On port-Hamiltonian modeling and structure-preserving model reduction

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)

Dem Fachbereich IV der Universität Trier
vorgelegt von

Björn Liljegren-Sailer

Trier, im August 2020

Berichterstatter:	Prof. Dr. Nicole Marheineke (Betreuerin)
	Prof. Dr. Volker Mehrmann
Verteidigung am:	14.07.2020

Abstract

In this thesis we study structure-preserving model reduction methods for the efficient and reliable approximation of dynamical systems. A major focus is the approximation of a nonlinear flow problem on networks, which can, e.g., be used to describe gas network systems. Our proposed approximation framework guarantees so-called port-Hamiltonian structure and is general enough to be realizable by projection-based model order reduction combined with complexity reduction. We divide the discussion of the flow problem into two parts, one concerned with the linear damped wave equation and the other one with the general nonlinear flow problem on networks.

The study around the linear damped wave equation relies on a Galerkin framework, which allows for convenient network generalizations. Notable contributions of this part are the profound analysis of the algebraic setting after space-discretization in relation to the infinite dimensional setting and its implications for model reduction. In particular, this includes the discussion of differential-algebraic structures associated to the network-character of our problem and the derivation of compatibility conditions related to fundamental physical properties. Amongst the different model reduction techniques, we consider the moment matching method to be a particularly well-suited choice in our framework.

The Galerkin framework is then appropriately extended to our general nonlinear flow problem. Crucial supplementary concepts are required for the analysis, such as the partial Legendre transform and a more careful discussion of the underlying energy-based modeling. The preservation of the port-Hamiltonian structure after the model-order- and complexity-reduction-step represents a major focus of this work. Similar as in the analysis of the model order reduction, compatibility conditions play a crucial role in the analysis of our complexity reduction, which relies on a quadrature-type ansatz. Furthermore, energy-stable time-discretization schemes are derived for our port-Hamiltonian approximations, as structure-preserving methods from literature are not applicable due to our rather unconventional parametrization of the solution.

Apart from the port-Hamiltonian approximation of the flow problem, another topic of this thesis is the derivation of a new extension of moment matching methods from linear systems to quadratic-bilinear systems. Most system-theoretic reduction methods for nonlinear systems rely on multivariate frequency representations. Our approach instead uses univariate frequency representations tailored towards user-defined families of inputs. Then moment matching corresponds to a one-dimensional interpolation problem rather than to a multi-dimensional interpolation as for the multivariate approaches, i.e., it involves fewer interpolation frequencies to be chosen. The notion of signal-generator-driven systems, variational expansions of the resulting autonomous systems as well as the derivation of convenient tensor-structured approximation conditions are the main ingredients of this part. Notably, our approach allows for the incorporation of general input relations in the state equations, not only affine-linear ones as in existing system-theoretic methods.

Zusammenfassung

In dieser Arbeit werden strukturerhaltende Modellreduktionsmethoden zur effizienten Simulation dynamischer Systeme untersucht. Große Teile der Arbeit fokussieren sich auf ein nichtlineares Strömungsproblem auf Netzen, welches unter anderem in der Simulation von Gasnetzen Anwendung findet, und dessen Approximation durch port-Hamiltonsche Systeme. Die Behandlung des Strömungsproblems ist in zwei Teile aufgeteilt, einen für den Spezialfall der linearen gedämpften Wellengleichung und den anderen für das allgemeine Strömungsproblem.

Die Diskussion bezüglich der gedämpften Wellengleichung stützt sich wesentlich auf einen Galerkin-Ansatz, der sich besonders gut auf Netze verallgemeinern lässt. Im Fokus unserer Analyse stehen die (differential-)algebraischen Strukturen, die sich nach Ortsdiskretisierung ergeben, sowie deren Bedeutung im Funktionenraum-Setting und für die Modellreduktion. Insbesondere werden Kompatibilitätsbedingungen für reduzierte Modelle hergeleitet, die den Erhalt physikalischer Strukturen garantieren, und deren Umsetzung für die Moment-Matching Methode diskutiert.

Unserem Ansatz für das allgemeine Strömungsproblem liegt eine geeignete Verallgemeinerung des Galerkin-Ansatzes zugrunde. Für die Untersuchungen des nichtlinearen Falls sind zusätzliche Konzepte vonnöten, allen voran die partielle Legendre-Transformation und eine tiefergreifende energiebasierte Modellierung. Die Erhaltung von port-Hamiltonscher Struktur nach Modellordnungs- und Komplexitätsreduktion stellen einen wesentlichen Fokus dieser Arbeit dar. Ähnlich wie bei der Modellordnungsreduktion, spielen bei der Analyse der Komplexitätsreduktion, die wir durch empirische Quadratur umsetzen, Kompatibilitätsbedingungen eine entscheidende Rolle. Wegen der nichtlinearen Parametrisierung, die unseren port-Hamiltonschen Approximationen zugrunde liegt, sind strukturerhaltende Methoden aus der Literatur zur Zeitdiskretisierung nicht anwendbar. Wir greifen auch diesen Punkt auf und schlagen neuartige energiestabile Methoden vor.

Losgelöst von der port-Hamiltonschen Approximation des Strömungsproblems, wird in dieser Arbeit außerdem eine Verallgemeinerung des Moment-Matching Verfahrens auf quadratisch-bilineare Systeme vorgeschlagen.

Die Mehrzahl der bisher behandelten systemtheoretischen Methoden für nichtlineare Systeme verwendet multivariate Frequenzdarstellungen. Unser Ansatz basiert hingegen auf univariaten Frequenzdarstellungen, die für geeignet gewählte Klassen von Inputs hergeleitet werden. Der resultierende Interpolationsansatz, der sich aus dem Moment-Matching ergibt, benötigt demzufolge auch nur die Wahl von Parametern in einem eindimensionalen Frequenzraum, was ihn von den multivariaten Ansätzen unterscheidet. Die Grundlage für die vorgeschlagene Methodik sind spezielle Signalgenerator-getriebene Systeme, eine variationelle Entwicklung dieser Systeme und geeignete Approximationsansätze in Tensorräumen. Wir diskutieren außerdem, wie sich unsere Methodik für Systeme mit allgemeinen Input-Abhängigkeiten, die über typischerweise behandelten affin-linearen Abhängigkeiten hinausgehen, erweitern lässt.

Acknowledgment

I would like to express my gratitude and very sincere appreciation to various people for their contributions to this research work.

First and foremost, I am indebted to my supervisor Prof. Dr. Nicole Marheineke for her dedicated mentorship throughout my graduate and post-graduate studies. I am very grateful, not only for her continuous support, valuable advice and patient guidance but also for the large degree of freedom I enjoyed while developing this project and which cannot be taken for granted.

Furthermore, I would like to thank Prof. Dr. Volker Mehrmann for taking on the task as reviewer. I am also very grateful to him for very much good advice and guidance that I received on every occasion I had the pleasure of seeing him.

I would also like to address my thanks to my colleague Alexander Vibe and to two other former colleagues, Dr. Stefan Schießl and Dr. Yi Lu. All three of them were warm-hearted congenial colleagues. A very special thank goes to my present colleagues, Kevin Tolle, Nadine Stahl, Sarah-Alexa Hauschild and Marc Harmening for their support, help and not least for their helpful feedback on this thesis.

Moreover, I would like to acknowledge my financial support by the German Research Foundation (DFG) via SFB/Transregio 154 ”Mathematical modeling, simulation and optimization using the example of gas networks” and the Federal Ministry for Economic Affairs and Energy (BMWI) via MathEnergy ”Mathematical key techniques for energy networks in a state of change”. These projects gave me the opportunity to work together with many distinguished researchers, without whom I would not have been able to complete this thesis. Here, I would like to single out the interaction and cooperation with Prof. Dr. Herbert Egger and Dr. Thomas Kugler with its great impact on this research work.

Last but not least, I would like to thank friends and my parents, Annika and Walter Sailer, for their unconditional support and belief in me.

Contents

Preface

The challenges in model reduction are various, and in this thesis different aspects are discussed in three different parts. Each part is written as a separate unit with its own structure and numbering. They are especially intended to be readable independently of each other. As Part I.A and Part I.B share the same general goal, namely the structure-preserving approximation of a flow problem on networks described by port-Hamiltonian models, they share an elementary introduction and have a common conclusion. Part II stands on its own. Let me very briefly characterize the parts and relate them to my scientific contributions in my time working on this dissertation:

- Part I.A is restricted to the linear damped wave equation, as a simple representative of the flow problem. Furthermore, the analysis is focused on the Galerkin approximation steps in space, which are the space discretization and model order reduction. While the study of the linear problem is an interesting topic on its own, it also serves as a viable preparation for Part I.B. The discussion of the linear problem largely relies on the works [EKLS$^+$18], [LSM17].

- Part I.B is concerned with an approximation procedure for our general nonlinear flow problem on networks. Considerable parts of the linear analysis can be reused in this part. Whenever this is the case, we are brief in the exposition and make a reference to the respective results from Part I.A. However, crucial new concepts have to be taken into account for the adaption to the nonlinear case, such as the partial Legendre transform, complexity reduction and formulation-adapted time discretization methods. The models we are able to cover by our analysis of Part I.B include the barotropic Euler equations on networks. We employ them to simulate gas network systems. In the context of this part, [EKLS20], [LSM19] have been published.

- Part II takes up the classical question of approximation quality. As an abstract approximation property, a new generalization of the notion of moment matching from the linear to the nonlinear case is proposed, and a related model reduction procedure for quadratic-bilinear systems is derived. The main ingredients are the notion of a system to be driven by a signal generator, variational expansions of the resulting autonomous systems, as well as the derivation of convenient tensor-structured approximation conditions. The contributions [LSM18], [SLSM19] provide the basis for this part.

Notation

Throughout this thesis, matrices/tensors, vectors and scalars are indicated by capital boldfaced, small boldfaced and small normal letters, respectively. The unit matrix and the zero matrix are denoted by $\mathbf{I}_N \in \mathbb{R}^{N,N}$ and $\mathbf{0}_{M,N} \in \mathbb{R}^{M,N}$, where the sub-index of the dimensions is omitted if they are clear from the context. The column- and row-wise concatenation of matrices or vectors is abbreviated with a comma and a semicolon, respectively. That is

$$\begin{aligned}
[\mathbf{A}, \mathbf{B}] &= \begin{bmatrix} \mathbf{A} & \mathbf{B} \end{bmatrix} \in \mathbb{R}^{M,N+N_B} \\
[\mathbf{A}; \mathbf{C}] &= \begin{bmatrix} \mathbf{A} \\ \mathbf{C} \end{bmatrix} \in \mathbb{R}^{M+M_C,N}, && \text{for } \mathbf{A} \in \mathbb{R}^{M,N},\ \mathbf{B} \in \mathbb{R}^{M,N_B},\ \mathbf{C} \in \mathbb{R}^{M_C,N} \\
[\mathbf{a}; \mathbf{b}] &= \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \in \mathbb{R}^{N+N_B}, && \text{for } \mathbf{a} \in \mathbb{R}^{N},\ \mathbf{b} \in \mathbb{R}^{N_B}.
\end{aligned}$$

The kernel and image of $\mathbf{A}$ are denoted as $ker(\mathbf{A})$ and $im(\mathbf{A})$, respectively. We write $[\mathbf{A}]_{i,j}$ for the entry of $\mathbf{A}$ in the i-th row and j-th column, and $[\mathbf{a}]_i$ for i-th entry of $\mathbf{a}$. The Euclidean scalar product for two vectors $\mathbf{a}, \mathbf{b}$ of the same length is written as $\mathbf{a} \cdot \mathbf{b}$.

Tensors are approached by matricization strategies as in [Rug81], [KT10], [Hac12]: The Kronecker-tensor product $\otimes$ is defined by

$$\mathbf{P} \otimes \mathbf{Q} = \begin{bmatrix} [\mathbf{P}]_{1,1}\mathbf{Q} & [\mathbf{P}]_{1,2}\mathbf{Q} & \dots & [\mathbf{P}]_{1,N}\mathbf{Q} \\ \dots & \dots & \dots & \dots \\ [\mathbf{P}]_{M,1}\mathbf{Q} & [\mathbf{P}]_{M,2}\mathbf{Q} & \dots & [\mathbf{P}]_{M,N}\mathbf{Q} \end{bmatrix} \quad \text{for } \mathbf{P}, \mathbf{Q} \in \mathbb{R}^{M,N}.$$

We abbreviate $\mathbf{P}^{\textcircled{2}} = \mathbf{P} \otimes \mathbf{P}$, $\mathbf{P}^{\textcircled{3}} = \mathbf{P} \otimes \mathbf{P} \otimes \mathbf{P}$. Additionally, we introduce the notation

$$\begin{aligned}
\textcircled{2}_{\mathbf{P}}\mathbf{Q} &= \mathbf{Q} \otimes \mathbf{P} + \mathbf{P} \otimes \mathbf{Q} && \in \mathbb{R}^{M^2,N^2} \\
\textcircled{3}_{\mathbf{P}}\mathbf{Q} &= \textcircled{2}_{\mathbf{P}}\mathbf{Q} \otimes \mathbf{P} + \mathbf{P}^{\textcircled{2}} \otimes \mathbf{Q} \\
&= \mathbf{Q} \otimes \mathbf{P} \otimes \mathbf{P} + \mathbf{P} \otimes \mathbf{Q} \otimes \mathbf{P} + \mathbf{P} \otimes \mathbf{P} \otimes \mathbf{Q} && \in \mathbb{R}^{M^3,N^3}.
\end{aligned}$$

The expressions $\mathbf{P}^{\textcircled{i}}$ and $\textcircled{i}_{\mathbf{P}}\mathbf{Q}$ are defined analogously for $i > 3$. The Kronecker product has precedence to matrix multiplications, thus the relations

$$\begin{aligned}
(\mathbf{A} \otimes \mathbf{B})\mathbf{C} &= \mathbf{A} \otimes \mathbf{B}\mathbf{C} \\
(\mathbf{A}\mathbf{B}) \otimes (\mathbf{C}\mathbf{D}) &= (\mathbf{A} \otimes \mathbf{C})\,(\mathbf{B} \otimes \mathbf{D}) = \mathbf{A} \otimes \mathbf{C}\,\mathbf{B} \otimes \mathbf{D}
\end{aligned}$$

hold matrices of appropriate dimension. From the definition of $\otimes$ it follows directly

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \otimes \mathbf{P} = \begin{bmatrix} \mathbf{A} \otimes \mathbf{P} & \mathbf{B} \otimes \mathbf{P} \\ \mathbf{C} \otimes \mathbf{P} & \mathbf{D} \otimes \mathbf{P} \end{bmatrix}.$$

Regarding nonlinear transformations and derivatives, we use the following notational conventions: Coordinate transformations in $\mathbb{R}^n$ are consistently marked by a hat (e.g. $\hat{\mathbf{a}}, \hat{\mathbf{z}} : \mathbb{R}^n \to \mathbb{R}^n$). Given a scalar field, $h : \mathbb{R}^n \to \mathbb{R}$ and a partitioning into vector-components $\mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2]$ with $\mathbf{z}_i \in \mathbb{R}^{n_i}$, $n_1 + n_2 = n$, we write $\nabla_{\mathbf{z}} h$ and $\nabla_{\mathbf{z}_i} h$ for the gradient and its respective sub-blocks, i.e.,

$$\nabla_{\mathbf{z}} h : \mathbb{R}^n \to \mathbb{R}^n, \quad \text{and} \quad \nabla_{\mathbf{z}} h = \begin{bmatrix} \nabla_{\mathbf{z}_1} h \\ \nabla_{\mathbf{z}_2} h \end{bmatrix}.$$

Moreover, $\nabla_{\mathbf{z}\mathbf{z}} h$ denotes the Hessian, and the respective sub-blocks $\nabla_{\mathbf{z}_i \mathbf{z}_j} h$ for $i, j = 1, 2$ are given by

$$\nabla_{\mathbf{z}\mathbf{z}} h : \mathbb{R}^n \to \mathbb{R}^{n,n}, \quad \text{and} \quad \nabla_{\mathbf{z}\mathbf{z}} h = \begin{bmatrix} \nabla_{\mathbf{z}_1\mathbf{z}_1} h & \nabla_{\mathbf{z}_1\mathbf{z}_2} h \\ \nabla_{\mathbf{z}_2\mathbf{z}_1} h & \nabla_{\mathbf{z}_2\mathbf{z}_2} h \end{bmatrix}.$$

We allow for derivatives w.r.t. coordinate-transformations, e.g., $\nabla_{\mathbf{z}} h(\hat{\mathbf{z}}(\mathbf{a}))$ for $\mathbf{a} \in \mathbb{R}^n$ should be read as

$$\nabla_{\mathbf{z}} h(\hat{\mathbf{z}}(\mathbf{a})) = \nabla_{\mathbf{z}} h(\tilde{\mathbf{z}})_{|\tilde{\mathbf{z}} = \hat{\mathbf{z}}(\mathbf{a})}.$$

Function-valued vector spaces (e.g. $\mathcal{V}_i, \mathcal{L}^2$) are distinguished in their typesetting from real-valued spaces and sets (e.g. $\mathbb{R}^N, \mathbb{A}$). The solutions of partial differential equations in this thesis depend on time t and space x, e.g. $z_i(t, x) \in \mathbb{R}$ for $i = 1, 2$. When it is convenient, we instead interpret them as functions in time with values in a function space, i.e., $z_i(t) \in \mathcal{V}_i$ for some function space $\mathcal{V}_i$. In either case, the concatenation of solution components in a vector is underscored to distinguish it from an ordinary vector, i.e.,

$$\underline{\mathbf{z}}(t, x) = [z_1(t, x); z_2(t, x)] \in \mathbb{R}^2 \qquad \text{or} \qquad \underline{\mathbf{z}}(t) = [z_1(t); z_2(t)] \in \mathcal{V}_1 \times \mathcal{V}_2.$$

The application of real-valued functions onto the solution components (e.g. $h(\underline{\mathbf{z}})$) has to be understood pointwise in space and time. This holds in particular for the derivative-expressions from above. For example, the expression $\nabla_{\mathbf{z}} h(\underline{\mathbf{z}})$ is defined by

$$\nabla_{\mathbf{z}} h(\underline{\mathbf{z}}(t, x)) = \nabla_{\mathbf{z}} h(\tilde{\mathbf{z}})_{|\tilde{\mathbf{z}} = \underline{\mathbf{z}}(t,x)}.$$

Respective expressions $\nabla_{z_i} h(\underline{\mathbf{z}})$ and $\nabla_{z_i z_j} h(\underline{\mathbf{z}})$ are similarly defined as the pointwise application of $\nabla_{z_i} h$ and $\nabla_{z_i z_j} h$ onto $\underline{\mathbf{z}}$ for $i, j = 1, 2$.

The acronyms FOM and ROM are used to abbreviate full order model and reduced order model. The latter is typically a low-dimensional approximation of the former. Quantities related to ROMs are marked by a subscript $_r$.

Finally, note that the parts are independently structured and have their own numbering. Cross-references from one part to another are indicated by a prefix, e.g., when we want to refer to Chapter 1 of Part I.A outside of this part, we write Part I.A-Chapter 1.

Part I

Port-Hamiltonian approximation for flow problems on networks

Introduction

Model problem

Part I of this thesis deals with the approximation of a prototypical partial differential equation on networks. The network is described by a directed graph. Each edge ω of the graph can be identified with an interval $(0, l^\omega)$ of length l^ω. Given the Hamiltonian density $h : \mathbb{R}^2 \to \mathbb{R}$ and a friction term $r : \mathbb{R}^2 \to \mathbb{R}^+$, the edgewise states $\underline{\mathbf{z}}^\omega = [z_1^\omega; z_2^\omega] : [0,T] \times (0, l^\omega) \to \mathbb{R}^2$ are governed by

$$\begin{aligned}\partial_t z_1^\omega(t,x) &= -\partial_x \nabla_{z_2} h(\underline{\mathbf{z}}^\omega(t,x))\\ \partial_t z_2^\omega(t,x) &= -\partial_x \nabla_{z_1} h(\underline{\mathbf{z}}^\omega(t,x)) - r(\underline{\mathbf{z}}^\omega(t,x))\nabla_{z_2} h(\underline{\mathbf{z}}^\omega(t,x)),\end{aligned}$$

with short-hand notation $\nabla_{z_i} h$ defined by

$$\nabla_{z_i} h(\underline{\mathbf{z}}^\omega(t,x)) = \nabla_{z_i} h([z_1; z_2])_{|[z_1,z_2]=\underline{\mathbf{z}}^\omega(t,x)}, \qquad \text{for } i = 1, 2.$$

In the upcoming, the above equations are also written in the more compact form

$$\partial_t \underline{\mathbf{z}}^\omega(t,x) = \begin{bmatrix} & -\partial_x \\ -\partial_x & -r(\underline{\mathbf{z}}^\omega(t,x)) \end{bmatrix} \nabla_{\mathbf{z}} h(\underline{\mathbf{z}}^\omega(t,x)). \tag{1a}$$

The edgewise systems are interconnected at inner nodes ν of the graph by the two coupling conditions

$$\sum_{\omega \in \mathcal{E}(\nu)} n^\omega[\nu] \nabla_{z_2} h(\underline{\mathbf{z}}^\omega(t,\nu)) = 0, \qquad \nabla_{z_1} h(\underline{\mathbf{z}}^\omega(t,\nu)) = \nabla_{z_1} h(\underline{\mathbf{z}}^{\tilde\omega}(t,\nu)) \quad \text{for } \omega, \tilde\omega \in \mathcal{E}(\nu), \tag{1b}$$

with $n^\omega[\nu] \in \{-1, 0, 1\}$ describing the direction of the edge and $\mathcal{E}(\nu)$ denoting the set of all edges incident to ν, see Part I.A-Chapter 1 for details. The system is complemented by appropriate boundary- and initial-conditions.

