SPSD stiffness matrix, $\boldsymbol{q}(t) \in \mathbb{R}^N$ the flexible coordinates describing the elastic deformation of the mechanical system in time $t \in T_{\text{cont}} := [0, t_{\text{end}}]$ with initial values $\boldsymbol{q}_0, \dot{\boldsymbol{q}}_0 \in \mathbb{R}^N$, and $\boldsymbol{u}(t) \in \mathbb{R}^{n_\text{i}}$ a time-dependent input with input matrix $\boldsymbol{B} \in \mathbb{R}^{N\times n_\text{i}}$. The output $\boldsymbol{y}(t) \in \mathbb{R}^{n_\text{o}}$ is a linear combination of the state $\boldsymbol{q}(t)$ described by the output matrix $\boldsymbol{C} \in \mathbb{R}^{n_\text{o}\times N}$. We will omit the time dependence in the following if it is clear from the context, i.e., we may write $\boldsymbol{q}$ instead of $\boldsymbol{q}(t)$. The formulation (1) usually arises form discretizing a continuum mechanical system in space with the finite element method (FEM). Many finite element programs in engineering like ANSYS Mechanical, Abaqus FEA, LS-DYNA, etc. allow to extract the system matrices $\boldsymbol{M}$, $\boldsymbol{D}$, and $\boldsymbol{K}$.

The second-order, ordinary differential equation (ODE) given in (1) can be transformed to the state-space representation given by the first-order ODE

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \tilde{\boldsymbol{B}}\boldsymbol{u}(t)\,, \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0\,, \\ \boldsymbol{y}(t) &= \tilde{\boldsymbol{C}}\boldsymbol{x}(t) \end{aligned} \tag{2}$$

with

$$\begin{aligned} \boldsymbol{x} &= \begin{pmatrix} \boldsymbol{q} \\ \dot{\boldsymbol{q}} \end{pmatrix},\ \boldsymbol{x}_0 = \begin{pmatrix} \boldsymbol{q}_0 \\ \dot{\boldsymbol{q}}_0 \end{pmatrix} \in \mathbb{R}^{2N}\,, \\ \boldsymbol{A} &= \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{M}^{-1}\boldsymbol{K} & -\boldsymbol{M}^{-1}\boldsymbol{D} \end{pmatrix} \in \mathbb{R}^{2N\times 2N}\,, \\ \tilde{\boldsymbol{B}} &= \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{M}^{-1}\boldsymbol{B} \end{pmatrix} \in \mathbb{R}^{2N\times n_\text{i}}, \tilde{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{n_\text{o}\times 2N}\,. \end{aligned} \tag{3}$$

2.2 | Linear Model Order Reduction

Model Order Reduction (MOR) transforms the systems (1) and (2) of dimensions $2N$ and N to systems of the same form with much smaller dimensions $2n \ll 2N$ and $n \ll N$, respectively. The goal consists in having the same system behavior for the reduced-order system as for the full-order system. Despite some special cases, the reduced-order system always constitutes an approximation rather than an exact replacement of the full-order system.

Most model order reduction methods use the Petrov-Galerkin ansatz described in the following for the second-order system (1). Assuming a so-called reduction matrix $\boldsymbol{V} \in \mathbb{R}^{N\times n}$, the state $\boldsymbol{q}$ is approximated by $\tilde{\boldsymbol{q}} := \boldsymbol{V}\overline{\boldsymbol{q}}$, where $\overline{\boldsymbol{q}} \in \mathbb{R}^n$ with $n \ll N$ is the so-called reduced state and $\tilde{\boldsymbol{q}} \in \mathbb{R}^N$ the reconstructed state. It is clear that $\tilde{\boldsymbol{q}} = \boldsymbol{q}$ is only possible for $\boldsymbol{q}$ in the column span of $\boldsymbol{V}$. Simply substituting $\boldsymbol{q}$ with $\tilde{\boldsymbol{q}} = \boldsymbol{V}\overline{\boldsymbol{q}}$ in (1) therefore introduces a residual

$$\boldsymbol{r}(t) := \boldsymbol{M}\boldsymbol{V}\ddot{\overline{\boldsymbol{q}}}(t) + \boldsymbol{D}\boldsymbol{V}\dot{\overline{\boldsymbol{q}}}(t) + \boldsymbol{K}\boldsymbol{V}\overline{\boldsymbol{q}}(t) - \boldsymbol{B}\boldsymbol{u}(t)$$

between the left and right hand side. Therefore, a second reduction matrix $\boldsymbol{W}^\mathsf{T} \in \mathbb{R}^{n\times N}$ is multiplied to the left of the equation $\boldsymbol{r}(t) = \boldsymbol{0}$ yielding the reduced system

$$\begin{aligned} \boldsymbol{W}^\mathsf{T}\boldsymbol{M}\boldsymbol{V}\ddot{\overline{\boldsymbol{q}}}(t) + \boldsymbol{W}^\mathsf{T}\boldsymbol{D}\boldsymbol{V}\dot{\overline{\boldsymbol{q}}}(t) + \boldsymbol{W}^\mathsf{T}\boldsymbol{K}\boldsymbol{V}\overline{\boldsymbol{q}}(t) &= \boldsymbol{W}^\mathsf{T}\boldsymbol{B}\boldsymbol{u}(t)\,, \\ \tilde{\boldsymbol{y}}(t) &= \boldsymbol{C}\boldsymbol{V}\overline{\boldsymbol{q}}(t)\,. \end{aligned} \tag{4}$$

The reduced matrices $\overline{\boldsymbol{M}} := \boldsymbol{W}^\mathsf{T}\boldsymbol{M}\boldsymbol{V}$, $\overline{\boldsymbol{D}} := \boldsymbol{W}^\mathsf{T}\boldsymbol{D}\boldsymbol{V}$, $\overline{\boldsymbol{K}} := \boldsymbol{W}^\mathsf{T}\boldsymbol{K}\boldsymbol{V}$, $\overline{\boldsymbol{B}} := \boldsymbol{W}^\mathsf{T}\boldsymbol{B}$, and $\overline{\boldsymbol{C}} := \boldsymbol{C}\boldsymbol{V}$ can be pre-computed. The system (4) is now solved for $\overline{\boldsymbol{q}}$, which yields the approximated state $\tilde{\boldsymbol{q}} = \boldsymbol{V}\overline{\boldsymbol{q}} \approx \boldsymbol{q}$ and approximated output $\tilde{\boldsymbol{y}}(t) \approx \boldsymbol{y}(t)$. It only depends on the system dimension n instead of N, which is expected to yield a faster integration in time.

An equivalent reduction for the first-order system (2) can be defined[6] with the biorthogonal ($\boldsymbol{W}_\mathrm{s}^\mathsf{T}\boldsymbol{V}_\mathrm{s} = \boldsymbol{I}_{2n\times 2n}$) reduction matrices

$$\boldsymbol{W}_\mathrm{s}^\mathsf{T} = \begin{pmatrix} \left(\boldsymbol{W}^\mathsf{T}\boldsymbol{V}\right)^{-1}\boldsymbol{W}^\mathsf{T} & \boldsymbol{0} \\ \boldsymbol{0} & \left(\boldsymbol{W}^\mathsf{T}\boldsymbol{M}\boldsymbol{V}\right)^{-1}\boldsymbol{W}^\mathsf{T}\boldsymbol{M} \end{pmatrix},$$
$$\boldsymbol{V}_\mathrm{s} = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V} \end{pmatrix}$$

yielding the reduced, first-order system

$$\begin{aligned} \dot{\overline{\boldsymbol{x}}}(t) &= \boldsymbol{W}_\mathrm{s}^\mathsf{T}\boldsymbol{A}\boldsymbol{V}_\mathrm{s}\overline{\boldsymbol{x}}(t) + \boldsymbol{W}_\mathrm{s}^\mathsf{T}\tilde{\boldsymbol{B}}\boldsymbol{u}(t), \\ \tilde{\boldsymbol{y}}(t) &= \tilde{\boldsymbol{C}}\boldsymbol{V}_\mathrm{s}\overline{\boldsymbol{x}}(t). \end{aligned} \tag{5}$$

The approximated outputs $\tilde{\boldsymbol{y}}(t)$ of (4) and (5) are the same since $\overline{\boldsymbol{x}}(t) = \begin{pmatrix} \overline{\boldsymbol{q}}(t) \\ \dot{\overline{\boldsymbol{q}}}(t) \end{pmatrix}$ holds for the states.

The choice of the reduction matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ are an important characteristic of each reduction method. Since the error estimator discussed later can be used with any reduction matrix, we only mention a few common choices for linear, second-order mechanical systems:

- modal reduction, which truncates eigenmodes above a certain frequency;[9]
- model reduction based on Krylov subspaces, which are used to match certain moments in the frequency response of the model;[10]
- balanced truncation, which removes states that are badly observable and controllable;[11]
- the Craig-Bampton method[12] and its extension CMS-Gram[13], which both combine different types of component modes;
- Proper Orthogonal Decomposition (POD), which takes the most important principal components of simulation snapshots as basis.[14]

The error of the state and output is now defined as the difference between original and approximated state, i.e.,

$$\begin{aligned} \boldsymbol{e}(t) &:= \boldsymbol{q}(t) - \tilde{\boldsymbol{q}}(t), \\ \boldsymbol{e}_\mathrm{out}(t) &:= \boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t). \end{aligned} \tag{6}$$

It is more feasible to analyze the error as a scalar. Therefore, a suitable norm has to be chosen.

Lemma 1. Let $\boldsymbol{S}$ be an SPD matrix. Then $(\boldsymbol{S}^{\frac{1}{2}})^{-1}$ exists as SPD matrix and it holds $(\boldsymbol{S}^{\frac{1}{2}})^{-1} = (\boldsymbol{S}^{-1})^{\frac{1}{2}}$. Therefore, we can simply write $\boldsymbol{S}^{-\frac{1}{2}}$ for this matrix.

Proof. Since $\boldsymbol{S}$ is symmetric and positive definite, there exists exactly one symmetric and positive definite square root, which we will call $\boldsymbol{S}^{\frac{1}{2}}$.[15] The inverse $(\boldsymbol{S}^{\frac{1}{2}})^{-1}$ exists and is symmetric since $\boldsymbol{S}^{\frac{1}{2}}$ is positive definite and symmetric. With similar arguments, the matrix $(\boldsymbol{S}^{-1})^{\frac{1}{2}}$ exists uniquely. We will now show that $(\boldsymbol{S}^{\frac{1}{2}})^{-1} = (\boldsymbol{S}^{-1})^{\frac{1}{2}}$, i.e., the matrix $(\boldsymbol{S}^{\frac{1}{2}})^{-1}$ should be the square root of the inverse of $\boldsymbol{S}$. In other words, squaring $(\boldsymbol{S}^{\frac{1}{2}})^{-1}$ should act like the inverse of $\boldsymbol{S}$:

$$\left((\boldsymbol{S}^{\frac{1}{2}})^{-1}\right)^2 \boldsymbol{S} = (\boldsymbol{S}^{\frac{1}{2}})^{-1}(\boldsymbol{S}^{\frac{1}{2}})^{-1}\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{S}^{\frac{1}{2}} = \boldsymbol{I}$$
$$\boldsymbol{S}\left((\boldsymbol{S}^{\frac{1}{2}})^{-1}\right)^2 = \boldsymbol{S}^{\frac{1}{2}}\boldsymbol{S}^{\frac{1}{2}}(\boldsymbol{S}^{\frac{1}{2}})^{-1}(\boldsymbol{S}^{\frac{1}{2}})^{-1} = \boldsymbol{I}$$

□

Definition 1 (Weighted Norm)**.** For a symmetric, positive definite weighting matrix $\boldsymbol{G} \in \mathbb{R}^{N\times N}$, the $\boldsymbol{G}$-weighted norm of a vector $\boldsymbol{z} \in \mathbb{C}^N$ and the induced, $\boldsymbol{G}$-weighted norm of a matrix $\boldsymbol{Z} \in \mathbb{C}^{N\times N}$ are defined as

$$\begin{aligned} \|\boldsymbol{z}\|_{\boldsymbol{G}} &:= \sqrt{\boldsymbol{z}^\mathsf{H}\boldsymbol{G}\boldsymbol{z}} = \left\|\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{z}\right\|_2, \\ \|\boldsymbol{Z}\|_{\boldsymbol{G}} &:= \sup_{\|\boldsymbol{z}\|_{\boldsymbol{G}}=1} \|\boldsymbol{Z}\boldsymbol{z}\|_{\boldsymbol{G}} = \left\|\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Z}\boldsymbol{G}^{-\frac{1}{2}}\right\|_2, \end{aligned} \tag{7}$$

respectively. The norm $\|\cdot\|_2$ is the Euclidean norm for vectors and the induced spectral norm for matrices.

Proof. The last equality in (7) is worth a proof:

$$\sup_{\|z\|_{\boldsymbol{G}}=1} \|\boldsymbol{Z}z\|_{\boldsymbol{G}} = \sup_{z\in\mathbb{C}^N} \frac{\|\boldsymbol{Z}z\|_{\boldsymbol{G}}}{\|z\|_{\boldsymbol{G}}} = \sup_{z\in\mathbb{C}^N} \frac{\left\|\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Z}z\right\|_2}{\left\|\boldsymbol{G}^{\frac{1}{2}}z\right\|_2} = \sup_{w=\boldsymbol{G}^{\frac{1}{2}}z\in\mathbb{C}^N} \frac{\left\|\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Z}\boldsymbol{G}^{-\frac{1}{2}}w\right\|_2}{\|w\|_2} = \left\|\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Z}\boldsymbol{G}^{-\frac{1}{2}}\right\|_2$$

The substitution $\boldsymbol{w} = \boldsymbol{G}^{\frac{1}{2}}\boldsymbol{z}$ is bijective since $\boldsymbol{G}$ – and therefore $\boldsymbol{G}^{\frac{1}{2}}$ – is regular, see Lemma 1. □

A common choice for the weighting matrix $\boldsymbol{G}$ in Definition 1 is the mass matrix $\boldsymbol{M}$ for finite element models to compensate for different units of the displacement and rotation in the state $\boldsymbol{q}$.[6] We will make use of this choice most of the time in this work but keep the results as general as possible. It is interesting to note that some results only hold for $\boldsymbol{G} = \boldsymbol{M}$, the most common choice. It will be clear from the context whether we mean the scalar $\|\boldsymbol{e}(t)\|_{\boldsymbol{G}}$ or the vector $\boldsymbol{e}(t)$ when we are talking about the error.

In order to prevent the inversion of $\boldsymbol{G}^{\frac{1}{2}}$ during the computation of the weighted matrix norm $\|\boldsymbol{Z}\|_{\boldsymbol{G}}$ from Definition 1, we instead solve a generalized eigenvalue problem.[8]

Proposition 1. With the assumptions from Definition 1, the weighted norm $\|\boldsymbol{Z}\|_{\boldsymbol{G}}$ is the square root of the largest eigenvalue of the generalized eigenvalue problem

$$\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z}z = \lambda\boldsymbol{G}z\,.$$

Proof. As already proven for Definition 1, $\|\boldsymbol{Z}\|_{\boldsymbol{G}} = \|\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Z}\boldsymbol{G}^{-\frac{1}{2}}\|_2$. The spectral norm of a matrix is its largest singular value, which is the square root of the largest eigenvalue of the matrix transposed and multiplied by itself:

$$\|\boldsymbol{Z}\|_{\boldsymbol{G}} = \sigma_{\max}\left(\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{Z}\boldsymbol{G}^{-\frac{1}{2}}\right) = \sqrt{\lambda_{\max}\left(\boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z}\boldsymbol{G}^{-\frac{1}{2}}\right)} = \sqrt{\lambda_{\max}(\boldsymbol{G}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z})}$$

In the last step, we used that the eigenvalue of similar matrices – in this case transformed with $\boldsymbol{G}^{\frac{1}{2}}$ – are identical. Symmetry of $\boldsymbol{G}$ and $\boldsymbol{G}^{\frac{1}{2}}$ was also used. By definition, $\lambda_{\max}(\boldsymbol{G}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z})$ is the largest eigenvalue of the problem

$$\begin{aligned}\boldsymbol{G}^{-1}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z}z &= \lambda z\\ \Leftrightarrow\ \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{Z}z &= \lambda\boldsymbol{G}z\,,\end{aligned}$$

which proves the proposition. □

Still, the error cannot be computed directly with (6) since the computation of $\boldsymbol{q}(t)$ for every t is equivalent to solving the full-order system; but obviously, there is no need for a reduced-order system or error once the full-order solution is known. Hence, we will learn how to estimate the error without computing the full-order system in the next section.

2.3 | A-posteriori Error Estimator

We now present the error estimator from Ruiner et al.[6], which gives upper bounds to the errors $\|\boldsymbol{e}(t)\|_{\boldsymbol{G}}$ and $\|\boldsymbol{e}_{\text{out}}(t)\|_2$ of (6). It is originally based on an error estimator for first-order systems by Haasdonk and Ohlberger.[7] Simply transforming the second-order system (1) to the equivalent first-order system (2) and applying this error estimator may lead to a high overestimation of the original error by the so-called hump phenomenon.[6] This is circumvented by writing the error estimator in a second-order form, which does not contain the term that is responsible for the high overestimation.

We will only give a short summary of how the error estimator is derived. First, a differential equation for the error $\begin{pmatrix}\boldsymbol{e}(t)\\ \dot{\boldsymbol{e}}(t)\end{pmatrix}$ of the first-order system is found by differentiating (6) with respect to t and substituting (2) and the residual of (5). Given initial conditions $\boldsymbol{e}_0 = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{W}^{\mathsf{T}})\boldsymbol{q}_0$ for the error in the state and $\dot{\boldsymbol{e}}_0 = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{W}^{\mathsf{T}})\dot{\boldsymbol{q}}_0$ for the error in the velocity, the differential equation can be solved analytically. Taking only the upper part representing $\boldsymbol{e}(t)$ yields the explicit solution

$$\boldsymbol{e}(t) = \boldsymbol{\Phi}_{11}(t)\boldsymbol{e}_0 + \boldsymbol{\Phi}_{12}(t)\dot{\boldsymbol{e}}_0 + \int_0^t \boldsymbol{\Phi}_{12}(t-\tau)\tilde{\boldsymbol{r}}(\tau)\,d\tau \tag{8}$$

for the error with the mass-inverted residual

$$\tilde{\boldsymbol{r}}(\tau) := -\boldsymbol{M}^{-1}\boldsymbol{r}(t) = \boldsymbol{M}^{-1}\boldsymbol{B}\boldsymbol{u}(t) - \boldsymbol{V}\ddot{\overline{\boldsymbol{q}}}(t) - \boldsymbol{M}^{-1}\boldsymbol{D}\boldsymbol{V}\dot{\overline{\boldsymbol{q}}}(t) - \boldsymbol{M}^{-1}\boldsymbol{K}\boldsymbol{V}\overline{\boldsymbol{q}}(t) \tag{9}$$

and the fundamental matrix

$$\boldsymbol{\Phi}(t) := \begin{pmatrix} \boldsymbol{\Phi}_{11} & \boldsymbol{\Phi}_{12} \\ \boldsymbol{\Phi}_{21} & \boldsymbol{\Phi}_{22} \end{pmatrix} := \exp(\boldsymbol{A}t) \tag{10}$$

of the system $\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t)$. The block matrices $\boldsymbol{\Phi}_{ij}$ of $\boldsymbol{\Phi}$ are of size $N \times N$. Taking the norm $\|\cdot\|_{\boldsymbol{G}}$ on both sides of (8), using the triangle inequality of norms as well as the definition of the matrix norm, and taking the norm inside the integral yields

$$\|\boldsymbol{e}(t)\|_{\boldsymbol{G}} \leq \Delta_{\boldsymbol{q}}(t) := C_{11}(t)\|\boldsymbol{e}_0\|_{\boldsymbol{G}} + C_{12}(t)\|\dot{\boldsymbol{e}}_0\|_{\boldsymbol{G}} + \int_0^t C_{12}(t-\tau)\|\tilde{\boldsymbol{r}}(\tau)\|_{\boldsymbol{G}}\, d\tau \tag{11}$$

with

$$C_{ij}(t) := \|\boldsymbol{\Phi}_{ij}(t)\|_{\boldsymbol{G}}, \quad 1 \leq i, j \leq 2\,. \tag{12}$$

The values $C_{11}(t)$ and $C_{12}(t)$ will play an important role in the following. We call them constants even though they depend on time t since they are independent of the reduced simulation result, i.e., online values. After these constants and the initial errors $\|\boldsymbol{e}_0\|_{\boldsymbol{G}}$, $\|\dot{\boldsymbol{e}}_0\|_{\boldsymbol{G}}$ are pre-computed in an offline step, the error bound $\Delta_{\boldsymbol{q}}(t)$ for the state can be computed by inserting the reduced simulation result into the residual $\tilde{\boldsymbol{r}}$ and integrating over time according to (11).

An error estimator $\Delta_{\boldsymbol{y}}(t) := \|\boldsymbol{C}\|_{\boldsymbol{G},2}\Delta_{\boldsymbol{q}}(t)$ for the output follows directly from

$$\begin{aligned} \boldsymbol{e}_{\text{out}} &= \boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t) = \boldsymbol{C}\left(\boldsymbol{q}(t) - \tilde{\boldsymbol{q}}(t)\right)\,, \\ \|\boldsymbol{e}_{\text{out}}\|_2 &\leq \|\boldsymbol{C}\|_{\boldsymbol{G},2}\|\boldsymbol{q}(t) - \tilde{\boldsymbol{q}}(t)\|_{\boldsymbol{G}} \leq \|\boldsymbol{C}\|_{\boldsymbol{G},2}\Delta_{\boldsymbol{q}}(t) \end{aligned}$$

with $\|\boldsymbol{C}\|_{\boldsymbol{G},2} := \sup_{\|\boldsymbol{q}\|_{\boldsymbol{G}}=1}\|\boldsymbol{C}\boldsymbol{q}\|_2$. It depends linearly on the error estimator of the state. Therefore, we will focus only on the error estimation of the state in the following. Note that improved output error bounds are possible by a suitable adjoint problem and output correction term.[7]

The described error estimator has several advantages. Obviously, it is a-posteriori since it only depends on the reduced state and its derivative. Most noticeably the error estimator may be applied for any Petrov-Galerkin reduction independently of the chosen reduction matrices. This allows for an approach to find the optimal reduction where even the reduction method can be varied without changing the error estimator. It can be implemented in an offline / online fashion. The constants $C_{11}(t)$ and $C_{12}(t)$ need to be pre-computed for a specific full-order system. Then the error can be estimated by (11) for an arbitrary reduction method during the simulation since the error bound at time t does not depend on quantities of future time instances.

2.4 | Analysis of Scalability

We will now roughly analyze the computational complexity and scalability for $N \to \infty$ for the error bound (11). Section 5 will cover this topic in greater detail. Assuming that the system matrices and initial values of (1) as well as the projection matrices $\boldsymbol{V}$, $\boldsymbol{W}$ are known, the initial errors $\boldsymbol{e}_0 = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{W}^\mathsf{T})\boldsymbol{q}_0$ and $\dot{\boldsymbol{e}}_0 = (\boldsymbol{I} - \boldsymbol{V}\boldsymbol{W}^\mathsf{T})\dot{\boldsymbol{q}}_0$ are easy to compute. Even though these computations scale with the full-order dimension N, they only involve matrix multiplications and are performed only once for each reduced simulation run with changed initial values.

During the online phase, the residual $\tilde{\boldsymbol{r}}$ needs to be computed for each discretized time instance of the simulation run. The residual itself is a quantity in $\mathbb{R}^N$ but only its norm is saved. The vector $\boldsymbol{B}\boldsymbol{u}(t)$ is usually also sparse. The left-multiplication by $\boldsymbol{M}^{-1}$ is performed by solving the equivalent linear system with the sparse matrix $\boldsymbol{M}$ and does not need the explicit inversion of $\boldsymbol{M}$. The integral in (11) can be approximated by any numerical integration method that only needs evaluations of the integrand available during simulation.