For the set of all edges $\mathcal{E}$, the total mass $\mathcal{M}$ and the Hamiltonian $\mathcal{H}$ are defined as

$$\mathcal{M}(\underline{\mathbf{z}}) = \sum_{\omega \in \mathcal{E}} \int_{(0,l^\omega)} z_1^\omega dx \qquad \text{and} \qquad \mathcal{H}(\underline{\mathbf{z}}) = \sum_{\omega \in \mathcal{E}} \int_{(0,l^\omega)} h(\underline{\mathbf{z}}^\omega) dx.$$

Conservation of mass and dissipation of the Hamiltonian (energy dissipation) up to exchange with the boundary are fundamental properties of our model problem. They are also central for our analysis and proposed structure-preserving approximation procedure.

Next, let us shortly go through the representatives of (1), which take a special role in this thesis. For convenience, we drop the superscript 'ω' and consider the one-edge problem for now. Throughout the main part we treat the network case.

Damped wave equation

The damped wave equation, cf. [AH14], [EK18], [SMH19], is obtained for our model problem by the choice

$$\underline{\mathbf{z}} = \begin{bmatrix} \alpha p \\ \beta m \end{bmatrix}, \qquad h([z_1; z_2]) = \frac{1}{2}\left(\frac{1}{\alpha} z_1^2 + \frac{1}{\beta} z_2^2\right).$$

The parameters $\alpha, \beta > 0$ may depend on the spatial variable x. When r is constant, this yields a linear system reading

$$\begin{aligned} &\alpha \partial_t p(t,x) + \partial_x m(t,x) = 0 \\ &\beta \partial_t m(t,x) + \partial_x p(t,x) = -r m(t,x), \qquad \text{for } x \in (0, l^\omega),\ t \in [0,T]. \end{aligned}$$

The whole Part I.A is devoted to this representative.

Euler equations

The Euler equations take an important role in many applications, see [Che05], [LeV02], [LeV90], [BGH11], [LM18], [Dom11]. They consist of the following set of equations characterizing density ρ, velocity v and a potential energy P,

$$\partial_t \rho + \partial_x (\rho v) = 0 \tag{2a}$$

$$\partial_t(\rho v) + \partial_x \left(\rho v^2 + p\right) = -\frac{\lambda}{2D}\rho |v| v \tag{2b}$$

$$\partial_t h + \partial_x (vh + vp) = -K_e, \qquad h = \rho v^2/2 + P \tag{2c}$$

for $x \in (0, l^\omega)$, $t \in [0,T]$. A constitutive law for the prescription of the pressure p has to be added, and the energy dissipation term K_e and the friction term λ may depend on the states like $K_e = K_e(\rho, v, m, P)$ and $\lambda = \lambda(\rho v)$ in general. Note that (2c) particularly describes the energy dissipation for the energy functional $\int h([\rho; v])dx$.

Barotropic Euler equations

For all our discussions, we take barotropic pressure models [Che05], meaning that we assume p to be a function only depending on the density, i.e., $p = p(\rho)$. All in all, the first two equations can then be solved independent of the third one, and ρ, v are characterized by

$$\partial_t \rho + \partial_x (\rho v) = 0 \tag{3a}$$

$$\partial_t(\rho v) + \partial_x \left(\rho v^2 + p(\rho)\right) = -\frac{\lambda}{2D}\rho |v| v. \tag{3b}$$

Still, the system inherits dissipation of the energy functional

$$\mathcal{H}([\rho; v]) = \int_{(0,l^\omega)} h([\rho; v])dx, \qquad \text{with} \quad h([\rho; v]) = \rho\frac{v^2}{2} + P(\rho), \tag{3c}$$

which we refer to as Hamiltonian. Note that P is now a function of density instead of a dependent variable. Given a pressure law p, it necessarily holds

$$P''(\rho) = \frac{1}{\rho} p'(\rho) \tag{3d}$$

due to compatibility reasons with (2). Clearly, several choices for P may be possible. This is not surprising, when one keeps in mind that in the full Euler equations (2) additional initial- and boundary-conditions have to be prescribed for P. Defining $\underline{\mathbf{z}} = [\rho; v]$, we can formally rewrite (3) in our abstract standard form

$$\partial_t \underline{\mathbf{z}}(t,x) = \begin{bmatrix} & -\partial_x \\ -\partial_x & -r(\underline{\mathbf{z}}(t,x)) \end{bmatrix} \nabla_{\mathbf{z}} h(\underline{\mathbf{z}}(t,x)), \qquad \text{with } r(\underline{\mathbf{z}}) = \frac{\lambda}{2D} \frac{|v|}{\rho}.$$

Let us list particular choices for pressure laws p and related consistent potential energy functions P, cf. [BGH11], [MLD19], [DHLT17]. We refer to the references for a physical interpretation of the constants and derivations of the models.

- Isentropic: $p(\rho) = c_p \rho^\gamma$, $c_p > 0$, $\gamma > 1$
  $P(\rho) = \frac{c_p}{\gamma - 1} \rho^\gamma$
- Isothermal: $p(\rho) = RT \frac{\rho}{1 - RT\alpha\rho}$, $RT > 0$, $\alpha < 0$
  $P(\rho) = RT\rho \log\left(\rho_{sc} \frac{1 - RT\alpha\rho}{\rho}\right)$
- Ideal isothermal gas: $p(\rho) = c_s^2 \rho$, $c_s > 0$
  $P(\rho) = c_s^2 \rho \log\left(\frac{\rho}{\rho_{sc}}\right)$

The factor ρ_{sc} above is needed for a non-dimensionalization and is one, when SI-units are taken.

Simplified barotropic Euler equations

In many applications it is acceptable to use model simplifications on the Euler equations. Motivated by [BGH11], [MLD19], we consider the simplification obtained by dropping the term $\partial_x(\rho v^2)$ in (3b). By introducing the mass flow $m = \rho v$, the system can then be re-expressed in ρ and m as

$$\partial_t \rho + \partial_x m = 0 \tag{4a}$$

$$\partial_t m + \partial_x p(\rho) = -\frac{\lambda}{2D} \frac{1}{\rho} |m| m. \tag{4b}$$

To apply our framework, we define the Hamiltonian

$$\mathcal{H}([\rho; m] = \int_{(0,l^\omega)} h([\rho; m]) dx, \qquad \text{with} \qquad h([\rho; m]) = \frac{m^2}{2} + \tilde{P}(\rho) \tag{4c}$$

with potential energy function $\tilde{P}$ chosen such that

$$\tilde{P}''(\rho) = p'(\rho). \tag{4d}$$

Again, several choices for $\tilde{P}$ are possible from a mathematical point of view. Equations (4) can then be rewritten in our abstract model standard form (1) with

$$\underline{\mathbf{z}} = \begin{bmatrix} \rho \\ m \end{bmatrix} \qquad \text{and} \qquad r(\underline{\mathbf{z}}) = \frac{\lambda}{2D}\frac{|m|}{\rho}.$$

Let us also list choices for potential energy functions $\tilde{P}$ for the pressure laws as above:

- Isentropic: $p(\rho) = c_p \rho^\gamma$, $c_p > 0$, $\gamma > 1$.
  $\tilde{P}(\rho) = \frac{c_p}{\gamma+1}\rho^{(\gamma+1)}$
- Isothermal: $p(\rho) = RT\frac{\rho}{1-RT\alpha\rho}$, $RT > 0$, $\alpha < 0$.
  $\tilde{P}(\rho) = \frac{-1}{RT\alpha^2}\left(\log\left(1 - RT\alpha\rho\right) + RT\alpha\rho\right)$
- Ideal isothermal gas: $p(\rho) = c_s^2\rho$.
  $\tilde{P}(\rho) = \frac{1}{2}c_s^2\rho^2$

Note that for the ideal isothermal gas model, the simplified barotropic Euler equations coincide with the damped wave equations with a nonlinear damping term.

In Part I.B the approximation framework is derived for the general abstract model problem (1). The realization by the barotropic Euler equations takes a special role in the last two chapters of that part: A well-posedness result for its approximation in our framework is derived, and all numerical examinations are performed for the isothermal Euler equations and their simplification.

Outline

The main theme of Part I is the systematic derivation of a structure-preserving approximation procedure for nonlinear flow problems on networks of the form (1) and its realization by model reduction. Special emphasis lies on the preservation of fundamental properties underlying our model problem, such as mass conservation, energy dissipation and the network structure. To approach this aim in a systematic manner, we built our approximation procedure around energy-based modeling concepts, specifically the notion of so-called port-Hamiltonian systems.

Structure of parts

We divide the discussion in the two separate parts Part I.A and Part I.B. Their basic outlines follow the same pattern: In Chapter 1 preliminary tools necessary for the formulation of the respective model problem are gathered, and a formal introductory discussion is given. The abstract approximation procedure is derived in Chapter 2, where also resulting coordinate representations are treated. Subsequently, the realization of the approximation procedure by model reduction methods is addressed in Chapter 3. The remainder is then reserved to the application of the derived approximation- and model reduction-methods. For Part I.A, the application solely consists of the presentation of numerical results (Part I.A-Chapter 4), whereas in Part I.B, the abstract model is first instanced at the example of the barotropic Euler equations (Part I.B-Chapter 4), and afterwards numerical results are discussed (Part I.B-Chapter 5).

The substructure of the chapters is guided by the hierarchy underlying our approximation procedure.

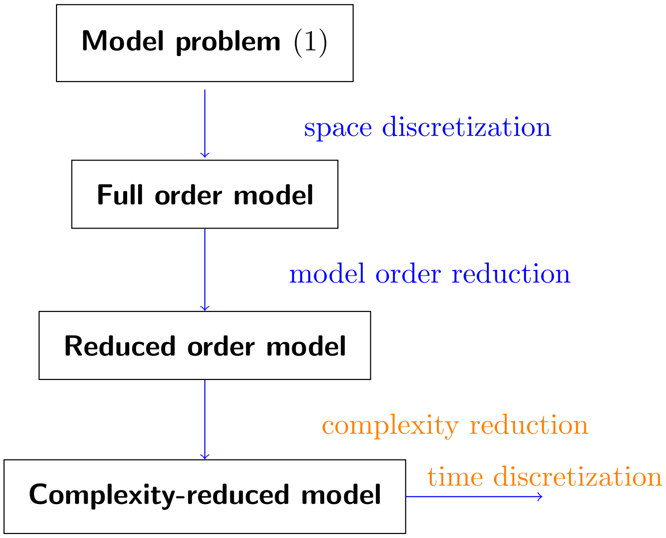

Figure 1: Sketch for all steps in our approximation procedure. The analysis of Part I.A is restricted to the linear damped wave equation and the steps marked in blue. In Part I.B, the general model problem (1) is addressed, and all approximation steps are considered.

Outline of the approximation procedure

Our approximation procedure includes all steps necessary in the construction of low-order and online-efficient reduced models. The individual stages of our approximation procedure, see Fig. 1, read as follows:

Given an instance of model problem (1), we first discretize it in space to obtain the full order model. Afterwards, a model order reduction is performed leading to the reduced (order) model. These first two approximation steps correspond to a Galerkin approximation in space. Although the reduced order model is typically of much lower dimension as the full order model, it is not necessarily more efficient to solve in the nonlinear case. That is, because the evaluation of reduced nonlinearities in general requires a lifting step to the space of the underlying full order model for each evaluation. To avoid this lifting-bottleneck, another approximation step is included, the complexity reduction. This step is not of Galerkin-type and, therefore, has to be analyzed separately. All our proposed approximation steps in space are such that they yield port-Hamiltonian structure. This structure also guides our choice of time discretization, which is the last approximation step in our hierarchy. We present our results in two separate parts that emphasize different challenges:

In Part I.A the discussion is restricted to the linear damped wave equation. The focus lies here on the handling of the network structure and the role of compatibility conditions related to the convectional terms in the model equations. Only the Galerkin approximation steps are discussed in detail here, i.e., the space discretization and the model order reduction. For the linear case, complexity reduction is not needed and appropriate structure-preserving time discretization schemes have already been discussed in literature. Let us also note that a standard port-Hamiltonian form can be recovered by linear coordinate transformations in this case.

In Part I.B we present a procedure for the general nonlinear model problem (1). While the incorporation of the network-aspect is by no means trivial for most approximation schemes, our ansatz is chosen such that an adaption of the framework presented in Part I.A is almost immediate. Moreover, our ansatz carefully takes into account the underlying structure of the model problem on all stages of the approximation. In contrast to the discussion in Part I.A, nonlinear state transformations have to be considered here. To deal with them, we make use of some tools from convex analysis, first and foremost of the (partial) Legendre transform. With the help of these tools, we are able to prove port-Hamiltonian structure for our approximations in the nonlinear case. In particular, it is shown that a reformulation of the space-discrete systems into a standard port-Hamiltonian form is possible. The involved nonlinear transformation is, however, not necessarily well-suited for practical implementation. Hence, time discretization schemes for port-Hamiltonian systems from literature are not applicable in Part I.B. New structure-preserving time discretization methods, accounting for our natural parametrization, are therefore proposed.

Overview on port-Hamiltonian approximation

Naive approximation methods for dynamical systems can in general fail to respect the physical and geometric structure of the system at hand. Needless to say that this is undesired on its own, but it is also a major source of instability for approximation methods. A remedy consists in the usage of structure-preserving methods. Their analysis has grown to a wide field of research, see e.g., [LeV02], [CMKO11], [CGM$^+$12] for selected overviews.

We concentrate here on the port-Hamiltonian framework. This energy-based modeling approach has undergone a tremendous development since its introduction by the authors of [MvdS92], [MvdS01]. It originates from the analysis of finite-dimensional interconnected systems [vdSM13], [vdSJ14], but also has been extended to the infinite-dimensional setting of partial differential equations [BMBM18], [GTvdSM04], and has been systematically generalized to constrained dynamical differential-algebraic systems [BMXZ17], [vdSJ14]. In what follows, we first touch upon finite-dimensional port-Hamiltonian systems, and then give a selected and concise overview on structure-preserving model reduction and discretization. Parallels between methods from literature and our proposed approximation procedure are drawn.

Finite-dimensional port-Hamiltonian systems

Central in the port-Hamiltonian approach is the Hamiltonian. In the finite-dimensional setting, it is a functional $H : \mathbb{R}^N \to \mathbb{R}$ and typically assumed to be convex. Often it relates to an energy or entropy in applications. We additionally presume it to be twice continuously differentiable here. Let $\mathbf{z} \in \mathbb{R}^N$ denote a state, and an anti-symmetric matrix $\bar{\mathbf{J}}(\mathbf{z})$ and a symmetric positive semi-definite matrix $\bar{\mathbf{R}}(\mathbf{z})$ be given, i.e.,

$$\bar{\mathbf{J}}(\mathbf{z}) = -\bar{\mathbf{J}}(\mathbf{z})^T \in \mathbb{R}^{N,N} \qquad \text{and} \qquad \bar{\mathbf{R}}(\mathbf{z}) = \bar{\mathbf{R}}(\mathbf{z})^T \in \mathbb{R}^{N,N}, \qquad \text{for all considered } \mathbf{z}.$$

Further introducing a matrix $\mathbf{K} \in \mathbb{R}^{N,p}$, we can formulate a prototype for port-Hamiltonian systems of central importance in this thesis:

System 0.1. *Find* $\mathbf{z} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$, $\mathbf{e}_B \in \mathcal{C}^0([0,T];\mathbb{R}^p)$ *and* $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^p)$ *such that*

$$\begin{aligned} \frac{d}{dt}\mathbf{z}(t) &= \left(\bar{\mathbf{J}}(\mathbf{z}) - \bar{\mathbf{R}}(\mathbf{z})\right) \nabla_{\mathbf{z}} H(\mathbf{z}(t)) + \mathbf{K}\mathbf{e}_B(t) \\ \mathbf{f}_B(t) &= \mathbf{K}^T \nabla_{\mathbf{z}} H(\mathbf{z}(t)). \end{aligned}$$

Given input $\mathbf{u} : [0,T] \to \mathbb{R}^p$, the system can, e.g., be closed by conditions of either type,

$$\textit{Type 1: } \mathbf{e}_B(t) = \mathbf{u}(t), \qquad \textit{Type 2: } \mathbf{f}_B(t) = \mathbf{u}(t), \qquad\qquad t \in [0,T],$$

together with appropriate initial conditions. In the port-Hamiltonian framework, $\mathbf{z}$ is called the energy variable, $\mathbf{e} = \nabla_{\mathbf{z}} H(\mathbf{z})$ the effort variable, and $\mathbf{e}_B$ and $\mathbf{f}_B$ the boundary effort and boundary flow, respectively. The system structure readily implies a fundamental property, which we refer to as the energy dissipation equality.

Lemma 0.2. *A solution of System 0.1 fulfills the energy dissipation equality*

$$\frac{d}{dt} H(\mathbf{z}(t)) = \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \nabla_{\mathbf{z}} H(\mathbf{z}(t))^T \bar{\mathbf{R}}(\mathbf{z}) \nabla_{\mathbf{z}} H(\mathbf{z}(t)) \leq \mathbf{e}_B(t) \cdot \mathbf{f}_B(t).$$

Proof. By application of the chain-rule, we obtain

$$\begin{aligned} \frac{d}{dt} H(\mathbf{z}) &= \frac{d}{dt}\mathbf{z} \cdot \nabla_{\mathbf{z}} H(\mathbf{z}) = \left[\left(\bar{\mathbf{J}}(\mathbf{z}) - \bar{\mathbf{R}}(\mathbf{z}) \right) \nabla_{\mathbf{z}} H(\mathbf{z}) + \mathbf{K}\mathbf{e}_B \right] \cdot \nabla_{\mathbf{z}} H(\mathbf{z}) \\ &= \nabla_{\mathbf{z}} H(\mathbf{z})^T \bar{\mathbf{J}}(\mathbf{z}) \nabla_{\mathbf{z}} H(\mathbf{z}) - \nabla_{\mathbf{z}} H(\mathbf{z})^T \bar{\mathbf{R}}(\mathbf{z}) \nabla_{\mathbf{z}} H(\mathbf{z}) + (\mathbf{K}\mathbf{e}_B) \cdot \nabla_{\mathbf{z}} H(\mathbf{z}). \end{aligned}$$

By employing the structural properties of the system, we see that the first term on the right side is equal to zero, the second one is bounded by zero from above, and the third one can be rewritten as $(\mathbf{K}\mathbf{e}_B) \cdot \nabla_{\mathbf{z}} H(\mathbf{z}) = \mathbf{e}_B \cdot (\mathbf{K}^T \nabla_{\mathbf{z}} H(\mathbf{z})) = \mathbf{e}_B \cdot \mathbf{f}_B$. □

As will be discussed in Part I.A in detail, closing conditions *Type 2* make System 0.1 a differential algebraic system. With closing conditions of *Type 1*, on the other hand, the system simplifies to an ordinary differential equation. We assume to be in the latter case for the remainder of this section. In accordance with standard notations, we define the output vector $\mathbf{y}(t) = \mathbf{f}_B(t)$ and the matrix $\mathbf{B} = \mathbf{K}$. Then System 0.1 specializes as follows.

System 0.3. *Find* $\mathbf{z} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$ *such that*

$$\begin{aligned} \frac{d}{dt}\mathbf{z}(t) &= \left(\bar{\mathbf{J}}(\mathbf{z}(t)) - \bar{\mathbf{R}}(\mathbf{z}(t)) \right) \nabla_{\mathbf{z}} H(\mathbf{z}(t)) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}^T \nabla_{\mathbf{z}} H(\mathbf{z}(t)), \qquad\qquad \mathbf{z}(0) = \mathbf{z}_0 \end{aligned}$$

for given input $\mathbf{u} \in \mathcal{C}^0([0,T];\mathbb{R}^p)$ *and initial conditions* $\mathbf{z}_0 \in \mathbb{R}^N$.

Model order reduction of port-Hamiltonian systems

The application of standard model reduction onto System 0.3 can be summarized as follows: One seeks high-fidelity bases $\mathbf{V}, \mathbf{W} \in \mathbb{R}^{N,n}$ with $n \ll N$ and defines the reduced model of System 0.3 as

$$\begin{aligned} (\mathbf{W}^T\mathbf{V}) \frac{d}{dt}\mathbf{z}_r(t) &= \mathbf{W}^T \left[\left(\bar{\mathbf{J}}(\mathbf{z}) - \bar{\mathbf{R}}(\mathbf{z}) \right) \nabla_{\mathbf{z}} H(\mathbf{z}) \right]_{|\mathbf{z}=\mathbf{V}\mathbf{z}_r(t)} + \mathbf{W}^T\mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}^T \nabla_{\mathbf{z}} H(\mathbf{z})_{|\mathbf{z}=\mathbf{V}\mathbf{z}_r(t)}, \qquad\qquad \mathbf{z}_r(0) = \mathbf{z}_{r,0}. \end{aligned} \tag{5}$$

Without any further doing, this reduced model most certainly does not inherit the port-Hamiltonian structure and not even the energy dissipation of Lemma 0.2. There have been a few attempts in literature for structure-preserving model reduction, which yield reduced models of port-Hamiltonian structure, i.e., of the form

$$\begin{aligned} \frac{d}{dt}\mathbf{z}_r(t) &= \left(\bar{\mathbf{J}}_r(\mathbf{z}_r(t)) - \bar{\mathbf{R}}_r(\mathbf{z}_r(t))\right) \nabla_{\mathbf{z}_r} H_r(\mathbf{z}_r(t)) + \mathbf{B}_r \mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}_r^T \nabla_{\mathbf{z}_r} H_r(\mathbf{z}_r(t)), \qquad\qquad \mathbf{z}_r(0) = \mathbf{z}_{r,0}, \end{aligned} \tag{6}$$

with $\bar{\mathbf{J}}_r(\mathbf{z}_r)$ anti-symmetric and $\bar{\mathbf{R}}_r(\mathbf{z}_r)$ symmetric positive semi-definite. These attempts seem mostly to fall in one of the following three categories: 1. Additional approximation of the Hamiltonian, 2. compatible reduction bases and 3. reparametrization.