As proven in Proposition 1, the computation of the matrix norm $\|\cdot\|_{\boldsymbol{G}}$ is equivalent to solving for the largest eigenvalue of a generalized eigenvalue problem. This is usually accomplished with the Lanczos method for Hermitian matrices as will be discussed later in Section 5. The computation of the weighted norm of a vector involves by definition only matrix multiplications and the usually sparse matrix $\boldsymbol{G}$.

Only the constants $C_{11}(t)$ and $C_{12}(t)$ as part of the offline stage remain to be analyzed. One problem lies in the matrix $\boldsymbol{A}t$ of which we need to take the matrix exponential according to (10). The matrix $\boldsymbol{A}$ needs to be computed directly. Its definition in (3) involves the inverse of $\boldsymbol{M}$. While $\boldsymbol{M}$ itself is a sparse matrix coming from an FE discretization, its inverse $\boldsymbol{M}$ is usually a dense matrix scaling like N^2 in terms of memory consumption. Even for sparse, e.g., diagonal $\boldsymbol{M}$, the matrix exponential is usually a dense matrix. Additionally, the matrix exponential (10) needs to be computed for each discretized time instance of t leading

to a high computational complexity. All together, the computation of these constants in the offline stage is infeasible for large systems, which will be discussed in greater detail in Section 5 and demonstrated in the numerical examples of Section 6.

The remainder of this work will largely improve the computational complexity to compute the constants $C_{11}(t)$ and $C_{12}(t)$, which constitutes the main novelty of this work. With little to no additional inaccuracy, they will be computable even for large systems within decent time and memory limitations. First, we will rewrite the constants as the solution of a one-dimensional problem involving generalized eigenvalues.

3 | ALTERNATIVE FORMS OF THE CONSTANTS

As has been argued in the last section, the computation of the constants $C_{11}(t)$ and $C_{12}(t)$ is a bottleneck for large systems. Instead of computing $\exp(\boldsymbol{A}t)$ directly and taking the norm $\|\exp(\boldsymbol{A}t)\|_{\boldsymbol{G}}$, we will derive an alternative form of this norm as supremum over eigenvalues of a specific eigenvalue problem. The computation of the norm is then reduced to a one-dimension problem once these eigenvalues are known. It will be discussed in Section 4 how we are able to calculate the supremum efficiently without knowing all N eigenvalues.

The damped case turns out the be more challenging than the undamped case. Therefore, we will sometimes look at both systems separately. Additionally, we will restrict ourselves to proportional or Rayleigh damping, i.e., $\boldsymbol{D} = \alpha\boldsymbol{M} + \beta\boldsymbol{K}$ with the damping parameters $\alpha, \beta \geq 0$.[16] The choice $\alpha = \beta = 0$ constitutes an undamped system. For damped systems, we assume $\beta > 0$ and $\alpha\beta \leq 1$ in order to circumvent some special cases.

Until now, we only worked with symmetric matrices. This will change and it is needed to extend the common definition of positive definiteness to non-Hermitian matrices.

Definition 2 (Definiteness)**.** For arbitrary $\boldsymbol{Z} \in \mathbb{K}^{N\times N}$ with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, we say that $\boldsymbol{Z}$ is positive definite ($\boldsymbol{Z} > \boldsymbol{0}$) or positive semi-definite ($\boldsymbol{Z} \geq \boldsymbol{0}$) if

$$\forall \boldsymbol{z} \in \mathbb{K}^N \setminus \{\boldsymbol{0}\} : \operatorname{Re}\langle \boldsymbol{Z}\boldsymbol{z}, \boldsymbol{z}\rangle > 0 \text{ or}$$
$$\forall \boldsymbol{z} \in \mathbb{K}^N \setminus \{\boldsymbol{0}\} : \operatorname{Re}\langle \boldsymbol{Z}\boldsymbol{z}, \boldsymbol{z}\rangle \geq 0\,,$$

respectively. Negative definiteness and negative semi-definiteness are defined analogously.

We define the matrices

$$\boldsymbol{F} := \boldsymbol{M}^{-1}\boldsymbol{K}\,, \tag{13}$$
$$\boldsymbol{E} := \frac{1}{2}\boldsymbol{M}^{-1}\boldsymbol{D} = \frac{\alpha}{2}\boldsymbol{I} + \frac{\beta}{2}\boldsymbol{F} \tag{14}$$

as mass-inverted forms of $\boldsymbol{K}$ and $\boldsymbol{D}$ for easier notation. This yields the new notation of the state matrix (3)

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \\ -\boldsymbol{F} & -2\boldsymbol{E} \end{pmatrix}. \tag{15}$$

Remark 1. We use the notation

$$\varphi\left(\tilde{A}\right) := \left\{\varphi(a) : a \in \tilde{A}\right\} \subseteq \operatorname{rg}(\varphi) \subseteq \boldsymbol{B}$$

for the image of the subset $\tilde{A}$ of the domain A of any arbitrary mapping φ with codomain $\boldsymbol{B}$.

In the remainder of this work, we will write $\sigma(\boldsymbol{Z}) \subseteq \mathbb{C}$ for the spectrum of a matrix $\boldsymbol{Z}$, i.e., the set of all its complex eigenvalues. The spectral theorem, see e.g. Reed and Simon[17, Theorems VI.6 and VII.1], will play a crucial part in the following. We will only need a variant for the special case of power series.

Theorem 1 (Spectral Theorem for Power Series)**.** For any Hermitian matrix $\boldsymbol{H}$ and power series

$$p(x) = \sum_{k=0}^{\infty} a_k x^k$$

with real coefficients $a_k \in \mathbb{R}$ for all $k \in \mathbb{N}_0$ and infinite radius of convergence, it holds:

(a) $\|\boldsymbol{H}\|_2 = \max|\sigma(\boldsymbol{H})|$,

(b) $p(\boldsymbol{H})$ is a Hermitian matrix in $\mathbb{C}^{N\times N}$,

(c) $\sigma(p(\boldsymbol{H})) = p(\sigma(\boldsymbol{H}))$ with $\bigoplus_{\lambda\in p^{-1}(\mu)} \mathcal{E}_\lambda(\boldsymbol{H}) = \mathcal{E}_\mu(p(\boldsymbol{H}))$ for the corresponding eigenspaces for every $\mu \in \sigma(p(\boldsymbol{H}))$,

(d) $\|p(\boldsymbol{H})\|_2 = \max|p(\sigma(\boldsymbol{H}))|$,

i.e., the spectral norm of a power series of a Hermitian matrix can be easily computed from its eigenvalues.

Proof.

(a) By definition, the spectral norm of $\boldsymbol{H} \in \mathbb{C}^{N\times N}$ is its largest singular value. A largest singular value of $\boldsymbol{H}$ corresponds to the square root of the largest eigenvalue of $\boldsymbol{H}^\mathsf{H}\boldsymbol{H}$. Since $\boldsymbol{H}$ is Hermitian, we have $\boldsymbol{H}^\mathsf{H}\boldsymbol{H} = \boldsymbol{H}^2$. Taking the square root of the largest eigenvalue of this positive semi-definite matrix is the same as taking the largest eigenvalue of $\boldsymbol{H}$ in absolute value, i.e., $\max|\sigma(\boldsymbol{H})|$.

(b) Let $p_K(x) := \sum_{k=0}^{K} a_k x^k$ be the partial sum of the power series p to the K-th order. The polynomial $p_K(\boldsymbol{H})$ in $\boldsymbol{H}$ is then defined as

$$p_K(\boldsymbol{H}) := \sum_{k=0}^{K} a_k \boldsymbol{H}^k \in \mathbb{C}^{N\times N}$$

with $\boldsymbol{H}^0 := \boldsymbol{I}$. It is Hermitian due to the properties of the conjugate transpose and the coefficients a_k being real.

First we note that the Cauchy criterion holds for all Banach spaces. Therefore, we only have to check the Cauchy criterion for the sequence $p_K(\boldsymbol{H})$ in the Banach algebra $(\mathbb{C}^{N\times N}, \|\cdot\|_2)$ for the convergence of $p(\boldsymbol{H})$. This immediately follows from the inequality

$$\left\| \sum_{k=K}^{L} a_k \boldsymbol{H}^k \right\|_2 \leq \sum_{k=K}^{L} |a_k| \|\boldsymbol{H}\|_2^k$$

and that the power series $p(x)$ converges in $\mathbb{R}$, i.e., it fulfills the Cauchy criterion for each x since we assumed an infinite convergence radius. Therefore, $p(\boldsymbol{H})$ is well-defined.

It still needs to be shown that $p(\boldsymbol{H})$ is Hermitian, i.e., a self-adjoint operator. The scalar product in $\mathbb{C}^N$ induces the matrix norm $\|\cdot\|_2$, which was used to define $p(\boldsymbol{H})$. Since each partial sum $p_K(\boldsymbol{H})$ is Hermitian and the scalar product is continuous, it follows immediately that $p(\boldsymbol{H})$ is also Hermitian.

(c) Let $\lambda \in \sigma(\boldsymbol{H})$ and define $\mathcal{E}_\lambda(\boldsymbol{H})$ as the eigenspace of $\boldsymbol{H}$ to λ. Then for every $\boldsymbol{v} \in \mathcal{E}_\lambda(\boldsymbol{H})$, we have $\boldsymbol{H}\boldsymbol{v} = \lambda\boldsymbol{v}$. It follows immediately $p_K(\boldsymbol{H})\boldsymbol{v} = p_K(\lambda)\boldsymbol{v}$ and by taking the limit on both sides also $p(\boldsymbol{H})\boldsymbol{v} = p(\lambda)\boldsymbol{v}$. Therefore, not only $p(\lambda) \in \sigma(p(\boldsymbol{H}))$ but also $\mathcal{E}_\lambda(\boldsymbol{H}) \subseteq \mathcal{E}_{p(\lambda)}(p(\boldsymbol{H}))$. Since $N = \sum_{\lambda\in\sigma(\boldsymbol{H})} \dim \mathcal{E}_\lambda(\boldsymbol{H}) = \sum_{\mu\in\sigma(p(\boldsymbol{H}))} \dim \mathcal{E}_\mu(p(\boldsymbol{H}))$ for the Hermitian matrices $\boldsymbol{H}$ and $p(\boldsymbol{H})$, and since eigenvectors to distinct eigenvalues are linearly independent, it follows $\bigoplus_{\lambda\in p^{-1}(\mu)} \mathcal{E}_\lambda(\boldsymbol{H}) = \mathcal{E}_\mu(p(\boldsymbol{H}))$ for every $\mu \in \sigma(p(\boldsymbol{H}))$ as well as $\sigma(p(\boldsymbol{H})) = p(\sigma(\boldsymbol{H}))$ by the pigeonhole principle. The union of the eigenspaces is necessary since p may not be injective.

(d) follows immediately from (a) and (c). □

The goal is now to extract $\boldsymbol{\Phi}_{11}$ and $\boldsymbol{\Phi}_{12}$ from the power series of $\exp(\boldsymbol{A}t)$, see Equations (10) and (15). We define the matrix

$$\boldsymbol{R} := \boldsymbol{E}^2 - \boldsymbol{F} = \frac{1}{4}\left(\beta^2\boldsymbol{F}^2 + (2\alpha\beta - 4)\boldsymbol{F} + \alpha^2\boldsymbol{I}\right) \tag{16}$$

for easier notation. It is easy to see that $\boldsymbol{E}$, $\boldsymbol{F}$, and $\boldsymbol{R}$ commute pairwisely.

Theorem 2 (Power Series of $\boldsymbol{\Phi}$)**.** The upper left and upper right block matrices of $\boldsymbol{\Phi}$ defined in Equation (10) have a power series expansion of

$$\boldsymbol{\Phi}_{11}(t) = \exp(-\boldsymbol{E}t)\left(p_\mathrm{c}(\boldsymbol{R};t) + \boldsymbol{E}p_\mathrm{s}(\boldsymbol{R};t)\right) , \tag{17}$$

$$\boldsymbol{\Phi}_{12}(t) = \exp(-\boldsymbol{E}t)p_\mathrm{s}(\boldsymbol{R};t) \tag{18}$$

w.r.t. $\boldsymbol{F}$ with

$$p_\mathrm{s}(x;t) := \frac{0}{0!} + \frac{t}{1!} + \frac{0}{2!} + \frac{xt^3}{3!} + \frac{0}{4!} + \frac{x^2t^5}{5!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k t^{2k+1}}{(2k+1)!} = \sinh\left(x^{\frac{1}{2}}t\right) x^{-\frac{1}{2}} , \tag{19}$$

$$p_\mathrm{c}(x;t) := \frac{1}{0!} + \frac{0}{1!} + \frac{xt^2}{2!} + \frac{0}{3!} + \frac{x^2t^4}{4!} + \frac{0}{5!} + \cdots = \sum_{k=0}^{\infty} \frac{x^k t^{2k}}{(2k)!} = \cosh\left(x^{\frac{1}{2}}t\right) . \tag{20}$$

The power series in Equations (17) and (18) have an infinite radius of convergence. We define $0^0 := 1$ in the sum notation. The terms $\sinh\left(x^{\frac{1}{2}}t\right)x^{-\frac{1}{2}}$ and $\cosh\left(x^{\frac{1}{2}}\right)$ are to be understood only symbolically for now.

Proof. Since $\boldsymbol{E}$, $\boldsymbol{F}$, and $\boldsymbol{R}$ commute pairwisely, one can set $N = 1$ w.l.o.g. to obtain the series expansions (17) and (18) symbolically from

$$\exp(\boldsymbol{A}t) = \sum_{k=0}^{\infty} \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}t \\ -\boldsymbol{F}t & -2\boldsymbol{E}t \end{pmatrix}^k \frac{1}{k!}\,.$$

The technical calculations are left to the reader.

Since the hyperbolic sine and cosine have a convergence radius of infinity, this is also true for $p_{\mathrm{s}}(x)$ and $p_{\mathrm{c}}(x)$. The matrices $\boldsymbol{E}$ and $\boldsymbol{R}$ can be written as polynomials in $\boldsymbol{F}$, see (14) and (16). Inserting a polynomial into a power series with radius of convergence of infinity makes it again a power series with the same radius of convergence. The same is true for the multiplication of two power series, e.g., $\exp(-\boldsymbol{E}t)$ and $p_{\mathrm{s}}(\boldsymbol{R};t)$. This shows that each of the Equations (17) and (18) can be indeed written as a power series of $\boldsymbol{F}$ with radius of convergence of infinity. □

These power series will help us in Section 4 to compute the constants $C_{11}(t)$ and $C_{12}(t)$. For this, we need to take the weighted norm of $\boldsymbol{\Phi}_{11}$ and $\boldsymbol{\Phi}_{12}$, which will eventually transform the power series of Equations (17) and (18) w.r.t. $\boldsymbol{F}$ to a power series w.r.t. $\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{F}\boldsymbol{G}^{-\frac{1}{2}}$. In order to apply the Spectral Theorem 1, we need to show that this matrix is Hermitian. This is in general only true for $\boldsymbol{G} = \boldsymbol{M}$ since $\boldsymbol{F}$ is usually not symmetric, see Remark 5.

Lemma 2. For $\boldsymbol{G} = \boldsymbol{M}$, the matrices

$$\boldsymbol{Q} := \boldsymbol{G}^{\frac{1}{2}}\boldsymbol{E}\boldsymbol{G}^{-\frac{1}{2}}\,,$$
$$\boldsymbol{P} := \boldsymbol{G}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{G}^{-\frac{1}{2}}$$

are symmetric. Additionally, they are polynomial in $\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{F}\boldsymbol{G}^{-\frac{1}{2}}$, which is also symmetric.

Proof. First, we look at

$$\boldsymbol{Q} = \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{E}\boldsymbol{M}^{-\frac{1}{2}} = \frac{\alpha}{2}\boldsymbol{I} + \frac{\beta}{2}\boldsymbol{M}^{\frac{1}{2}}(\boldsymbol{M}^{-1}\boldsymbol{K})\boldsymbol{M}^{-\frac{1}{2}} = \frac{\alpha}{2}\boldsymbol{I} + \frac{\beta}{2}\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}\,.$$

The product $\boldsymbol{G}^{\frac{1}{2}}\boldsymbol{F}\boldsymbol{G}^{-\frac{1}{2}} = \boldsymbol{M}^{\frac{1}{2}}(\boldsymbol{M}^{-1}\boldsymbol{K})\boldsymbol{M}^{-\frac{1}{2}} = \boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}$ is symmetric since it reads the same from right to left as from left to right and each factor is symmetric, see Lemma 1. Therefore, $\boldsymbol{Q}$ is also symmetric. Similar calculations can be done for

$$\begin{aligned} 4\boldsymbol{P} &= 4\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{R}\boldsymbol{M}^{-\frac{1}{2}} = \boldsymbol{M}^{\frac{1}{2}}\left(\beta^2\boldsymbol{F}^2 + (2\alpha\beta - 4)\boldsymbol{F} + \alpha^2\boldsymbol{I}\right)\boldsymbol{M}^{-\frac{1}{2}} = \beta^2\boldsymbol{M}^{\frac{1}{2}}(\boldsymbol{M}^{-1}\boldsymbol{K})^2\boldsymbol{M}^{-\frac{1}{2}} + (2\alpha\beta - 4)\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}} + \alpha^2\boldsymbol{I} \\ &= \beta^2(\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}})^2 + (2\alpha\beta - 4)\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}} + \alpha^2\boldsymbol{I}\,. \end{aligned}$$

We already know that $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}$ is symmetric. Additionally, the square root of a symmetric matrix is again symmetric. □

All these preparations allow us to write the constants $C_{11}(t)$ and $C_{12}(t)$ in an alternative form that only depends on the solution of the generalized eigenvalue problem $\boldsymbol{K}\boldsymbol{v} = \lambda\boldsymbol{M}\boldsymbol{v}$.

Theorem 3 (Alternative Representation of $C_{11}(t)$ and $C_{12}(t)$)**.** For $\boldsymbol{G} = \boldsymbol{M}$, the constants $C_{11}(t)$ and $C_{12}(t)$ can be computed as

$$C_{11}(t) = \max_{f\in\sigma(\boldsymbol{F})} \left|\exp(-e(f)t)\left(\cosh\left(s(f)t\right) + e(f)s(f)^{-1}\sinh\left(s(f)t\right)\right)\right|\,, \tag{21}$$

$$C_{12}(t) = \max_{f\in\sigma(\boldsymbol{F})} \left|\exp(-e(f)t)\sinh\left(s(f)t\right)s(f)^{-1}\right| \tag{22}$$

with

$$e(f) := \frac{\alpha}{2} + \frac{\beta}{2}f\,, \tag{23}$$

$$r(f) := \frac{1}{4}\left(\beta^2 f^2 + (2\alpha\beta - 4)f + \alpha^2\right)\,, \tag{24}$$

$$s(f) := \sqrt{r(f)} = \frac{1}{2}\sqrt{\beta^2 f^2 + (2\alpha\beta - 4)f + \alpha^2}\,. \tag{25}$$

Proof. The time t is assumed to be fixed. First, we will incorporate the weighting matrix $\boldsymbol{G}$ in the power series of Theorem 2 with the help of Definition 1. This needs $\boldsymbol{G}^{\frac{1}{2}} \exp(\boldsymbol{Z}) \boldsymbol{G}^{-\frac{1}{2}} = \exp(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{Z} \boldsymbol{G}^{-\frac{1}{2}})$ and $\boldsymbol{I} = \boldsymbol{G}^{-\frac{1}{2}} \boldsymbol{G}^{\frac{1}{2}}$ in the power series.

$$\begin{aligned} C_{11}(t) &= \|\boldsymbol{\Phi}_{11}(t)\|_{\boldsymbol{G}} = \|\exp(-\boldsymbol{E}t) \left(p_{\mathrm{c}}(\boldsymbol{R};t) + \boldsymbol{E} p_{\mathrm{s}}(\boldsymbol{R};t)\right)\|_{\boldsymbol{G}} \\ &= \left\| \exp(-\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{E} \boldsymbol{G}^{-\frac{1}{2}} t) \left(p_{\mathrm{c}}(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{R} \boldsymbol{G}^{-\frac{1}{2}}; t) + \boldsymbol{G}^{\frac{1}{2}} \boldsymbol{E} \boldsymbol{G}^{-\frac{1}{2}} p_{\mathrm{s}}(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{R} \boldsymbol{G}^{-\frac{1}{2}}; t) \right) \right\|_2 \\ &= \left\| \exp(-\boldsymbol{Q}t) \left(p_{\mathrm{c}}(\boldsymbol{P};t) + \boldsymbol{Q} p_{\mathrm{s}}(\boldsymbol{P};t) \right) \right\|_2 , \end{aligned} \tag{26}$$

$$\begin{aligned} C_{12}(t) &= \|\boldsymbol{\Phi}_{12}(t)\|_{\boldsymbol{G}} = \|\exp(-\boldsymbol{E}t) p_{\mathrm{s}}(\boldsymbol{R};t)\|_{\boldsymbol{G}} = \left\| \exp(-\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{E} \boldsymbol{G}^{-\frac{1}{2}} t) p_{\mathrm{s}}(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{R} \boldsymbol{G}^{-\frac{1}{2}}; t) \right\|_2 \\ &= \left\| \exp(-\boldsymbol{Q}t) p_{\mathrm{s}}(\boldsymbol{P};t) \right\|_2 . \end{aligned} \tag{27}$$

We already know from Theorem 2 that $\exp(-\boldsymbol{E}t) \left(p_{\mathrm{c}}(\boldsymbol{R};t) + \boldsymbol{E} p_{\mathrm{s}}(\boldsymbol{R};t)\right)$ and $\exp(-\boldsymbol{E}t) p_{\mathrm{s}}(\boldsymbol{R};t)$ are indeed power series w.r.t. $\boldsymbol{F}$. Substituting $\boldsymbol{E}$ by $\boldsymbol{Q}$ and $\boldsymbol{R}$ by $\boldsymbol{P}$ and taking the spectral norm leads to (26) and (27). Since we know from Lemma 2 that $\boldsymbol{Q}$ and $\boldsymbol{P}$ are polynomial in $\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}$, (26) and (27) can indeed be written as spectral norm of a power series – called $p_{11}(x;t)$ and $p_{12}(x;t)$ for now – w.r.t. $\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}$, which is symmetric. This allows the application of the Spectral Theorem 1 for fixed t with $\boldsymbol{H} = \boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}$ and $p = p_{11}$ or $p = p_{12}$, respectively.

$$C_{11}(t) = \left\| \exp(-\boldsymbol{Q}t) \left(p_{\mathrm{c}}(\boldsymbol{P};t) + \boldsymbol{Q} p_{\mathrm{s}}(\boldsymbol{P};t)\right) \right\|_2 = \left\| p_{11}\left(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}; t\right) \right\|_2 = \max \left| p_{11}\left(\sigma(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}); t\right) \right| = \max \left| p_{11}\left(\sigma(\boldsymbol{F}); t\right) \right| ,$$

$$C_{12}(t) = \left\| \exp(-\boldsymbol{Q}t) p_{\mathrm{s}}(\boldsymbol{P};t) \right\|_2 = \left\| p_{12}\left(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}; t\right) \right\|_2 = \max \left| p_{12}\left(\sigma(\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}); t\right) \right| = \max \left| p_{12}\left(\sigma(\boldsymbol{F}); t\right) \right| .$$

The last equality in each line uses that the spectrum of the similar matrices $\boldsymbol{G}^{\frac{1}{2}} \boldsymbol{F} \boldsymbol{G}^{-\frac{1}{2}}$ and $\boldsymbol{F}$ are the same.