1. Additional approximation of the Hamiltonian Taking a closer look at the reduced model (5), it can be recognized that the fundamental structure, which the model is in general not able to capture, is the gradient structure present in System 0.3. More precisely, we cannot identify $\nabla_{\mathbf{z}} H(\mathbf{z})_{|\mathbf{z}=\mathbf{V}\mathbf{z}_r}$ with a reduced Hamiltonian acting on a reduced state. A possible remedy is to replace this expression by an approximation, for which there exists an underlying reduced Hamiltonian. Following the ideas of [CBG16], this can be done by the approximation ansatz

$$\nabla_{\mathbf{z}} H(\mathbf{z})_{|\mathbf{z}=\mathbf{V}\mathbf{z}_r} \approx \mathbf{W}\left(\mathbf{V}^T\mathbf{W}\right)^{-1}\mathbf{V}^T \nabla_{\mathbf{z}} H(\mathbf{z})_{|\mathbf{z}=\mathbf{V}\mathbf{z}_r} = \mathbf{W}\mathbf{V}^T \nabla_{\mathbf{z}} H(\mathbf{z})_{|\mathbf{z}=\mathbf{V}\mathbf{z}_r}, \tag{7}$$

where $\mathbf{V}^T\mathbf{W} = \mathbf{I}_n$ is assumed for the last equality. The latter assumption is not restrictive, as it can be reached by rescaling, given $im(\mathbf{V})$ and $im(\mathbf{W})$ have generic orientations to each other. The reason for the special choice of approximation (7) is that it may be interpreted as replacing the Hamiltonian H by an approximate one reading

$$\tilde{H}(\mathbf{z}) := H(\mathbf{V}\mathbf{W}^T\mathbf{z}), \qquad \text{as it holds} \qquad \nabla_{\mathbf{z}}\tilde{H}(\mathbf{z}) = \mathbf{W}\mathbf{V}^T\nabla_{\mathbf{z}} H(\mathbf{z}) \quad \text{for } \mathbf{z} \in im(\mathbf{V}).$$

With that replacement, the reduced model can be recast in the port-Hamiltonian structure (6) with

$$\begin{aligned} H_r(\mathbf{z}_r) &= \tilde{H}(\mathbf{V}\mathbf{z}_r) = H(\mathbf{V}\mathbf{z}_r), & \mathbf{B}_r &= \mathbf{W}^T\mathbf{B} \\ \bar{\mathbf{J}}_r(\mathbf{z}_r) &= \mathbf{W}^T\bar{\mathbf{J}}(\mathbf{V}\mathbf{z}_r)\mathbf{W}, & \bar{\mathbf{R}}_r(\mathbf{z}_r) &= \mathbf{W}^T\bar{\mathbf{R}}(\mathbf{V}\mathbf{z}_r)\mathbf{W}. \end{aligned}$$

Apart from the model order- and complexity reduction-method proposed in [CBG16], also, e.g., the discretization schemes [Kot13], [GTvdSM04], [PAvdS12], [CML19], [MM14] can be considered to be of this category.

Remark 0.4. *The above interpretation of approximation step* (7) *by a new Hamiltonian* $\tilde{H}$ *is not the only one. Another, possibly more natural one for someone familiar with basic Riemannian geometry, is that we replaced the Euclidean gradient* $\nabla_{\mathbf{z}}$ *by a Riemannian gradient acting on the subspace* $\mathbf{V}$*, cf.[AMS08, pp. 60].*

2. Compatible reduction bases By the additional approximation step (7), the upper approach does not yield a pure Galerkin projection of the full order model but includes an additional approximation of an operator. This may of course be undesirable. Apart from the additional error, it also complicates the analysis, as such operator approximations do not necessarily have a useful interpretation in terms of approximating partial differential equations. The reason for the latter is that orthogonal projections between infinite dimensional Hilbert spaces are not necessarily bounded operators. To avoid these difficulties related to the replacement (7), one can in certain situations construct compatible Petrov-Galerkin approximations instead. For this to be practicable, we need the matrix operators $\bar{\mathbf{J}}$ and $\bar{\mathbf{R}}$ to be of sufficiently simple structure. For given orthogonal reduction basis $\mathbf{V} \in \mathbb{R}^{N,n}$ we need to be able to construct a compatible reduction basis $\mathbf{W} \in \mathbb{R}^{N,n}$ such that

$$\mathbf{W}^T \bar{\mathbf{J}}(\mathbf{V}\mathbf{z}_r) = \bar{\mathbf{J}}_r(\mathbf{z}_r)\mathbf{V}^T \qquad \text{and} \qquad \mathbf{W}^T \bar{\mathbf{R}}(\mathbf{V}\mathbf{z}_r) = \bar{\mathbf{R}}_r(\mathbf{z}_r)\mathbf{V}^T$$

for an anti-symmetric $\bar{\mathbf{J}}_r(\mathbf{z}_r) \in \mathbb{R}^{n,n}$ and a symmetric positive semi-definite $\bar{\mathbf{R}}_r(\mathbf{z}_r) \in \mathbb{R}^{n,n}$. Given this is possible, the reduced model may be rewritten in the port-Hamiltonian form (6) with $H_r(\mathbf{z}_r) = H(\mathbf{V}\mathbf{z}_r)$. Note that in contrast to the first approach, the substitution (7) is not explicitly added here as an additional approximation step but appears naturally through the compatibility between $\mathbf{V}$ and $\mathbf{W}$. Such compatible projections have been constructed in [PM16], [AH17], [AH19] for the special case of $\bar{\mathbf{R}}$ equal to zero and $\bar{\mathbf{J}}$ being a constant symplectic matrix. The approach there goes by the name *symplectic model reduction*. A finite element method using similar ideas can, e.g., be found in [SMH19].

3. Reparametrization It may also be worthwhile to reconsider the parametrization of the solution. An important case is the reformulation of the system in the effort variable $\mathbf{e} = \nabla_{\mathbf{z}} H(\mathbf{z})$. It is well-known that this generally nonlinear change of variable is connected to the so-called Legendre transform of H, which we denote by G. It can be characterized by

$$\nabla_{\mathbf{e}} G(\mathbf{e})_{|\mathbf{e}=\nabla_{\mathbf{z}} H(\mathbf{z})} = \mathbf{z}, \qquad \text{for all } \mathbf{z}.$$

For $\mathbf{e} = \nabla_{\mathbf{z}} H(\mathbf{z})$, let us define the matrix-valued functions

$$\mathbf{Q}(\mathbf{e}) = \nabla_{\mathbf{ee}} G(\mathbf{e}), \qquad \tilde{\mathbf{J}}(\mathbf{e})_{|\mathbf{e}=\nabla_{\mathbf{z}} H(\mathbf{z})} = \bar{\mathbf{J}}(\mathbf{z}), \qquad \tilde{\mathbf{R}}(\mathbf{e})_{|\mathbf{e}=\nabla_{\mathbf{z}} H(\mathbf{z})} = \bar{\mathbf{R}}(\mathbf{z}).$$

The reparametrized version of System 0.3 then reads: Find $\mathbf{e} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$ solving

$$\begin{aligned} \mathbf{Q}(\mathbf{e}(t))\frac{d}{dt}\mathbf{e}(t) &= \left(\tilde{\mathbf{J}}(\mathbf{e}(t)) - \tilde{\mathbf{R}}(\mathbf{e}(t))\right)\mathbf{e}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}^T\mathbf{e}(t), \qquad\qquad \mathbf{e}(0) = \mathbf{e}_0. \end{aligned}$$

When applying a Galerkin projection on this reparametrization with $\mathbf{W} = \mathbf{V}$, one obtains the reduced model

$$\begin{aligned} \mathbf{Q}_r(\mathbf{e}_r(t))\frac{d}{dt}\mathbf{e}_r(t) &= \left(\tilde{\mathbf{J}}_r(\mathbf{e}(t)) - \tilde{\mathbf{R}}_r(\mathbf{e}_r(t))\right)\mathbf{e}_r(t) + \mathbf{B}_r\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}_r^T\mathbf{e}(t), \qquad\qquad \mathbf{e}_r(0) = \mathbf{e}_{r,0} \end{aligned}$$

for the reduced effort $\mathbf{e}_r \in \mathcal{C}^1([0,T];\mathbb{R}^n)$, $\mathbf{B}_r = \mathbf{V}^T\mathbf{B}$ and respective reduced matrix functions

$$\mathbf{Q}_r(\mathbf{e}_r) = \mathbf{V}^T\mathbf{Q}(\mathbf{V}\mathbf{e}_r)\mathbf{V}, \qquad \tilde{\mathbf{J}}_r(\mathbf{e}_r) = \mathbf{V}^T\tilde{\mathbf{J}}(\mathbf{V}\mathbf{e}_r)\mathbf{V}, \qquad \tilde{\mathbf{R}}_r(\mathbf{e}_r) = \mathbf{V}^T\tilde{\mathbf{R}}(\mathbf{V}\mathbf{e}_r)\mathbf{V}.$$

The anti-symmetry of $\tilde{\mathbf{J}}_r(\mathbf{e}_r)$ and the positive semi-definiteness of $\tilde{\mathbf{R}}_r(\mathbf{e}_r)$ easily follow. By construction, the reduced model also inherits a reduced Legendre-transformed

$$G_r(\mathbf{e}_r) = G(\mathbf{V}\mathbf{e}_r), \qquad \text{for which} \qquad \mathbf{Q}_r(\mathbf{e}_r) = \nabla_{\mathbf{e}_r\mathbf{e}_r} G_r(\mathbf{e}_r)$$

and by that the full port-Hamiltonian structure. To see the latter, one can rearrange the system in standard form as in System 0.3 by performing a Legendre transform on G_r and making use of the related state transformation, cf. Part I.B for details. In the linear case, the situation significantly simplifies, see [GPBvdS09], [WLEK10] for model reduction methods and [FKJ$^+$13] for a finite-element discretization. For the nonlinear case, see, e.g., [Egg19].

Relation to our approach Which, if any of the above structure-preserving approximation ansatzes is suitable, depends on the underlying problem. Especially when the underlying problem is linear, the third ansatz, the reparametrization, seems to be the most straight-forward, as it allows for direct Galerkin-approximation. We pursue this approach in Part I.A, where we are faced with a linear model.

On the other hand, when considering the general nonlinear model problem (1), the full reparametrization may be unfeasible. This is, e.g., the case for the barotropic Euler equations (3). Our approximation procedure in Part I.B instead relies on ideas of both, the second approach by compatible reduction bases and the third one by reparametrization. More precisely, we reparametrize parts of the variables and enforce compatibility conditions to handle the remaining variables.

Our complexity reduction method follows the idea of empirical quadrature [HCF17], [Jam08], [FACC14], but we include additional compatibility conditions in order to guarantee port-Hamiltonian structure. To the best of the author's knowledge, the latter has not been proposed before in the context of structure-preserving approximation of port-Hamiltonian systems.

Subpart I.A

Linear damped wave equation

Chapter 1

Preliminaries

1.1 Motivation

Part I.A is devoted to the linear damped wave equation on networks. For a motivation, let us briefly discuss the equations with spatial domain chosen as the interval $(0, l^\omega)$ instead. Given $T > 0$, the pressure and the mass flow, $p, m : [0, T] \times (0, l^\omega) \to \mathbb{R}$, are assumed to be governed by

$$\alpha \partial_t p(t, x) + \partial_x m(t, x) = 0 \tag{1.1a}$$

$$\beta \partial_t m(t, x) + \partial_x p(t, x) = -r m(t, x), \qquad \text{for } x \in (0, l^\omega),\ t \in [0, T]. \tag{1.1b}$$

The parameters α, β, r may depend on the spatial variable x, but we suppress that in our notation. We assume there exist constants $c_{low}, c_{up} > 0$ such that $c_{low} \le \alpha(x), \beta(x), r(x) \le c_{up}$ for $x \in (0, l^\omega)$. This system is a basic instance of our abstract model problem. To rewrite it in the abstract form, let us define the energy variable $\underline{\mathbf{z}}$ and the Hamiltonian density $h : \mathbb{R}^2 \mapsto \mathbb{R}$, as

$$\underline{\mathbf{z}} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \alpha p \\ \beta m \end{bmatrix}, \qquad h([z_1; z_2]) = \frac{1}{2} \left(\frac{1}{\alpha} z_1^2 + \frac{1}{\beta} z_2^2 \right).$$

Defining the effort variable and the constant Hessian of the Hamiltonian density as

$$\underline{\mathbf{e}} = \nabla_{\mathbf{z}} h(\underline{\mathbf{z}}) = [p; m], \qquad \underline{\mathbf{Q}} = \nabla_{\mathbf{zz}} h(\underline{\mathbf{z}}) = \begin{bmatrix} \frac{1}{\alpha} & \\ & \frac{1}{\beta} \end{bmatrix},$$

one can see that (1.1) corresponds to a formulation in effort variables, i.e.,

$$\underline{\mathbf{Q}}^{-1} \partial_t \underline{\mathbf{e}}(t, x) = \begin{bmatrix} & -\partial_x \\ -\partial_x & -r \end{bmatrix} \underline{\mathbf{e}}(t, x).$$

Next, let us sketch the weak form underlying our approximation schemes. The function spaces $\mathcal{L}^2((0, l^\omega))$ and $\mathcal{H}^1((0, l^\omega))$ are defined as the Sobolev space of square integrable functions on $(0, l^\omega)$ and the Sobolev space with additionally square integrable weak derivatives, respectively. The $\mathcal{L}^2$-scalar product and norm for $b, \tilde{b} \in \mathcal{L}^2((0, l^\omega))$ is written as

$$\langle b, \tilde{b} \rangle = \int_{(0, l^\omega)} b(x) \tilde{b}(x) dx \qquad \text{and} \qquad ||b|| = \sqrt{\langle b, b \rangle}.$$

Given $b \in \mathcal{H}^1((0, l^\omega))$, we can associate boundary values $b[\nu] \in \mathbb{R}$ to $\nu \in \{0, l^\omega\}$ by means of the trace theorem, [Bra07]. By that, we can define a linear bounded trace operator

$$\mathcal{T} : \mathcal{H}^1((0, l^\omega)) \to \mathbb{R}^2, \qquad \mathcal{T}b = \begin{bmatrix} -b[l^\omega] \\ b[0] \end{bmatrix},$$

as recalled later in Definition 2.1. In anticipation of the port-Hamiltonian formulations, we introduce additional variables acting on the boundaries of the spatial domain, the so-called boundary effort and boundary flow

$$\mathbf{e}_B = \begin{bmatrix} e_B[l^\omega] \\ e_B[0] \end{bmatrix}, \qquad \mathbf{f}_B = \begin{bmatrix} f_B[l^\omega] \\ f_B[0] \end{bmatrix}, \qquad \mathbf{e}_B(t), \mathbf{f}_B(t) \in \mathbb{R}^2, \text{ for } t \in [0, T].$$

These degrees of freedom are chosen such that $\mathbf{e}_B = [p[l^\omega]; p[0]]$ and $\mathbf{f}_B = [-m[l^\omega]; m[0]]$ for smooth solutions. By testing (1.1) with appropriate test-functions, taking the integral and using integration by parts once on the second equation, we obtain the following weak form:

Find $p \in \mathcal{C}^1([0,T]; \mathcal{L}^2((0, l^\omega)))$, $m \in \mathcal{C}^1([0,T]; \mathcal{H}^1(0, l^\omega))$ fulfilling

$$\begin{aligned} \langle \alpha \partial_t p(t), b_1 \rangle &= -\langle \partial_x m(t), b_1 \rangle \\ \langle \beta \partial_t m(t), b_2 \rangle &= \langle p(t), \partial_x b_2 \rangle - \langle r m(t,x), b_2 \rangle + \mathbf{e}_B(t) \cdot \mathcal{T} b_2(t) \\ \mathbf{f}_B(t) &= \mathcal{T} m(t) \end{aligned}$$

for all $b_1 \in \mathcal{L}^2((0, l^\omega))$, $b_2 \in \mathcal{H}^1((0, l^\omega))$. The boundary term in the second equation originates from partial integration. To close the system, data $p_0, m_0 : (0, l^\omega) \to \mathbb{R}$ prescribing the initial conditions $p(0) = p_0$ and $m(0) = m_0$ as well as boundary data $u^\nu : [0, T] \to \mathbb{R}$ for $\nu \in \{0, l^\omega\}$ are assumed to be given. One boundary condition per ν is prescribed given by either one of the following types:

$$\textit{Type 1: } e_B[\nu] = u^\nu, \qquad \textit{Type 2: } f_B[\nu] = n[\nu] u^\nu, \qquad\qquad t \in [0, T],$$

with $n[\nu]$ describing the inner normal vector, i.e., $n[0] = 1$, $n[l^\omega] = -1$. Type 1 relates to pressure boundary conditions and is weakly imposed in our weak form, whereas Type 2 has the interpretation of a mass-inflow boundary condition and takes the role of a so-called essential or Dirichlet-boundary condition for our weak form. Next, we introduce the necessary concepts for a formulation of the network problem. First the description of the spatial topology, the network graph, is discussed, and then the Sobolev spaces are generalized to the network case. For this purpose we employ the framework derived in [EK18], [Kug19].

1.2 Network topology

Let a set of nodes $\mathcal{N} = \{\nu_1, \dots, \nu_l\}$ and edges $\mathcal{E} = \{\omega_1, \dots, \omega_k\} \subset \mathcal{N} \times \mathcal{N}$ be given. To every edge ω, we additionally associate a positive length l^ω, and collect all lengths in the set $l = \{l^{\omega_1}, \dots, l^{\omega_k}\}$. The tuple $\mathcal{G} = (\mathcal{N}, \mathcal{E}, l)$ describes a directed graph. For $\omega \in \mathcal{E}$ and $\nu \in \mathcal{N}$, the incidence mapping $n^\omega[\nu]$ is defined by

$$n^\omega[\nu] = \begin{cases} 1 & \text{for } \omega = (\nu, \bar{\nu}) \text{ for some } \bar{\nu} \in \mathcal{N} \\ -1 & \text{for } \omega = (\bar{\nu}, \nu) \text{ for some } \bar{\nu} \in \mathcal{N} \\ 0 & \text{else.} \end{cases}$$

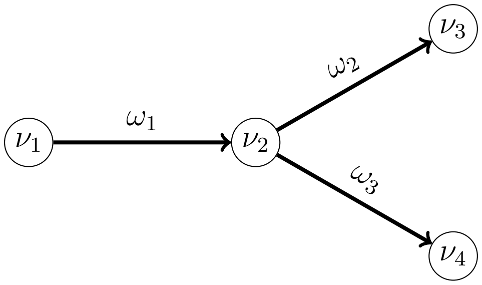

Figure 1.1: Graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ with vertices $\mathcal{N} = \{\nu_1, \nu_2, \nu_3, \nu_4\}$ and edges $\mathcal{E} = \{\omega_1, \omega_2, \omega_3\}$ defined by $\omega_1 = (\nu_1, \nu_2)$, $\omega_2 = (\nu_2, \nu_3)$ and $\omega_3 = (\nu_2, \nu_4)$. Consequently, $\mathcal{N}_0 = \{\nu_2\}$, $\mathcal{N}_\partial = \{\nu_1, \nu_3, \nu_4\}$ and $\mathcal{E}(\nu_2) = \{\omega_1, \omega_2, \omega_3\}$. Furthermore, $n^{\omega_1}[\nu_1] = n^{\omega_2}[\nu_2] = n^{\omega_3}[\nu_2] = 1$ and $n^{\omega_1}[\nu_2] = n^{\omega_2}[\nu_3] = n^{\omega_3}[\nu_4] = -1$.

Remark 1.1. *The matrix $\mathbf{N} \in \mathbb{R}^{l,k}$ with components $[\mathbf{N}]_{i,j} = -n^{\omega_j}[\nu_i]$ is known as the incidence matrix of the graph, cf. [LMT13], [DHLT17].*

The set of all edges adjacent to the node ν is denoted by

$$\mathcal{E}(\nu) = \{\omega \in \mathcal{E} : \omega = (\nu, \bar{\nu}), \text{ or } \omega = (\bar{\nu}, \nu)\}.$$

Furthermore, the nodes are grouped into interior nodes $\mathcal{N}_0 \subset \mathcal{N}$ and boundary nodes $\mathcal{N}_\partial = \mathcal{N} \backslash \mathcal{N}_0$. The network notations are illustrated in Fig. 1.1. Throughout the manuscript, we make the following additional assumptions on the network topology:

Assumption 1.2.

A1) The graph $(\mathcal{N}, \mathcal{E})$ is connected, finite and directed.

A2) It holds $|\mathcal{E}(\nu)| = 1$ for all boundary nodes $\nu \in \mathcal{N}_\partial$.

1.3 Function spaces

Let us now introduce the function spaces defined on the network topology. They consist of compositions of standard Sobolev spaces for every edge. Note that every edge $\omega \in \mathcal{E}$ can be naturally identified with the interval $(0, l^\omega)$ with l^ω being its length, which we tacitly do for all upcoming integral-expressions. The spatial domain Ω consists then of the union of edges,

$$\Omega = \{x : x \in \omega, \text{ for } \omega \in \mathcal{E}\}.$$

The space of square-integrable functions on $\mathcal{E}$ is then defined as

$$\mathcal{L}^2(\mathcal{E}) = \left\{b : \Omega \to \mathbb{R} \text{ with } b_{|\omega} \in \mathcal{L}^2(\omega) \text{ for all } \omega \in \mathcal{E}\right\}.$$

Given $b, \tilde{b} \in \mathcal{L}^2(\mathcal{E})$, the scalar product and norm then read

$$\langle b, \tilde{b}\rangle_{\mathcal{E}} = \sum_{\omega \in \mathcal{E}} \langle b_{|\omega}, \tilde{b}_{|\omega}\rangle_\omega \qquad \text{and} \qquad ||b||_{\mathcal{E}} = \sqrt{\langle b, b\rangle_{\mathcal{E}}}.$$

The subscript for the domain will typically be suppressed, when it is clear from the context. Further, a broken derivative operator $\partial_x{}'$ is introduced as the edgewise weak derivative

$$(\partial_x{}'b)_{|\omega} = \partial_x b_{|\omega}, \qquad \text{for all } \omega \in \mathcal{E}.$$

For convenience, we simplify the notation $\partial_x{}'$ to ∂_x, as the broken derivative operator is the only derivative operator we use on the network domain. We then define the space of functions with square-integrable weak broken derivative on $\mathcal{E}$ by

$$\mathcal{H}^1_{pw}(\mathcal{E}) = \left\{ b \in \mathcal{L}^2(\mathcal{E}) : \partial_x b \in \mathcal{L}^2(\mathcal{E}) \right\}.$$

A natural identification of the two introduced spaces with product spaces is given by

$$\mathcal{L}^2(\mathcal{E}) \cong \prod_{\omega\in\mathcal{E}} \mathcal{L}^2(\omega) \cong \prod_{\omega\in\mathcal{E}} \mathcal{L}^2((0,l^\omega)) \qquad \text{and} \qquad \mathcal{H}^1_{pw}(\mathcal{E}) \cong \prod_{\omega\in\mathcal{E}} \mathcal{H}^1(\omega) \cong \prod_{\omega\in\mathcal{E}} \mathcal{H}^1((0,l^\omega)).$$

The function space of piecewise smooth functions are defined accordingly for $k \geq 0$,

$$\mathcal{C}^k_{pw}(\mathcal{E}) = \left\{ b : \Omega \to \mathbb{R} \text{ with } b_{|\omega} \in \mathcal{C}^k(\omega) \text{ for all } \omega \in \mathcal{E} \right\}.$$

Finally, note that the spatial domains for boundary- and coupling-conditions are the sets of nodes $\mathcal{N}_\partial$ and $\mathcal{N}_0$. We tacitly identify the function space acting on these sets by the Euclidean vector space $\mathbb{R}^{|\mathcal{N}_\partial|}$ and $\mathbb{R}^{|\mathcal{N}_0|}$ equipped with the standard Euclidean scalar product and norm. By means of the trace theorem, [Bra07], we can define an operator for the values of the boundary nodes as follows.

Definition 1.3. *For $b \in \mathcal{H}^1_{pw}(\mathcal{E})$ and a boundary node $\nu \in \mathcal{N}_\partial = \{\nu_1, \ldots, \nu_p\}$, we denote by $b[\nu]$ the evaluation of the function at the boundary in the sense of the trace theorem. The boundary trace operator $\mathcal{T} : \mathcal{H}^1_{pw}(\mathcal{E}) \to \mathbb{R}^p$ is then introduced componentwise as*

$$\left[\mathcal{T}b\right]_i = n^\omega[\nu_i] b_{|\omega}[\nu_i] \qquad \text{for } \omega \in \mathcal{E}(\nu_i), \quad i = 1, \ldots, p \quad \text{and} \quad b \in H^1_{pw}(\mathcal{E}).$$

Further, boundary efforts and boundary flows $\mathbf{e}_B, \mathbf{f}_B \in \mathbb{R}^p$ associated to all boundary nodes of the network are introduced with components

$$\mathbf{e}_B = \left[e_B[\nu_1]; \ldots; e_B[\nu_p]\right], \qquad \mathbf{f}_B = \left[f_B[\nu_1]; \ldots; f_B[\nu_p]\right].$$

The coupling conditions for the network generalization consist of conservation of mass and continuity of pressure on the inner nodes,

$$\sum_{\omega\in\mathcal{E}(\nu)} n^\omega[\nu] m_{|\omega}[\nu] = 0 \qquad \text{for } \nu \in \mathcal{N}_0, \qquad p_{|\omega}[\nu] = p_{|\tilde{\omega}}[\nu] \qquad \text{for } \omega, \tilde{\omega} \in \mathcal{E}(\nu),\ \nu \in \mathcal{N}_0.$$

To incorporate them, it is useful to introduce an additional operator acting on the node values.

Definition 1.4. *For $\omega = (\nu, \nu') \in \mathcal{E}$ and $\phi \in \mathcal{H}^1(\omega)$, let $\phi[\nu]$, $\phi[\nu']$ denote the evaluations of ϕ in terms of the trace theorem. Let further, $\mathcal{N}_0 = \{\nu_1, \ldots, \nu_q\}$. Then $\mathcal{T}^{\mathcal{N}_0} : \mathcal{H}^1_{pw}(\mathcal{E}) \to \mathbb{R}^q$ is defined componentwise by*

$$\left[\mathcal{T}^{\mathcal{N}_0} b\right]_i = \sum_{\omega\in\mathcal{E}(\nu_i)} n^\omega[\nu_i] b_{|\omega}[\nu_i] \qquad \text{for } i = 1, \ldots, q, \qquad b \in H^1_{pw}(\mathcal{E}).$$

Instead of incorporating the coupling conditions explicitly by the operator from Definition 1.4, one can also include them in the ansatz spaces directly. Following [EK18], we define

$$\mathcal{H}^1_{div}(\mathcal{E}) = \left\{ b \in \mathcal{H}^1_{pw}(\mathcal{E}) : \sum_{\omega \in \mathcal{E}(\nu)} n^\omega[\nu] b_{|\omega}[\nu] = 0, \text{ for } \nu \in \mathcal{N}_0 \right\}.$$

Remark 1.5. *By construction, it holds* $\mathcal{H}^1_{div}(\mathcal{E}) = \left\{ b \in \mathcal{H}^1_{pw}(\mathcal{E}) : \mathcal{T}^{\mathcal{N}_0} b = \mathbf{0} \right\}$

Chapter 2

Approximation procedure

The approximation framework for the damped wave equation on networks is presented in this chapter. In Section 2.1 the network problem is introduced. Its weak form and Galerkin approximation are then addressed in Section 2.2 and Section 2.3, respectively. The latter can, e.g., be realized by finite elements, cf. Section 2.4. The remaining two sections are devoted to the structure of the Galerkin approximation on the level of coordinate representations, Section 2.5, and the construction of formulations, in which the algebraic equations relating to coupling- and boundary-conditions are fully or partly eliminated, Section 2.6.

2.1 Network problem

Let a directed graph $(\mathcal{N}, \mathcal{E}, l)$ and parameters $\alpha, \beta, r : \Omega \to \mathbb{R}^+ \backslash \{0\}$ be given with $c_{low} \leq \alpha(x), \beta(x), r(x) \leq c_{up}$ for $x \in \Omega$ and constants $c_{low}, c_{up} > 0$. The edgewise pressures and mass flows, $p^\omega, m^\omega : [0,T] \times \omega \to \mathbb{R}$ for $\omega \in \mathcal{E}$, are assumed to be governed by the damped wave equation

$$\alpha \partial_t p^\omega(t,x) + \partial_x m^\omega(t,x) = 0 \tag{2.1a}$$

$$\beta \partial_t m^\omega(t,x) + \partial_x p^\omega(t,x) = -r m^\omega(t,x), \qquad \text{for } x \in \omega, \quad \omega \in \mathcal{E} \text{ and } t \in [0,T]. \tag{2.1b}$$

They are interconnected at $\nu \in \mathcal{N}_0$ by the coupling conditions

$$\sum_{\omega \in \mathcal{E}(\nu)} n^\omega[\nu] m^\omega(t,\nu) = 0 \qquad \text{for } \nu \in \mathcal{N}_0, \qquad p^\omega(t,\nu) = p^{\tilde{\omega}}(t,\nu) \quad \text{for } \omega, \tilde{\omega} \in \mathcal{E}(\nu), \tag{2.1c}$$

and one boundary condition for each boundary node $\nu \in \mathcal{N}_\partial$ is set, each of one of the types

$$\textit{Type 1: } p^\omega(t,\nu) = u^\nu(t), \qquad \textit{Type 2: } n^\omega[\nu] m^\omega(t,\nu) = u^\nu(t), \qquad \text{for } \omega \in \mathcal{E}(\nu), \tag{2.1d}$$

for $t \in [0,T]$. To set up a first weak formulation, we introduce auxiliary edgewise boundary operators and boundary ports.

Definition 2.1 (Edgewise boundary ports)**.** *Let $\omega = (\nu, \tilde{\nu}) \in \mathcal{E}$ and $\phi \in \mathcal{H}^1(w)$. Let $\phi[\nu]$ and $\phi[\tilde{\nu}]$ denote the evaluations of ϕ at the respective boundary node in the sense of the trace theorem. Then we define*

$$\mathcal{T}^\omega : \mathcal{H}^1(\omega) \to \mathbb{R}^2, \qquad \mathcal{T}^\omega \phi = \begin{bmatrix} -\phi[\tilde{\nu}] \\ \phi[\nu] \end{bmatrix}.$$

Furthermore, boundary efforts and boundary flows $\mathbf{e}_B^\omega(t), \mathbf{f}_B^\omega(t) \in \mathbb{R}^2$ *for* $t \in [0,T]$*, are introduced with components* $\mathbf{e}_B^\omega = [e_B^\omega[\tilde{\nu}]; e_B^\omega[\nu]]$ *and* $\mathbf{f}_B^\omega = [f_B^\omega[\tilde{\nu}]; f_B^\omega[\nu]]$ *associated to the nodes* $\nu, \tilde{\nu}$*.*

The weak form of the network system then reads:
Find $p \in \mathcal{C}^1([0,T];\mathcal{L}^2(\mathcal{E}))$, $m \in \mathcal{C}^1([0,T];\mathcal{H}^1_{pw}(\mathcal{E}))$, $\mathbf{e}_B^\omega \in C^0([0,T];\mathbb{R}^2)$, $\mathbf{f}_B^\omega \in \mathcal{C}^1([0,T];\mathbb{R}^2)$ and $\lambda^\nu \in C^0([0,T];\mathbb{R})$ for all $\omega \in \mathcal{E}$, $\nu \in \mathcal{N}_0$ fulfilling for each $\omega \in \mathcal{E}$

$$\langle \alpha \partial_t p(t), b_1 \rangle_{\mathcal{E}} = -\langle \partial_x m(t), b_1 \rangle_{\mathcal{E}} \tag{2.2a}$$

$$\langle \beta \partial_t m(t), b_2 \rangle_{\mathcal{E}} = \langle p(t), \partial_x b_2 \rangle_{\mathcal{E}} - \langle r m(t), b_2 \rangle_{\mathcal{E}} + \sum_{\omega \in \mathcal{E}} \mathbf{e}_B^\omega(t) \cdot \mathcal{T}^\omega b_{2|\omega} \tag{2.2b}$$

$$\mathbf{f}_B^\omega(t) = \mathcal{T}^\omega m_{|\omega}(t), \qquad \text{for all } \omega \in \mathcal{E}, \tag{2.2c}$$

for all $b_1 \in \mathcal{L}^2(\mathcal{E})$, $b_2 \in \mathcal{H}^1_{pw}(\mathcal{E})$, and $t \in [0,T]$, and the coupling conditions

$$\sum_{\omega \in \mathcal{E}(\nu)} f_B^\omega[\nu] = 0 \quad \text{for } \nu \in \mathcal{N}_0, \qquad e_B^\omega[\nu] = \lambda^\nu \quad \text{for } \omega \in \mathcal{E}(\nu),\ \nu \in \mathcal{N}_0. \tag{2.2d}$$

The system then has to be complemented with one boundary condition per boundary node and initial conditions.

Lemma 2.2. *Let* $p^\omega, m^\omega \in \mathcal{C}^1([0,T];\mathcal{C}^1(\omega))$ *for* $\omega \in \mathcal{E}$ *be a solution of* (2.1)*. Then it also fulfills the variational principle* (2.2) *with*

$$\mathbf{e}_B^\omega = \begin{bmatrix} p^\omega[\nu'] \\ p^\omega[\nu] \end{bmatrix}, \qquad \mathbf{f}_B^\omega = \mathcal{T}^\omega m^\omega = \begin{bmatrix} -m^\omega[\nu'] \\ m^\omega[\nu] \end{bmatrix}, \qquad \textit{for } \omega = (\nu, \nu').$$

Conversely, any sufficiently smooth solution of the weak problem describes a solution of equation (2.1)*.*

Proof. Let $p, m \in \mathcal{C}^1([0,T];\mathcal{C}^1(\mathcal{E}))$ be defined edgewise by the solution components p^ω, m^ω of (2.1) for $\omega \in \mathcal{E}$. Let $\omega = (\nu, \nu') \in \mathcal{E}$. Testing (2.1b) with $b_2 \in \mathcal{H}^1(\omega)$ and integration by parts then yields

$$\begin{aligned} \langle \beta \partial_t m(t), b_2 \rangle_\omega &= -\langle \partial_x p(t), b_2 \rangle_\omega - \langle r m(t), b_2 \rangle_\omega = \\ &= \langle p(t), \partial_x b_2 \rangle_\omega - \langle r m(t), b_2 \rangle_\omega - \big(p^\omega[\nu'](t) b_2[\nu'] - p^\omega[\nu](t) b_{2|\omega}[\nu]\big) \\ &= \langle p(t), \partial_x b_2 \rangle_\omega - \langle r m(t), b_2 \rangle_\omega + \mathbf{e}_B^\omega(t) \cdot \mathcal{T}^\omega b_{2|\omega}, \end{aligned}$$

i.e., equation (2.2b). Equation (2.2a) and (2.2c) follow immediately from the strong form. Summarizing, this shows that a strong solution fulfills the variational principle. On the other hand, given a solution of (2.2) with $p, m \in \mathcal{C}^1([0,T];\mathcal{C}^1(\mathcal{E}))$, we can go through the reverse steps to show the converse result. □

2.2 Weak form

Throughout, we assume that at each boundary node $\nu \in \mathcal{N}_\partial$ one of the following boundary conditions is prescribed, as well as initial conditions given as follows:

$$\begin{aligned} &\textit{Type 1: } e_B[\nu] = u^\nu \in \mathcal{C}([0,\infty),\mathbb{R}), \qquad \textit{Type 2: } f_B[\nu] = u^\nu \in \mathcal{C}^1([0,\infty),\mathbb{R}) \\ &\underline{\mathrm{e}}(0) = \underline{\mathrm{e}}_0, \quad u^\nu : [0,T] \to \mathbb{R} \quad \text{given such that } \begin{bmatrix} \mathcal{T} \\ \mathcal{T}^{\mathcal{N}_0} \end{bmatrix} e_2(0) = \begin{bmatrix} \mathbf{f}_B(0) \\ \mathbf{0} \end{bmatrix}, \\ &\text{and at least one } \nu \in \mathcal{N}_\partial \text{ with boundary conditions of Type 1.} \end{aligned} \tag{2.3}$$

An equivalent, more compact formulation of (2.2) reads as follows.

System 2.3. *Find $p \in \mathcal{C}^1([0,T];\mathcal{L}^2(\mathcal{E}))$, $m \in \mathcal{C}^1([0,T];\mathcal{H}^1_{pw}(\mathcal{E}))$, $\mathbf{e}_B \in C^0([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$, $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ and $\boldsymbol{\lambda} \in C^0([0,T];\mathbb{R}^{|\mathcal{N}_0|})$ fulfilling*

$$\begin{aligned} \langle \alpha\partial_t p(t), b_1\rangle_{\mathcal{E}} &= -\langle \partial_x m(t), b_1\rangle_{\mathcal{E}} \\ \langle \beta\partial_t m(t), b_2\rangle_{\mathcal{E}} &= \langle p(t), \partial_x b_2\rangle_{\mathcal{E}} - \langle rm(t), b_2\rangle_{\mathcal{E}} + \boldsymbol{\lambda}(t)\cdot\mathcal{T}^{\mathcal{N}_0} b_2 + \mathbf{e}_B(t)\cdot\mathcal{T} b_2 \\ \mathbf{0} &= \mathcal{T}^{\mathcal{N}_0} m(t) \\ \mathbf{f}_B(t) &= \mathcal{T} m(t) \end{aligned}$$

for all $b_1 \in \mathcal{L}^2(\mathcal{E})$, $b_2 \in \mathcal{H}^1_{pw}(\mathcal{E})$, and $t \in [0,T]$. Closing conditions (2.3) *are posed.*

The Hamiltonian and the total mass of the network system read

$$\begin{aligned} \mathcal{H}([p;m]) &= \left\langle \frac{1}{2}(\alpha p^2 + \beta m^2), 1 \right\rangle_{\mathcal{E}} = \sum_{\omega\in\mathcal{E}} \int_\omega \frac{1}{2}(\alpha p^2 + \beta m^2)dx \\ \mathcal{M}([p;m]) &= \langle \alpha p, 1\rangle_{\mathcal{E}} = \sum_{\omega\in\mathcal{E}} \int_\omega \alpha p dx. \end{aligned}$$

As we show now, the variational principle of System 2.3 readily implies dissipation of the Hamiltonian and mass conservation.

Theorem 2.4 (Energy-dissipation equality)**.** *For any solution of the weak form System 2.3, it holds for $t \in [0,T]$*

$$\frac{d}{dt}\mathcal{H}([p(t);m(t)]) = \mathbf{e}_B(t)\cdot\mathbf{f}_B(t) - \langle rm(t), m(t)\rangle_{\mathcal{E}} < \mathbf{e}_B(t)\cdot\mathbf{f}_B(t).$$

Proof. By elementary calculations and the variational principle of System 2.3, it holds

$$\begin{aligned} \frac{d}{dt}\frac{1}{2}\int_\omega \alpha p^2 + \beta m^2 dx &= \int_\omega (\alpha\partial_t p)p + (\beta\partial_t m)m dx \\ &= \int_\omega (-\partial_x m)p + (p\partial_x m - rm^2)dx + \mathbf{e}_B^\omega\cdot\mathcal{T}^\omega m_{|\omega} \\ &= -\int_\omega rm^2 dx + \mathbf{e}_B^\omega\cdot\mathcal{T}^\omega m_{|\omega}. \end{aligned}$$

Summing over all edges $\omega \in \mathcal{E}$, using the coupling conditions of System 2.3 and the definition of the Hamiltonian $\mathcal{H}$, we get

$$\begin{aligned}\frac{d}{dt}\mathcal{H}([p;m]) &= \sum_{\omega\in\mathcal{E}} \frac{d}{dt}\frac{1}{2}\int_\omega (\alpha p^2 + \beta m^2) dx = \sum_{\omega\in\mathcal{E}} \int_\omega -rm^2 dx + \mathbf{e}_B^\omega \cdot \mathcal{T}^\omega m_{|\omega} \\ &= -\left\langle r, m^2 \right\rangle_{\mathcal{E}} + \mathbf{e}_B \cdot \mathcal{T} m = -\left\langle r, m^2 \right\rangle_{\mathcal{E}} + \mathbf{e}_B \cdot \mathbf{f}_B < \mathbf{e}_B \cdot \mathbf{f}_B.\end{aligned}$$

□

Theorem 2.5 (Global mass conservation). *For any solution of the weak form, System 2.3, it holds for $t \in [0,T]$,*

$$\frac{d}{dt}\mathcal{M}([p(t); m(t)]) = \sum_{\bar\nu \in \mathcal{N}_\partial} f_B[\bar\nu](t).$$

Proof. The change of mass in time on one edge $\omega = (\nu, \tilde\nu)$ is given as

$$\frac{d}{dt}\int_\omega \alpha p \, dx = -\int_\omega \partial_x m \, dx = -m[\tilde\nu] + m[\nu] = f_B^\omega[\tilde\nu] + f_B^\omega[\nu] = \mathbf{f}_B^\omega \cdot [1;1].$$

By summing over all edges $\omega \in \mathcal{E}$ and utilizing the coupling conditions of System 2.3, we can then conclude

$$\frac{d}{dt}\mathcal{M}([p;m]) = \frac{d}{dt}\langle \alpha p, 1\rangle = \sum_{\omega\in\mathcal{E}} \int_\omega \partial_t(\alpha p) dx = \sum_{\omega\in\mathcal{E}} \mathbf{f}_B^\omega \cdot [1;1] = \sum_{\bar\nu\in\mathcal{N}_\partial} f_B[\bar\nu].$$

□

Moreover, existence of unique steady states as well as exponential decay to steady states for constant boundary data have been shown in [EK18]. The respective results are shown by variational arguments, which carry over to appropriate Galerkin approximation and are uniform with respect to the discretization parameters. We therefore only state them on this level, cf. Lemma 2.12 and Lemma 2.13. Let us mention that also a local mass conservation property holds for the continuous solution.

Lemma 2.6. *Let $\omega \in \mathcal{E}$, and $[w_1, w_2] \subset \omega$ be a subpart of the edge. Then it holds for solutions of System 2.3 that*

$$\frac{d}{dt}\int_{[w_1,w_2]} \alpha p(t) \, dx = m(t)[w_1] - m(t)[w_2].$$

Proof. The assertion follows by testing the first equation of the weak form with $b_1 = \chi_{[w_1,w_2]}$, the indicator function of the domain $[w_1, w_2]$. □

2.3 Galerkin approximation

Space discretizations as well as reduced models are constructed by Galerkin approximation of System 2.3. The main tool for the derivation of the theoretical results are the following compatibility conditions on the ansatz spaces, see [EKLS+18].