For $f \in \sigma(\boldsymbol{F})$, we can now reorder the power series p_{11} and p_{22} again to have a form like (17) and (18), respectively. The difference is that we are not inserting the matrices but their eigenvalues into the power series. They can be interchanged freely according to Theorem 1(c) and since $e(f)$ and $r(f)$ were defined such that $\boldsymbol{E} = e(\boldsymbol{F})$ and $\boldsymbol{R} = r(\boldsymbol{F})$.

$$\begin{aligned} C_{11}(t) &= \max_{f \in \sigma(\boldsymbol{F})} \left| p_{11}(f;t) \right| = \max_{f \in \sigma(\boldsymbol{F})} \left| \exp(-e(f)t) \left(p_{\mathrm{c}}(r(f);t) + e(f) p_{\mathrm{s}}(r(f);t)\right) \right| \\ &= \max_{f \in \sigma(\boldsymbol{F})} \left| \exp(-e(f)t) \left(\cosh\left(r(f)^{\frac{1}{2}} t\right) + e(f) r(f)^{-\frac{1}{2}} \sinh\left(r(f)^{\frac{1}{2}} t\right) \right) \right| , \\ C_{12}(t) &= \max_{f \in \sigma(\boldsymbol{F})} \left| p_{12}(f;t) \right| = \max_{f \in \sigma(\boldsymbol{F})} \left| \exp(-e(f)t) p_{\mathrm{s}}(r(f);t) \right| \\ &= \max_{f \in \sigma(\boldsymbol{F})} \left| \exp(-e(f)t) \sinh\left(r(f)^{\frac{1}{2}} t\right) r(f)^{-\frac{1}{2}} \right| . \end{aligned}$$

□

Theorem 3 constitutes one of the main results of this work since it allows to compute the constants $C_{11}(t)$ and $C_{12}(t)$ as a maximum over the eigenvalues of $\boldsymbol{F}$. We will see in Section 4 how to find this maximum quickly without computing the full spectrum of $\boldsymbol{F}$. At some occasions, it will be more convenient to use a variant of this theorem which depends on the spectrum of $\boldsymbol{E}$ instead of $\boldsymbol{F}$. This is the statement of the following corollary.

Corollary 1 (Alternative Representation w.r.t. $\sigma(\boldsymbol{E})$)**.** For $\boldsymbol{G} = \boldsymbol{M}$ and $\beta \neq 0$, the constants $C_{11}(t)$ and $C_{12}(t)$ can be computed as

$$C_{11}(t) = \max_{e \in \sigma(\boldsymbol{E})} \left| c_{11}(e;t) \right| , \tag{28}$$

$$C_{12}(t) = \max_{e \in \sigma(\boldsymbol{E})} \left| c_{12}(e;t) \right| \tag{29}$$

with

$$r(e) := e^2 - \frac{2}{\beta} e + \frac{\alpha}{\beta} , \tag{30}$$

$$s(e) := \sqrt{r(e)} = \sqrt{e^2 - \frac{2}{\beta} e + \frac{\alpha}{\beta}} , \tag{31}$$

$$c_{11}(e;t) = \exp(-et) \left(\cosh\left(s(e)t\right) + \frac{e}{s(e)} \sinh\left(s(e)t\right) \right) , \tag{32}$$

$$c_{12}(e;t) = \exp(-et) \frac{\sinh\left(s(e)t\right)}{s(e)} . \tag{33}$$

It will be clear from the context whether $s(\cdot)$ refers to (25) or (31). The same holds for $r(\cdot)$.

Proof. The proof is skipped since it is almost the same as the proof of Theorem 3. □

Remark 2. While the terms $\sinh\left(x^{\frac{1}{2}}t\right)x^{-\frac{1}{2}}$ and $\cosh\left(x^{\frac{1}{2}}\right)$ were only to be understood symbolically in Theorem 2, they are now valid and $s(f) \neq 0$ and $s(e) \neq 0$ can be assumed in Theorem 3 and Corollary 1. See the next remark for the case $s(f) = s(e) = 0$. Due to their structure, they yield the same result for any square root of $r(f)$ and $r(e)$, respectively. For easier notation, we will define $s(f)$ and $s(e)$ to be the unique principal value of the square root of $r(f)$ and $r(e)$ in the rest of this work, respectively.

Remark 3. It is allowed to divide by zero in Theorem 3 and Corollary 1 since

$$\lim_{s\to 0}\frac{\sinh(st)}{s} = \lim_{s\to 0}\cosh(st)t = t$$

according to L'Hôpital's rule in $\mathbb{C}$, which is obviously the same value as inserting $s = 0$ into the corresponding power series (19).

With Theorem 3 and Corollary 1, we got rid of the direct inversion of $\boldsymbol{M}$ and are left with a one-dimensional maximization problem of the spectrum of $\boldsymbol{F}$. The question remains if the full spectrum of $\boldsymbol{F}$ needs to be computed in order to find the maximum. We will find a way to circumvent this in the next section.

4 | FAST CONSTANT APPROXIMATION

This section constitutes the second main contribution of this work. The goal is to find the maximum in Theorem 3 and Corollary 1 with low computational costs since computing the full spectrum of $\boldsymbol{F}$ is infeasible. The computational costs will be discussed in detail in Section 5. For now it is enough to assume that the $m \ll N$ lowest / highest eigenvalues of $\boldsymbol{F} = \boldsymbol{M}^{-1}\boldsymbol{K}$ can be efficiently computed by solving the generalized eigenvalue problem $\boldsymbol{K}\boldsymbol{x} = \lambda\boldsymbol{M}\boldsymbol{x}$ with the Lanczos method. Therefore, we aim at using as few eigenvalues of $\boldsymbol{F}$ as possible.

For the remainder of this work, we assume $\boldsymbol{G} = \boldsymbol{M}$ since we are building upon Theorem 3 and Corollary 1. We can focus on the calculation of $C_{11}(t)$ and $C_{12}(t)$ for positive t since $\boldsymbol{\Phi}_{11}(0) = \boldsymbol{I}_{N\times N}$ and $\boldsymbol{\Phi}_{12}(0) = \boldsymbol{0}_{N\times N}$ according to (15) yielding $C_{11}(0) = 1$ and $C_{12}(0) = 0$. Therefore, we assume w.l.o.g. $t > 0$ in the following.

4.1 | Undamped Systems

Throughout this section, we assume an undamped system, i.e., $\alpha = \beta = 0$, which leads to $\boldsymbol{D} = \boldsymbol{E} = \boldsymbol{Q} = \boldsymbol{0}$ as well as $\boldsymbol{R} = -\boldsymbol{F}$. It is of importance for the following results to analyze the spectrum of $\boldsymbol{F}$ before applying Theorem 3.

Lemma 3. If $\boldsymbol{K}$ is positive semi-definite, then $\boldsymbol{F}$ is also a positive semi-definite, real matrix, i.e., $\sigma(\boldsymbol{F}) \subseteq \mathbb{R}_{\geq 0}$.

Proof. $\boldsymbol{M}^{-1}$ is real, symmetric, and positive definite since $\boldsymbol{M}$ is real, symmetric, and positive definite. Therefore, $\boldsymbol{M}^{-1}$ has a real, symmetric, and positive definite square root, which we will call $\boldsymbol{M}^{-\frac{1}{2}}$, see Lemma 1.

As already shown for the proof of Lemma 2, $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}$ is real and symmetric. It is also positive semi-definite: For arbitrary $\boldsymbol{z} \in \mathbb{C}^N$,

$$\langle \boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{z}, \boldsymbol{z}\rangle = \langle \boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{z}, \boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{z}\rangle \geq 0$$

since $\boldsymbol{K}$ is positive semi-definite.

Therefore, all eigenvalues of $\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K}\boldsymbol{M}^{-\frac{1}{2}}$ are non-negative, as is true for $\boldsymbol{M}^{-1}\boldsymbol{K} = \boldsymbol{M}^{-\frac{1}{2}}(\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{K})$ since both matrices are equivalent. Thus, $\boldsymbol{F} = \boldsymbol{M}^{-1}\boldsymbol{K}$ is positive semi-definite and real as a factor of two real matrices. □

Remark 4. For positive definite $\boldsymbol{K}$, it can be shown with only small changes to the above proof that $\boldsymbol{F}$ is also positive definite.

Remark 5. $\boldsymbol{F}$ is in general not symmetric. Choose, e.g., $\boldsymbol{M} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and any 2×2 SPSD matrix for $\boldsymbol{K}$ as counterexample.

In the undamped case, the constants $C_{11}(t)$ and $C_{12}(t)$ have a much easier form, which will be seen in the following corollary following directly from Theorem 3 and the last lemma:

Corollary 2. Under the assumptions of this section, it holds

$$C_{11}(t) = \max_{f\in\sigma(\boldsymbol{F})} |\cosh(s(f)t)| = \max_{f\in\sigma(\boldsymbol{F})} |\cos(\sqrt{f}t)|\,, \tag{34}$$

$$C_{12}(t) = \max_{f\in\sigma(\boldsymbol{F})} \left|\sinh(s(f)t)\, s(f)^{-1}\right| = \max_{f\in\sigma(\boldsymbol{F})} \frac{\left|\sin(\sqrt{f}t)\right|}{\sqrt{f}}\,. \tag{35}$$

Proof. The first equalities in (34) and (35) follow directly by setting $\alpha = \beta = 0$ in Theorem 3.

According to Lemma 3, $s(f) = \sqrt{-f} = i\sqrt{f}$ is purely imaginary having w.l.o.g. non-negative imaginary part, i.e., $s(f) \in i\mathbb{R}_{\geq 0}$, see Remark 2. Dividing by zero is again defined with the limit (Remark 3). Due to known properties of the hyperbolic sine and cosine, we can simplify (34) and (35) even further. We use known trigonometric identities for the hyperbolic sine and cosine:

$$\begin{aligned}
|\cosh(s(f)t)| &= \left|\cosh\left(i\sqrt{f}t\right)\right| = \left|\cos\left(\sqrt{f}t\right)\right|\,,\\
\left|\sinh(s(f)t)\, s(f)^{-1}\right| &= \left|\sinh\left(i\sqrt{f}t\right)\left(i\sqrt{f}\right)^{-1}\right| = \left|i\sin\left(\sqrt{f}t\right)\left(i\sqrt{f}\right)^{-1}\right| = \left|\sin\left(\sqrt{f}t\right)\right|\left(\sqrt{f}\right)^{-1}\,.
\end{aligned}$$

□

Since it is assumed that we cannot compute the full spectrum of $\boldsymbol{F}$, we are unable to predict the value of (34) since there is always the possibility that one eigenvalue f of $\boldsymbol{F}$ fulfills $\sqrt{f}t \in \pi\mathbb{Z}$ leading to the maximum absolute value of the cosine. But we can use this to our advantage: Since $\sigma(\boldsymbol{F})$ has usually around $N \gg 1$ distinct eigenvalues scattered above several orders of magnitude (see the examples in Section 6), there is a high probability that the maximum will be near 1. Therefore, we are fine using 1 as an upper bound for $C_{11}(t)$ in the undamped case. The estimation for $C_{12}(t)$ will be a little bit more complicated.

Theorem 4 (Upper bounds for $C_{11}(t)$ and $C_{12}(t)$ for undamped systems)**.**

(a) $C_{11}(t) \leq 1 =: C_{11,\text{bound}}(t)$ for all $t \geq 0$ and $\boldsymbol{K}$ positive definite. For $0 \in \sigma(\boldsymbol{K})$, even $C_{11}(t) = 1$ holds.

(b) Let $f_1 \leq f_2 \leq \ldots \leq f_N$ be the increasingly ordered elements of the spectrum of $\boldsymbol{F}$. Additionally, define $\mu_j := \sqrt{f_j}$. If $\mu_1 = 0$, then $C_{12}(t) = t$ for all $t \geq 0$. Otherwise,

$$C_{12}(t) \leq C^{(m)}_{12,\text{bound}}(t) := \max\left\{\frac{|\sin(\mu_1 t)|}{\mu_1}, \frac{|\sin(\mu_2 t)|}{\mu_2}, \ldots, \frac{|\sin(\mu_{m-1}t)|}{\mu_{m-1}}, \min\left\{\frac{1}{\mu_m}, t\right\}\right\} \tag{36}$$

for arbitrary $1 \leq m \leq N$ and all $t \geq 0$.

Proof.

(a) This is clear from Corollary 2 and $|\cos(\cdot)| \leq 1$.

(b) Due to the monotonicity of the square root, the μ_j's are also ordered monotonically increasing. They are also non-negative according to Lemma 3. From Corollary 2, we know

$$C_{12}(t) = \max_{1\leq j\leq N} \frac{\left|\sin(\mu_j t)\right|}{\mu_j}\,. \tag{37}$$

If $\mu_1 = 0$, then $|\sin(\mu_1 t)|\mu_1^{-1}$ is defined by the limit

$$\lim_{\mu\to 0+} \frac{\sin(\mu t)}{\mu} = \lim_{\mu\to 0+} t\cos(\mu t) = t$$

with L'Hôpital's rule (Remark 3). Together with $C_{12}(t) \leq 1$, this leads to $C_{12}(t) = 1$ in the case $\mu_1 = 0$.

Assume now $\mu_1 > 0$. Let μ_l for $1 \leq l \leq N$ be the index for which this maximum is reached. We differentiate two cases:

- If $1 \leq l \leq m-1$, then (37) is in the set from which the maximum is taken in (36).
- If $l \geq m$, then we can bound the sine by 1:

$$\frac{|\sin(\mu_l t)|}{\mu_l} \leq \frac{1}{\mu_l} \leq \frac{1}{\mu_m}\,.$$

We insert $x := \mu_j t \geq 0$ into the known equation $|\sin(x)| \leq x$, which holds for $x \geq 0$. It immediately follows $|\sin(\mu_j t)|\mu_j^{-1} \leq t$ for all $1 \leq j \leq N$ and therefore $C_{12}(t) \leq t$ with Equation (37). Both bounds can be combined to $\min\left\{\frac{1}{\mu_m}, t\right\}$, which constitutes the last value in the set of (36).

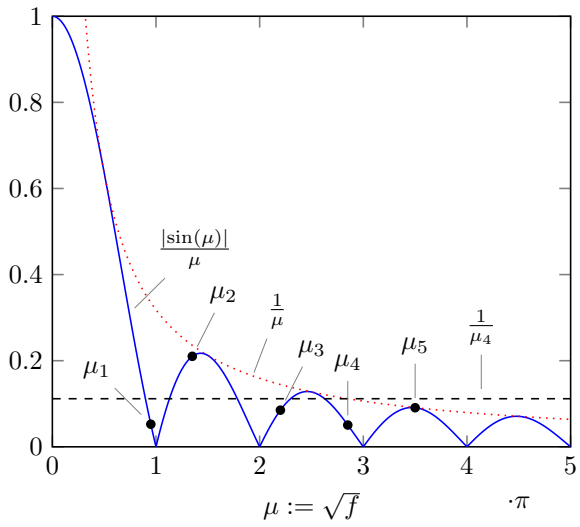

FIGURE 1 This figure helps to understand the motivation behind the proof of Theorem 4 for $t = 1$, $m = 4$, and positive definite $\boldsymbol{K}$. The blue function $|\sin(\mu)|\mu^{-1}$ is not monotonically decreasing. We want to find the maximum function value over all μ_j. It may happen that it has a small value for the first few μ_j like for μ_1 in this figure. Therefore, we are taking the maximum at the points $\mu_1, \dots, \mu_3$ – which is at μ_2 in this example – and approximate the function value at the remaining $\mu_4, \dots, \mu_N$ by μ_4^{-1}, which is the black, dashed line.

Both cases yield (36). Figure 1 depicts this idea in more detail. □

Remark 6. The bound $C_{12,\text{bound}}^{(m)}(t)$ of $C_{12}(t)$ is sharp in the sense that $C_{12,\text{bound}}^{(m)}(t) \to C_{12}(t)$ monotonically decreasing for $m \to N$.

The case $0 \in \sigma(\boldsymbol{K})$ does not need any algorithmic consideration since $C_{11}(t) = 1$ and $C_{12}(t) = t$ holds for every t according to Theorem 4. For the case $0 \notin \sigma(\boldsymbol{K})$, the results of Theorem 4 still allow us to find an easy algorithm which calculates the bounds $C_{11,\text{bound}}(t)$ and $C_{12,\text{bound}}^{(m)}(t)$ for $C_{11}(t)$ and $C_{12}(t)$, respectively. These bounds are then used in the definition of the error estimator in (11) instead of the original $C_{11}(t)$ and $C_{12}(t)$. The algorithm is given as Algorithm 1 for a discrete time set T_{disc} in the setting of a time-discretized system. The computational complexity of computing the m smallest eigenvalues of $\boldsymbol{F}$ will be analyzed in Section 5.

Algorithm 1 Bounds for undamped systems ($0 \notin \sigma(\boldsymbol{K})$)

Input: $\boldsymbol{M}$, $\boldsymbol{K}$, time set T_{disc}, $1 \le m \le N$

Require: $0 \notin \sigma(\boldsymbol{K})$

1: $\{f_1, \dots, f_m\} \leftarrow m$ smallest eigenvalues of $\boldsymbol{K}\boldsymbol{v} = f\boldsymbol{M}\boldsymbol{v}$

2: **for all** $t \in T_{\text{disc}}$ **do**

3: $\quad C_{11,\text{bound}}(t) \leftarrow 1$

4: $\quad C_{12,\text{bound}}^{(m)}(t) \leftarrow \max\left\{ \frac{\left|\sin\left(\sqrt{f_1}t\right)\right|}{\sqrt{f_1}}, \frac{\left|\sin\left(\sqrt{f_2}t\right)\right|}{\sqrt{f_2}}, \dots, \frac{\left|\sin\left(\sqrt{f_{m-1}}t\right)\right|}{\sqrt{f_{m-1}}}, \min\left\{ \frac{1}{\sqrt{f_m}}, t \right\} \right\}$

5: **end for**

Output: $C_{11,\text{bound}}(t)$, $C_{12,\text{bound}}^{(m)}(t)$ for all $t \in T_{\text{disc}}$

Corollary 2 states in Equation (35) that

$$C_{12}(t) = \max_{f \in \sigma(\boldsymbol{F})} \frac{\left|\sin\left(\sqrt{f}t\right)\right|}{\sqrt{f}}.$$

Since $\frac{1}{\sqrt{f}}$ is a monotonically decreasing bound of $\frac{\left|\sin\left(\sqrt{f}t\right)\right|}{\sqrt{f}}$ w.r.t. f, we are able to modify Algorithm 1 such that $\boldsymbol{C}_{12}(t)$ can be calculated precisely with only a few eigenvalues.

Algorithm 2 $\boldsymbol{C}_{12}(t)$ and bound of $\boldsymbol{C}_{11}(t)$ for undamped systems ($0 \notin \sigma(\boldsymbol{K})$)

Input: $\boldsymbol{M}$, $\boldsymbol{K}$, time set T_{disc}
Require: $0 \notin \sigma(\boldsymbol{K})$
1: **for all** $t \in T_{\text{disc}}$ **do**
2: $\quad \boldsymbol{C}_{11,\text{bound}}(t) \leftarrow 1$
3: $\quad$ best $\leftarrow 0$
4: $\quad$ **for all** $f \in \sigma(\boldsymbol{F})$ **do** ▷ loop in increasing order
5: $\quad\quad$ **if** best $\geq \frac{1}{\sqrt{f}}$ **then** ▷ see Theorem 4
6: $\quad\quad\quad \boldsymbol{C}_{12}(t) \leftarrow$ best
7: $\quad\quad\quad$ **continue** with next t ▷ no higher value possible
8: $\quad\quad$ **end if**
9: $\quad\quad$ **if** $\frac{\left|\sin\left(\sqrt{f}t\right)\right|}{\sqrt{f}} \geq$ best **then**
10: $\quad\quad\quad$ best $\leftarrow \frac{\left|\sin\left(\sqrt{f}t\right)\right|}{\sqrt{f}}$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: $\quad \boldsymbol{C}_{12}(t) \leftarrow$ best ▷ all eigenvalues were processed
14: **end for**
Output: $\boldsymbol{C}_{11,\text{bound}}(t)$, $\boldsymbol{C}_{12}(t)$ for all $t \in T_{\text{disc}}$

The loop in line 4 of Algorithm 2 finds the maximum of $\frac{\left|\sin(\sqrt{f}t)\right|}{\sqrt{f}}$ by evaluating it for each eigenvalue of $\boldsymbol{F}$. Therefore, the algorithm finishes for sure with $\boldsymbol{C}_{12}(t)$. If the highest value so far is larger than the monotonically decreasing bound $f^{-\frac{1}{2}}$ to the right of the next eigenvalue f (see line 5), then no higher value can be found and the algorithm finishes earlier in line 7. Usually, only a few eigenvalues of $\boldsymbol{F}$ need to be computed before the precise value of $\boldsymbol{C}_{12}(t)$ is known. We will analyze this in great detail in Sections 5 and 6 for similar algorithms for the damped case.

4.2 | Damped Systems

For damped systems, i.e., $\beta > 0$ and $\alpha\beta \leq 1$, we will work in the context of Corollary 1, i.e., everything is written with respect to the spectrum of $\boldsymbol{E}$.

It will also be of great benefit to separate $\sigma(\boldsymbol{E})$ in four parts depending on the monotonicity and sign of the real function r from (30). Therefore, we start by collecting some properties of this function.

Lemma 4. The function

$$r : \mathcal{D}_r := \left[\frac{\alpha}{2}, \infty\right) \to \mathbb{R}, \quad e \mapsto e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta} = \left(e - \frac{1}{\beta}\right)^2 + \frac{\alpha}{\beta} - \frac{1}{\beta^2} \tag{38}$$

has the following properties.

(a) r has its only local and global minimum at $e_{\min r} := \frac{1}{\beta}$ with minimum value $\frac{\alpha}{\beta} - \frac{1}{\beta^2} \leq 0$ and zero only for $\alpha = \beta = 1$.

(b) r is monotonically decreasing for $e \leq e_{\min r}$ and monotonically increasing for $e \geq e_{\min r}$.

(c) r has the only zeros at

$$e_{0,\text{left}} := \frac{1}{\beta} - \frac{1}{\beta}\sqrt{1-\alpha\beta}\,, \tag{39}$$

$$e_{0,\text{right}} := \frac{1}{\beta} + \frac{1}{\beta}\sqrt{1-\alpha\beta}\,. \tag{40}$$

(d) The domain $\mathcal{D}_r$ can be divided in the four connected sets

$$\begin{aligned}
\Omega_{+,\text{left}} &:= \{e \in \mathcal{D}_r : e \le e_{0,\text{left}}\}\,,\\
\Omega_{-,\text{left}} &:= \{e \in \mathcal{D}_r : e_{0,\text{left}} \le e \le e_{\min r}\}\,,\\
\Omega_{-,\text{right}} &:= \{e \in \mathcal{D}_r : e_{\min r} \le e \le e_{0,\text{right}}\}\,,\\
\Omega_{+,\text{right}} &:= \{e \in \mathcal{D}_r : e_{0,\text{right}} \le e\}\,.
\end{aligned}$$

As can be seen in Figure 2 , these sets are chosen such that the first subscript describes the sign of $r(e)$ and the second subscript the orientation to the minimum, i.e., the monotonicity of r in this set. They may intersect at the border for easier notation. We may also use $\Omega_- := \Omega_{-,\text{left}} \cup \Omega_{-,\text{right}}$, $\Omega_{\text{left}} := \Omega_{+,\text{left}} \cup \Omega_{-,\text{left}}$, etc. in the following.

(e) $r(e) \le e^2$ and $s(e) \le e$ for $e \in \Omega_+$.

Proof. The left border of $\mathcal{D}_r$ was chosen as the lowest possible eigenvalue of $\boldsymbol{E}$. It is $\frac{\alpha}{2}$ since $\boldsymbol{F}$ is positive semi-definite (see Lemma 3) and since (14) holds. The assumption $\alpha\beta \le 1$ for the damping parameters will be used several times.

(a) Obviously, $e - \beta^{-1}$ is the smallest for $e = e_{\min r}$, which is in the domain $\mathcal{D}_r$ because $e_{\min r} = \beta^{-1} \ge \alpha \ge \frac{\alpha}{2}$. Since r is a parabola, this local minimum is also global. The minimum value $\frac{\alpha}{\beta} - \frac{1}{\beta^2} = \frac{1}{\beta}(\alpha - \beta^{-1})$ is non-positive since $\alpha - \beta^{-1} \le 0$ is fulfilled by the condition $\alpha\beta \le 1$ on the damping coefficients.