Assumption 2.7 (Compatibility of spaces for System 2.9)**.** *Let $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, and let $\mathcal{V}_1 \subset \mathcal{L}^2(\mathcal{E})$ and $\mathcal{V}_2 \subset \mathcal{H}^1_{pw}(\mathcal{E})$ be finite dimensional subspaces fulfilling the compatibility conditions*

A1) $\mathcal{V}_1 = \partial_x \mathcal{V}_2, \qquad$ with $\partial_x \mathcal{V}_2 = \{\xi : \text{It exists } \zeta \in \mathcal{V}_2 \text{ with } \partial_x \zeta = \xi\}$

A2) $1_{\mathcal{E}} \in \mathcal{V}_1$, with $1_{\mathcal{E}} : \Omega \to \mathbb{R}$ and $1_{\mathcal{E}}(x) = 1$ for $x \in \Omega$

A3) $\{b_2 \in \mathcal{H}^1_{pw}(\mathcal{E}) : \partial_x b_2 = 0\} \subset \mathcal{V}_2$.

Remark 2.8. *The space $\{b_2 \in \mathcal{H}^1_{pw}(\mathcal{E}) : \partial_x b_2 = 0\}$ in Assumption 2.7-(A3) is spanned by the edgewise constant functions, i.e., its dimension is equal to the number of edges.*

We are now in the position to set up the Galerkin approximation for the network system.

System 2.9 (Galerkin-approximation of System 2.3)**.** *Find $p \in \mathcal{C}^1([0,T];\mathcal{V}_1)$, $m \in \mathcal{C}^1([0,T];\mathcal{V}_2)$, $\mathbf{e}_B \in C^0([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$, $\mathbf{f}_B^\omega \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ and $\boldsymbol{\lambda} \in C^0([0,T];\mathbb{R}^{|\mathcal{N}_0|})$ fulfilling*

$$\begin{aligned}
\langle \alpha \partial_t p(t), b_1 \rangle_{\mathcal{E}} &= -\langle \partial_x m(t), b_1 \rangle_{\mathcal{E}} \\
\langle \beta \partial_t m(t), b_2 \rangle_{\mathcal{E}} &= \langle p(t), \partial_x b_2 \rangle_{\mathcal{E}} - \langle r m(t), b_2 \rangle_{\mathcal{E}} + \boldsymbol{\lambda}(t) \cdot \mathcal{T}^{\mathcal{N}_0} b_2 + \mathbf{e}_B(t) \cdot \mathcal{T} b_2 \\
\mathbf{0} &= \mathcal{T}^{\mathcal{N}_0} m(t) \\
\mathbf{f}_B(t) &= \mathcal{T} m(t)
\end{aligned}$$

for all $b_1 \in \mathcal{V}_1$, $b_2 \in \mathcal{V}_2$ and $t \in [0,T]$ and closing conditions (2.3)*. Assumption 1.2 and Assumption 2.7 are supposed to hold.*

With essentially the same arguments as on the continuous level, we can shown energy dissipation and global mass conservation. We therefore state the results without repeating the proofs.

Theorem 2.10 (Energy-dissipation equality)**.** *For any solution of System 2.9, it holds for $t \in [0,T]$,*

$$\frac{d}{dt}\mathcal{H}([p(t);m(t)]) = \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r m(t), m(t) \rangle_{\mathcal{E}} < \mathbf{e}_B(t) \cdot \mathbf{f}_B(t).$$

Theorem 2.11 (Global mass conservation)**.** *For any solution of System 2.9 it holds for $t \in [0,T]$,*

$$\frac{d}{dt}\mathcal{M}([p(t);m(t)]) = \frac{d}{dt}\sum_{\omega \in \mathcal{E}} \int_\omega p(t)dx = \sum_{\bar{\nu} \in \mathcal{N}_\partial} f_B[\bar{\nu}](t).$$

Note that the global mass conservation is ensured by Assumption 2.7-(A2). Local mass conservation can, however, in general not be shown in the same way as done on the continuous level, cf. Lemma 2.6, and does not necessarily hold. On the other hand, the assumptions are sufficient to show exponential decay of the energy with constants uniform in the discretization parameters.

Lemma 2.12 (Uniform exponential stability)**.** *Let all boundary conditions of System 2.9 be zero, i.e.,* $u^\nu \equiv 0$ *for all* $\nu \in \mathcal{N}_\partial$*. Then it holds for the solution*

$$\mathcal{H}([p(t); m(t)]) \leq C \exp(-\gamma(t-s))\mathcal{H}([p(s); m(s)]), \qquad \textit{for } t \geq s \geq 0,$$

with constants $C, \gamma > 0$*, which are independent of the discretization parameters and the closing conditions* (2.3)*.*

The result has been stated in [EKLS$^+$18, Lemma A.2.6] and [EK18, Theorem 7.4]. By the help of exponential decay, one can also derive the following, see [EKLS$^+$18, Lemma A.2.7] and [EK18, Theorem 6.3].

Lemma 2.13 (Unique steady states)**.** *Let all boundary conditions of System 2.9 be chosen constant in time, i.e.,* $u^\nu \equiv$ *constant for all* $\nu \in \mathcal{N}_\partial$*. Then the system omits a unique steady state, and the solution of the system converges to this steady state for any choice of initial conditions in* (2.3)*.*

Remark 2.14. *With the help of the uniform exponential stability, uniform a-priori error bounds have been derived for the Galerkin approximation, System 2.9, see [EK18, Theorem 7.5] and [EK18, Theorem 8.5]. Given the continuous solution, i.e., the solution of System 2.3, is smooth enough, the error of the Galerkin-approximation can be bounded pointwise in time by appropriate norms of the continuous solution and its time derivative.*

2.4 Finite element spaces

As a particular choice of approximation spaces fulfilling Assumption 2.7, we employ mixed finite element methods. Given an interval $(0, l^\omega)$, a natural choice of ansatz spaces are piecewise polynomial functions. As the identification of edges ω with intervals $(0, l^\omega)$ is possible, we can similarly construct ansatz functions, which are piecewise polynomial on each edge. Given a uniform partitioning $T_\omega = \{T_{\omega,k} : k = 1, \ldots, K_\omega\}$ for the edge ω, where $T_{\omega,k}$ can be identified with sub-intervals of $(0, l^\omega)$, we define

$$\mathcal{Q}_q(T_{\omega,k}; \mathbb{R}) = \left\{ \phi : T_{\omega,k} \to \mathbb{R} : \ \phi(x) = \sum_{j=0}^{q} x^j \xi_j, \ \text{for } x \in T_{\omega,k}, \ \text{with } \xi_j \in \mathbb{R} \right\}$$
$$\mathcal{Q}_q(T_\omega) = \left\{ \phi : \omega \to \mathbb{R} : \ \phi_{|T_{\omega,k}} \in \mathcal{Q}_q(T_{\omega,k}; \mathbb{R}), \ \text{for } k = 1 \ldots, K_\omega \right\}.$$

For the partition of all edges $T_\mathcal{E} = \{T_\omega : \omega \in \mathcal{E}\}$ we define spaces of piecewise polynomial ansatz functions on the network

$$\mathcal{Q}_q(T_\mathcal{E}) = \left\{ \phi : \Omega \to \mathbb{R} : \ \phi_{|\omega} \in \mathcal{Q}_q(T_\omega), \ \text{for } \omega \in \mathcal{E} \right\}, \qquad \mathcal{P}_q(T_\mathcal{E}) = \mathcal{Q}_q(T_\mathcal{E}) \cap \mathcal{H}^1_{pw}(\mathcal{E}).$$

Note that $\mathcal{Q}_q(T_\mathcal{E})$ is by construction a subspace of $\mathcal{L}^2(\mathcal{E})$, but not of $\mathcal{H}^1_{pw}(\mathcal{E})$. The latter only contains functions continuous along each edge, cf. [Bra07]. Our particular choice of ansatz spaces for the space discretization is

$$\mathcal{V}_1 = \mathcal{Q}_0(T_\mathcal{E}), \qquad \mathcal{V}_2 = \mathcal{P}_1(T_\mathcal{E}), \tag{2.4}$$

which clearly fulfill the Assumptions 2.7.

Remark 2.15. *Similarly, edgewise higher-order polynomial ansatz spaces fulfilling Assumption 2.7 can be conveniently implemented by employing edgewise Lagrange polynomials for the pressure and edgewise Legendre polynomials for the mass flow, as done in [EKLS20].*

2.5 Coordinate representations

We can characterize the Galerkin-approximation by differential-algebraic equations in $\mathbb{R}^N$ by considering coordinate representations.

Definition 2.16. *Let $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, and let bases $b_i^1, \ldots, b_i^{N_i}$ of $\mathcal{V}_i$ be fixed for $i = 1, 2$. For $\underline{e} = [p; m] \in \mathcal{V}$ we define the coordinate representation $\mathbf{e} = [\mathbf{e}_1; \mathbf{e}_2] \in \mathbb{R}^N$, with $N = N_1 + N_2$, by*

$$p(t) = \sum_{l=1}^{N_1} b_1^l \mathfrak{e}_1^l(t), \qquad m(t) = \sum_{l=1}^{N_1} b_2^l \mathfrak{e}_2^l(t), \qquad \mathbf{e}_i = [\mathfrak{e}_i^1; \ldots; \mathfrak{e}_i^{N_i}] \in \mathbb{R}^{N_i}.$$

The transformation from the coordinate representation $\mathbf{e} \in \mathbb{R}^N$ to the function $\underline{e} = [p; m] \in \mathcal{V}$ is defined as

$$\Psi : \mathbb{R}^N \to \mathcal{V}, \qquad \Psi(\mathbf{e}) = \begin{bmatrix} \sum_{l=1}^{N_1} b_1^l \mathfrak{e}_1^l(t) \\ \sum_{l=1}^{N_2} b_2^l \mathfrak{e}_2^l(t) \end{bmatrix} = \begin{bmatrix} p \\ m \end{bmatrix}.$$

The system matrices are defined by

$$\begin{aligned}
\mathbf{M}_1 &= [\langle \alpha b_1^n, b_1^m \rangle_{\mathcal{E}}]_{m,n=1,\ldots,N_1}, & \mathbf{M}_2 &= [\langle \beta b_2^n, b_2^m \rangle_{\mathcal{E}}]_{m,n=1,\ldots,N_2}, \quad \mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \\ & \mathbf{M}_2 \end{bmatrix} \\
\mathbf{R} &= [\langle r b_2^n, b_2^m \rangle_{\mathcal{E}}]_{m,n=1,\ldots,N_2}, & \mathbf{J} &= [\langle -\partial_x b_2^n, b_1^m \rangle_{\mathcal{E}}]_{m=1,\ldots,N_1, n=1,\ldots,N_2} \\
\mathbf{L}_2 &= \left[\mathcal{T} b_2^1, \ldots, \mathcal{T} b_2^{N_2}\right]^T, & \mathbf{K}_2 &= \left[\mathcal{T}^{\mathcal{N}_0} b_2^1, \ldots, \mathcal{T}^{\mathcal{N}_0} b_2^{N_2}\right]^T \\
\mathbf{L} &= \begin{bmatrix} \mathbf{0}_{N_1,p} \\ \mathbf{L}_2 \end{bmatrix} \in \mathbb{R}^{N,p}, & \mathbf{K} &= \begin{bmatrix} \mathbf{0}_{N_1,q} \\ \mathbf{K}_2 \end{bmatrix} \in \mathbb{R}^{N,q},
\end{aligned}$$

with $p = |\mathcal{N}_\partial|$ and $q = |\mathcal{N}_0|$.

System 2.17 (Coordinate representation of System 2.9)**.** *Find $\mathbf{e} \in \mathcal{C}^1([0,T]; \mathbb{R}^N)$, $\mathbf{e}_B \in C^0([0,T]; \mathbb{R}^p)$, $\mathbf{f}_B \in \mathcal{C}^1([0,T]; \mathbb{R}^p)$, $\boldsymbol{\lambda} \in C^0([0,T]; \mathbb{R}^q)$ with*

$$\mathbf{M} \frac{d}{dt} \mathbf{e}(t) = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R} \end{bmatrix} \mathbf{e}(t) + \mathbf{K} \boldsymbol{\lambda}(t) + \mathbf{L} \mathbf{e}_B(t) \tag{2.5a}$$

$$\mathbf{0} = \mathbf{K}^T \mathbf{e}(t) \tag{2.5b}$$

$$\mathbf{f}_B = \mathbf{L}^T \mathbf{e}(t) \tag{2.5c}$$

and closing conditions as in (2.3).

With the help of Assumption 2.7-(A3), the following can be shown, see [EKLS$^+$18].

Lemma 2.18. *Let the system matrices of System 2.17 be given. Let $\tilde{\mathbf{L}}_2 \in \mathbb{R}^{N_2, p-1}$ be obtained by removing one column in $\mathbf{L}_2$. Then the matrices $[-\mathbf{J}^T, \mathbf{K}_2, \tilde{\mathbf{L}}_2]$ and $[\mathbf{K}_2, \mathbf{L}_2]$ have both full column rank.*

As a direct consequence of Lemma 2.18, one can show that System 2.17 is a differential-algebraic system with differentiation index 2 in so-called Hessenberg form. The lemma can also be used to proof Lemma 2.13 on the level of coordinate representations.

2.6 Elimination of Lagrange multipliers

For an analysis of differential-algebraic systems, it is useful to decouple the equations into their underlying ordinary differential equations and algebraic equations, cf. [LMT13], [KM06]. In our systems, the algebraic equations stem from coupling- and boundary-conditions. We derive more compact formulations, which incorporate the coupling-conditions into the ansatz spaces for all stages, i.e., for the weak formulation, the Galerkin approximation with respective compatibility conditions, and for the coordinate representation. Finally, the underlying ordinary differential equation is constructed for the coordinate representation.

Weak form

Recall from Remark 1.5 that $\mathcal{H}^1_{div}(\mathcal{E}) = \{b \in \mathcal{H}^1_{pw}(\mathcal{E}) : \mathcal{T}^{\mathcal{N}_0} b = \mathbf{0}\}$. As an immediate consequence, System 2.3 can be reduced to the upcoming more compact formulation.

System 2.19. *Find $p \in \mathcal{C}^1([0,T];\mathcal{L}^2(\mathcal{E}))$, $m_D \in \mathcal{C}^1([0,T];\mathcal{H}^1_{div}(\mathcal{E}))$, $\mathbf{e}_B \in C^0([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$, $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ fulfilling*

$$\begin{aligned}\langle \alpha\partial_t p(t), b_1\rangle_{\mathcal{E}} &= -\langle \partial_x m_D(t), b_1\rangle_{\mathcal{E}}\\ \langle \beta\partial_t m_D(t), b_2\rangle_{\mathcal{E}} &= \langle p(t), \partial_x b_2\rangle_{\mathcal{E}} - \langle r m_D(t), b_2\rangle_{\mathcal{E}} + \mathbf{e}_B(t)\cdot \mathcal{T} b_2(t)\\ \mathbf{f}_B(t) &= \mathcal{T} m_D(t)\end{aligned}$$

for all $b_1 \in \mathcal{L}^2(\mathcal{E})$, $b_2 \in \mathcal{H}^1_{div}(\mathcal{E})$ and $t \in [0,T]$ with one closing condition for each boundary node, each of one of the following types

$$\text{Type 1: } e_B[\nu] = u^\nu, \qquad \text{Type 2: } f_B[\nu] = u^\nu, \qquad \text{for } \nu \in \mathcal{N}_\partial.$$

Galerkin approximation

As the Lagrange multipliers λ^ν for $\nu \in \mathcal{N}_0$ have been removed in System 2.19, Assumption 2.7 has to be adapted.

Assumption 2.20 (Compatibility of spaces for System 2.21)**.** *Let $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, and let $\mathcal{V}_1 \subset \mathcal{L}^2(\mathcal{E})$ and $\mathcal{V}_2 \subset \mathcal{H}^1_{div}(\mathcal{E})$ be finite dimensional subspaces fulfilling the compatibility conditions*

A1) $\mathcal{V}_1 = \partial_x \mathcal{V}_2$, *with $\partial_x \mathcal{V}_2 = \{\xi :$ It exists $\zeta \in \mathcal{V}_2$ with $\partial_x \zeta = \xi\}$*

A2) $1_{\mathcal{E}} \in \mathcal{V}_1$, *with $1_{\mathcal{E}} : \Omega \to \mathbb{R}$ and $1_{\mathcal{E}}(x) = 1$ for $x \in \Omega$*

A3) $\{b_2 \in \mathcal{H}^1_{div}(\mathcal{E}) :\ \partial_x b_2 = 0\} \subset \mathcal{V}_2$.

Note that the space $\{b_2 \in \mathcal{H}^1_{div}(\mathcal{E}) :\ \partial_x b_2 = 0\}$ is finite dimensional and in particular a subspace of the space appearing in Assumption 2.7-(A3).

System 2.21 (Galerkin approximation of System 2.19)**.** *Find $p \in \mathcal{C}^1([0,T];\mathcal{V}_1)$, $m_D \in \mathcal{C}^1([0,T];\mathcal{V}_2)$, $\mathbf{e}_B \in C^0([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ fulfilling*

$$\begin{aligned}\langle \alpha\partial_t p(t), b_1\rangle_{\mathcal{E}} &= -\langle \partial_x m_D(t), b_1\rangle_{\mathcal{E}}\\ \langle \beta\partial_t m_D(t), b_2\rangle_{\mathcal{E}} &= \langle p(t), \partial_x b_2\rangle_{\mathcal{E}} - \langle r m_D(t), b_2\rangle_{\mathcal{E}} + \mathbf{e}_B(t)\cdot \mathcal{T} b_2(t)\\ \mathbf{f}_B(t) &= \mathcal{T} m_D(t)\end{aligned}$$

for all $b_1 \in \mathcal{V}_1$, $b_2 \in \mathcal{V}_2$ and closing conditions (2.3). *Assumption 1.2 and Assumption 2.20 are supposed to hold.*

The results of Lemma 2.13, Lemma 2.12 and Remark 2.14 still can be shown to hold true for the solutions of System 2.21, see [EK18]. In fact, an equivalence of solutions of System 2.9 and System 2.21 and can be derived. This becomes evident by the upcoming discussion on condensed coordinate representations.

Coordinate representation

Formulations with (partially) removed algebraic equations are constructed here on the algebraic level by post-processing the coordinate representation System 2.17. In particular, Corollary 2.23 states the coordinate representation for System 2.19. The underlying ordinary differential equations are constructed in Lemma 2.22 for a special case and in Lemma 2.25 for the general case, respectively. The latter highlights the different roles of boundary conditions of *Type 1* and *Type 2* in (2.3).

Let us note that our algebraic construction is similar to the one in [GSW13], [HSS08], where the decoupling of differential-algebraic equations with Hessenberg-index-2-structure is explained. We also refer to [BH15], where the algebraic structure stemming from the discretization of essential boundary conditions is discussed.

Lemma 2.22. *Let in System 2.17 all boundary conditions in* (2.3) *be of Type 1, i.e.,* $\mathbf{e}_B = \mathbf{u}$ *for* $\mathbf{u} : [0, T] \to \mathbb{R}^p$*, and the system may be rewritten as*

$$\mathbf{M}\frac{d}{dt}\mathbf{e}(t) = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R} \end{bmatrix} \mathbf{e}(t) + \mathbf{K}\boldsymbol{\lambda}(t) + \mathbf{L}\mathbf{u}(t)$$
$$\mathbf{0} = \mathbf{K}^T\mathbf{e}(t).$$

Let $\mathbf{W}_2 \in \mathbb{R}^{N_2, N_2 - |\mathcal{N}_0|}$ *as a basis matrix of ker*$(\mathbf{K}_2^T)$*, such that*

$$\mathbf{W} = \begin{bmatrix} \mathbf{I}_{N_1} & \\ & \mathbf{W}_2 \end{bmatrix} \qquad \textit{fulfills } \mathbf{K}^T\mathbf{W} = \mathbf{0}.$$

Let further

$$\mathbf{J}_D = \mathbf{J}\mathbf{W}_2, \qquad \mathbf{R}_D = \mathbf{W}_2^T\mathbf{R}\mathbf{W}_2, \qquad \mathbf{M}_D = \mathbf{W}^T\mathbf{M}\mathbf{W}, \qquad \mathbf{L}_D = \mathbf{W}^T\mathbf{L}.$$

Then the system can be equivalently characterized by the underlying ordinary differential

$$\mathbf{M}_D\frac{d}{dt}\mathbf{e}_D(t) = \begin{bmatrix} & \mathbf{J}_D \\ -\mathbf{J}_D^T & -\mathbf{R}_D \end{bmatrix} \mathbf{e}_D(t) + \mathbf{L}_D\mathbf{u}(t),$$

together with the algebraic equations

$$\mathbf{e}(t) = \mathbf{W}\mathbf{e}_D(t), \qquad \boldsymbol{\lambda}(t) = -\left(\mathbf{K}^T\mathbf{M}^{-1}\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{M}^{-1}\left(\begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R} \end{bmatrix} \mathbf{W}\mathbf{e}_D(t) + \mathbf{L}\mathbf{u}(t)\right).$$

Proof. By (2.5b), it follows $\mathbf{e}(t) \in \mathit{ker}(\mathbf{K}^T)$. Thus, for $\mathbf{W}^+$ being the Moore-Penrose pseudoinverse of $\mathbf{W}$, and $\mathbf{e}_D(t) := \mathbf{W}^+\mathbf{e}(t)$, it holds

$$\mathbf{e}(t) = \mathbf{W}\mathbf{W}^+\mathbf{e}(t) = \mathbf{W}\mathbf{e}_D(t).$$

By pre-multiplying (2.5a) with $\mathbf{W}^T$, and inserting the latter relation for $\mathbf{e}$, we obtain the underlying ordinary differential equation for $\mathbf{e}_D$, which shows the first part.

On the other hand, when (2.5a) is pre-multiplied by $\mathbf{K}^T\mathbf{M}^{-1}$, the equation can be solved for $\boldsymbol{\lambda}$, as $\mathbf{K}$ has full column rank, cf. Lemma 2.18. This gives

$$\boldsymbol{\lambda}(t) = \left(\mathbf{K}^T\mathbf{M}^{-1}\mathbf{K}\right)^{-1}\left(\frac{d}{dt}(\mathbf{K}^T\mathbf{e}(t) - \mathbf{K}^T\mathbf{M}^{-1}\begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R}\end{bmatrix}\mathbf{e}(t) - \mathbf{K}^T\mathbf{M}^{-1}\mathbf{L}\mathbf{u}(t)\right).$$

The first summand in the brackets is zero, as seen by differentiating (2.5b). Inserting again the representation $\mathbf{e}(t) = \mathbf{W}\mathbf{e}_D(t)$ then finishes the proof. □

The algebraic counterpart of System 2.21 in the general case can be derived in the same manner. But as can be seen below, there remain algebraic coupling conditions relating to the boundary terms.

Corollary 2.23 (Coordinate representation of System 2.21)**.** *Let System 2.17 be given, and let* $\mathbf{W}$*,* $\mathbf{J}_D$*,* $\mathbf{R}_D$*,* $\mathbf{M}_D$ *and* $\mathbf{L}_D$ *be defined as in Lemma 2.22. Then the solution component* $\mathbf{e} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$ *of System 2.17 is equally characterized by* $\mathbf{e} = \mathbf{W}\mathbf{e}_D$*, closing conditions* (2.3) *and*

$$\begin{aligned}\mathbf{M}_D\frac{d}{dt}\mathbf{e}_D(t) &= \begin{bmatrix} & \mathbf{J}_D \\ -\mathbf{J}_D^T & -\mathbf{R}_D\end{bmatrix}\mathbf{e}_D(t) + \mathbf{L}_D\mathbf{e}_B(t)\\ \mathbf{f}_B(t) &= \mathbf{L}_D^T\mathbf{e}_D(t).\end{aligned}$$

Remark 2.24. *A basis matrix* $\mathbf{W}$*, as utilized in Lemma 2.22 and Corollary 2.23, can typically be constructed analytically in very effective manner for our systems. This is, because every algebraic equation in* (2.5b) *relates to one coupling condition at one inner node.*

To derive the underlying ordinary differential equation in the general case, not only algebraic manipulations, but also a differentiation step is involved. Therefore, the solution then can depend on the derivative of the boundary data. Just to simplify the notation, we assume the boundary nodes to be grouped into *Type 1*- and *Type 2*-boundary conditions, respectively, in the upcoming result.

Lemma 2.25. *Let System 2.17 be given with* p_a *boundary conditions of Type 1,* p_b *boundary conditions of Type 2, where* $p_a + p_b = p$*, and let the boundary conditions* $\mathbf{u}$ *be grouped as*

$$\mathbf{u}(t) = \begin{bmatrix}\mathbf{u}_a(t)\\ \mathbf{u}_b(t)\end{bmatrix}, \qquad \mathbf{u}_a(t) \in \mathbb{R}^{p_a}, \quad \mathbf{u}_b(t) \in \mathbb{R}^{p_b}.$$

Let

$$\mathbf{S}_a = \begin{bmatrix}\mathbf{I}_{p_a}\\ \mathbf{0}_{p_b,p_a}\end{bmatrix}, \quad \mathbf{S}_b = \begin{bmatrix}\mathbf{0}_{p_a,p_b}\\ \mathbf{I}_{p_b}\end{bmatrix} \qquad \text{and} \qquad \mathbf{L}_b = \mathbf{L}\mathbf{S}_b.$$

Let $\mathbf{W}_2 \in \mathbb{R}^{N_2, N_2-(|\mathcal{N}_0|+p_b)}$ *be a basis matrix of ker*$([\mathbf{K}_2, \mathbf{L}_2\mathbf{S}_b]^T)$*, such that*

$$\mathbf{W} = \begin{bmatrix} \mathbf{I}_{N_1} & \\ & \mathbf{W}_2 \end{bmatrix} \qquad \textit{fulfills } [\mathbf{K}, \mathbf{L}_b]^T\mathbf{W} = \mathbf{0}.$$

Define further,

$$\mathbf{J}_D = \mathbf{J}\mathbf{W}_2, \qquad \mathbf{R}_D = \mathbf{W}_2^T\mathbf{R}\mathbf{W}_2, \qquad \mathbf{M}_D = \mathbf{W}^T\mathbf{M}\mathbf{W}$$
$$\mathbf{L}_D = \mathbf{W}^T\mathbf{L}\mathbf{S}_a, \qquad \mathbf{F} = \mathbf{L}_b\left(\mathbf{L}_b^T\mathbf{L}_b\right)^{-1}\mathbf{S}_b^T.$$

Then the system can be equivalently characterized by the underlying ordinary differential equation

$$\begin{aligned}\mathbf{M}_D\frac{d}{dt}\mathbf{e}_D(t) = &\begin{bmatrix} & \mathbf{J}_D \\ -\mathbf{J}_D^T & -\mathbf{R}_D \end{bmatrix}\mathbf{e}_D(t) + \mathbf{L}_D\mathbf{u}_a(t) \\ &- \mathbf{W}^T\begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R} \end{bmatrix}\mathbf{F}\mathbf{u}_b(t) + \mathbf{W}^T\mathbf{M}\mathbf{F}\frac{d}{dt}\mathbf{u}_b(t),\end{aligned}$$

together with the algebraic equations

$$\begin{aligned}&\mathbf{e}_G(t) = \mathbf{F}\mathbf{u}_b(t), \qquad \mathbf{e}(t) = \mathbf{W}\mathbf{e}_D(t) + \mathbf{e}_G(t), \qquad \mathbf{e}_B(t) = [\mathbf{u}_a(t); \boldsymbol{\mu}(t)] \\ &\boldsymbol{\lambda}(t) = -\left(\mathbf{K}^T\mathbf{M}^{-1}\mathbf{K}\right)^{-1}\mathbf{K}^T\mathbf{M}^{-1}\left(\begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R} \end{bmatrix}\mathbf{e}(t) + \mathbf{L}\mathbf{e}_B(t)\right) \\ &\boldsymbol{\mu}(t) = -\left(\mathbf{L}_b^T\mathbf{M}^{-1}\mathbf{L}_b\right)^{-1}\mathbf{L}_b^T\mathbf{M}^{-1}\left(\begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R} \end{bmatrix}\mathbf{e}(t) + \mathbf{L}\mathbf{e}_B(t)\right) + \left(\mathbf{L}_b^T\mathbf{M}\mathbf{L}_b\right)^{-1}\mathbf{F}\frac{d}{dt}\mathbf{u}_b(t).\end{aligned}$$

The proof is similar to the one of Lemma 2.22. We refer to [GSW13], [HSS08] and [KM06] for detailed discussions on the construction of underlying ordinary differential equations for differential-algebraic equations in Hessenberg-form.

Remark 2.26. *By construction, the boundary conditions in Lemma 2.25 can be expressed by the two equations*

$$\mathbf{S}_a^T\mathbf{e}_B(t) = \mathbf{u}_a(t) \qquad \textit{and} \qquad \mathbf{S}_b^T\mathbf{f}_B(t) = \mathbf{u}_b(t), \qquad \textit{for } t > 0.$$

Chapter 3

Realization of model reduction

The starting point for this section is a high-fidelity, but also high-order space discretization by finite elements, Section 2.4, which we will refer to as *full order model.* The subsequent structured model reduction of the full order model is the focus here. The upcoming derivations are mainly made in the algebraic setting of coordinate representations, and the results are framed in the wording of model reduction, cf. [BBF14], [MS05], [Ant05]. The full order model is described by System 2.17, where we for ease of presentation assume all boundary conditions to be of *Type 1* in (2.3). We relate to the boundary conditions as input $\mathbf{u}$, and take the boundary flows as output $\mathbf{y}$. By that, the input and output fully characterize the exchange of the system over the boundaries.

System 3.1 (Full order model)**.** *The full order model, describing the input-output relation* $\mathbf{u} \mapsto \mathbf{y}$*, reads*

$$\begin{aligned} \mathbf{E}\frac{d}{dt}\mathbf{x}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{B}^T\mathbf{x}(t), \qquad\qquad \mathbf{x}(0) = \mathbf{x}_0, \end{aligned}$$

with $\mathbf{x} = [\mathbf{e}_1; \mathbf{e}_2; \boldsymbol{\lambda}]$ *and* $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{N+q,N+q}$, $\mathbf{B} \in \mathbb{R}^{N+q,p}$ *defined by the matrices before System 2.17 as*

$$\mathbf{E} = \begin{bmatrix} \mathbf{M}_1 & & \\ & \mathbf{M}_2 & \\ & & \mathbf{0}_q \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} & \mathbf{J} & \\ -\mathbf{J}^T & -\mathbf{R} & \mathbf{K}_2 \\ & \mathbf{K}_2^T & \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{L} \\ \mathbf{0}_{q,p} \end{bmatrix}.$$

Let $\Psi : \mathbb{R}^N \to \mathcal{V}_f$ *be the transformation from coordinate representation to the related function space* $\mathcal{V} = \mathcal{V}_f$ *as in Definition 2.16. We suppose* $\Psi(\mathbb{R}^N)$ *fulfills Assumption 2.7 and* $\mathbf{u} \in \mathcal{C}([0, T]; \mathbb{R}^p)$*.*

Remark 3.2. *The initial conditions for the Lagrange-multiplier* λ *in System 3.1 are assumed to be chosen consistent to the characterization in Lemma 2.22 throughout this part, as other choices would lead to inconsistencies.*

It is evident from our discussion throughout this part that System 3.1 is the coordinate representation of a Galerkin approximation as in System 2.9 with an ansatz space $\mathcal{V}_f$ fulfilling Assumption 2.7. The first crucial observation for us is that compatibility conditions of the ansatz space $\mathcal{V}_f$ for the full order model can be transferred to reduced order space $\mathcal{V}_r$ by imposing compatibility conditions only depending on these two spaces.

Lemma 3.3. *Let Assumption 2.7 hold for* $\mathcal{V}_f = \mathcal{V}_{f,1} \times \mathcal{V}_{f,2}$. *Let* $\mathcal{V}_r = \mathcal{V}_{r,1} \times \mathcal{V}_{r,2} \subset \mathcal{V}_f$ *fulfill*

1. $\mathcal{V}_{r,1} = \partial_x \mathcal{V}_{r,2}$, *with* $\partial_x \mathcal{V}_{r,2} = \{\xi : \text{It exists } \zeta \in \mathcal{V}_{r,2} \text{ with } \partial_x \zeta = \xi\}$
2. $1_{\mathcal{E}} \in \mathcal{V}_{r,1}$, *with* $1_{\mathcal{E}} : \Omega \to \mathbb{R}$ *and* $1_{\mathcal{E}}(x) = 1$ *for* $x \in \Omega$
3. $\{b_2 \in \mathcal{V}_{2,f} : \partial_x b_2 = 0\} \subset \mathcal{V}_{r,2}$.

Then $\mathcal{V}_r$ *fulfills Assumption 2.7.*

The proof of the latter result is straight-forward, and therefore omitted.

Remark 3.4. *The Lagrange multiplier* $\boldsymbol{\lambda}(t) \in \mathbb{R}^q$ *is not reduced in the model reduction step. If desired, it can be eliminated, partly or fully, beforehand in the full order model, cf. Corollary 2.23. This still gives an input-output system as in System 3.1, but with a lower-dimensional Lagrange multiplier.*

Typically it holds for $N = dim(\mathcal{V}_f)$ and $n = dim(\mathcal{V}_r)$ that $n \ll N$. We will characterize the transition from the space $\mathcal{V}_f$ to the space $\mathcal{V}_r$ by a reduction basis $\mathbf{V} \in \mathbb{R}^{N+q,n+q}$ in the algebraic setting. Then the assumptions on $\mathcal{V}_r$ of Lemma 3.3 can directly be recast as assumptions on $\mathbf{V}$.

Assumption 3.5 (Compatibility of reduction basis)**.** *For System 3.1 given, let the reduction basis* $\mathbf{V}$ *have block structure*

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & & \\ & \mathbf{V}_2 & \\ & & \mathbf{I}_q \end{bmatrix}, \qquad \mathbf{V}_1 \in \mathbb{R}^{N_1,n_1},\ \mathbf{V}_2 \in \mathbb{R}^{N_2,n_2}, \quad n = n_1 + n_2.$$

Let further, $\mathbf{o}_1 \in \mathbb{R}^{N_1,1}$ *be the coordinate representation of* $1_{\mathcal{E}} : \Omega \to \mathbb{R}$, $1_{\mathcal{E}}(x) = 1$ *for* $x \in \Omega$. *Then* $\mathbf{V}$ *is assumed to fulfill*

A1) $im(\mathbf{M}_1\mathbf{V}_1) = im(\mathbf{J}\mathbf{V}_2)$

A2) $\mathbf{o}_1 \in im(\mathbf{V}_1)$

A3) $ker(\mathbf{J}) \subset im(\mathbf{V}_2)$.

System 3.6 (Reduced order model)**.** *Given the full order model in System 3.1 and a reduction basis* $\mathbf{V}$ *fulfilling Assumption 3.5, the reduced order model is defined by*

$$\begin{aligned} \mathbf{E}_r \frac{d}{dt}\mathbf{x}_r(t) &= \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{B}_r\mathbf{u}(t) \\ \mathbf{y}_r(t) &= \mathbf{B}_r^T\mathbf{x}_r(t), \qquad\qquad \mathbf{x}_r(0) = \mathbf{V}^\dagger\mathbf{x}_0, \end{aligned}$$

with $\mathbf{x}_r = [\mathbf{e}_{r,1}; \mathbf{e}_{r,2}; \boldsymbol{\lambda}]$ *and* $\mathbf{E}_r, \mathbf{A}_r \in \mathbb{R}^{n+q,n+q}$, $\mathbf{B}_r \in \mathbb{R}^{n+q,p}$. *Further,* $\mathbf{V}^\dagger$ *denotes the pseudo-inverse of* $\mathbf{V}$ *w.r.t. the energy scalar product and*

$$\mathbf{E}_r = \begin{bmatrix} \mathbf{M}_{r,1} & & \\ & \mathbf{M}_{r,2} & \\ & & \mathbf{0}_q \end{bmatrix} = \mathbf{V}^T\mathbf{E}\mathbf{V}, \qquad \mathbf{A}_r = \begin{bmatrix} & \mathbf{J}_r & \\ -\mathbf{J}_r^T & -\mathbf{R}_r & \mathbf{K}_{r,2} \\ & \mathbf{K}_{r,2}^T & \end{bmatrix} = \mathbf{V}^T\mathbf{A}\mathbf{V}$$

$$\mathbf{B}_r = \begin{bmatrix} \mathbf{L}_r \\ \mathbf{0}_{q,p} \end{bmatrix} = \mathbf{V}^T\mathbf{B}, \qquad \mathbf{V}^\dagger = \left(\mathbf{V}^T \begin{bmatrix} \mathbf{M}_1 & & \\ & \mathbf{M}_2 & \\ & & \mathbf{I}_q \end{bmatrix} \mathbf{V} \right)^{-1} \mathbf{V}^T \begin{bmatrix} \mathbf{M}_1 & & \\ & \mathbf{M}_2 & \\ & & \mathbf{I}_q \end{bmatrix}.$$

Remark 3.7. *The initial conditions as chosen in System 3.6 by the pseudo-inverse* $\mathbf{V}^\dagger$ *minimizes the projection-error in the energy-scalar product, i.e., for*

$$\tilde{\mathbf{x}}_0 = \mathbf{V}\left(\mathbf{V}^\dagger \mathbf{x}_0\right) \qquad \text{it holds} \quad \tilde{\mathbf{x}}_0 = \operatorname*{argmin}_{\mathbf{x}\in im(\mathbf{V})} (\mathbf{x}_0 - \mathbf{x})^T \begin{bmatrix} \mathbf{M}_1 & & \\ & \mathbf{M}_2 & \\ & & \mathbf{I}_q \end{bmatrix} (\mathbf{x}_0 - \mathbf{x})\,.$$

As is evident from Lemma 3.3, System 3.6 is again a coordinate representation of a compatible Galerkin approximation, i.e., of System 2.9 fulfilling Assumptions 2.7. Therefore, all results can be transferred to the algebraic level. Let us state the algebraic counterpart to Theorem 2.10 and Theorem 2.11.

Corollary 3.8 (Energy dissipation equality and mass conservation)**.** *Let for System 3.6 the Hamiltonian* $H_r : \mathbb{R}^n \to \mathbb{R}$ *and the total mass* $M_r : \mathbb{R}^n \to \mathbb{R}$ *be defined by*

$$H_r(\mathbf{e}_r) = \frac{1}{2}\mathbf{e}_r^T \mathbf{M}_r \mathbf{e}_r, \qquad M_r(\mathbf{e}_r) = \mathbf{o}_{r,1}^T \mathbf{M}_{r,1} \mathbf{e}_{r,1},$$

with $\mathbf{o}_{r,1} = \mathbf{V}_1^\dagger \mathbf{o}_1 \in \mathbb{R}^{n_1}$ *the reduced coordinate representation of* $1_{\mathcal{E}} \in \mathcal{V}_1$*, cf. Assumption 3.5. Then it holds*

$$\frac{d}{dt} H_r(\mathbf{e}_r(t)) = \mathbf{u}(t)\cdot \mathbf{y}_r(t) - \mathbf{e}_{r,2}^T(t)\mathbf{R}_r \mathbf{e}_{r,2}(t) < \mathbf{u}(t)\cdot \mathbf{y}_r(t)$$
$$\frac{d}{dt} M_r(\mathbf{e}_r(t)) = \mathbf{o}_{r,1}\cdot \mathbf{y}_r(t)$$

Proof. The dissipation equality for the Hamiltonian can be derived conveniently from the algebraic representation by basic calculations,

$$\begin{aligned} \frac{d}{dt} H_r(\mathbf{e}_r(t)) &= \frac{d}{dt}\mathbf{e}_r(t)\cdot \nabla_{\mathbf{e}_r} H_r(\mathbf{e}_r(t)) = \left(\mathbf{M}_r \frac{d}{dt}\mathbf{e}_r(t)\right)\cdot \mathbf{e}_r(t) \\ &= \left(\begin{bmatrix} & \mathbf{J}_r \\ -\mathbf{J}_r^T & -\mathbf{R}_r \end{bmatrix} \mathbf{e}_r(t) + \mathbf{L}_r \mathbf{u}(t)\right)\cdot \mathbf{e}_r(t) = -\mathbf{e}_{r,2}(t)^T \mathbf{R}_r \mathbf{e}_{r,2}(t) + \mathbf{u}(t)\cdot \mathbf{y}_r(t). \end{aligned}$$

The mass conservation is most appropriately shown by changing from coordinate representations to the function space setting. We refer to Theorem 2.11 and [EKLS$^+$18, Lemma A3.3]. □

Clearly, the existence of unique steady states and the exponential stability with stability constants uniform in the reduction parameters also hold by Lemma 2.12 and Lemma 2.13.

3.1 Compatible basis

For our approach, we consider the construction of the reduced subspace $\mathcal{V}_r = \mathcal{V}_{r,1} \times \mathcal{V}_{r,2}$ in the form

$$\mathcal{V}_{r,1} = \mathcal{W}_1 + \mathcal{Z}_1, \qquad \mathcal{V}_{r,2} = \mathcal{W}_2 + \mathcal{Z}_2.$$

A standard model reduction procedure is applied to construct the high-fidelity spaces $\mathcal{W} \subset \mathcal{V}_f$, from which we extract separate bases $\mathcal{W}_i \subset \mathcal{V}_{f,i}$ for $i = 1, 2$. The reduction spaces are complemented by spaces $\mathcal{Z}_i$ to guarantee the compatibility conditions of Lemma 3.3. Our procedure thus consists of several steps.

Stable splitting with partial compatibility

Starting from unstructured bases, our first aim is to extract block bases $\mathbf{W}_i \in \mathbb{R}^{N_i,\bar{n}_i}$ with $\bar{n}_i \ll N_i$ for $i = 1, 2$, fulfilling Assumption 3.5-(A1). Defining $\mathcal{W}_i = im(\mathbf{W}_i)$, this may be expressed as

$$im(\mathbf{M}_1\mathbf{W}_1) = im(\mathbf{J}\mathbf{W}_2), \qquad \text{or equivalently} \qquad \mathbf{M}_1\mathcal{W}_1 = \mathbf{J}\mathcal{W}_2. \tag{3.1}$$

Here, we tacitly identified the matrices $\mathbf{M}_1, \mathbf{J}$ with their related linear functions. We consider two different strategies to reach this aim.