(b) This follows directly from $r'(e) = 2(e - \beta^{-1})$.

(c) The zeros can be computed with the p-q-formula for quadratic equations applied to (38). While it is obvious that $e_{0,\text{right}} \in \mathcal{D}_r$ with part (a), it still needs to be shown that $e_{0,\text{left}} \in \mathcal{D}_r$, i.e.,

$$e_{0,\text{left}} \ge \frac{\alpha}{2} \Leftrightarrow \frac{1}{\beta} - \frac{1}{\beta}\sqrt{1-\alpha\beta} \ge \frac{\alpha}{2} \Leftrightarrow \frac{1}{\beta} - \frac{\alpha}{2} \ge \frac{1}{\beta}\sqrt{1-\alpha\beta} \ge 0 \overset{(\cdot)^2}{\Leftrightarrow} \frac{1}{\beta^2} - \frac{\alpha}{\beta} + \frac{\alpha^2}{4} \ge \frac{1}{\beta^2}(1-\alpha\beta) \Leftrightarrow \frac{\alpha^2}{4} \ge 0\,.$$

The last inequality is obviously true, therefore, $e_{0,\text{left}} \in \mathcal{D}_r$.

(d) The sets are connected as intersections of intervals in $\mathbb{R}$. The rest follows from (a), (b), and (c).

(e) $r(e)$ can be rewritten to $r(e) = e^2 - \frac{2}{\beta}\left(e - \frac{\alpha}{2}\right)$. The statements follows immediately from $e \ge \frac{\alpha}{2}$. □

We will now show the following properties of $|c_{11}(\cdot;t)|, |c_{12}(\cdot;t)| : \mathcal{D}_r \to \mathbb{R}$ defined in (32) and (33), which will be formulated in detail and proved in the next section:

- On $\Omega_{+,\text{left}}$, $|c_{12}(\cdot;t)|$ is monotonically decreasing and $|c_{11}(\cdot;t)|$ has a monotonically decreasing bound, see Theorem 5.
- $|c_{11}(\cdot;t)|$ and $|c_{12}(\cdot;t)|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$, see Theorem 6.
- There exists a bound for $|c_{12}(\cdot;t)|$ on Ω_+, see Theorem 7.
- There exist monotonically decreasing bounds for $|c_{11}(\cdot;t)|$ and $|c_{12}(\cdot;t)|$ on $\Omega_{+,\text{right}}$, see Theorem 8.
- $c_{11}(\cdot;t)$ is monotonically decreasing in $\Omega_{+,\text{right}}$ under certain conditions, see Theorem 9
- $c_{11}(\cdot;t)$ is monotonically increasing in $\Omega_{+,\text{right}}$ under certain conditions, see Theorem 10.
- There exists a monotonically decreasing bound for $|c_{11}(\cdot;t)|$ and $|c_{12}(\cdot;t)|$ on Ω_-, see Theorem 11.

These properties together with the theoretical result of Corollary 1 are then leveraged in Algorithm 3 and Algorithm 4 to efficiently compute $\boldsymbol{C}_{12}(t)$ and $\boldsymbol{C}_{11}(t)$, respectively. The proofs of the properties listed above are given in the appendix.

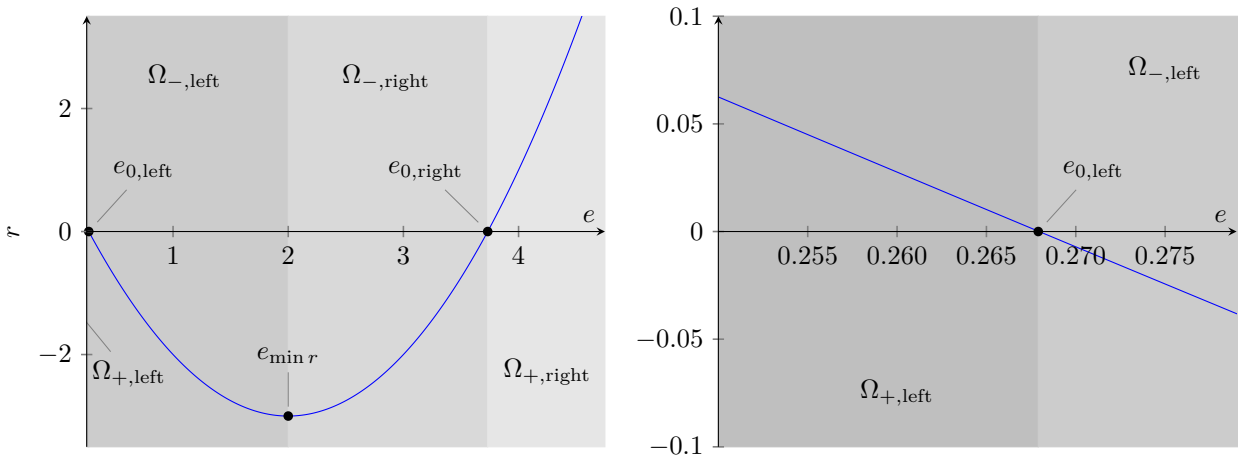

FIGURE 2 Notation from Lemma 4 for the function $r(e) = e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta}$ exemplified for $\alpha = \beta = \frac{1}{2}$. On the right, there is a zoomed view in order to see the area $\Omega_{+,\text{left}}$.

4.2.1 | Algorithms

The properties of $c_{11}(t)$ and $c_{12}(t)$ stated in the last section and proven in the appendix are now utilized to construct Algorithms 3 and 4 to compute $C_{12}(t)$ and $C_{11}(t)$ fast and memory efficiently. See Section 5 for a discussion about the computational complexity. The difference to Algorithm 1 (and partly Algorithm 2) for undamped systems is that $C_{12}(t)$ and $C_{11}(t)$ are calculated precisely for the damped system instead of using only sharp bounds. Again, we assume a time-discretized system with n_T discrete time steps T_{disc}.

The condition in line 5 of Algorithm 3 needs some explanation: Recall that $C_{12}(t)$ is defined in (29) as the maximum value of $|c_{12}(\cdot;t)|$ over the spectrum of $\boldsymbol{E}$. The function $|c_{12}(\cdot;t)|$ has its largest value in $\Omega_{+,\text{left}} \cap \sigma(\boldsymbol{E})$ at $e_{\text{min,left}}$ since $c_{12}(\cdot;t)$ is monotonically decreasing and positive in $\Omega_{+,\text{left}}$, see Theorem 5. Additionally, $|c_{12}(\cdot;t)|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$ (Theorem 6) which makes $c_{12}(e_{\text{min,left}};t)$ the largest value on $(\Omega_{+,\text{left}} \cup \Omega_-) \cap \sigma(\boldsymbol{E})$. If it is also larger than $\min\left\{b_{12,+}(t), b_{12,+,\text{right}}(e_{\text{min,right}};t)\right\}$, an upper bound on $\Omega_{+,\text{right}} \cap \sigma(\boldsymbol{E})$ according to Theorems 7 and 8, then $c_{12}(e_{\text{min,left}};t)$ is the largest value of $|c_{12}(\cdot;t)|$ on $\sigma(\boldsymbol{E})$, which is $C_{12}(t)$ by definition.

If the last condition is not satisfied to finish the algorithm right at the beginning, we have to loop over all e in the spectrum of $\boldsymbol{E}$ in increasing order in line 10. This guarantees that the algorithm finds $C_{12}(t)$ in the worst case by looping over all eigenvalues in the spectrum. But this is rarely needed due to the theory from the appendix: For each newly calculated e, we check in line 19 if any higher value for $|c_{12}(e;t)|$ is possible for the remaining eigenvalues larger than e. Since $c_{12}(\cdot;t)$ is monotonically decreasing in $\Omega_{+,\text{left}}$, no higher eigenvalue is possible from this area. We already discussed above about the highest possible value in $\Omega_{+,\text{right}}$. If $e \in \Omega_{+,\text{right}}$, then we can update $\text{bound}_{12,+,\text{right}}$ (see line 17) since $b_{12,+,\text{right}}(\cdot;t)$ is monotonically decreasing (see Theorem 8). It only remains to be discussed how large values of $|c_{12}(\cdot;t)|$ can get on Ω_- right of e. Only if $\sigma(\boldsymbol{E}) \cap \Omega_{+,\text{right}} = \emptyset$, the bound from Theorem 11 needs to be checked. Otherwise, we know from Theorem 6 that no eigenvalues from Ω_- will achieve a higher value for $|c_{12}(\cdot;t)|$ than already saved in `best`. In this case, the bound is deactivated by setting $\text{bound}_{12,-}$ to $-\infty$.

Algorithm 4 for computing $C_{11}(t)$ is quite similar to Algorithm 3. There are a few differences in the theory which lead to alterations in the algorithm: Theorem 5 does not state that $c_{11}(\cdot;t)$ is monotonically decreasing in $\Omega_{+,\text{left}}$ but only a monotonically decreasing bound is given in Theorem 5(b). Therefore, skipping the loop over $\sigma(\boldsymbol{E})$ as in line 5 of Algorithm 3 is not possible anymore. Additionally, we need to take the bound on $\Omega_{+,\text{left}}$ into account at line 31 of Algorithm 4. Theorem 10 allows to finish the algorithm early if the conditions of increasing monotonicity are matched, see line 20, or if we know that a sufficiently large area of decreasing monotonicity is to the right of e, see line 24.

Remark 7. Algorithms 3 and 4 are not given in full detail to keep them understandable. They can be optimized even further.

- Some variables may not exist, e.g., $e_{\text{min,left}}$ if $\Omega_{+,\text{left}} = \emptyset$. In these cases, the corresponding parts of the algorithm need to be skipped.

Algorithm 3 C_{12} for damped systems

Input: $\boldsymbol{M}$, $\boldsymbol{K}$, α, β, time set T_{disc}

1: $e_{\text{min,left}} \leftarrow \min(\sigma(\boldsymbol{E}) \cap \Omega_{+,\text{left}})$
2: $e_{\text{min,right}} \leftarrow \min(\sigma(\boldsymbol{E}) \cap \Omega_{+,\text{right}})$
3: **for all** $t \in T_{\text{disc}}$ **do**
4: $\quad \text{bound}_{12,+,\text{right}} \leftarrow \min\left\{b_{12,+}(t), b_{12,+,\text{right}}(e_{\text{min,right}}; t)\right\}$ ▷ see (A.7) and (A.10) from Theorems 7 and 8
5: $\quad$ **if** $c_{12}(e_{\text{min,left}}; t) \geq \text{bound}_{12,+,\text{right}}$ **then**
6: $\quad\quad C_{12}(t) \leftarrow c_{12}(e_{\text{min,left}}; t)$
7: $\quad\quad$ **continue** with next t
8: $\quad$ **end if**
9: $\quad \text{best} \leftarrow 0$
10: $\quad$ **for all** $e \in \sigma(\boldsymbol{E})$ **do** ▷ loop in increasing order
11: $\quad\quad$ **if** $e \in \Omega_{-}$ **then**
12: $\quad\quad\quad \text{bound}_{12,-} \leftarrow b_{12,-}(e; t)$ ▷ see (A.17) from Theorem 11
13: $\quad\quad$ **else**
14: $\quad\quad\quad \text{bound}_{12,-} \leftarrow -\infty$ ▷ deactivate this bound
15: $\quad\quad$ **end if**
16: $\quad\quad$ **if** $e \in \Omega_{+,\text{right}}$ **then**
17: $\quad\quad\quad \text{bound}_{12,+,\text{right}} \leftarrow \min\left\{b_{12,+}(t), b_{12,+,\text{right}}(e; t)\right\}$ ▷ update bound for $\Omega_{+,\text{right}}$ from Theorems 7 and 8
18: $\quad\quad$ **end if**
19: $\quad\quad$ **if** $\text{best} \geq \max\left\{\text{bound}_{12,-}, \text{bound}_{12,+,\text{right}}\right\}$ **then**
20: $\quad\quad\quad C_{12}(t) \leftarrow \text{best}$
21: $\quad\quad\quad$ **continue** with next t ▷ no higher value possible
22: $\quad\quad$ **end if**
23: $\quad\quad$ **if** $|c_{12}(e; t)| \geq \text{best}$ **then**
24: $\quad\quad\quad \text{best} \leftarrow |c_{12}(e; t)|$
25: $\quad\quad$ **end if**
26: $\quad$ **end for**
27: $\quad C_{12}(t) \leftarrow \text{best}$ ▷ all eigenvalues were processed
28: **end for**

Output: $C_{12}(t)$ for all $t \in T_{\text{disc}}$

- The algorithms put a lot of effort into finding the precise values for $C_{12}(t)$ and $C_{11}(t)$. If a small additional error is tolerable, then the bounds can be used to break of the algorithm early with a little bit higher value. Choose any given $\text{tol} > 0$, then lines 19 and following in Algorithm 3 can be changed to:

  if $\text{best} \cdot (1 + \text{tol}) \geq \max\left\{\text{bound}_{12,-}, \text{bound}_{12,+,\text{right}}\right\}$ **then**
  $\quad C_{12,\text{bound}}(t) \leftarrow \max\left\{\text{best}, \text{bound}_{12,-}, \text{bound}_{12,+,\text{right}}\right\}$
  $\quad$ **continue** with next t ▷ tolerance met
  end if

  This modified algorithm guarantees that $C_{12}(t) \leq C_{12,\text{bound}}(t) \leq (1 + \text{tol}) \cdot C_{12}(t)$. Similar modifications can be done for Algorithm 4 to trade accuracy for speed. The relative increase in $\Delta_{\boldsymbol{q}}(t)$ is at most as high as the tolerance since $\Delta_{\boldsymbol{q}}(t)$ in (11) depends linearly on the constants $C_{11}(t)$ and $C_{12}(t)$. An example will be given in Section 6.

- Leveraging the monotonicity in certain areas allows to skip certain eigenvalues in line 10 of Algorithm 3 and line 6 of Algorithm 4.

Algorithm 4 C_{11} for damped systems

Input: $\boldsymbol{M}$, $\boldsymbol{K}$, α, β, time set T_{disc}

$e_{\text{min,right}} \leftarrow \min(\sigma(\boldsymbol{E}) \cap \Omega_{+,\text{right}})$
$e_{\max} \leftarrow \max(\sigma(\boldsymbol{E}))$
for all $t \in T_{\text{disc}}$ **do**
  $\text{bound}_{11,+,\text{right}} \leftarrow b_{11,+,\text{right}}(e_{\text{min,right}}; t)$ ▷ see (A.11) from Theorem 8
  best $\leftarrow 0$
  for all $e \in \sigma(\boldsymbol{E})$ **do** ▷ loop in increasing order
    if $e \in \Omega_{+,\text{left}}$ **then**
      $\text{bound}_{11,+,\text{left}} \leftarrow b_{11,+,\text{left}}(e; t)$ ▷ see (A.1) from Theorem 5
    else
      $\text{bound}_{11,+,\text{left}} \leftarrow -\infty$ ▷ deactivate this bound
    end if
    if $e \in \Omega_{-}$ **then**
      $\text{bound}_{11,-} \leftarrow b_{11,-}(e; t)$ ▷ see (A.19) from Theorem 11
    else
      $\text{bound}_{11,-} \leftarrow -\infty$ ▷ deactivate this bound
    end if
    if $e \in \Omega_{+,\text{right}}$ **then**
      $\text{bound}_{11,+,\text{right}} \leftarrow b_{11,+,\text{right}}(e; t)$ ▷ update bound for $\Omega_{+,\text{right}}$ from Theorem 8
    end if
    if e in monotonically increasing area of Theorem 10 **then**
      $C_{11}(t) \leftarrow c_{11}(e_{\max}; t)$
      continue with next t ▷ no higher value possible
    end if
    if $e \in \Omega_{+,\text{right}}$ and $\frac{\partial}{\partial e} c_{11}(e; t) \leq 0$ **then**
      $[e, \bar{e}]$:= monotonically decreasing area from Remark 8
      if $\bar{e} \geq e_{\max}$ or best $\geq b_{11,+,\text{right}}(\bar{e}; t)$ **then**
        $C_{11}(t) \leftarrow$ best
        continue with next t ▷ no higher value possible
      end if
    end if
    if best $\geq \max\left\{\text{bound}_{11,+,\text{left}}, \text{bound}_{11,-}, \text{bound}_{11,+,\text{right}}\right\}$ **then**
      $C_{11}(t) \leftarrow$ best
      continue with next t ▷ no higher value possible
    end if
    if $|c_{11}(e; t)| \geq$ best **then**
      best $\leftarrow |c_{11}(e; t)|$
    end if
  end for
  $C_{11}(t) \leftarrow$ best ▷ all eigenvalues were processed
end for

Output: $C_{11}(t)$ for all $t \in T_{\text{disc}}$

5 | IMPLEMENTATION AND COMPUTATIONAL COMPLEXITY

In this section, we discuss the implementation and computational complexity of the Algorithms 1, 2, 3, and 4, which will be called *improved methods* in the following. The *direct method* as formulated in Equations (3), (10), and (12) instead constitutes in computing first $\boldsymbol{A}$, then $\boldsymbol{\Phi}(t)$ as matrix exponential, and finally $C_{11}(t)$, $C_{12}(t)$ by taking the $\boldsymbol{G}$-weighted norm of parts of $\boldsymbol{\Phi}(t)$. We focus on discussing the implementation in MATLAB by MathWorks since MATLAB is used later for the examples in Section 6.

In the improved method, we have broken down the computation of the constants to basic, one-dimensional computations over the spectrum of $\sigma(\boldsymbol{F})$ and $\sigma(\boldsymbol{E})$. Therefore, the question remains how to efficiently compute the spectrum of $\sigma(\boldsymbol{F})$ and $\sigma(\boldsymbol{E})$. Due to the relation $\boldsymbol{E} = \frac{\alpha}{2}\boldsymbol{I} + \frac{\beta}{2}\boldsymbol{F}$, the spectrum of $\boldsymbol{E}$ is easily computed from the spectrum of $\boldsymbol{F}$, i.e., $\sigma(\boldsymbol{E}) = \left\{ \frac{\alpha}{2} + \frac{\beta}{2} f \,:\, f \in \sigma(\boldsymbol{F}) \right\}$. The spectrum of $\boldsymbol{F} = \boldsymbol{M}^{-1}\boldsymbol{K}$ on the other hand is equal to all eigenvalues of the generalized eigenvalue problem

$$\boldsymbol{K}\boldsymbol{v} = f\,\boldsymbol{M}\boldsymbol{v}\,. \tag{41}$$

In all algorithms, we are only interested in computing a handful of these eigenvalues with the smallest / largest absolute value or nearest to a given value. The examples will show indeed that only a small quantity of all eigenvalues is needed. This can be accomplished by the iterative Arnoldi (for non-Hermitian matrices) or Lanczos (for Hermitian matrices) method based on Krylov subspaces, see Sorensen[18]. These algorithms are more advanced than the power iteration and faster than computing all eigenvalues, see Saad[19]. Since the matrices $\boldsymbol{K}$ and $\boldsymbol{M}$ are symmetric, the Lanczos method – or to be more precise, the Implicitly Restarted Lanczos Method (IRLM) – is discussed in the following. One famous implementation for IRLM is ARPACK (ARnoldi PACKage)[20], which heavily uses LAPACK (Linear Algebra PACKage)[21] and BLAS (Basic Linear Algebra Subprograms, see Dongarra et al.[22] and the references therein). MATLAB used ARPACK for `eigs` until version 2017a and its own implementation beginning with version 2017b. The latter is used in this work. Other software for sparse eigenvalue problems is listed in Hernández et al.[23]

ARPACK needs only $N\mathcal{O}(m) + \mathcal{O}(m^2)$ storage to compute a subset of $m \ll N$ eigenvalues. Since m is negligible compared to N, we can assume a storage complexity of $\mathcal{O}(N)$. Unfortunately, it is not possible to know in advance the steps until the iterative IRLM will converge to the desired subset of eigenvalues.[20] But the examples in Section 6 will show a generally fast computation.

Sometimes single eigenvalues like $e_{\text{min,right}} = \min(\sigma(\boldsymbol{E}) \cap \Omega_{+,\text{right}})$ in Algorithm 4 need to be computed. This is no problem with the IRLM since the subset of eigenvalues can be defined by an arbitrary condition but certain implementations like `expm` in MATLAB only allow to get the m smallest / largest eigenvalues or the m nearest eigenvalues to a specific value. By definition, $e_{\text{min,right}}$ is the one positive eigenvalue of $\boldsymbol{E}$ nearest to $e_{0,\text{right}}$. Since it is easier to work with $\boldsymbol{F}$, which prevents the introduction of additional numerical error of terms with $\boldsymbol{D}$, we are equivalently interested in the positive eigenvalue of $\boldsymbol{F}$ nearest to $f_{\text{min,right}} = \frac{2}{\beta} e_{\text{min,right}} - \frac{\alpha}{\beta}$. The positive eigenvalue nearest to $f_{\text{min,right}}$ of the problem

$$\boldsymbol{F}\boldsymbol{v} = \lambda\boldsymbol{v}$$

is equivalent to finding the largest positive eigenvalue of the problem

$$\left(\boldsymbol{F} - f_{\text{min,right}}\boldsymbol{I}\right)^{-1}\boldsymbol{v} = \mu\boldsymbol{v} \;\Leftrightarrow\; \boldsymbol{M}\boldsymbol{v} = \mu\left(\boldsymbol{K} - f_{\text{min,right}}\boldsymbol{M}\right)\boldsymbol{v} \tag{42}$$

with the transformation $\lambda = \mu^{-1} + f_{\text{min,right}}$. Here, we simply used that λ is near $f_{\text{min,right}}$ if and only if $\lambda - f_{\text{min,right}}$ is near zero if and only if $\mu := \left(\lambda - f_{\text{min,right}}\right)^{-1}$ is large. The right eigenvalue problem of (42) can be computed again with common implementations of IRLM. Similar spectral transformations can be used to compute $e_{\text{min,left}}$ and the next larger eigenvalue in $\sigma(\boldsymbol{F})$ needed in line 10 of Algorithm 3 and line 6 of Algorithm 4.

When assessing the direct computation of the constants $\boldsymbol{C}_{11}(t)$, $\boldsymbol{C}_{12}(t)$ with Equations (3), (10), and (12), we need to make a fair comparison to the IRLM:

- Computing $\boldsymbol{A}$ with Equation (3) involves computing $\boldsymbol{F} = \boldsymbol{M}^{-1}\boldsymbol{K}$. Since $\boldsymbol{M}$ and $\boldsymbol{K}$ are banded matrices coming from the finite element method, the linear solver `SGBTRS` from LAPACK for general banded matrices can be used. Its version without partial pivoting has a time complexity of $\mathcal{O}(N^2 b)$ with $b \ll N$ being the bandwidth of $\boldsymbol{M}$.[24] During the forward and back substitution stage, the algorithm does not take into account that $\boldsymbol{K}$ is also banded. Modifying the algorithm in this way should then lead to a complexity of $\mathcal{O}(Nb^2)$. In general – with a diagonal $\boldsymbol{M}$ as an exception – $\boldsymbol{F} = \boldsymbol{M}^{-1}\boldsymbol{K}$ is a dense matrix leading to a storage complexity of $\mathcal{O}(N^2)$.

- Taking the matrix exponential of the $2N \times 2N$ matrix $\boldsymbol{A}t$ in (10) is the bottle neck of the direct computation. The current implementation of `expm` in MATLAB 2017b uses the scaling and squaring algorithm based on a Padé approximation of order 13. It needs 6 matrix multiplications and involves powers of up to $\boldsymbol{A}^{13}$, see Higham[25]. Even for sparse $\boldsymbol{A}$ in the case of diagonal $\boldsymbol{M}$, this power easily gets dense, i.e., having a storage complexity of $\mathcal{O}(N^2)$: For a simple beam model with $N = 2457$ degrees of freedom, diagonal mass matrix, and banded matrices $\boldsymbol{K}$ and $\boldsymbol{F}$ with bandwidth of 89, the power $\boldsymbol{A}^{13}$ only has 44 % zero elements and $\exp(\boldsymbol{A})$ has even no zero elements at all. Therefore, the time complexity depends on the matrix multiplication, which scales with $\mathcal{O}(N^3)$.

- Computing the $\boldsymbol{G}$-weighted norm in (12) needs neither the inverse of $\boldsymbol{G} = \boldsymbol{M}$ nor the matrix square root. Instead, we proved in Proposition 1 that the largest eigenvalue of a generalized eigenvalue problem needs to be found. We already discussed that this can be accomplished relatively easily with the IRLM.

In summary, we have argued that the storage complexity of the algorithms presented in this work is only $\mathcal{O}(N)$ compared to $\mathcal{O}(N^2)$ of the direct computation. Since the computation of $C_{11}(t)$ and $C_{12}(t)$ is part of the offline step, scaling with the dimension N of the full-order system is acceptable. We have argued that the direct computation has a time complexity of $\mathcal{O}(N^3)$. While the time complexity of the iterative algorithm IRLM is unpredictable and the number m of eigenvalues computed with the IRLM was assumed to be negligible compared to N, we will find numerical evidence in Section 6 to support this claim and an overall time complexity of $\mathcal{O}(N^{1.8})$ of the Algorithms 3 and 4.

The discussion so far only involved one time step $t \in T_{\text{disc}}$. Another advantage of the improved method over the direct computation of the constants lies in the almost linear scaling with the number n_T of time steps with unchanged T_{cont}. The direct computation involves recomputing $\boldsymbol{\Phi}(t)$ in (10) and taking the $\boldsymbol{G}$-weighted norm in (12) for every t, which is only trivial for $T_{\text{disc}} \subseteq \mathbb{Z}$. Thus, the direct method has computational complexity of $\mathcal{O}(N^3 n_T)$.

Algorithm 1 only calls the IRLM once for a subset of $m \ll N$ eigenvalues. The computation of the constants for each time step are then only one-dimensional computations based on the m eigenvalues computed before. Even though this scales with the number of time steps, it can be easily parallelized and each step is very cheap to compute. Algorithms 2, 3, and 4 have the same structure: an outer loop over all time steps and an inner loop over all eigenvalues of $\boldsymbol{F}$ or $\boldsymbol{E}$. On a first look, one suspects a high computational complexity but these algorithms are designed in a smart way. The two loops are almost independent since the spectrum does not depend on the current time step. Again, only one-dimensional computations are performed – with the computation of the spectrum as an exception. But the spectrum does not need to be computed fully. Instead we use the IRLM to compute one eigenvalue after the other. These time-independent eigenvalues will be saved and used for other time steps at no additional cost and the design of the algorithms uses as much knowledge of $c_{11}(\cdot; t)$ and $c_{12}(\cdot; t)$ to abort the loop over the spectrum as early as possible. As will be seen in Section 6, only a handful of eigenvalues need to be computed at all. The design of the algorithm based on the behavior of the continuous functions $c_{11}(e; t)$ and $c_{12}(e; t)$ w.r.t. t guarantees that the same amount of eigenvalues need to be computed independently of the discretization detail of the time interval $T_{\text{cont}} = [0, t_{\text{end}}]$. For few time steps n_T, the IRLM will dominate the computation. Since a finer time discretization usually does not change the small amount of eigenvalues needed and these eigenvalues are reused in all time steps, the improved method experimentally scales like $\mathcal{O}(N^{1.8}) + \mathcal{O}(n_T)$ compared to $\mathcal{O}(N^3 n_T)$ of the direct method.

Algorithms 2, 3, and 4 are implemented together with the error estimator described in Section 2 in the software package CCMOR. They consist of one single MATLAB class, which can be plugged non-invasively into any ODE solver of MATLAB through the callback function `OutputFcn` if T_{disc} is fixed. This way, any existing model reduction based on Petrov-Galerkin reduction can easily be extended by this error estimator.

6 | EXAMPLES

This section presents numerical examples for different finite element models to support the claims of Section 5 with regards to time and storage complexity. For brevity, Algorithms 1 and 2 for the undamped case are not investigated since the damped case is more challenging. The authors like to emphasize that the error estimator itself as described in Section 2 including its online performance and overestimation is not the topic of this work. Instead we solely focus on the fast and well-scalable computation of the constants $C_{12}(t)$ and $C_{11}(t)$ in the offline step. Other properties of this error estimator are presented in the references Ruiner et al.[6] and Fehr et al.[8]

Algorithms 3 and 4 discussed in this section are designed to return the exact values of $C_{12}(t)$ and $C_{11}(t)$, respectively. A comparison to the direct computation with Equations (3), (10), and (12) will be performed whenever possible to validate the results. We refrain from giving errors since the direct computation is not a good candidate for the truth of these values. In fact, the direct computation is deemed to give a less precise result due to numerical errors of the inversion of $\boldsymbol{M}$. Instead, we will note that the results are validated when the relative difference of $C_{12}(t)$ and $C_{11}(t)$ for each t is at most 1 %. In many examples, the relative difference will be orders of magnitude smaller. Additionally, the algorithms have been validated with random matrices, which is not part of this work.

TABLE 1 Summary of the models used in Section 6 as examples.

	N	Description	$\sigma(\boldsymbol{M})$	$\sigma(\boldsymbol{K})$	Notes	Reference
Model A	114	Stabilization linkage of a car front suspension	10^{-7} to 10^{9}	10^{1} to 10^{9}	–	[26]
Model B	2 457	0.1 m × 0.5 m × 1.2 m aluminium beam	10^{-2} to 10^{-1}	10^{-1} to 10^{5}	$\boldsymbol{M}$ is diagonal	–
Model C	4 131	Two stiffly coupled beams	10^{-4} to 10^{-2}	10^{3} to 10^{11}	–	[27]
Model D	7 338	Piston rod of a crank drive from a combustion engine	10^{-6} to 10^{-3}	10^{5} to 10^{10}	Very small $\boldsymbol{C}_{12}(t)$, $\boldsymbol{C}_{11}(t)$	[28]
Model E	22 680	1 m×2 m×0.4 m aluminium block	10^{-3} to 10^{-1}	10^{4} to 10^{10}	N is adjustable	[29]
Model F	50 784	Support in the shape of a wishbone	10^{-7} to 10^{-4}	10^{1} to 10^{9}	–	[30]

TABLE 2 Single time step comparison for $t = 0.01$.

	N	Loop iter. improv.		Total time [s]		Speedup	Validated
		$\boldsymbol{C}_{11}(t)$	$\boldsymbol{C}_{12}(t)$	Improv.	Direct		
Model A	114	4	2	0.0848	0.159	1.88	Yes
Model B	2 457	14	1	1.12	62.9	56.2	Yes
Model C	4 131	1	1	1.34	3 420	2 553	Yes
Model D	7 338	1	1	1.11	–	–	–
Model E	22 680	2	1	32.9	–	–	–
Model F	50 784	1	1	24.2	–	–	–

6.1 | Models and Settings

Models A to F given in Table 1 will serve as examples. They are all clamped having a positive definite stiffness matrix. The models differ by many orders in the system dimension N and the range of the spectra of $\boldsymbol{M}$ and $\boldsymbol{K}$. All examples are given in SI units, which are omitted most of the time in the following.

If not stated otherwise, we use $\alpha = 5 \cdot 10^{-3}$ and $\beta = 10^{-3}$ as realistic damping parameters and the $n_T = 500$ time steps $T_{\text{disc}} = \{0.01, 0.02, \ldots, 5.00\}$ for time discretization. The time step $t = 0$ is not part of this discrete time set since the values for $\boldsymbol{C}_{11}(t)$ and $\boldsymbol{C}_{12}(t)$ are clear for $t = 0$, see the beginning of Section 4. The parameter $\delta = 2$ used in Algorithm 4 is chosen for Theorem 10. The weighting matrix is set to be the mass matrix, i.e., $\boldsymbol{G} = \boldsymbol{M}$.

All computations are performed on a Dell Precision T1600 personal computer with an Intel Xeon E3-1245 v1 CPU and 16 GB RAM running a Debian 9 Stretch operating system. MATLAB 2017b is used for all calculations, which has approx. 12.5 GB RAM available on the system. This number will be important for calculations that abort because of insufficient memory.

6.2 | Results

In Section 5, we argued that the improved computation with Algorithms 3 and 4 has a large advantage for many time steps T_{disc}, i.e., large n_T compared to the direct computation with Equations (3), (10), and (12). Hence, we first investigate the performance for both computational approaches for only one time step $t = 0.01$.

Table 3 shares the same structure as Table 2 . Hence, the meaning of each column is only described here once. The first and second column state the name of the example models and their system size N. The next two columns give the number of iterations the improved Algorithms 4 and 3 spend in the loops beginning in lines 6 and 10, respectively. Tables covering several time steps will give the minimum, mean, and maximum loop iterations over all time steps as triple. The total computation time of the improved method of this work and the direct method are shown in columns 5 and 6, respectively. They are followed by the

TABLE 3 Comparison for all six example models for $T_{\text{disc}} = \{0.01, 0.02, \dots, 5.00\}$.

	N	Loop iter. improv. (min / mean / max)		Total time [s]		Speedup	Validated
		$C_{11}(t)$	$C_{12}(t)$	Improv.	Direct		
Model A	114	(1, 1.80, 4)	(1, 1.71, 2)	0.919	49.9	54.3	Yes
Model B	2 457	(3, 5.51, 15)	(1, 1, 1)	1.45	40 900	28 300	Yes
Model C	4 131	(1, 1, 1)	(1, 1, 1)	1.57	–	–	–
Model D	7 338	(1, 1, 1)	(1, 1.20, 3)	1.68	–	–	–
Model E	22 680	(1, 1.00, 2)	(1, 1, 1)	34.0	–	–	–
Model F	50 784	(1, 1, 1)	(1, 1, 1)	23.7	–	–	–

speedup of the improved method over the direct method and whether the results of the improved method are validated against the direct method in the sense defined above. The three last columns do not show any values if and only if the direct method needs more memory than available on the computer described in Section 6.1 or the computation takes more than one day. All values are given for three significant digits.

Already the single time step comparison in Table 2 shows the superiority of the improved method over the direct computation. Since most of the time only one loop iteration is needed in lines 10 and 6 of Algorithms 3 and 4, only a small fraction of the spectrum needs to be computed. This leads to a vast speedup of Model B to F. The improved method shows a great performance boost in comparison to the direct method. This will be investigated later in greater detail.

Table 3 and Figure 3 show the results for the default time discretization $T_{\text{disc}} = \{0.01, 0.02, \dots, 5.00\}$ defined in Section 6.1. The direct method already becomes impractical for Model B with a computation time of over 11 hours and uncomputable for any larger model due to the memory restrictions of a standard PC. The improved method shows computation times roughly the same for these $n_T = 500$ time steps as for the single time step computations in Table 2 .

As already discussed in Section 5, the IRLM has a computational complexity which cannot be quantified easily. Hence, we will analyze the scalability with respect to the system dimension N. The aluminium block of Model E is modified to be 1 m in length, width, and height. In each of these dimensions, it is discretized with n_{el} solid elements with 8 nodes having 3 degrees of freedom each. Since one side of the aluminium block is fixed, the system dimension can be computed by $N = 3n_{\text{el}}(n_{\text{el}} + 1)^2$. The computation times for the direct and improved method can be seen in Figure 4 with n_{el} varied from 1 to 31 leading to system dimensions N between 12 and 95 232. Only the single time step $t = 0.01$ is analyzed. As already discussed above, the performance improvement will be much higher for the multi-timestep scenario.

The direct method runs out of memory on the testing computer for $n_{\text{el}} \geq 10$ ($N \geq 3\,630$) and the improved method only for $n_{\text{el}} \geq 32$ ($N \geq 104\,544$). This limit can be increased by using better hardware or using out-of-core algorithms. Both methods yield the same results for $1 \leq n_{\text{el}} \leq 5$. For $6 \leq n_{\text{el}} \leq 9$, the direct method is unable to compute a result for C_{11} but yields the same result as the improved method for C_{12}. The improved method is much faster than the direct method with a maximum speedup of 719 for $n_{\text{el}} = 9$. If the direct method would not run out of memory on the testing computer, then much higher speedups would be possible. With a time budget of 12 min, the improved method is able to compute C_{11} and C_{12} for a system with 35 times higher N (95 232 instead of 2 700).

The data points covering several orders of system dimensions allow an extrapolation shown as dashed lines in Figure 4 . The computation time of the direct method scales approximately with $1.70 \cdot 10^{-8} \cdot N^{2.94}$. This polynomial dependency with order 3 was already predicted in Section 5. The improved method instead scales with $1.21 \cdot 10^{-6} \cdot N^{1.76}$ on the testing computer, i.e., one order less than the direct method.

The improvement of one order in time complexity was only analyzed for one time step. As already discussed in Section 5, the improved method also scales very well with the number of time steps n_T for a constant time interval T_{cont}. This will now be shown numerically for Model A and C for $T_{\text{cont}} = [0, 5]$ and $T_{\text{disc}} = \{\Delta_t, 2\Delta_t, \dots, 5\}$ with varying time step size $\Delta_t \in \{10^0, 10^{-1}, 10^{-2}, \dots, 10^{-6}\}$. Figure 5 shows that the total computation time is first dominated by the IRLM, which needs more time for Model C due to the 40 times larger system dimension as can be seen in Table 2 . Beginning with around 10 000 time steps, the one-dimensional computations dominate because they have to be repeated 10 000 times. From this moment, the computation time scales linearly in the number of time steps independent of N, which confirms the time complexity of $\mathcal{O}(N^{1.8}) + \mathcal{O}(n_T)$. The one-dimensional computations can be accelerated by using parallelization. It has to be noted that the loop iterations over the spectrum (lines 6 and 10 of Algorithms 4 and 3, respectively) increases from 2 and 1 to 11 and 3 for Model A

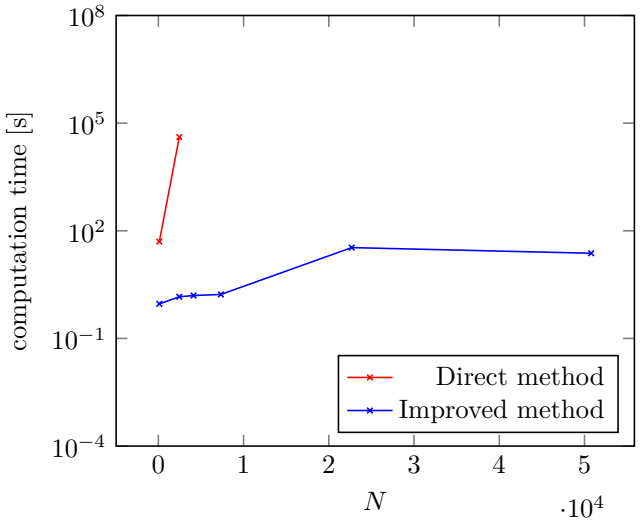

FIGURE 3 Computation times for all six models for $T_{\text{disc}} = \{0.01, 0.02, \dots, 5.00\}$. The data is taken from Table 3 . Missing data points mean that the testing computer runs out of memory before finishing the computation.

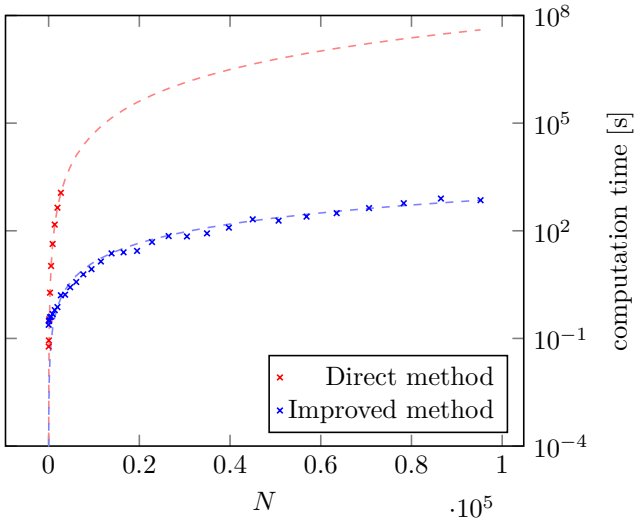

FIGURE 4 Computational results for a modified Model E with varying number of elements and $t = 0.01$. Crosses: Measured computation time. Missing data points mean that the testing computer runs out of memory before finishing the computation. Dashed lines: Estimated, polynomial dependence of the computation time from the system dimension N. Red: Direct method. Blue: Improved method.

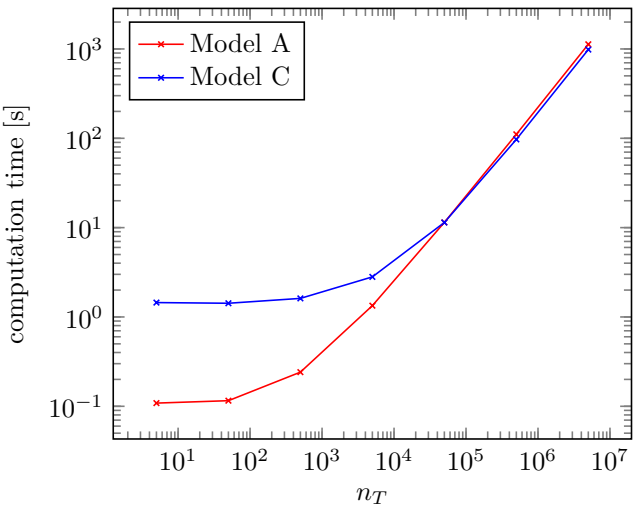

FIGURE 5 Scalability of the improved algorithms w.r.t. the number of time steps n_T for Model A (red) and Model C (blue).

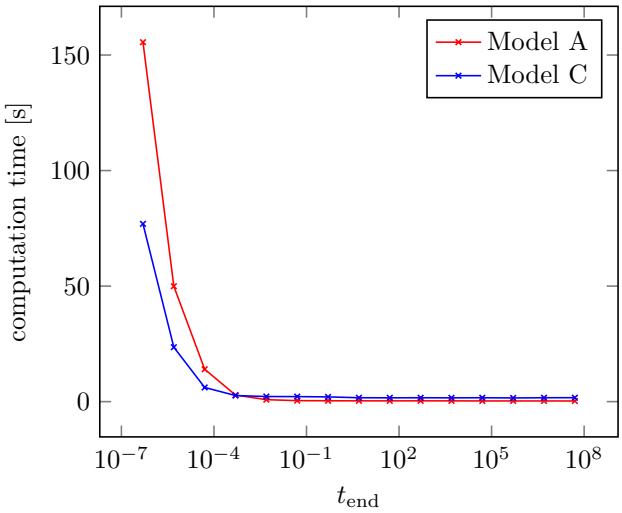

FIGURE 6 Scalability of the improved algorithms w.r.t. to t_{end} for Model A and C.

and Model C, respectively, with 10^6 times the amount of time steps. While this also means that the IRLM needs to be called 11 times more often, this number is comparable small to the increase in time steps and therefore negligible.

Now, t_{end} is varied from $5 \cdot 10^{-7}$ to $5 \cdot 10^7$ with a fixed number of time steps $n_T = 500$ for the improved algorithms. The results are depicted in Figure 6 for Model A and C as examples. For small t, an increase in computational time can be observed. This is due to a more and more smaller derivative of $c_{11}(\cdot; t)$ in absolute value for decreasing t, which makes it difficult for Algorithm 4 to leverage certain monotonicity properties. This leads to a higher amount of eigenvalues needed, here 46 (out of 114) and 21

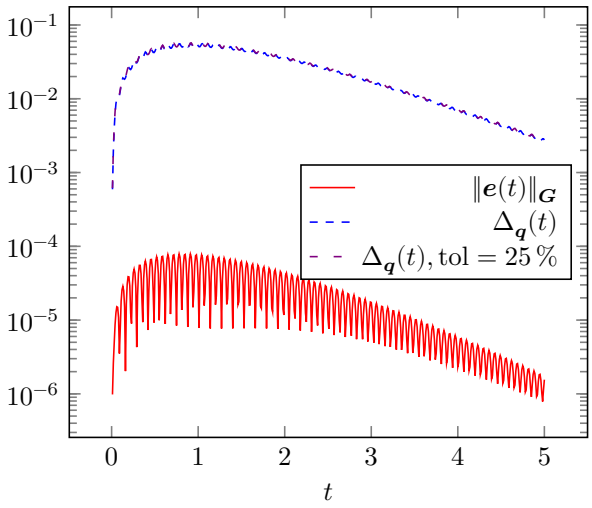

FIGURE 7 The true error $\|e(t)\|_G$ of the state in Model A is compared to the error estimator of the state $\Delta_q(t)$ without and with a tolerance of 25 % set, see Remark 7. This way, a quarter of the loop iterations for the calculation of $C_{11}(t)$ in Algorithm 4 were saved with only an 11 % increase of the error bound.

TABLE 4 Summary of the differences between direct and improved method for the damped case.

	Direct method	Improved method
Implementation	Equations (3), (10), and (12)	Algorithms 3 and 4
Time complexity	$\mathcal{O}(N^3 n_T)$	$\mathcal{O}(N^{1.8}) + \mathcal{O}(n_T)$
Storage complexity	$\mathcal{O}(N^2)$	$\mathcal{O}(N)$

(out of 4131) with $t_{\text{end}} = 5 \cdot 10^{-7}$ compared to only 1 and 1 with $t_{\text{end}} = 5$ for Model A and Model C, respectively. This effect would be even higher without the check for decreasing monotonicity in line 24 of Algorithm 4. It diminishes with an increased number of time steps for fixed t_{end} as discussed above.

In Figure 7 , we apply the error estimator to Model A with a force of $f(t) = 100 \cdot \sin(2\pi t)$ applied to node 10 of the stabilization linkage in z-direction. A second-order, Gramian matrix reduction technique described in Fehr et al.[4] is used with $n = 10$. Ten POD snapshots in the frequency range of $[0, 1500]$ are used. This example shows that the relative increase of $\Delta_q(t)$ when applied with a tolerance described in Remark 7 is at most as high as the tolerance since $\Delta_q(t)$ in (11) depends linearly on the constants $C_{11}(t)$ and $C_{12}(t)$. In this case, $\Delta_q(t)$ increased only by 11 % for a predefined tolerance of 25 %. The true error is given as a reference and shows the same time evolution but an effectivity $\Delta_q(t)/\|e(t)\|_G$ of 3 orders for this example. Effectivities of around 1 have also been observered.[8]

The results of this section are summarized in Table 4 for damped systems.

7 | CONCLUSIONS AND OUTLOOK

After recalling the error estimator of Ruiner et al.[6] and discussing its computational downsides, we were able to derive terms for the constants $C_{11}(t)$ and $C_{12}(t)$ – the bottleneck of the error estimator – only depending on the spectrum of $\boldsymbol{M}^{-1}\boldsymbol{K}$ and one-dimensional computations in Section 3. These terms were then analyzed in detail in the appendix for certain properties like monotonicity in the spectrum, which were then leveraged to develop Algorithms 1 and 2 for undamped and Algorithms 3 and 4 for (proportionally) damped systems. These algorithms try to use as few eigenvalues as possible such that only a fraction of the spectrum needs to be computed with the fast Implicitly Restarted Lanczos Method. The time and storage complexity of the algorithms for the more difficult damped case were then analyzed in Section 5 and confirmed by numerical experiments with models of largely different system dimensions in Section 6.

It has been shown that the improved method has a storage complexity of $\mathcal{O}(N)$ instead of $\mathcal{O}(N^2)$ and time complexity of $\mathcal{O}(N^{1.8}) + \mathcal{O}(n_T)$ instead of $\mathcal{O}(N^3 n_T)$ with N being the system dimension and n_T the number of time steps in the discretized setting. This constitutes a vast improvement and makes it possible for the first time to use the error estimator of Ruiner et al.[6] with non-academic models since the offline step is not a bottleneck anymore in terms of time and storage. We refrained from analyzing the quality of the error estimator and referenced other scientific contributions instead.