Remark 3.9. *Notably, it will turn out that the situation is simpler in the linear case, when moment matching is applied. Nevertheless, we keep the discussion more general here, as we will need it later in Part I.B, but also if other model reduction procedures for the linear model are applied.*

Strategy 1:

One possibility is to base the whole construction on an *unstructured* high-fidelity basis $\tilde{\mathbf{W}}_1 \in \mathbb{R}^{N_1,\tilde{n}_1}$ relating to the first component $\mathbf{e}_1$ only. Let us introduce a pseudo-inverse of $[\mathbf{J}; \mathbf{K}_2^T]$ as

$$\begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix}^{\dagger} = \mathbf{M}_2^{-1} \begin{bmatrix} \mathbf{J}^T & \mathbf{K}_2 \end{bmatrix} \left(\begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix} \mathbf{M}_2^{-1} \begin{bmatrix} \mathbf{J}^T & \mathbf{K}_2 \end{bmatrix} \right)^{-1}. \tag{3.2}$$

By Lemma 2.18, one can see that it is well-defined. The pseudo-inverse is related to a minimization problem. Given $\mathbf{g}_1 \in \mathbb{R}^{N_1}$, it holds

$$\mathbf{g}_2 = \begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix}^{\dagger} \mathbf{g}_1, \qquad \text{solves} \qquad \begin{aligned} \mathbf{g}_2 = &\operatorname*{argmin}_{\tilde{\mathbf{o}}_2 \in \mathbb{R}^{N_2}} \left(\tilde{\mathbf{o}}_2^T \mathbf{M}_2 \tilde{\mathbf{o}}_2 \right) \\ &\text{s.t.} \begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix} \tilde{\mathbf{o}}_2 = \begin{bmatrix} \mathbf{M}_1\mathbf{g}_1 \\ \mathbf{0}_{q,1} \end{bmatrix}. \end{aligned}$$

The bases fulfilling the compatibility condition (3.1) are then chosen as

$$\mathbf{W}_1 := \tilde{\mathbf{W}}_1 \qquad \text{and} \qquad \mathbf{W}_2 := \begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix}^{\dagger} \mathbf{W}_1.$$

The main issue in this strategy is that we do not directly control the second solution component $\mathbf{e}_2$ by a high-fidelity basis, but need to somehow encode enough fidelity in $\tilde{\mathbf{W}}_1$ for both components. We follow this approach in Part I.B.

Strategy 2:

Another possibility is to base the procedure on a high-fidelity basis $\tilde{\mathbf{W}} \in \mathbb{R}^{\tilde{N},\tilde{n}}$ relating to both solution components $\mathbf{e}_1, \mathbf{e}_2$. Then we first need to extract basis matrices $\tilde{\mathbf{W}}_i$ for the sub-components $\mathbf{e}_i$ for $i = 1, 2$. The naive splitting $\tilde{\mathbf{W}} = [\tilde{\mathbf{W}}_1; \tilde{\mathbf{W}}_2]$, followed by a separate orthonormalization, is known to be very sensitive to numerical errors. We therefore apply the cosine-sine splitting instead, [GVL96], [VL85]. For the matrices $\tilde{\mathbf{W}}_1$, $\tilde{\mathbf{W}}_2$, the related cosine-sine decomposed is given as

$$\begin{bmatrix} \tilde{\mathbf{W}}_1 \\ \tilde{\mathbf{W}}_2 \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{W}}_1 & \\ & \bar{\mathbf{W}}_2 \end{bmatrix} \begin{bmatrix} \mathbf{C} \\ \mathbf{S} \end{bmatrix} \mathbf{X}^T,$$

with $\mathbf{U}_1, \mathbf{U}_2$ and $\mathbf{X}$ orthogonal matrices, and $\mathbf{C}$, $\mathbf{S}$ diagonal matrices with $[\mathbf{C}]_{ii} + [\mathbf{S}]_{ii} = 1$. Note that the cosine-sine decomposition can be computed in a numerical efficient and stable manner. We present in Algorithm 3.10 an implementation, which takes into account non-standard scalar products.

Algorithm 3.10 (Stable splitting via cosine-sine decomposition)**.**

```
function [Wb1,Wb2]=split(W1,W2,M1,M2,tolerance)
      % compute cholesky factorizations Mi=Ri*Ri'
      R1 = chol(M1); R2 = chol(M2);
      % compute generalized svd
      [U1,U2,X,C,S] = gsvd(R1*W1,R2*W2,0);  % 0: economic version
      % eliminate dependent columns
      kc = find(diag(C)>tolerance);
      ks = find(diag(S)>tolerance);
      Wb1 = R1\U1(:,kc); Wb2 = R2\U2(:,ks);
end
```

Having constructed block bases $\bar{\mathbf{W}}_i \in \mathbb{R}^{N_i,\bar{n}_i}$ for $i = 1,2$ by Algorithm 3.10, let $\bar{\mathcal{W}}_i = im\big(\bar{\mathbf{W}}_i\big)$. Next, a post-processing step is needed to guarantee the compatibility condition (3.1). The easiest choice is to set

$$\mathcal{W}_1 := \bar{\mathcal{W}}_1 + \mathbf{M}_1^{-1}\mathbf{J}\bar{\mathcal{W}}_2, \qquad \mathcal{W}_2 := \bar{\mathcal{W}}_2 + \begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix}^{\dagger} \bar{\mathcal{W}}_1.$$

We refer to [EKLS20], where this has been applied.

Remark 3.11. *Strategy 2 seems not optimal for truncation-based methods like proper orthogonal decomposition in terms of reducing the order most efficiently. Let us therefore point towards a related discussion for a slightly simpler related problem. Given a Hamiltonian system in the following canonical form*

$$\frac{d}{dt}\mathbf{x}(t) = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{N}} \\ -\mathbf{I}_{\bar{N}} & \mathbf{0} \end{bmatrix} \nabla_{\mathbf{x}} H(\mathbf{x}(t)), \qquad \textit{with } H : \mathbb{R}^{2\bar{N}} \to \mathbb{R},$$

the so-called symplectic model order reduction aims for a symplectic reduction basis $\mathbf{V} \in \mathbb{R}^{2\bar{N},2\bar{n}}$ *with* $\bar{n} \ll \bar{N}$*. This means, one seeks for a reduction basis fulfilling*

$$\mathbf{V}^T \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{N}} \\ -\mathbf{I}_{\bar{N}} & \mathbf{0} \end{bmatrix} \mathbf{V} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{\bar{n}} \\ -\mathbf{I}_{\bar{n}} & \mathbf{0} \end{bmatrix}.$$

Both greedy and singular value decomposition based approaches have been proposed, see [PM16], [AH17], [AH19], to accomplish that. The adaption of the ideas from symplectic model reduction to our problem of finding a basis fulfilling (3.1) *seems possible. We, however, do not pursue this direction here.*

Complementing spaces and compatibility

Let us assume high-fidelity spaces $\mathcal{W}_i$, $i = 1, 2$, fulfilling the compatibility condition (3.1), have been constructed by either of the former strategies. Let $\mathbf{o}_1$ be as in Assumption 3.5, and set

$$\mathbf{o}_2 = \begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix}^\dagger \mathbf{o}_1, \qquad \text{with } \begin{bmatrix} \mathbf{J} \\ \mathbf{K}_2^T \end{bmatrix}^\dagger \text{ as in (3.2).}$$

We then define the complementing spaces $\mathcal{Z}_i$ and the overall reduction space $\mathcal{V}_{r,i}$ by

$$\begin{aligned} \mathcal{Z}_1 &= span\{\mathbf{o}_1\}, & \mathcal{Z}_2 &= span\{\mathbf{o}_2\} + ker(\mathbf{J}) \\ \mathcal{V}_{r,1} &= \mathcal{W}_1 + \mathcal{Z}_1, & \mathcal{V}_{r,2} &= \mathcal{W}_2 + \mathcal{Z}_2 \\ &\mathbf{V}_i \text{ as basis matrix to } \mathcal{V}_{r,i} \text{ for } i = 1, 2. \end{aligned} \tag{3.3}$$

The orthogonalization is done w.r.t. the related energy scalar products $(\mathbf{a}, \mathbf{b}) \mapsto \mathbf{a} \cdot (\mathbf{M}_i \mathbf{b})$ for $i = 1, 2$ throughout, see Algorithm 3.12.

Algorithm 3.12 (Orthogonalization)**.**

```
function Vo=ortho(W,U,Ms,tolerance)
% constructs orthogonal basis of 'W', orthogonal to im(U), in scalar
% product induced by 'Ms'
      for k = 1:size(W,2)
          % orthogonalize to Uj
          for r=1:2 % use re-orthonormalization
              for j = 1:size(U,2)
                  hk1j = U(:, j)' * Ms * W(:, k);
                  W(:, k) = W(:, k) - U(:, j) * hk1j;
              end
          end
          % orthogonalize to previous Wj
          for r=1:2
              for j = 1:k-1
                  if d(j)<tolerance, continue; end
                  hk1j = W(:, j)' * Ms * W(:, k);
                  W(:, k) = W(:, k) - W(:, j) * hk1j;
              end
          end
          % normalize
          d(k) = sqrt(W(:,k)' * Ms * W(:,k));
          if d(k)>=tolerance,
             W(:, k) = W(:, k) / d(k);
          end
      end
      % only keep relevant vectors
      Vo = W(:,find(d>tolerance));
end
```

By construction, $\mathbf{V}_1, \mathbf{V}_2$ then fulfill Assumption 3.5. Algorithm 3.13 summarizes their construction from bases $\mathbf{W}_1, \mathbf{W}_2$ fulfilling the partial compatibility (3.1).

Algorithm 3.13 (Modifications).

```
function [V1,V2]=modify(W1,W2,M1,M2,J,K2,o1,KernelJ,tolerance)
      V1 = ortho([W1,M1\(J*W2),o1],[],M1,tolerance);
      S = [J',K2];
      V2 = M2\(S*((S'*(M2\S))\[M1*V1;zeros(size(K2,2),size(V1,2))]));
      V2 = ortho([KernelJ,V2],[],M2,tolerance);
end
```

3.2 Moment matching

Next, we discuss the construction of $\mathcal{W}_1, \mathcal{W}_2$ by moment matching, cf. [Ant05], [Gri97]. Given the transfer function $s \mapsto (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$, the moments $\mathbf{k}_j$ around an expansion frequency $s_0 \in \mathbb{C}$ read

$$\mathbf{k}_j = -\left[\mathbf{A}_{s_0}^{-1}\mathbf{E}\right]^j \mathbf{A}_{s_0}^{-1}\mathbf{B}, \qquad \text{with } \mathbf{A}_{s_0} = \mathbf{A} - s_0\mathbf{E}, \quad \text{for } j \geq 0.$$

Let us separate these moments in block structure according to System 3.1, as

$$\mathbf{k}_j = \begin{bmatrix} \boldsymbol{\mu}_j \\ \boldsymbol{\eta}_j \\ \boldsymbol{\nu}_j \end{bmatrix}, \qquad \text{with } \boldsymbol{\mu}_j \in \mathbb{R}^{N_1,p},\ \boldsymbol{\eta}_j \in \mathbb{R}^{N_2,p},\ \boldsymbol{\nu}_j \in \mathbb{R}^{q,p}.$$

The zeroth moment is then characterized by

$$\begin{aligned} -s_0\mathbf{M}_1\,\boldsymbol{\mu}_0 \qquad\quad + \quad \mathbf{J}\,\boldsymbol{\eta}_0 \qquad\qquad &= \mathbf{0} \\ -s_0\mathbf{M}_2\,\boldsymbol{\eta}_0 - \mathbf{J}^T\boldsymbol{\mu}_0 - \mathbf{R}\,\boldsymbol{\eta}_0 + \mathbf{K}_2\,\boldsymbol{\nu}_0 &= -\mathbf{L}_2 \\ \mathbf{K}_2^T\,\boldsymbol{\eta}_0 \qquad\qquad &= \mathbf{0}. \end{aligned} \tag{3.4a}$$

The recursion for $j \geq 1$ reads accordingly

$$\begin{aligned} -s_0\mathbf{M}_1\,\boldsymbol{\mu}_j \qquad\quad + \quad \mathbf{J}\,\boldsymbol{\eta}_j \qquad\qquad &= \mathbf{M}_1\,\boldsymbol{\mu}_{j-1} \\ -s_0\mathbf{M}_2\,\boldsymbol{\eta}_j - \mathbf{J}^T\boldsymbol{\mu}_j - \mathbf{R}\,\boldsymbol{\eta}_j + \mathbf{K}_2\,\boldsymbol{\nu}_j &= \mathbf{M}_2\,\boldsymbol{\eta}_{j-1} \\ \mathbf{K}_2^T\,\boldsymbol{\eta}_j \qquad\qquad &= \mathbf{0}. \end{aligned} \tag{3.4b}$$

As a direct consequence of Lemma 2.18, one can see that $\mathbf{A}_{s_0}$ as above for $s_0 \geq 0$ is regular, which shows the well-posedness of all moments.

Lemma 3.14. *The recursion* (3.4) *is well-defined for any $s_0 \geq 0$ and $j \geq 0$.*

The moment matching method is a particularly suitable choice, as in this case the splitting into subspaces $\mathcal{W}_1$, $\mathcal{W}_2$ fulfilling (3.1) is almost immediate.

Lemma 3.15. *Let $\boldsymbol{\mu}_j$ and $\boldsymbol{\eta}_j$ for $j = 0, \ldots, L$, be defined by the recursion* (3.4)*. Let further,*

$$\mathcal{W}_1^L = \mathit{span}\{\boldsymbol{\mu}_0, \ldots, \boldsymbol{\mu}_L\}, \qquad \mathcal{W}_2^L = \mathit{span}\{\boldsymbol{\eta}_0, \ldots, \boldsymbol{\eta}_L\}.$$

Then for $s_0 > 0$, it holds $\mathbf{M}_1\mathcal{W}_1^L = \mathbf{J}\mathcal{W}_2^L$, whereas for $s_0 = 0$ it is instead $\mathbf{M}_1\mathcal{W}_1^{L-1} = \mathbf{J}\mathcal{W}_2^L$. Moreover, $\mathbf{K}_2^T\,\boldsymbol{\eta}_j = 0$ for any choice of $s_0 \geq 0$ and $j \geq 0$.

Theorem 3.16. *Given $s_0 \geq 0$, let $\mathcal{W}_1^L$ and $\mathcal{W}_2^L$ be as in Lemma 3.15. Let further,*

$$\mathcal{W}_1 := \mathcal{W}_1^L, \qquad \mathcal{W}_2 := \begin{cases} \mathcal{W}_2^L & \text{for } s_0 > 0 \\ \mathcal{W}_2^{L+1} & \text{for } s_0 = 0. \end{cases}$$

Then $\mathbf{V}_1, \mathbf{V}_2$ defined as in (3.3), fulfill Assumption 3.5. Moreover, the related reduced model, System 3.6, matches the first $2(L+1)$ moments at s_0 of System 3.1.

Proof. Most parts of Theorem 3.16 conclude the upper discussion and therefore have already been shown. The only assertion not yet discussed, is the improved moment matching result. We claim to match $2(L+1)$ moments, whereas in the general case of Galerkin-type moment matching, only half the amount can be guaranteed, cf. [Ant05, Gri97]. The improved result can be deduced from the special symmetries present in our system, which directly follows from the results of [Fre08]. □

Chapter 4

Numerical results

The numerical results of this chapter have already been published in [EKLS$^+$18]. The focus of the upcoming studies lies on the performance of our structure preserving model order reduction approach. To ensure negligible errors in all other discretization steps, those are carried out with sufficiently high resolution. If not stated otherwise, space discretization is realized by the finite element method with 1000 elements per edge and the ansatz spaces $Q_0(T_\mathcal{E})$, $\mathcal{P}_1(T_\mathcal{E})$ from (2.4). The time discretization employs 1000 steps with uniform time step size Δ_t, and is based on the θ-scheme with $\theta = \frac{1}{2} + \Delta_t$, [EK18], [Kug19].

The compatible reduced models in all numerical experiments are constructed in accordance to Theorem 3.16 with $s_0 = 0$ as chosen expansion frequency. Given the moment matching spaces $\mathcal{W}_1^L$, $\mathcal{W}_2^L$ and complementing spaces $\mathcal{Z}_1$, $\mathcal{Z}_2$ as in (3.3), our reduced models employ the ansatz spaces

$$\mathcal{V}_{r,1} = \mathcal{W}_1^L \cup \mathcal{Z}_1, \qquad \mathcal{V}_{r,2} = \mathcal{W}_2^{L+1} \cup \mathcal{Z}_2, \qquad\qquad \text{for } L \geq 0. \tag{4.1}$$

4.1 Structure-preservation

Standard non-compatible reduced models can in general not cover the essential structural properties, if they are under-resolved. To showcase this, we compare standard ROMs constructed by the moment matching spaces

$$\mathcal{V}_{r,1} = \mathcal{W}_1^L, \qquad \mathcal{V}_{r,2} = \mathcal{W}_2^L, \qquad\qquad \text{for } L \geq 0,$$

against their compatible counterpart from (4.1). Clearly the standard model does not fulfill Assumption 3.5.

Our first numerical experiments are carried out on the most simple network consisting of one pipe, Fig. 4.1. Throughout, zero boundary conditions are chosen at the right node ν_2, and the related column in the input matrix $\mathbf{B}$ of FOM, System 3.1, is removed, which leads to a system

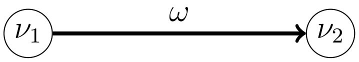

Figure 4.1: Network consisting of a single pipe of unit length $l^\omega = 1$. The model parameters are chosen as $a = b = r = 1$.

		exact	standard ROM			modified compatible ROM		
	L		0	2	9	0	2	9
projection	$M(0)$	1.000	0.750	0.902	0.949	1.000	1.000	1.000
	$H(0)$	0.500	0.375	0.451	0.475	0.500	0.500	0.500
mass constraint	$M(0)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$H(0)$	0.500	0.667	0.554	0.527	0.500	0.500	0.500

Table 4.1: Initial values of $M(0)$ and $H(0)$ for the mass and energy for full and the reduced order models obtained by projection in the energy norm with and without additional mass constraint for varying number of moments $L+1$.

with the single input $u^{\nu_1}(t)$. This setup is already appropriate to showcase the loss of essential structural properties for standard ROMs.

Remark 4.1. *Removing the columns relating to zero boundary conditions in the matrix* $\mathbf{B}$ *of FOM does not affect the simulation. It does, however, influence the moment matching procedure. This artificial simplification is only done for the tests using the one-edge network of Fig. 4.1.*

Errors in initial conditions

As discussed in Remark 3.7, our choice $\mathbf{x}_r(0) = \mathbf{V}^{\dagger}\mathbf{x}_0$ for the initial conditions of the reduced model provides the best approximation with respect to the initial Hamiltonian $H(0)$ for a given ansatz space $\mathcal{V}$. This alone, however, is not sufficient to guarantee mass conservation. Recall that Assumption 3.5-(A2) has been used to prove the conservation of mass in time. But in fact, it also guarantees that the reduced model reproduces the total mass $M(0)$ of the full order model at the initial time. The standard reduced model, on the other hand, may suffer from large deficiencies in the initial mass. As a remedy, which not employs our analysis, one could enforce mass conservation in the initial conditions as a constraint, similar to [CCS18].

For initial values of FOM chosen as

$$p(0,x) = 0, \qquad m(0,x) = 1, \qquad x \in \Omega, \tag{4.2}$$

the different strategies for the construction of initial values in the ROMs are compared in Table 4.1 for varying included moments L. For the standard non-compatible ROM, we observe a substantial miss-specification of the total mass at initial time if the initial conditions are chosen by the energy projection. Exact representation of the total mass via the constraint, on the other hand, leads to an artificial increase of the initial energy. The size of both defects can be reduced by increasing the approximation order L, which allows to approximate the initial conditions better and better. Our compatible ROM, on the other hand, satisfies Assumption 3.5 and therefore leads to the exact representation of the mass and a good approximation of the energy at the same time. For the problem under investigation, the energy can even be represented exactly.

Exponential stability

We next consider the influence of the basis construction on the exponential stability of the reduced models. We again set the initial values as in (4.2), and choose the input at the left boundary

	exact	standard ROM			modified compatible ROM		
$t \setminus L$		0	2	9	0	2	9
0	0.5000	0.3750	0.4512	0.4745	0.5000	0.5000	0.5000
1	0.1528	0.3750	0.1665	0.1514	0.1876	0.1552	0.1527
2	0.0512	0.3750	0.0708	0.0509	0.0690	0.0511	0.0511
3	0.0174	0.3750	0.0384	0.0173	0.0245	0.0173	0.0174
4	0.0059	0.3750	0.0273	0.0059	0.0084	0.0059	0.0059

Table 4.2: Energy decay $H(t)$ for the full order model, the standard non-compatible reduced model and the modified compatible reduced model for varying number $L+1$ of included moments.

conditions as $u^{\nu_1}(t) = 1$ for $t \geq 0$. In Table 4.2, we display the values of the energy for the full order and the reduced models for a sequence of time steps.

The modified compatible ROM yields uniform exponential decay of the energy in all cases. Already for $L = 2$, the energy is predicted accurately over the whole time interval. The standard reduced model, on the other hand, underestimates the initial energy and does not provide the correct decay rate for small L. For the smallest model with $L = 0$, we do not even observe any decay in energy at all.

4.2 Numerical robustness

Recall that our underlying Galerkin framework in Chapter 2 suggests the interpretation of the columns of the reduction bases $\mathbf{V}_1$ and $\mathbf{V}_2$ as orthogonal bases for the subspaces $\mathcal{V}_{r,1}$ and $\mathcal{V}_{r,2}$. In fact, any single moment corresponds to the solution of an elliptic boundary value problem, which explains why the basis functions are smooth. For the one-pipe network Fig. 4.1, which is again used throughout this section, the underlying spatial domain can be identified with the interval $[0,1]$. In Fig. 4.2, we display our basis functions (4.1) obtained for $L = 3$.