There are several possibilities to continue this work:

- The *hump phenomenon* described in Ruiner et al.[6] may be analyzed by looking at the power series expansion of $\|\boldsymbol{\Phi}_{21}(t)\|_{\boldsymbol{G}}$ in the same way as it was done for $\|\boldsymbol{\Phi}_{11}(t)\|_{\boldsymbol{G}}$ and $\|\boldsymbol{\Phi}_{12}(t)\|_{\boldsymbol{G}}$.
- An extension to parameter- and time-dependent system matrices may be possible. The algorithms presented in this work cannot be used efficiently without modifications since reuse of the eigenvalues is not possible anymore. If a change in parameter or time only perturbs the system matrix by a small quantity, then perturbation theory for eigenvalues may be applied, which studies the stability of eigenvalues under perturbation.
- For damped systems, we assumed proportional damping with damping parameters $\alpha, \beta \geq 0$. For technical reasons, we additionally assumed $\beta > 0$ (used, e.g., in Corollary 1) and $\alpha\beta \leq 1$ (used, e.g., in Lemma 4, Lemma 5(b), Lemma 6(a), Theorem 10, and Theorem 11), which may be dropped with only small modifications. A possible extension for a general SPSD damping matrix $\boldsymbol{D}$ is not known to the authors and deemed to be hard.
- The condition $\boldsymbol{G} = \boldsymbol{M}$ is needed for Theorem 3, which plays a crucial role in the overall proof. Maybe it is possible to prove similar results for different weighting matrices $\boldsymbol{G}$.
- The algorithms presented in this work can still be improved in terms of performance. These include the improvements listed in Remark 7, parallelization over T_{disc}, and a faster variant of the IRLM (see Hernández et al.[23] for an overview).

ACKNOWLEDGMENTS

The authors like to thank Benjamin Fröhlich, Darko Milaković, and Nadine Walker for providing the example models in Section 6 and Henrik Ebel for proofreading this work.

Author contributions

Bernard Haasdonk provided the idea for the first-order error estimator. Jörg Fehr adopted the error estimator to second-order systems as described in Section 2 and provided some of the models from Section 6. Dennis Grunert developed the speedup of the error estimator as illustrated in Sections 3 and 4. He also carried out the implementation (Section 5) as well as numerical examples (Section 6).

Financial disclosure

The authors would like to thank the German Research Foundation (DFG) for financial support FE 1583/2-1 and HA 5821/5-1 of the project at the University of Stuttgart.

Conflict of interest

The authors declare no potential conflict of interests.

SUPPORTING INFORMATION

No supporting information is available as part of the online article.

APPENDIX

The appendix covers the proofs of all properties of c_{11} and c_{12} listed in Section 4.2. Most of the following proofs will focus on c_{12} and then give a shorter proof for c_{11} when it is analogue. We start by showing that $|c_{12}(\cdot;t)|$ is monotonically decreasing and $|c_{11}(\cdot;t)|$ has a monotonically decreasing bound on $\Omega_{+,\text{left}}$. We will omit the absolute value bars for $e \in \Omega_+$ since then $r(e), s(e) \geq 0$, which yields $c_{11}(e;t), c_{12}(e;t) \geq 0$ with (32) and (33).

Theorem 5 (Monotonicity on $\Omega_{+,\text{left}}$)**.**

(a) $c_{12}(\cdot;t)$ is monotonically decreasing and positive on $\Omega_{+,\text{left}}$.

(b) $c_{11}(\cdot;t)$ is bound on $\Omega_{+,\text{left}}$ by

$$b_{11,+,\text{left}}(e;t) := \exp(-et)\cosh(s(e)t)(1+et)\,. \tag{A.1}$$

The bound $b_{11,+,\text{left}}$ is monotonically decreasing on $\Omega_{+,\text{left}}$.

Proof.

(a) $s(e) \geq 0$ for $e \in \Omega_{+,\text{left}}$ by definition of $\Omega_{+,\text{left}}$. First we note that $\frac{\sinh(x)}{x} \leq \cosh(x)$ for all $x \in \mathbb{R}$ by direct comparison of the power series

$$\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots\,,$$
$$\cosh(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots\,.$$

Dividing by zero is again possible with a limit argument, see Remark 3. From this, it follows immediately with $x = \tilde{s}t$

$$\forall \tilde{s} \geq 0 : \frac{\sinh(\tilde{s}t)}{\tilde{s}} \leq t\cosh(\tilde{s}t)\,. \tag{A.2}$$

Keeping this inequality in mind, we will now prove that $\frac{\sinh(\tilde{s}t)}{\tilde{s}}$ is monotonically decreasing in $\tilde{s} \geq 0$. For $\tilde{s} = 0$, this follows by the limit $\lim_{\tilde{s}\to 0}\frac{\sinh(\tilde{s}t)}{\tilde{s}} = t$ and Inequality (A.2) with $\cosh(\tilde{s}t) \leq 1$. For $\tilde{s} > 0$, we look at the derivative:

$$\frac{\partial}{\partial\tilde{s}}\frac{\sinh(\tilde{s}t)}{\tilde{s}} = \frac{1}{\tilde{s}}\left(t\cosh(\tilde{s}t) - \frac{\sinh(\tilde{s}t)}{\tilde{s}}\right) \overset{(A.2)}{\geq} 0$$

Since $s(e)$ is monotonically decreasing in $e \in \Omega_{+,\text{left}}$ by definition, we have $\frac{\sinh(s(e)t)}{s(e)}$ monotonically decreasing in e and positive by setting $\tilde{s} = s(e)$. The same holds for $\exp(-et)$. Therefore, $c_{12}(\cdot;t)$ is monotonically decreasing in $\Omega_{+,\text{left}}$.

(b) We use the upper bound (A.2) for bounding $c_{11}(\cdot;t)$:

$$c_{11}(e;t) = \exp(-et)\left(\cosh(s(e)t) + e\frac{\sinh(s(e)t)}{s(e)}\right) \leq \exp(-et)\cosh(s(e)t)(1+et) = b_{11,+,\text{left}}(e;t)$$

$b_{11,+,\text{left}}(e;t)$ is also monotonically decreasing in $e \in \Omega_{+,\text{left}}$: First, we note that $s(e)$ is monotonically decreasing for increasing $e \in \Omega_{+,\text{left}}$ by definition; so is $\cosh(s(e)t)$. For the remaining factor $\exp(-et)(1+et)$, we will show that its derivative is non-negative:

$$\frac{\partial}{\partial e}\exp(-et)(1+et) = -t\exp(-et)(1+et) + \exp(-et)t = -\exp(-et)et^2 \leq 0$$

for $e, t \geq 0$. Thus, $b_{11,+,\text{left}}(\cdot;t)$ is also monotonically decreasing on $\Omega_{+,\text{left}}$. □

The next theorem states that we cannot find any larger value for $|c_{11}(\cdot;t)|$ and $|c_{12}(\cdot;t)|$ on Ω_- than on $\Omega_{+,\text{left}}$.

Theorem 6 ($\Omega_{+,\text{left}}$ vs. Ω_-)**.**

(a) $|c_{12}(\cdot;t)|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$, i.e.,

$$|c_{12}(e_+;t)| \geq |c_{12}(e_-;t)| \text{ for all } e_+ \in \Omega_{+,\text{left}}, e_- \in \Omega_-\,.$$

(b) $|c_{11}(\cdot;t)|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$, i.e.,

$$|c_{11}(e_+;t)| \geq |c_{11}(e_-;t)| \text{ for all } e_+ \in \Omega_{+,\text{left}}, e_- \in \Omega_-\,.$$

Proof.

(a) We will first prove the following two bounds:

$$\forall x > 0 : \frac{|\sin(xt)|}{x} \leq t\,, \tag{A.3}$$

$$\forall x > 0 : \frac{\sinh(xt)}{x} \geq t\,. \tag{A.4}$$

The bound $\sin(xt) \leq xt$ is well-known for $xt \in [0, \pi/2]$. It obviously extends to all $x \geq 0$ and having the absolute value on the left hand side. Equation (A.3) follows directly when dividing by $x > 0$ on both sides.

The idea is similar for bound (A.4). Since $\frac{d}{dy}\sinh(y) = \cosh(y) \geq 1 = \frac{d}{dy}y$ for $y \geq 0$ and $\sinh(y)|_{y=0} = 0 = y|_{y=0}$, it follows $\sinh(y) \geq y$ for $y \geq 0$. The bound (A.4) follows with $y := xt$ and by dividing by $x > 0$ on both sides.

We now look closer at $|\sinh(s(e)t)\, s(e)^{-1}|$ as factor of $|c_{12}(e;t)|$, see (33). We need to distinguish three cases:

(i) $e \in \Omega_{+,\text{left}} \setminus \{e_{0,\text{left}}\}$: Then, $s(e) > 0$ since $r(e) > 0$ by definition of $\Omega_{+,\text{left}}$. It follows directly

$$|\sinh(s(e)t)\, s(e)^{-1}| = \sinh(s(e)t)\, s(e)^{-1} \geq t \tag{A.5}$$

according to (A.4).

(ii) $e \in \Omega_- \setminus \{e_{0,\text{left}}, e_{0,\text{right}}\}$: Then, $s(e) =: ix \in i\mathbb{R}_+$ since $r(e) < 0$. Due to $\sinh(s(e)t) = \sinh(ixt) = i\sin(xt)$, it follows from (A.3)

$$|\sinh(s(e)t)\, s(e)^{-1}| = |\sin(xt)|x^{-1} \leq t\,. \tag{A.6}$$

(iii) $e = e_{0,\text{left}}$ or $e = e_{0,\text{right}}$ and therefore $s(e) = 0$: We have seen in Remark 3 that $\lim_{s(e)\to 0}|\sinh(s(e)t)\, s(e)^{-1}| = t$.

Combining all three cases shows that $|\sinh(s(\cdot)t)\, s(\cdot)^{-1}|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$. Since $\exp(-et) > 0$ is monotonically decreasing in e and $\Omega_{+,\text{left}} \leq \Omega_-$, $|c_{12}(\cdot;t)|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$.

(b) The proof for c_{11} is very similar to the proof above for c_{12}. Again, we distinguish three cases:

(i) $e \in \Omega_{+,\text{left}} \setminus \{e_{0,\text{left}}\}$: We will again use the lower bound (A.5) and $\cosh(s(e)t) \geq 1$ for $s(e) \geq 0$:

$$|c_{11}(e;t)| = \exp(-et)\left(\cosh(s(e)t) + e\frac{\sinh(s(e)t)}{s(e)}\right) \geq \exp(-et)(1+et)$$

(ii) $e \in \Omega_- \setminus \{e_{0,\text{left}}, e_{0,\text{right}}\}$: Let us again write $s(e) =: ix \in i\mathbb{R}_+$. Using (A.6) and $|\cos(xt)| \leq 1$, we yield the upper bound

$$|c_{11}(e;t)| = \exp(-et)\left|\cos(xt) + ie\frac{\sin(xt)}{x}\right| \leq \exp(-et)\left(|\cos(xt)| + e\left|\frac{\sin(xt)}{x}\right|\right) \leq \exp(-et)(1+et)\,.$$

(iii) $e = e_{0,\text{left}}$ or $e = e_{0,\text{right}}$ and therefore $s(e) = 0$: We use $\cos(0) = 1$ and again $\lim_{s(e)\to 0}|\sinh(s(e)t)\, s(e)^{-1}| = t$ (Remark 3):

$$|c_{11}(e;t)| = \exp(-et)\,|1+et| = \exp(-et)(1+et)$$

Combining all three cases shows that $|c_{11}(e;t)|$ will not take larger values on Ω_- as on $\Omega_{+,\text{left}}$ since $\exp(-et)(1+et)$ is monotonically decreasing as shown in the proof of Theorem 5. □

On $\Omega_+ = \Omega_{+,\text{left}} \cup \Omega_{+,\text{right}}$, we can at least find a bound for $c_{12}(\cdot;t)$.

Theorem 7 (Bound on Ω_+)**.** $c_{12}(\cdot;t)$ is bound on Ω_+ by

$$b_{12,+}(t) := \frac{1-\exp(-2s_{\min,\mathrm{r}}t)}{2s_{\min,\mathrm{r}}} \tag{A.7}$$

with

$$s_{\min,\mathrm{r}} := \min_{e\in\Omega_+}\sqrt{r(e)} = \min(s(\sigma(\boldsymbol{E})) \cap \mathbb{R})$$

the lowest real value of $s(\sigma(\boldsymbol{E}))$. For $s_{\min,\mathrm{r}} = 0$, $b_{12,+}(t)$ is defined by the limit $s_{\min,\mathrm{r}} \to 0$ as usual.

Proof. Let $e \in \Omega_+$ and fix $t > 0$ arbitrarily. Note that $c_{12}(e;t) \geq 0$ since $s(e) \geq 0$. We first assume $s_{\min,\mathrm{r}} > 0$. Then,

$$\begin{aligned} c_{12}(e;t) &= \exp(-et)\sinh(s(e)t)\,s(e)^{-1} \leq \exp(-s(e)t)\sinh(s(e)t)\,s(e)^{-1} = \exp(-s(e)t)(2s(e))^{-1}\left(\exp(s(e)t) - \exp(-s(e)t)\right) \\ &= (2s(e))^{-1}\left(1 - \exp(-2s(e)t)\right) \end{aligned} \tag{A.8}$$

with Lemma 4(e) and the definition of the hyperbolic sine. The statement of the lemma follows immediately from the decreasing monotonicity of $q(x;t) := x^{-1}(1-\exp(-xt))$ in $x > 0$ for all $t > 0$. Then, we can bound (A.8) by $q(2s_{\min,\mathrm{r}};t)$ since $q(2s_{\min,\mathrm{r}};t) \geq q(2s(e);t) =$ (A.8) for every $e \in \Omega_+$ (equivalently every real $s(e)$) due to the monotonicity. For the decreasing monotonicity of $q(\cdot;t)$, we will show that its derivative

$$\frac{\partial}{\partial x} q(x;t) = \frac{t\exp(-xt)x - (1-\exp(-xt))}{x^2} = \frac{(1+xt)\exp(-xt) - 1}{x^2}$$

is non-positive, i.e., $g(x;t) := (1+xt)\exp(-xt) \leq 1$ for $x \geq 0$. We will show that $g(\cdot;t)$ defined on $\mathbb{R}$ has a global maximum at $x = 0$ with value $g(0;t) = 1$. Then, $g(x;t) \leq 1$ for all $x \geq 0$.

The first derivative $\frac{\partial}{\partial x} g(x;t) = -\exp(-xt)t^2 x$ delivers the only critical point $x_0 = 0$. For this critical point, the second derivative $\frac{\partial^2}{\partial x^2} g(x;t) = \exp(-xt)(-t^2 + t^3 x)$ is negative since $t > 0$. Therefore, $g(\cdot;t)$ has a local maximum at $x_0 = 0$ with value $g(0;t) = 1$. Since $\lim_{x\to\infty} g(x;t) = 0$ and $\lim_{x\to-\infty} g(x;t) = -\infty$, $x_0 = 0$ is also a global maximum for $g : \mathbb{R} \to \mathbb{R}$. Therefore, $g(x;t) \leq 1$ for all $x \geq 0$, $q(\cdot;t)$ is monotonically decreasing on $x > 0$, and the statement is proven for $s_{\min,\mathrm{r}} > 0$.

For $s_{\min,\mathrm{r}} = 0$, taking the limit $s_{\min,\mathrm{r}} \to 0$ in the definition of $b_{12,+}(t)$ results in

$$b_{12,+}(t) = \lim_{s_{\min,\mathrm{r}}\to 0} \frac{1-\exp(-2s_{\min,\mathrm{r}}t)}{2s_{\min,\mathrm{r}}} = \lim_{s_{\min,\mathrm{r}}\to 0} \frac{2t\exp(-2s_{\min,\mathrm{r}}t)}{2} = t$$

with L'Hôpital's rule. With the same limit argument, the decreasing monotonicity of $q(\cdot;t)$ can be extended to $[0,\infty)$. This proves the statement for $s_{\min,\mathrm{r}} = 0$. □

If we restrict ourselves to $\Omega_{+,\mathrm{right}}$ instead of Ω_+, we can even find a better bound which holds for every $e \geq \tilde{e}$ with $\tilde{e} \in \Omega_{+,\mathrm{right}}$. First, we need some preparations which are contained in a separate lemma.

Lemma 5.

(a) $c_{11}(e;t)$ can be rewritten to

$$c_{11}(e;t) = \frac{1}{2}\left[\left(1 + \frac{e}{s(e)}\right)\exp(t(s(e)-e)) + \left(1 - \frac{e}{s(e)}\right)\exp(-t(s(e)+e))\right]. \tag{A.9}$$

(b) $\frac{e}{s(e)}$ is monotonically decreasing for all $e \geq e_{0,\mathrm{right}}$.

Proof.

(a) We are using the definition of the hyperbolic sine and cosine:

$$\begin{aligned} c_{11}(e;t) &= \exp(-et)\left(\cosh(s(e)t) + \frac{e}{s(e)}\sinh(s(e)t)\right) \\ &= \exp(-et)\left(\frac{1}{2}\left(\exp(s(e)t) + \exp(-s(e)t)\right) + \frac{e}{2s(e)}\left(\exp(s(e)t) - \exp(-s(e)t)\right)\right) \\ &= \frac{1}{2}\left[\left(1 + \frac{e}{s(e)}\right)\exp(t(s(e)-e)) + \left(1 - \frac{e}{s(e)}\right)\exp(-t(s(e)+e))\right] \end{aligned}$$

(b) $\frac{e}{s(e)}$ is monotonically decreasing for all $e \geq \alpha$ since the derivative

$$\frac{d}{de}\frac{s(e)^2}{e^2} = \frac{\partial}{\partial e}\frac{e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta}}{e^2} = \frac{\left(2e - \frac{2}{\beta}\right)e^2 - \left(e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta}\right)2e}{e^4} = \frac{2}{\beta e^3}(e - \alpha)$$

is non-negative for $e \geq \alpha$. The inequality $e \geq e_{0,\mathrm{right}} \geq \beta^{-1} \geq \alpha$ holds due to Lemma 4(c), $\alpha\beta \leq 1$ and $s(e) \geq 0$. For $s(e) = 0$, the fraction $\frac{e}{s(e)}$ can be extended continuously with 1 as usual. □

With these preparations, we are now able to bound $c_{11}(\cdot;t)$ and $c_{12}(\cdot;t)$ by a monotonically decreasing bound on $\Omega_{+,\mathrm{right}}$.

Theorem 8 (Bounds on $\Omega_{+,\mathrm{right}}$)**.** Define $e_{\max} := \max\sigma(\boldsymbol{E})$.

(a) For arbitrary $\tilde{e} \in \Omega_{+,\text{right}}$, $c_{12}(\cdot;t)$ is bound on $[\tilde{e}, e_{\max}]$ by

$$b_{12,+,\text{right}}(\tilde{e};t) := \frac{1}{2s(\tilde{e})}\left(\exp\left(-\frac{t}{\beta}\left(1-\frac{\alpha}{2\tilde{e}}\right)\right) - \exp\left(-(s(e_{\max})+e_{\max})t\right)\right). \tag{A.10}$$

The bound $b_{12,+,\text{right}}(\tilde{e};t)$ is monotonically decreasing in $\tilde{e}$ and $b_{12,+,\text{right}}(e_{0,\text{right}};t) = \infty$.

(b) For arbitrary $\tilde{e} \in \Omega_{+,\text{right}}$, $c_{11}(\cdot;t)$ is bound on $[\tilde{e}, e_{\max}]$ by

$$b_{11,+,\text{right}}(\tilde{e};t) := \frac{1}{2}\left(1+\frac{\tilde{e}}{s(\tilde{e})}\right)\exp\left(-\frac{t}{\beta}\left(1-\frac{\alpha}{2\tilde{e}}\right)\right). \tag{A.11}$$

The bound $b_{11,+,\text{right}}(\tilde{e};t)$ is monotonically decreasing in $\tilde{e}$ and $b_{11,+,\text{right}}(e_{0,\text{right}};t) = \infty$.

Proof.

(a) Let $e \in [\tilde{e}, e_{\max}]$ with $\tilde{e} \in \Omega_{+,\text{right}}$. First, we observe that s is monotonically increasing on $\Omega_{+,\text{right}}$. We are first finding a bound for $s(e) - e$:

$$s(e) - e = \frac{s(e)^2 - e^2}{s(e)+e} \leq \frac{s(e)^2 - e^2}{2e} = \frac{-\frac{2}{\beta}\left(e - \frac{\alpha}{2}\right)}{2e} = -\frac{1}{\beta}\left(1-\frac{\alpha}{2e}\right) \tag{A.12}$$

where we used that $s(e) \leq e$ (Lemma 4(e)) and $s(e)^2 - e^2 \leq 0$. This bound is still negative since $e \geq \tilde{e} \geq e_{0,\text{right}} \geq \alpha/2$ (Lemma 4(c)). We now insert this bound into the definition of $c_{12}(\cdot;t)$ using the monotonicity of $s(\cdot)$ and $\exp(\cdot)$, which also proves the monotonicity of $b_{12,+,\text{right}}(\tilde{e};t)$ in $\tilde{e}$.

$$\begin{aligned} c_{12}(e;t) &= \exp(-et)\sinh(s(e)t)\,s(e)^{-1} = \frac{1}{2s(e)}\left(\exp((s(e)-e)t) - \exp(-(s(e)+e)t)\right) \\ &\leq \frac{1}{2s(e)}\left(\exp\left(-\frac{t}{\beta}\left(1-\frac{\alpha}{2e}\right)\right) - \exp(-(s(e)+e)t)\right) \\ &\leq \frac{1}{2s(\tilde{e})}\left(\exp\left(-\frac{t}{\beta}\left(1-\frac{\alpha}{2\tilde{e}}\right)\right) - \exp\left(-(s(e_{\max})+e_{\max})t\right)\right) = b_{12,+,\text{right}}(\tilde{e};t). \end{aligned}$$

(b) Let $e \in [\tilde{e}, e_{\max}]$ with $\tilde{e} \in \Omega_{+,\text{right}}$. Then, $\frac{e}{s(e)} \leq \frac{\tilde{e}}{s(\tilde{e})}$ due to Lemma 5(b). Together with Lemma 5(a), $\frac{e}{s(e)} \geq 1$, inequality (A.12), and the reasoning used in the first part, we are able to show

$$\begin{aligned} c_{11}(e;t) &= \frac{1}{2}\left[\left(1+\frac{e}{s(e)}\right)\exp((s(e)-e)t) + \left(1-\frac{e}{s(e)}\right)\exp(-(s(e)+e)t)\right] \\ &\leq \frac{1}{2}\left(1+\frac{e}{s(e)}\right)\exp((s(e)-e)t) \leq \frac{1}{2}\left(1+\frac{\tilde{e}}{s(\tilde{e})}\right)\exp\left(-\frac{t}{\beta}\left(1-\frac{\alpha}{2\tilde{e}}\right)\right) = b_{11,+,\text{right}}(\tilde{e};t). \end{aligned}$$

The monotonicity of $b_{11,+,\text{right}}(\cdot;t)$ is clear with Lemma 5(b). □

Even though we have already found a monotonically decreasing bound in Theorem 8, it will be of great benefit to know a stronger statement, the monotonicity of c_{11}, for a certain interval. The next theorem provides us with this statement.

Theorem 9 (Decreasing Monotonicity in $\Omega_{+,\text{right}}$)**.** Let $0 < \delta < 1$ and $z_1 \geq z_2$ the two largest real zeros of the polynomial

$$\begin{aligned} p_\delta(e) := (e-\alpha)\Bigg(&\left[t^2(\delta-1)\right]e^3 + \left[\left(\alpha(\delta+1) + \frac{2-4\delta}{\beta}\right)t^2 + 2\delta t\right]e^2 + \\ &\left[\left(4\frac{\delta}{\beta^2} - (3+2\delta)\frac{\alpha}{\beta} + \alpha^2\delta\right)t^2 - 4\frac{\delta}{\beta}t + \delta\right]e + \left[\left(\frac{\alpha^2}{\beta}(1-2\delta) + \alpha^3\delta\right)t^2 + 2\frac{\alpha}{\beta}\delta t - \alpha\delta\right]\Bigg). \end{aligned}$$

Then, $c_{11}(\cdot;t)$ is monotonically decreasing in $[z_2, z_1] \cap \left\{e \in \Omega_{+,\text{right}} : \tanh^2(s(e)t) \geq \delta\right\}$.

Proof. Let $e \in \Omega_{+,\text{right}}$ for the entire proof. We start by differentiating $c_{11}(e;t)$ w.r.t. e and simplifying the condition when it will be non-positive, i.e., monotonically decreasing.

$$\begin{aligned}
&\frac{\partial}{\partial e} c_{11}(e;t) = \frac{\partial}{\partial e}\left(\exp(-et)\left(\cosh(s(e)t) + \frac{e}{s(e)}\sinh(s(e)t)\right)\right)\\
&= -t\exp(-et)\left(\cosh(s(e)t) + \frac{e}{s(e)}\sinh(s(e)t)\right)\\
&\quad + \exp(-et)\left(\sinh(s(e)t)t\frac{e-\beta^{-1}}{s(e)} + \frac{s(e)^2 - e^2 + \beta^{-1}e}{s(e)^3}\sinh(s(e)t) + \frac{e}{s(e)}\cosh(s(e)t)t\frac{e-\beta^{-1}}{s(e)}\right) \le 0\\
&\Leftrightarrow t\cosh(s(e)t)\left(\frac{e^2-\beta^{-1}e}{s(e)^2} - 1\right) + \sinh(s(e)t)\left(t\frac{e-\beta^{-1}}{s(e)} + \frac{s(e)^2 - e^2 + \beta^{-1}e}{s(e)^3} - t\frac{e}{s(e)}\right) \le 0\\
&\Leftrightarrow \sinh(s(e)t)\left(\frac{t}{\beta s(e)} + \frac{e-\alpha}{\beta s(e)^3}\right) \ge \cosh(s(e)t)t\frac{e-\alpha}{\beta s(e)^2}\\
&\Leftrightarrow \tanh(s(e)t)\left(ts(e)^2 + e - \alpha\right) \ge ts(e)(e-\alpha) \qquad \text{(A.13)}
\end{aligned}$$

We used $s'(e) = \frac{e-\beta^{-1}}{s(e)}$. It is hard to solve Inequality (A.13) exactly. Instead, we restrict ourselves to all $e \in \Omega_{+,\text{right}}$ which satisfy $\tanh^2(s(e)t) \ge \delta$ for arbitrary $0 < \delta < 1$. This way, we can rewrite (A.13) to a stronger condition and therefore reduce the size of known decreasing monotonicity. The advantage is that tanh will not be involved anymore. First, we square Inequality (A.13), substitute $\tanh^2(s(e)t)$ by δ, and substitute the definition of $s(e)^2$. This then leads to the inequality

$$\delta\left(te^2 + \left(1 - \frac{2t}{\beta}\right)e + t\frac{\alpha}{\beta} - \alpha\right)^2 \ge t^2\left(e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta}\right)(e-\alpha)^2, \qquad \text{(A.14)}$$

which is stronger than (A.13) due to $\tanh^2(s(e)t) \ge \delta$. Inequality (A.14) can be rewritten to $p_\delta(e) \ge 0$ by sorting the coefficients by monomials of e and factoring out the zero α. This lengthy computation is omitted due to brevity. The polynomial p_δ has at least two real zeros since one zero is $\alpha \in \mathbb{R}$ and the remaining complex three zeros cannot be all purely imaginary. We will call the largest real zeros z_1 and z_2 with $z_1 \ge z_2$. Since $p_\delta(e) \to -\infty$ for $e \to \infty$, p_δ is non-negative in the interval $[z_2, z_1]$. Together with the condition $\tanh^2(s(e)t) \ge \delta$, which restricts e again, we have that Inequality (A.14) is satisfied at least for all $e \in [z_2, z_1] \cap \left\{e \in \Omega_{+,\text{right}} : \tanh^2(s(e)t) \ge \delta\right\}$. Since this inequality is stronger than (A.13), these e also satisfy Inquality (A.13), which is equivalent to $\frac{\partial}{\partial e} c_{11}(e;t) \le 0$, the condition for decreasing monotonicity of $c_{11}(\cdot;t)$. □

Remark 8. Even though Theorem 9 does not deliver all domains in $\Omega_{+,\text{right}}$ with decreasing monotonicity, it can be used iteratively to increase the area of known decreasing monotonicity. We start with arbitrary $\tilde{e} \in \Omega_{+,\text{right}}$ with $\frac{\partial}{\partial e} c_{11}(\tilde{e};t) \le 0$. Setting $\delta = \tanh^2(s(\tilde{e})t)$, we know that $\tanh^2(s(e)t) \ge \tanh^2(s(\tilde{e})t) = \delta$ for all $e \ge \tilde{e}$ due to the increasing monotonicity of s in $\Omega_{+,\text{right}}$. Applying Theorem 9 returns the interval $[\tilde{e}, z_1]$, in which $c_{11}(\cdot;t)$ is monotonically decreasing if $z_2 \le \tilde{e}$. We can now repeat this by setting $\tilde{e} = z_1$, defining $\delta = \tanh^2(s(\tilde{e})t)$ and applying Theorem 9 again. Since δ will now be larger as in the iteration before, another interval will be found, which can be appended to the already found interval if they overlap. This can be repeated until the interval cannot be extended anymore with help of Theorem 9. Each iteration is cheap to compute since only basic, one-dimensional computations are performed and the zeros of p_δ can be found by using the explicit solution of the cubic function since the zero α is already factored out.

For c_{11}, we will also show that $c_{11}(\cdot;t)$ is at least monotonically increasing in a certain interval under some conditions. Theorem 8 only showed the monotonicity for an upper bound. First, we need some preparations, which will be presented in a separate lemma:

Lemma 6.

(a) $s(e) \le e - \frac{1}{\beta}$ for $e \in \Omega_{+,\text{right}}$.

(b) For arbitrary $\delta > 1$, the inequality $s(e) \ge e - \frac{\delta}{\beta}$ holds for every $e \ge \max\left\{\frac{\delta^2 - \alpha\beta}{2\beta(\delta-1)}, e_{0,\text{right}}\right\}$.

(c) For arbitrary $\delta \ge 2$, the inequality $s(e) \ge e - \frac{\delta}{\beta}$ holds for every $e \in \Omega_{+,\text{right}}$.

Proof.

(a) $s(e) \geq 0$ since $e \in \Omega_{+,\text{right}}$. Additionally, $e \geq e_{0,\text{right}} \geq \frac{1}{\beta}$. Therefore, both sides of $s(e) \leq e - \frac{1}{\beta}$ are non-negative. The inequality to be shown is in fact equivalent to $\alpha\beta \leq 1$, which is a global assumption:

$$s(e) \leq e - \frac{1}{\beta} \Leftrightarrow s(e)^2 \leq \left(e - \frac{1}{\beta}\right)^2 \Leftrightarrow e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta} \leq e^2 - \frac{2}{\beta}e + \frac{1}{\beta^2} \Leftrightarrow \frac{\alpha}{\beta} \leq \frac{1}{\beta^2} \Leftrightarrow \alpha\beta \leq 1$$

(b) $s(e) \geq 0$ since $e \geq e_{0,\text{right}} \Leftrightarrow e \in \Omega_{+,\text{right}}$. For $e < \frac{\delta}{\beta}$, the inequality is fulfilled trivially since the left hand side is non-negative and the right hand side negative. From now on, we assume $e \geq \frac{\delta}{\beta}$. Then,

$$s(e) \geq e - \frac{\delta}{\beta} \Leftrightarrow s(e)^2 \geq \left(e - \frac{\delta}{\beta}\right)^2 \Leftrightarrow e^2 - \frac{2}{\beta}e + \frac{\alpha}{\beta} \geq e^2 - \frac{2\delta}{\beta}e + \frac{\delta^2}{\beta^2} \Leftrightarrow e \geq \frac{\delta^2 - \alpha\beta}{2\beta(\delta - 1)}.$$

(c) From the definition of $e_{0,\text{right}}$ in (40), we see $e_{0,\text{right}} \leq \frac{2}{\beta} \leq \frac{\delta}{\beta}$. First, we will show that the conditions from Lemma 6(b) are met for $\delta \geq 2$ and $e \geq \frac{\delta}{\beta}$:

$$\frac{\delta^2 - \alpha\beta}{2\beta(\delta - 1)} \leq \frac{\delta^2}{2\beta(\delta - 1)} = \frac{\delta}{\beta}\frac{\delta}{2(\delta - 1)} \leq \frac{\delta}{\beta}$$

Therefore, $e \geq \max\left\{\frac{\delta^2-\alpha\beta}{2\beta(\delta-1)}, e_{0,\text{right}}\right\}$ follows from $e \geq \frac{\delta}{\beta}$ and $\delta \geq 2$. For the second case $e_{0,\text{right}} \leq e < \frac{\delta}{\beta}$, the inequality is trivially fulfilled since the left hand side is non-negative and the right hand side negative. Hence, it holds for every $e \geq e_{0,\text{right}}$ and $\delta \geq 2$. □

With these preparations, we are now able to find an area in $\Omega_{+,\text{left}}$ in which $|c_{11}(\cdot; t)|$ is monotonically increasing.

Theorem 10 (Increasing Monotonicity in $\Omega_{+,\text{right}}$)**.** Let $\alpha\beta < 1$, $t > \frac{\beta}{1-\alpha\beta}$, and $\mathcal{Z}$ be the set of zeros of the polynomial

$$\left[2\beta^2 t(1 - \alpha\beta) - 2\beta^3\right] e^2 + \left[2\alpha\beta^3 + \beta^2 - 3\beta\delta t(1 - \alpha\beta)\right] e + \left[\delta^2 t(1 - \alpha\beta) - \alpha\beta^2\right]$$

in e for arbitrary $\delta \geq 2$. Then, $c_{11}(e; t)$ is monotonically increasing in e

- for $e \geq \max(\{e_{0,\text{right}}\} \cup \mathcal{Z})$ if $\mathcal{Z} \subseteq \mathbb{R}$, and
- for $e \geq e_{0,\text{right}}$ if $\mathcal{Z} \subseteq i\mathbb{R}$.

Proof. Using Lemma 5(a), we show for both summands of

$$c_{11}(e; t) = \frac{1}{2}\left[\left(1 + \frac{e}{s(e)}\right)\exp(t(s(e) - e)) + \left(1 - \frac{e}{s(e)}\right)\exp(-t(s(e) + e))\right]$$

that they are monotonically increasing under the conditions of the lemma.

We start with the second summand $\left(1 - \frac{e}{s(e)}\right)\exp(-t(s(e) + e))$. The function $\frac{e}{s(e)}$ is monotonically decreasing in e for all $e \geq e_{0,\text{right}}$ due to Lemma 5(b). Hence, $1 - \frac{e}{s(e)}$ is monotonically increasing and $1 - \frac{e}{s(e)}$ is non-positive since $s(e) \leq e$ with Lemma 4(e). Obviously, $\exp(-t(s(e) + e))$ is positive and monotonically decreasing. Both factors of the non-positive second summand $\left(1 - \frac{e}{s(e)}\right)\exp(-t(s(e) + e))$ are monotonically decreasing in absolute value. Therefore, this summand is non-positive and monotonically increasing.

The first summand takes some more work. Using $s'(e) = \frac{e-\beta^{-1}}{s(e)}$, we calculate the first derivative

$$\begin{aligned}\frac{\partial}{\partial e}\left[\left(1 + \frac{e}{s(e)}\right)\exp(t(s(e) - e))\right] &= \frac{s(e) - es'(e)}{s(e)^2}\exp(t(s(e) - e)) + \left(1 + \frac{e}{s(e)}\right)t(s'(e) - 1)\exp(t(s(e) - e) \\ &= \exp(t(s(e) - e))\left(\frac{1}{s(e)} - \frac{e(e - \beta^{-1})}{s(e)^3} + \left(1 + \frac{e}{s(e)}\right)t\left(\frac{e - \beta^{-1}}{s(e)} - 1\right)\right),\end{aligned}$$

which is non-negative if and only if it is non-negative when multiplied with $s(e)^3 \exp(-t(s(e)-e)) \geq 0$:

$$\begin{aligned}
&s(e)^2 - e\left(e-\beta^{-1}\right) + (s(e)+e)t\left(\left(e-\beta^{-1}\right)s(e) - s(e)^2\right)\\
&= s(e)^2 - e\left(e-\beta^{-1}\right) + t\left(e-\beta^{-1}\right)s(e)^2 + t\left(e-\beta^{-1}\right)es(e) - ts(e)^3 - tes(e)^2 \geq 0\\
&= s(e)^2 - e\left(e-\beta^{-1}\right) + t\left(s(e)^2 + es(e)\right)\left(\left(e-\beta^{-1}\right) - s(e)\right) \geq 0\\
&\overset{\cdot\left((e-\beta^{-1})+s(e)\right)\geq 0}{\Leftrightarrow} \left(\left(e-\beta^{-1}\right)+s(e)\right)\left(\alpha\beta^{-1} - \beta^{-1}e\right) + t(s(e)^2+es(e))\left(\left(e-\beta^{-1}\right)^2 - s(e)^2\right) \geq 0\\
&\overset{\cdot\beta}{\Leftrightarrow} \left(\left(e-\beta^{-1}\right)+s(e)\right)(\alpha-e) + s(e)t(s(e)+e)\left(\beta^{-1}-\alpha\right) \geq 0\\
&\Leftrightarrow (e+s(e))\left(\alpha - e + s(e)t\left(\beta^{-1}-\alpha\right)\right) + \beta^{-1}(e-\alpha) \geq 0\\
&\Leftrightarrow (e+s(e))s(e)t\left(\beta^{-1}-\alpha\right) + \beta^{-1}(e-\alpha) \geq (e+s(e))(e-\alpha)
\end{aligned}$$

So far, only equivalence transformations have been applied. Both sides are non-negative due to $e \geq \alpha$ and $\alpha\beta \leq 1$. We will now substitute $s(e)$ with the inequality of Lemma 6(c) with $\delta \geq 2$ on the left hand side and with the inequality of Lemma 4(e) on the right hand side, which yields a stronger statement instead of an equivalence transformation:

$$\begin{aligned}
&(e+s(e))s(e)t\left(\beta^{-1}-\alpha\right) + \beta^{-1}(e-\alpha) \geq (e+s(e))(e-\alpha) &\text{(A.15)}\\
&\Leftarrow \left(2e-\delta\beta^{-1}\right)\left(e-\delta\beta^{-1}\right)t\left(\beta^{-1}-\alpha\right) + \beta^{-1}(e-\alpha) \geq 2e(e-\alpha)\\
&\overset{\cdot\beta^3}{\Leftrightarrow} (2\beta e-\delta)(\beta e-\delta)t(1-\alpha\beta) + \beta^2(e-\alpha) - 2\beta^3 e(e-\alpha) \geq 0\\
&\Leftrightarrow \left[2\beta^2 t(1-\alpha\beta) - 2\beta^3\right]e^2 + \left[2\alpha\beta^3 + \beta^2 - 3\beta\delta t(1-\alpha\beta)\right]e + \left[\delta^2 t(1-\alpha\beta) - \alpha\beta^2\right] \geq 0 &\text{(A.16)}
\end{aligned}$$

Each solution e of (A.16) also fulfills (A.15) which is equivalent to the increasing monotonicity of the first summand of $c_{11}(\cdot;t)$. The left hand side of (A.16) is a parabola which is open to the top for $t > \frac{\beta}{1-\alpha\beta}$ since the coefficient $2\beta^2 t(1-\alpha\beta) - 2\beta^3$ of e^2 is positive in this case. For $\mathcal{Z} \subseteq i\mathbb{R}$, there exist no real zeros of the parabola. Since it is open to the top, it only takes positive values. For real zeros, the parabola takes positive values beginning from the right zero. This proves that also the first summand of c_{11} is monotonically increasing under the above conditions. □

Remark 9. The start of the interval with known monotonicity in Theorem 10 can be shifted when using Lemma 6(b) instead of Lemma 6(c) in the proof. Then, δ needs to be optimized such that $\frac{\delta^2-\alpha\beta}{2\beta(\delta-1)}$ is as low as possible under the condition that either only imaginary zeros or real zeros as small as possible exist for the parabola in Theorem 10.

Theorem 11 (Bounds on Ω_-)**.** Define

$$\begin{aligned}
\Delta &:= 1 - \alpha\beta - \frac{\beta^2}{t^2},\\
e_{\text{switch,left}} &:= \frac{1}{\beta} - \frac{1}{\beta}\sqrt{\Delta},\\
e_{\text{switch,right}} &:= \frac{1}{\beta} + \frac{1}{\beta}\sqrt{\Delta}.
\end{aligned}$$

(a) For arbitrary $\tilde{e} \in \Omega_- = [e_{0,\text{left}}, e_{0,\text{right}}]$, $|c_{12}(\cdot;t)|$ is bound on $[\tilde{e},\infty) \cap \Omega_- = [\tilde{e}, e_{0,\text{right}}]$ by

$$b_{12,-}(\tilde{e};t) := \begin{cases}
\exp(-\tilde{e}t)t & \text{if } \Delta < 0,\\
\exp(-\tilde{e}t)t & \text{if } \Delta \geq 0 \wedge e_{0,\text{left}} \leq \tilde{e} < e_{\text{switch,left}},\\
\max\left\{\frac{\exp(-\tilde{e}t)}{|s(\tilde{e})|}, \exp(-e_{\text{switch,right}}t)t\right\} & \text{if } \Delta \geq 0 \wedge e_{\text{switch,left}} \leq \tilde{e} \leq e_{\text{switch,right}},\\
\exp(-\tilde{e}t)t & \text{if } \Delta \geq 0 \wedge e_{\text{switch,right}} < \tilde{e} \leq e_{0,\text{right}}.
\end{cases} \quad \text{(A.17)}$$

The bound $b_{12,-}(\cdot;t)$ is monotonically decreasing on Ω_-.

(b) Let the real roots of the polynomial

$$t\beta e^3 - 2te^2 + (1+\alpha t)e - \alpha \quad \text{(A.18)}$$

in e be $z_1 \leq z_2 \leq z_3 \in \mathbb{R}$ in the case of three real roots, and $z_1 \in \mathbb{R}$ in the case of only one real root. We define $(e_{\text{mono,left}}, e_{\text{mono,right}}] := (-\infty, z_1]$ for one real root or if

$$\left|(-\infty, z_1] \cap [e_{\text{switch,left}}, e_{\text{switch,right}}]\right| \geq \left|[z_2, z_3] \cap [e_{\text{switch,left}}, e_{\text{switch,right}}]\right|$$

in the case of three real roots, and $[e_{\text{mono,left}}, e_{\text{mono,right}}] := [z_2, z_3]$ otherwise. Additionally,

$$\begin{aligned} e_{11,-,\text{left}} &:= \max\left\{e_{\text{switch,left}}, e_{\text{mono,left}}\right\}, \\ e_{11,-,\text{right}} &:= \min\left\{e_{\text{switch,right}}, e_{\text{mono,right}}\right\}. \end{aligned}$$

Then, for arbitrary $\tilde{e} \in \Omega_- = [e_{0,\text{left}}, e_{0,\text{right}}]$, $|c_{11}(\cdot;t)|$ is bound on $[\tilde{e}, \infty) \cap \Omega_- = [\tilde{e}, e_{0,\text{right}}]$ by

$$b_{11,-}(\tilde{e};t) := \begin{cases} \exp(-\tilde{e}t)(1+\tilde{e}t) & \text{if } \Delta < 0, \\ \exp(-\tilde{e}t)(1+\tilde{e}t) & \text{if } \Delta \geq 0 \wedge e_{0,\text{left}} \leq \tilde{e} < e_{11,-,\text{left}}, \\ \max\left\{\exp(-\tilde{e}t)\left(1+\frac{\tilde{e}}{|s(\tilde{e})|}\right), \exp(-e_{11,-,\text{right}}t)(1+e_{11,-,\text{right}}t)\right\} & \text{if } \Delta \geq 0 \wedge e_{11,-,\text{left}} \leq \tilde{e} \leq e_{11,-,\text{right}}, \\ \exp(-\tilde{e}t)(1+\tilde{e}t) & \text{if } \Delta \geq 0 \wedge e_{11,-,\text{right}} < \tilde{e} \leq e_{0,\text{right}}. \end{cases} \tag{A.19}$$

The bound $b_{11,-}(\cdot;t)$ is monotonically decreasing on Ω_-.

Proof. The time-dependence of many variables like Δ is omitted on purpose for a clearer notation. In Figure 8 , the notation of this theorem can be followed easier.

(a) Let $\tilde{e} \in \Omega_- = [e_{0,\text{left}}, e_{0,\text{right}}]$ and $\tilde{e} \leq e \leq e_{0,\text{right}}$, which results in $s(e) \in i\mathbb{R}_{\geq 0}$ since $e \in \Omega_-$. First we note that

$$\begin{aligned} |c_{12}(e;t)| = \exp(-et)\frac{|\sinh(s(e)t)|}{|s(e)|} = \exp(-et)\frac{|\sin(|s(e)|t)|}{|s(e)|} &\leq \exp(-et)t =: b_1(e;t), \\ &\leq \frac{\exp(-et)}{|s(e)|} =: b_2(e;t) \end{aligned}$$

are both bounds to $|c_{12}(\cdot;t)|$ due to $|\sin(xt)| \leq xt$ and $|\sin(xt)| \leq 1$ for $x = |s(e)| > 0$. The first bound $b_1(\cdot;t)$ is also valid for $s(e) = 0$ (see Remark 3). The second bound $b_2(\cdot;t)$ diverges to ∞ for $e \to e_{0,\text{left/right}} \Leftrightarrow s(e) \to 0$. Therefore, it is beneficial to combine both bounds by using $b_1(\cdot;t)$ near the boundaries of Ω_- and using $b_2(\cdot;t)$ otherwise. They both intersect when $t = |s(e)|^{-1} \Leftrightarrow t^{-2} = |s(e)|^2 = -e^2 + 2/\beta \cdot e - \alpha/\beta$ is satisfied. This quadratic equation has the solutions

$$\begin{aligned} e_{\text{switch,left}} &= \frac{1}{\beta} - \frac{1}{\beta}\sqrt{1-\alpha\beta-\frac{\beta^2}{t^2}}, \\ e_{\text{switch,right}} &= \frac{1}{\beta} + \frac{1}{\beta}\sqrt{1-\alpha\beta-\frac{\beta^2}{t^2}} \end{aligned}$$

for $\Delta = 1 - \alpha\beta - \frac{\beta^2}{t^2} \geq 0$ and no real solutions otherwise. In the latter case, we only use b_1 as upper bound, i.e., $b_{12,-}(e;t) := b_1(e;t)$ since $b_2(e;t)$ will be larger. Then, $b_{12,-}(e;t)$ is monotonically decreasing in e. Therefore, we use the lowest possible value for e, which is $\tilde{e}$, to achieve a bound which is independent of e:

$$|c_{12}(e;t)| \leq b_{12,-}(e;t) \leq b_{12,-}(\tilde{e};t)$$

for all $\tilde{e} \leq e \leq e_{0,\text{right}}$. This constitutes the first case in (A.17).

Let us now assume $\Delta \geq 0$. Then, $e_{\text{switch,left/right}} \in \Omega_-$ by direct comparison with (39) and (40). One could now define $b_{12,-}(\cdot;t)$ as $b_1(\cdot;t)$ on $[e_{0,\text{left}}, e_{\text{switch,left}}] \cup [e_{\text{switch,right}}, e_{0,\text{right}}]$ and as $b_2(\cdot;t)$ on $[e_{\text{switch,left}}, e_{\text{switch,right}}]$. But then $b_{12,-}(\cdot;t)$ would still depend on e, which will be a disadvantage. Instead, we seek a monotonically decreasing bound, which allows us to bound $c_{12}(e;t)$ independently of e.

The goal is to find an interval $[e_{\text{mono,left}}, e_{\text{mono,right}}]$ in which $b_2(\cdot;t)$ is monotonically decreasing. By taking the square and inverting, we find that $b_2(e;t) = \frac{\exp(-et)}{|s(e)|}$ is monotonically decreasing if and only if $-\exp(2et)s(e)^2$ is monotonically increasing. Its derivative is

$$\frac{\partial}{\partial e}\left(-\exp(2et)s(e)^2\right) = -2t\exp(2et)s(e)^2 - 2\exp(2et)s'(e)s(e) = -2\exp(2et)\left(ts(e)^2 + e - \beta^{-1}\right)$$

where we used $s'(e) = \frac{e-\beta^{-1}}{s(e)}$. This derivative is non-negative if and only if $ts(e)^2 + e - \beta^{-1}$ is non-positive:

$$\begin{aligned} ts(e)^2 + e - \beta^{-1} = te^2 - t\frac{2}{\beta}e + t\frac{\alpha}{\beta} + e - \frac{1}{\beta} &\leq 0 \\ \Leftrightarrow\ e^2 + \left(\frac{1}{t} - \frac{2}{\beta}\right)e - \frac{1}{\beta t} + \frac{\alpha}{\beta} &\leq 0. \end{aligned} \tag{A.20}$$

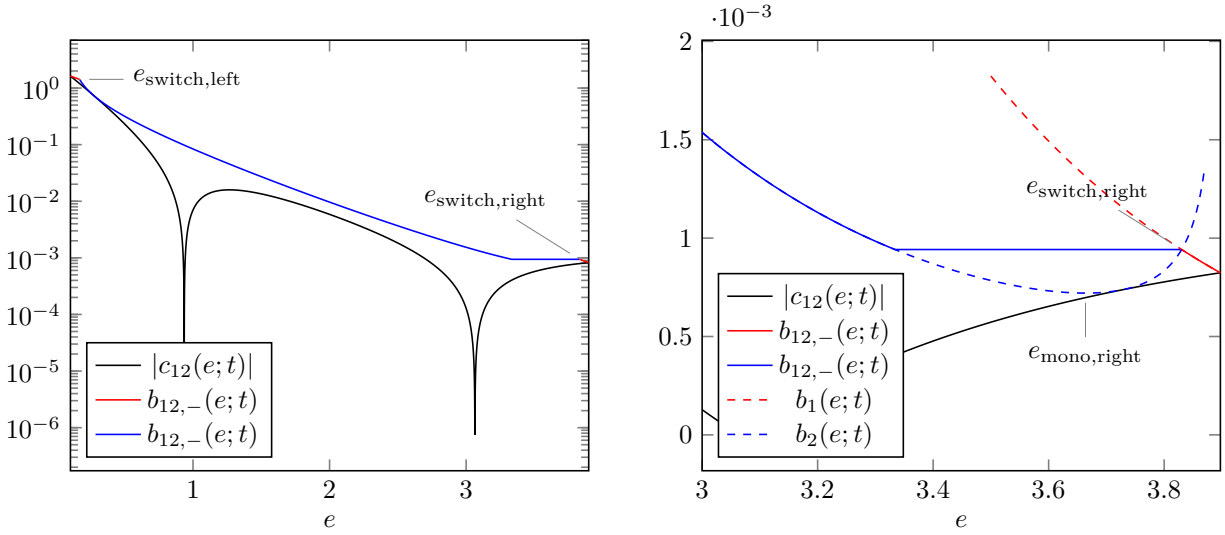

FIGURE 8 Left: Logarithmic plot of $|c_{12}(e;t)|$ (black) and the bound $b_{12,-}(e;t)$ (red for second and fourth case in (A.17) and blue for the third case in (A.17)) in the domain Ω_- for $\alpha = 0.2$, $\beta = 0.5$, and $t = 2$ (which results in $\Delta \geq 0$) to illustrate the proof of Theorem 11. Right: Zoomed view to illustrate the definition of $e_{\text{switch,right}}$ and $e_{\text{mono,right}}$.

This quadratic term has zeros at

$$e_{\text{mono,left}} := \frac{1}{\beta} - \frac{1}{\beta}\sqrt{1 - \alpha\beta + \frac{\beta^2}{4t^2}} - \frac{1}{2t},$$
$$e_{\text{mono,right}} := \frac{1}{\beta} + \frac{1}{\beta}\sqrt{1 - \alpha\beta + \frac{\beta^2}{4t^2}} - \frac{1}{2t}.$$

Both zeros are real due to $\alpha\beta \leq 1$. Since the coefficient in front of e^2 in (A.20) has a positive sign, $b_2(\cdot;t)$ is monotonically decreasing in $[e_{\text{mono,left}}, e_{\text{mono,right}}]$. This interval does not necessarily need to be a subset of Ω_-. By comparison with (39), we see that $e_{\text{mono,left}} \leq e_{0,\text{left}} \leq e_{\text{switch,left}}$.

The idea outlined in Figure 8 is as follows: We know the monotonicity of $b_2(\cdot;t)$ only up to $e_{\text{mono,right}}$. In the end, we want to reach the point $e_{\text{switch,right}}$ with a bound which is as low as possible but monotonically decreasing at the same time. This leads to the following definition:

$$b_{12,-}(e;t) := \begin{cases} b_1(e;t) & \text{if } e_{0,\text{left}} \leq e < e_{\text{switch,left}}, \\ \max\{b_2(e;t), b_2(e_{\text{switch,right}};t)\} & \text{if } e_{\text{switch,left}} \leq e \leq e_{\text{switch,right}}, \\ b_1(e;t) & \text{if } e_{\text{switch,right}} < e \leq e_{0,\text{right}}. \end{cases}$$

$b_{12,-}(\cdot;t)$ is monotonically decreasing in each of these three areas. It only remains to be checked that there are no positive jumps at $e_{\text{switch,left}}$ and $e_{\text{switch,right}}$. For $e = e_{\text{switch,left}}$, we have $\max\{b_2(e;t), b_2(e_{\text{switch,right}};t)\} = b_2(e_{\text{switch,left}};t) = b_1(e_{\text{switch,left}};t)$ since $b_2(e;t) = b_2(e_{\text{switch,left}};t) = b_1(e_{\text{switch,left}};t) \geq b_1(e_{\text{switch,right}};t) = b_2(e_{\text{switch,right}};t)$, i.e., no jump. For the other side $e = e_{\text{switch,right}}$, $\max\{b_2(e;t), b_2(e_{\text{switch,right}};t)\} \geq b_2(e_{\text{switch,right}};t) = b_1(e_{\text{switch,right}};t)$ holds by definition, i.e., no positive jump.

In summary, we know $|c_{12}(e;t)| \leq b_{12,-}(e;t)$ for $e \in \Omega_-$ and that $b_{12,-}(\cdot;t)$ is monotonically decreasing in Ω_-. Due to this monotonicity, we have $b_{12,-}(e;t) \leq b_{12,-}(\tilde{e};t)$ and therefore $|c_{12}(e;t)| \leq b_{12,-}(\tilde{e};t)$ for any $\tilde{e} \in \Omega_-$ and $\tilde{e} \leq e \leq e_{0,\text{right}}$, which proves the statement.

(b) The proof for c_{11} is similar to the proof of c_{12}. Therefore, we will only sketch the proof and note differences. We will redefine most of the variables used in the proof above if not already done in the formulation of the lemma (like $e_{\text{mono,left}}$).

Again, we have two bounds available:

$$\begin{aligned}|c_{11}(e;t)| &\leq \exp(-et)\left(|\cosh(s(e)t)| + \frac{e}{|s(e)|}|\sinh(s(e)t)|\right) = \exp(-et)\left(|\cos(|s(e)|t)| + e\frac{|\sin(|s(e)|t)|}{|s(e)|}\right)\\ &\leq \exp(-et)(1+et) =: b_1(e;t)\,,\\ &\leq \exp(-et)\left(1+\frac{e}{|s(e)|}\right) =: b_2(e;t)\,.\end{aligned}$$

The bound $b_1(\cdot;t)$ is monotonically decreasing by looking at its derivative. The meaning of $e_{\text{switch,left}}$ and $e_{\text{switch,right}}$ stays the same which already proves the first case in (A.19).

Lets now assume $\Delta \geq 0$. In order to find intervals with decreasing monotonicity for $b_2(\cdot;t)$, we look at the derivative of $\exp(-2et)\frac{e^2}{|s(e)|^2}$ as square of the second summand of $b_2(e;t)$:

$$\begin{aligned}&\frac{\partial}{\partial e}\left(\exp(-2et)\frac{e^2}{|s(e)|^2}\right) = -2t\exp(-2et)\frac{e^2}{|s(e)|^2} + \exp(-2et)\left(\frac{2e|s(e)|^2 - e^2(-2e+2\beta^{-1})}{|s(e)|^4}\right) \leq 0\\ \overset{\cdot|s(e)|^4}{\Leftrightarrow}\ &\exp(-2et)\left(\left(-2t|s(e)|^2 - \frac{2}{\beta}\right)e^2 + 2e^3 + 2e|s(e)|^2\right) \leq 0\\ \overset{:(2e\exp(-2et))}{\Leftrightarrow}\ &-t\left(-e^2+\frac{2}{\beta}e-\frac{\alpha}{\beta}\right)e - \frac{1}{\beta}e + e^2 + \left(-e^2+\frac{2}{\beta}e-\frac{\alpha}{\beta}\right) \leq 0\\ \overset{\cdot\beta}{\Leftrightarrow}\ &t\beta e^3 - 2te^2 + (1+\alpha t)e - \alpha \leq 0\end{aligned}$$

Since $\exp(-et)\frac{e}{|s(e)|}$ is positive, we have shown that $b_2(\cdot;t)$ is monotonically decreasing if the polynomial (A.18) is non-positive. If this polynomial only has one real root z_1, then it is non-positive in the interval $(-\infty, z_1]$ since the limit for $e \to -\infty$ is $-\infty$ and $+\infty$ for $e \to +\infty$. Due to the same limit argument, it is non-positive in the area $(-\infty, z_1] \cup [z_2, z_3]$ in the case of three real zeros $z_1 \leq z_2 \leq z_3$. For simplicity, we only take the interval with the largest intersection with $[e_{\text{switch,left}}, e_{\text{switch,right}}]$. This explains the definition of $[e_{\text{mono,left}}, e_{\text{mono,right}}]$.

This time, we do not know whether $e_{\text{mono,left}} \leq e_{\text{switch,left}}$. Therefore, we define

$$\begin{aligned}e_{11,-,\text{left}} &:= \max\left\{e_{\text{switch,left}}, e_{\text{mono,left}}\right\}\,,\\ e_{11,-,\text{right}} &:= \min\left\{e_{\text{switch,right}}, e_{\text{mono,right}}\right\}\,.\end{aligned}$$

Now, $e_{11,-,\text{left}} \geq e_{0,\text{left}}$ and $e_{11,-,\text{right}} \leq e_{0,\text{right}}$. It is still possible that $e_{11,-,\text{left}} \geq e_{0,\text{right}}$ or $e_{11,-,\text{right}} \leq e_{0,\text{left}}$ but theses cases are covered in the definition of $b_{11,-}$ by the second and fourth case in (A.19), respectively.

Instead of bounding $|c_{11}(\cdot;t)|$ by $b_2(\cdot;t)$ on $[e_{\text{switch,left}}, e_{\text{switch,right}}]$, we only take this bound on $[e_{11,-,\text{left}}, e_{11,-,\text{right}}] \subseteq [e_{\text{mono,left}}, e_{\text{mono,right}}]$ since monotonicity is guaranteed on this (maybe empty) interval. For the remaining areas $[e_{0,\text{left}}, e_{11,-,\text{left}}] \cup [e_{11,-,\text{right}}, e_{0,\text{right}}]$, we take the bound b_1. To summarize:

$$\hat{b}_{11,-}(e;t) := \begin{cases} b_1(e;t) & \text{if } e_{0,\text{left}} \leq e < e_{11,-,\text{left}}\,,\\ b_2(e;t) & \text{if } e_{11,-,\text{left}} \leq e \leq e_{11,-,\text{right}}\,,\\ b_1(e;t) & \text{if } e_{11,-,\text{right}} < e \leq e_{0,\text{right}}\,.\end{cases}$$

While $\hat{b}_{11,-}(\cdot;t)$ is monotonically decreasing restricted to any of the three intervals, it is not monotonically decreasing on Ω_- since there may be a positive jump at $e_{11,-,\text{right}}$. Instead we define

$$b_{11,-}(e;t) := \begin{cases} b_1(e;t) & \text{if } e_{0,\text{left}} \leq e < e_{11,-,\text{left}}\,,\\ \max\{b_2(e;t), b_1(e_{11,-,\text{right}};t)\} & \text{if } e_{11,-,\text{left}} \leq e \leq e_{11,-,\text{right}}\,,\\ b_1(e;t) & \text{if } e_{11,-,\text{right}} < e \leq e_{0,\text{right}}\,.\end{cases}$$

Obviously, $|c_{11}(\cdot;t)| \leq b_{11,-}(\cdot;t)$ on Ω_-. The bound $b_1(\cdot;t)$ restricted to $[e_{11,-,\text{left}}, e_{11,-,\text{right}}]$ is monotonically decreasing since $b_2(\cdot;t)$ is monotonically decreasing in this interval (see above) and $b_1(e_{11,-,\text{right}};t)$ is constant. Since $b_1(\cdot;t)$ is also monotonically decreasing, we only need to look at the boundaries of the intervals for the regular case $e_{0,\text{left}} \leq e_{11,-,\text{left}} \leq e_{11,-,\text{right}} \leq e_{0,\text{right}}$.

- Let $\hat{e} = e_{11,-,\text{left}}$. Then $b_1(\hat{e};t) \geq b_2(\hat{e};t)$ due to $\hat{e} \in [e_{\text{switch,left}}, e_{\text{switch,right}}]$ and the definition of $e_{\text{switch,left/right}}$. Additionally, $b_1(\hat{e};t) \geq b_1(e_{11,-,\text{right}};t)$ since $b_1(\cdot;t)$ is monotonically decreasing. Combined, we have $b_1(\hat{e};t) \geq \max\{b_2(\hat{e};t), b_1(e_{11,-,\text{right}};t)\}$. Thus $b_{11,-}(\cdot;t)$ has no positive jump at $e_{11,-,\text{left}}$.
- Let $\hat{e} = e_{11,-,\text{right}}$. Then

$$\max\left\{b_2(\hat{e};t), b_1(e_{11,-,\text{right}};t)\right\} \geq b_1(e_{11,-,\text{right}};t) = b_1(\hat{e};t)$$

by definition of $\hat{b}_{11,-}(e;t)$. Thus, $\hat{b}_{11,-}(\cdot;t)$ has no positive jump at $e_{11,-,\text{right}}$ and the decreasing monotonicity continues over this boundary.

In summary, we know $|c_{11}(e;t)| \leq b_{11,-}(e;t)$ for $e \in \Omega_-$ and that $b_{11,-}(\cdot;t)$ is monotonically decreasing in Ω_-. Due to this monotonicity, we have $b_{11,-}(e;t) \leq b_{11,-}(\tilde{e};t)$ and therefore $|c_{11}(e;t)| \leq b_{11,-}(\tilde{e};t)$ for any $\tilde{e} \in \Omega_-$ and $\tilde{e} \leq e \leq e_{0,\text{right}}$, which proves the statement. □

References

1. Schilders W, Vorst H, Rommes J, eds. *Model Order Reduction: Theory, Research Aspects and Applications*. Mathematics in Industry, vol. 13: Berlin: Springer; 2008.
2. Benner P, Gugercin S, Willcox K. A Survey of Projection-Based Model Reduction Methods for Parametric Dynamical Systems. *SIAM Rev.* 2015;57(4):483–531.
3. Fehr J. *Automated and Error-Controlled Model Reduction in Elastic Multibody Systems*. Dissertation, Schriften aus dem Institut für Technische und Numerische Mechanik der Universität Stuttgart, Vol. 21. Aachen: Shaker Verlag; 2011.
4. Fehr J, Fischer M, Haasdonk B, Eberhard P. Greedy-based Approximation of Frequency-weighted Gramian Matrices for Model Reduction in Multibody Dynamics. *Zeitschrift für angewandte Mathematik und Mechanik.* 2012;93(8):501–519.
5. Grimme EJ. Krylov Projection Methods for Model Reduction. PhD thesis. University of Illinois at Urbana-Champaign; 1997.
6. Ruiner T, Fehr J, Haasdonk B, Eberhard P. A-posteriori Error Estimation for Second Order Mechanical Systems. *Acta Mechanica Sinica.* 2012;28(3):854–862.
7. Haasdonk B, Ohlberger M. Efficient Reduced Models and A-Posteriori Error Estimation for Parametrized Dynamical Systems by Offline/Online Decomposition. *Mathematical and Computer Modelling of Dynamical Systems.* 2011;17(2):145–161.
8. Fehr J, Grunert D, Bhatt A, Haasdonk B. A Sensitivity Study of Error Estimation in Reduced Elastic Multibody Systems. In: Proceedings of the 9th Vienna International Conference on Mathematical Modelling. 202–207; 2018.
9. Bonvin D, Mellichamp DA. A Unified Derivation and Critical Review of Modal Approaches to Model Reduction. *International Journal of Control.* 1982;35:829-848.
10. Freund RW. Model Reduction Methods Based on Krylov Subspaces. *Acta Numerica.* 2003;12:267–319.
11. Moore BC. Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction. *IEEE Transactions on Automatic Control.* 1981;26(1):17–32.
12. Craig R. Coupling of Substructures for Dynamic Analyses: An Overview. In: Proceedings of the AIAA Dynamics Specialists Conference; 2000; Atlanta.
13. Holzwarth P, Eberhard P. SVD-Based Improvements for Component Mode Synthesis in Elastic Multibody Systems. *European Journal of Mechanics A/Solids.* 2015;49:408–418.
14. Pinnau R. Model Reduction via Proper Orthogonal Decomposition. In: Schilders WHA, Vorst HA, Rommes J, eds. *Model Order Reduction: Theory, Research Aspects and Applications*, Mathematics in Industry, vol. 13: Berlin: Springer 2008 (pp. 95–109).

15. Horn AR, Johnson CR. *Topics in Matrix Analysis*. Cambridge Univ. Press; 2010.

16. Clough RW, Penzien J. *Dynamics of Structures*. Berkeley: Computers and Structures; 2nd, rev. ed. 2010.

17. Reed M, Simon B. *Functional Analysis*. New York: Acad. Pr.; rev. and enl. ed., 8. ed. 1995.

18. Sorensen DC. *Implicitly Restarted Arnoldi/Lanczos Methods for Large Scale Eigenvalue Calculations.* Institute for Computer Applications in Science and Engineering (ICASE); 1996.

19. Saad J. *Iterative Methods for Sparse Linear Systems*. Philadelphia: SIAM; 2nd ed. 2003.

20. Lehoucq RB, Sorensen DC, Yang C. *ARPACK Users' Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. Philadelphia: SIAM; 1998.

21. Anderson E, Bai Z, Bischof CH, et al. *LAPACK Users' Guide*. Philadelphia: Society for Industrial and Applied Mathematics; 3rd ed. 1999.

22. Dongarra JJ, Du Croz J, Hammarling S, Duff IS. A Set of Level 3 Basic Linear Algebra Subprograms. *ACM Transactions on Mathematical Software (TOMS).* 1990;16:1–17.

23. Hernández V, Román JE, Tomás A, Vidal V. *A Survey of Software for Sparse Eigenvalue Problems.* Universitat Politècnica de València; 2009. SLEPc Technical Report STR-6.

24. Walker DW, Aldcroft T, Cisneros A, Fox GC, Furmanski W. LU Decomposition of Banded Matrices and the Solution of Linear Systems on Hypercubes. In: C3P, vol. 2: 1635–1655. ACM; 1988; New York.

25. Higham NJ. The Scaling and Squaring Method for the Matrix Exponential Revisited. *SIMA Review.* 2009;51(4):747–764.

26. Wallrapp O, Wiedemann S. Comparison of Results in Flexible Multibody Dynamics Using Various Approaches. *Nonlinear Dynamics.* 2003;34:189–206.

27. Holzwarth P. *Modellordnungsreduktion für substrukturierte mechanische Systeme (in German).* No. 51 in Dissertation, Schriften aus dem Institut für Technische und Numerische Mechanik der Universität Stuttgart, Aachen: Shaker Verlag; 2017.

28. Milaković D. *Leistungsevaluation eines Fehlerschätzers an komplexen Modellen in Neweul-M^2 (in German).* Student Thesis STUD-480: Institute of Engineering and Computational Mechanics, University of Stuttgart; 2018.

29. Fröhlich B, Gade J, Geiger F, Bischoff M, Eberhard P. Geometric Element Parameterization and Parametric Model Order Reduction in Finite Element Based Shape Optimization. *Computational Mechanics.* 2018.

30. Fröhlich B, Eberhard P. Shape Finding in Structural Optimization with Parametrically Reduced Finite Element Models. ScienceOpen Posters; 2018. doi: 10.14293/P2199-8442.1.SOP-MATH.DPYRVU.v1.

AUTHOR BIOGRAPHY

Dennis Grunert. After receiving the diploma in mathematics at the University of Stuttgart in 2014, Dennis Grunert started to pursue a PhD degree in mechanics at the Institute of Engineering and Computational Mechanics at the University of Stuttgart. His thesis will be about error-controlled, non-linear model order reduction for car crash simulations. During a 3-month research stay at the Sandia National Laboratories in Livermore, California, he worked on machine learning methods for non-linear model reduction. Dennis Grunert is also part of the Cluster of Excellence SimTech.

Jörg Fehr. Jun.-Prof. Dr.-Ing. Jörg Fehr studied Mechatronics in Stuttgart and Mechanical Engineering in Madison, Wisconsin, USA. In 2006 he finished his diploma as well as master studies. From 2006 till 2011 he was a research assistant of Prof. Dr.-Ing. Peter Eberhard at the University of Stuttgart. His research focus was the automated model reduction for elastic multibody systems, and he developed the model reduction program *Model Reduction of Elastic Multibody Systems* (Morembs) – one of the first non-modal reduction programs for elastic bodies. Furthermore, he was one of the first Ph.D. students of the Cluster of Excellence *SimTech*. Afterwards, as a simulation engineer at TRW Automotive GmbH, he was in charge of the development of new mechatronic vehicle safety systems. Since 2014 Jörg Fehr is Junior Professor at the Institute for Engineering and Computational Mechanics and Cluster of Excellence Simulation Technology (SRC SimTech) of the University of Stuttgart and since October 2015 he is a dean of the B. Sc. and M. Sc. study program of Mechatronics. One goal of his current research is the development of optimal human body models for the simulation of vehicle safety systems and the speedup of the simulations using linear and nonlinear model reduction methods.

Bernard Haasdonk. B. Haasdonk studied Physics, Mathematics and Computer Science at the University of Freiburg and obtained his PhD in Machine Learning in 2005. One particular focus of his work represents kernel methods and kernel design. He extended his focus to the field of model reduction of numerical simulation methods and joined the Applied Mathematics Institute at the University of Freiburg as a Postdoc, spent some months at the Massachusetts Institute of Technology and moved to the University of Münster in 2007. In 2009, he has joined the *Excellence Cluster in Simulation Technology SimTech* at the University of Stuttgart as Juniorprofessor. In 2014 B. Haasdonk was appointed a professorship on *Numerical Mathematics* at the Institute of Applied Analysis and Numerical Simulation of the University of Stuttgart. From 2014–2018 he served as a German representative in the Management Committee of the *European Model Reduction Network* funded by the European Union. He is a member of the IRTG *Soft-Tissue Robotics* and Cluster of Excellence *Data-integrated Simulation Sciences*.