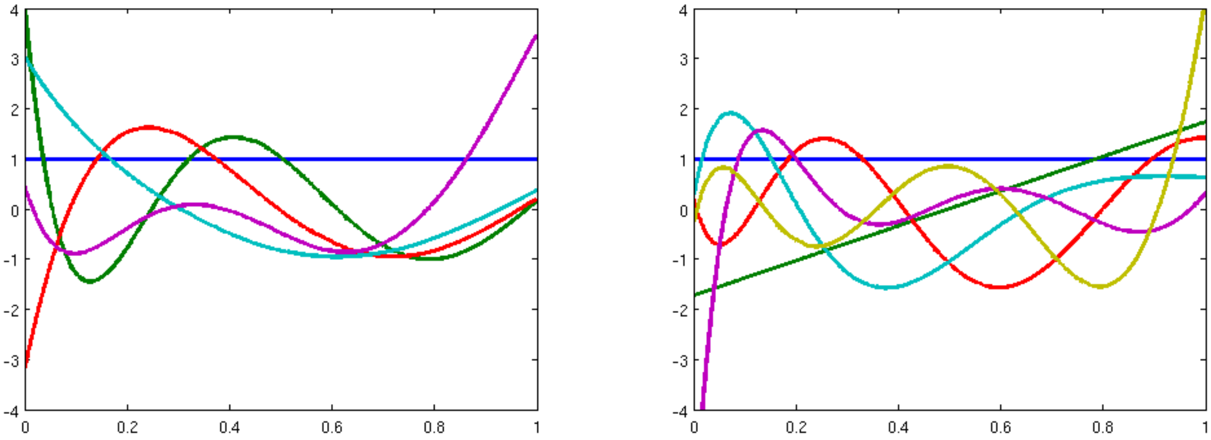

Figure 4.2: Bases for the subspaces $\mathcal{V}_{r,1} = \mathcal{W}_1 + \mathcal{Z}_1$ *(left)* and $\mathcal{V}_{r,2} = \mathcal{W}_2 + \mathcal{Z}_2$ *(right)* from (4.1) obtained for $L = 3$. The resulting dimensions are $\dim(\mathcal{V}_{r,1}) = 5$ and $\dim(\mathcal{V}_{r,2}) = 6$ here.

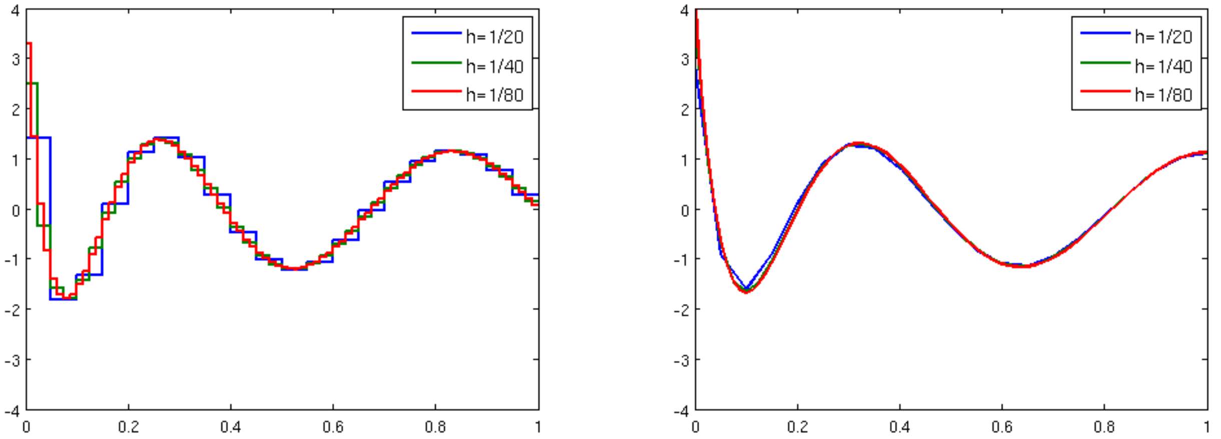

Figure 4.3: Basis functions computed with the same Krylov iteration but using different full order models obtained by finite element discretization with mesh sizes $h = \frac{1}{20}$, $\frac{1}{40}$ and $\frac{1}{80}$. *Left:* pressure; *Right:* mass flow.

Also note that the functions look similar to a sequence of orthogonal polynomials of increasing degree. This indicates that the projection onto the moment spaces leads to some sort of higher order approximation.

Mesh independence

In the formulation of our algorithms, we paid special attention to a construction that respects the underlying function space setting. As a consequence, the subspaces $\mathcal{V}_{r,i}$ and even the corresponding bases turn out to be almost independent of the underlying full order model. To illustrate this fact, we display in Fig. 4.3 one of the basis functions for pressure and mass flow computed with full order models resulting from discretization on different meshes. We choose rather coarse discretizations for the full order model here in order to visualize the differences clearly.

Note that the basis functions for different levels coincide almost perfectly, up to discretization errors. This clearly demonstrates the mesh independence of the proposed algorithms. As mentioned before, all algorithms could even be formulated directly for the infinite dimensional problem and the basis functions depicted in Fig. 4.3 thus correspond to approximations for the corresponding functions that would be obtained by the Krylov iteration in infinite dimensions.

Stability of the splitting step

In our basis construction we use the cosine-sine decomposition in order to improve the numerical stability of the splitting $\mathbf{W} = [\mathbf{W}_1; \mathbf{W}_2]$. A simple splitting with re-orthogonalization of $\mathbf{W}_1$ and $\mathbf{W}_2$, on the other hand, could be realized as follows.

```
function [W1,W2]=simplesplit(W1,W2,M1,M2,tolerance)
      W1 = ortho(W1,[],M1,tolerance);
      W2 = ortho(W2,[],M1,tolerance);
end
```

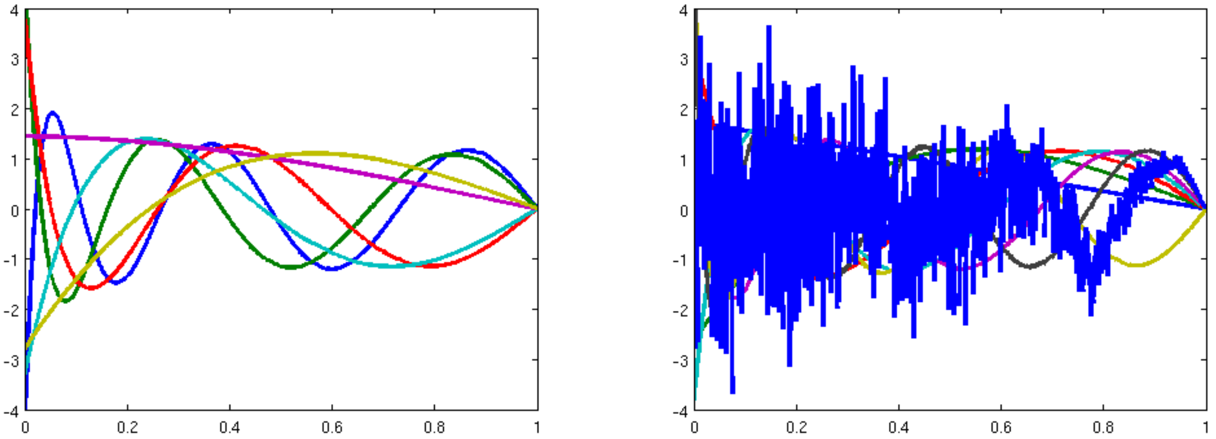

Figure 4.4: Basis functions for the pressure space $\mathcal{W}_1^L$ obtained after $L = 9$ Krylov iterations and splitting with the cosine-sine decomposition *(left)*, and the simple splitting and re-orthogonalization *(right)*.

In Fig. 4.4, we compare the results obtained by splitting the spaces $\mathcal{W}_i^L$ of moments by this simple strategy with those obtained by means of the cosine-sine decomposition.

Due to the possibility of interpreting the basis vectors as functions on the interval $[0, 1]$, one can easily conclude that already for relatively small dimensions, the standard splitting suffers from severe numerical instabilities. Let us emphasize that this is caused only by the instability of the splitting step and not by the employed Arnoldi iteration defining the moment spaces. The splitting via cosine-sine decomposition, on the other hand, does not suffer from these instabilities and should therefore always be preferred in practice. Let us refer to [GVL96] for further discussion.

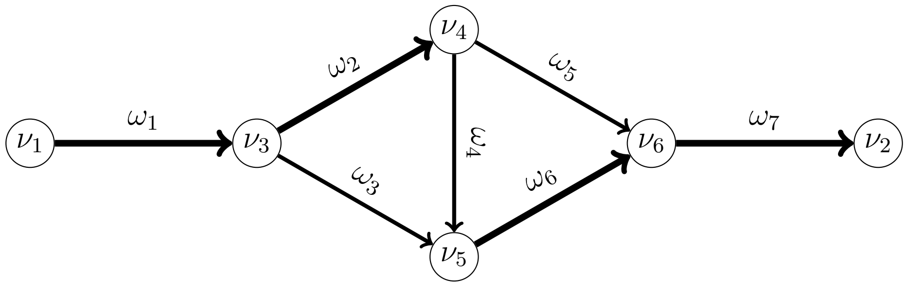

Figure 4.5: Small network example with edges of unit length. The model parameters for the edges are specified by the vectors $a = [4, 4, 1, 1, 1, 4, 4]$, $b = \left[\frac{1}{4}, \frac{1}{4}, 1, 1, 1, \frac{1}{4}, \frac{1}{4}\right]$ and $r = r_0 \cdot \left[\frac{1}{8}, \frac{1}{8}, 1, 1, 1, \frac{1}{8}, \frac{1}{8}\right]$ with constant damping parameter r_0 varying over the test cases.

4.3 Approximation of input-output behavior

By Theorem 3.16 we know that the reduced models obtained with our algorithms exactly match the first few moments of the transfer function, from which we can expect a good overall approximation

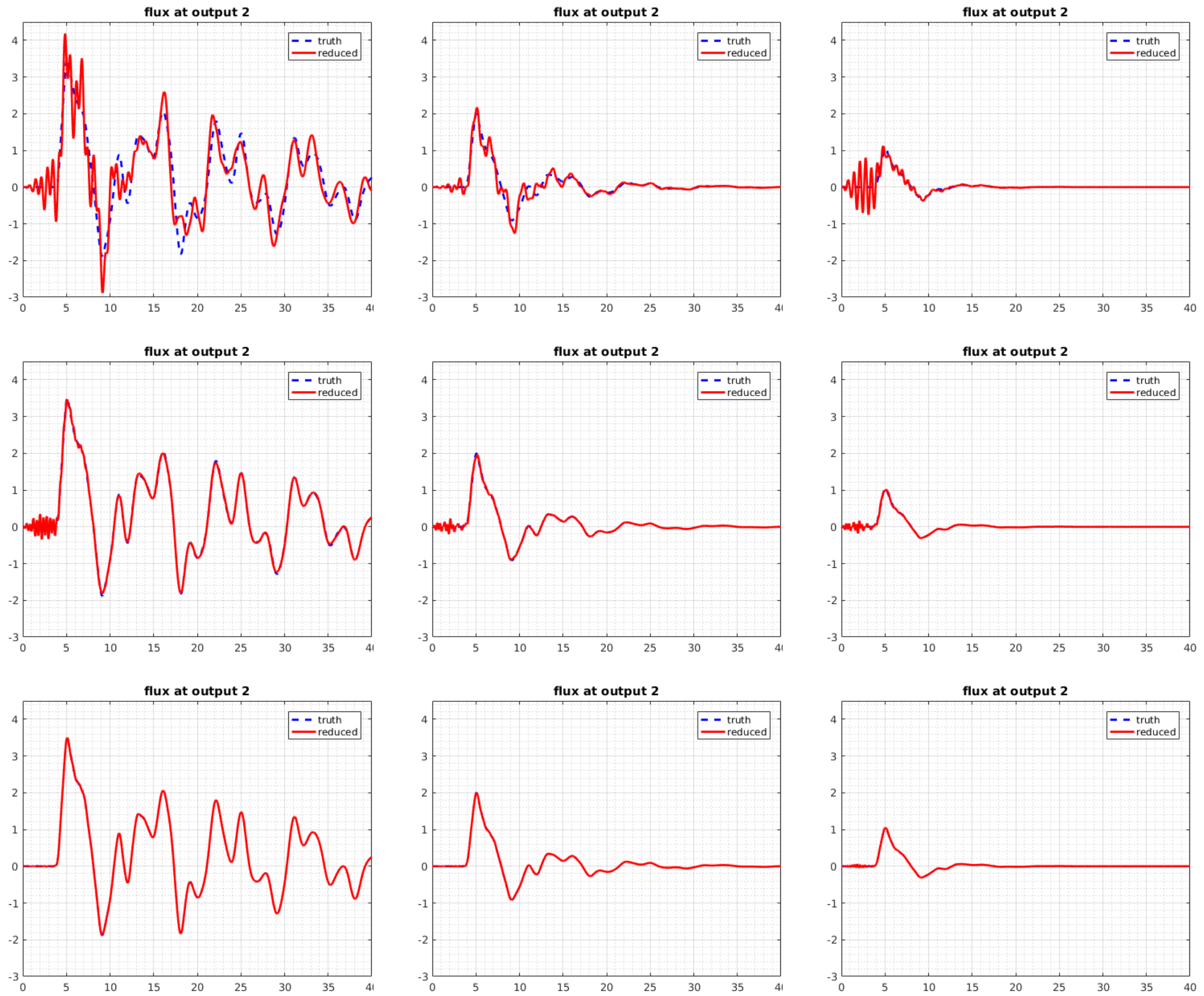

Figure 4.6: Mass inflow $y^{\nu_2}(t) = m(t, \nu_2)n[\nu_2]$ over boundary node ν_2 for the network test problem. Results are displayed for FOM (blue) and the ROM (red) for damping parameters $r_0 = 0.1, 0.5, 1$ (left to right) and $L = 4, 9, 19$ (top to bottom).

of the input-output behavior in frequency domain. With the following tests, we take a closer look also at the approximation in time domain. First, we consider the small network depicted in Fig. 4.5. The chosen edge-parameters correspond to the modeling of pipes with different cross-sectional areas.

The initial- and boundary-conditions are specified as

$$
\begin{aligned}
&p(0,x) = 0, \qquad m(0,x) = 0, && x \in \Omega \\
&u^{\nu_1}(t) = \begin{cases} t, & 0 \le t < 1 \\ 2 - t, & 1 \le t < 2 \\ 0, & t \ge 2 \end{cases} \qquad \text{and} \qquad u^{\nu_2}(t) = 0, && t \ge 0
\end{aligned}
\tag{4.3}
$$

For varying damping parameter r_0, we compare the mass inflow of the FOM and the ROM over time at the vertex ν_2 in Fig. 4.6. The resulting mass-flow-profile is relatively complex due to

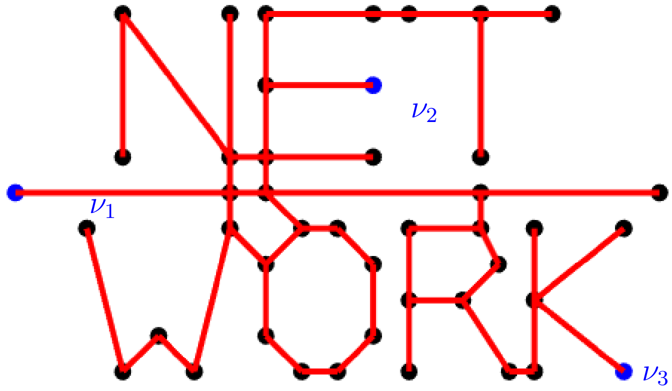

Figure 4.7: Large network example with edges of unit length. The model parameters are globally chosen as $a = b = 1$ and $r = 0.5$. The boundary nodes are depicted in blue.

multiple pathways through the network and possible reflections at the junctions. The oscillations in the initial phase of the output are due to a Gibbs phenomenon. This effect, however, becomes negligible when increasing the dimension of the reduced models. For $L = 19$, which here corresponds to $\dim(\mathcal{V}_1) = 37$ and $\dim(\mathcal{V}_2) = 44$, we already observe an almost perfect prediction of the input-output behavior. The plots displayed in Fig. 4.6 also illustrate the exponential decay of the output which becomes faster when the damping factor r_0 is increased.

4.4 Results for a larger network

To demonstrate the viability and efficiency of the proposed model reduction approach also for larger problems, we consider as a last test case the network depicted in Fig. 4.7. The full order model is of dimension 94.078 in this case.

In Fig. 4.8, we plot the mass fluxes over the three boundary nodes which result as response to the input $\mathbf{u} = [u^{\nu_1}; u^{\nu_2}; u^{\nu_3}]$ with component u^{ν_1} as defined in (4.3) and $u^{\nu_2} = u^{\nu_3} \equiv 0$.

For $L = 9$ Krylov iterations, we obtain a reduced model with $\dim(V_1) = 19$ and $\dim(V_2) = 66$ here, which is not capable to capture the relevant system behavior. Setting $L = 19$, results in a reduced model with $\dim(V_1) = 34$ and $\dim(V_2) = 81$, which already allows to predict the input-output behavior rather well. For $L = 39$, we obtain a reduced model with $\dim(V_1) = 61$ and $\dim(V_2) = 108$ which yields an almost perfect reproduction of the system output.

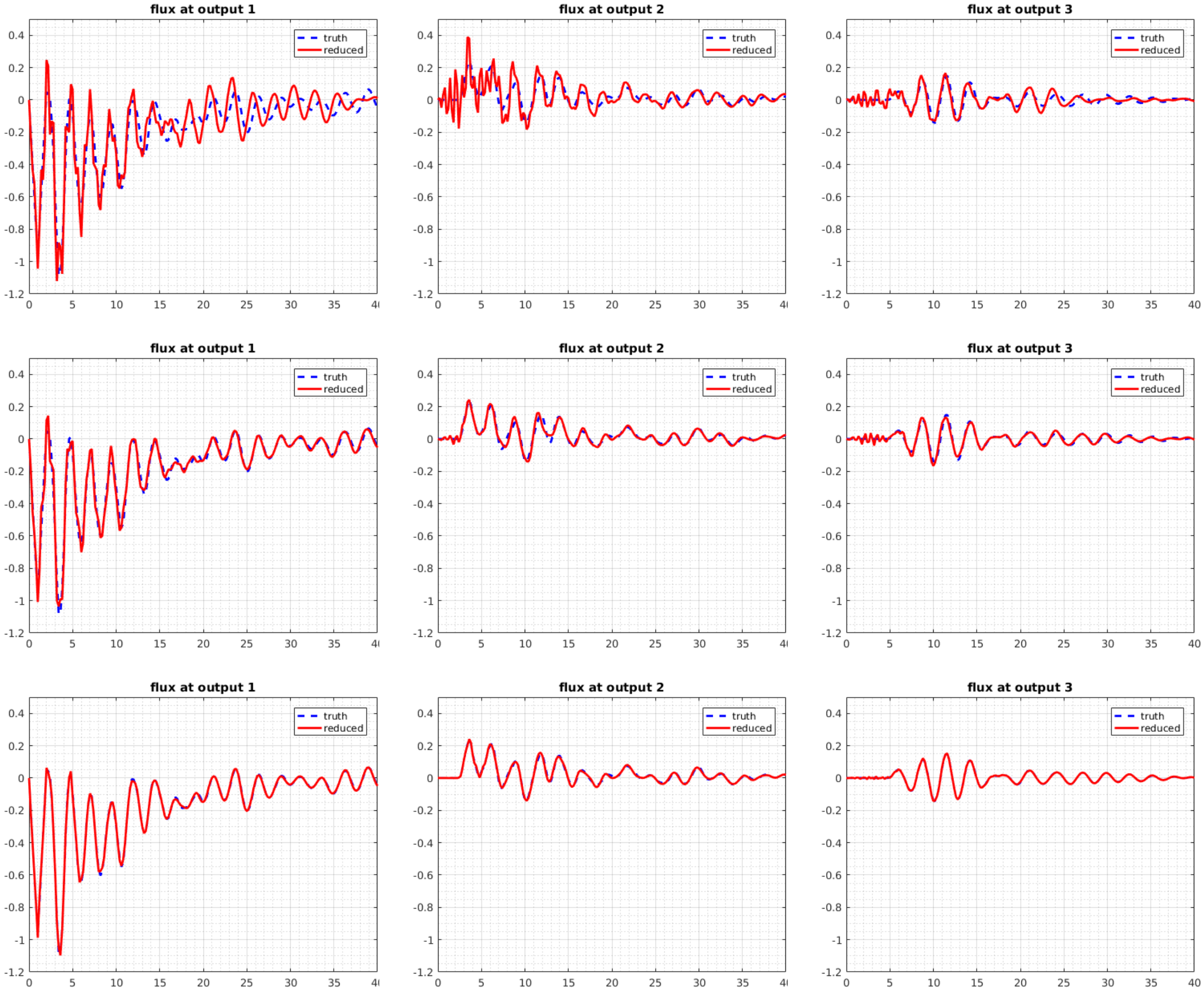

Figure 4.8: Mass-flux across the output ports ν_1, ν_2, ν_3 (left to right) obtained with the full order model (blue) and the reduced order models (red) with $L = 9, 19, 39$ (top to bottom).

Subpart I.B

Nonlinear flow problem

Chapter 1

Preliminaries

This part is concerned with the approximation of our general nonlinear flow problem on networks, see the prior introduction of Part I, or (2.1) in Section 2.2 below. In this chapter we shortly review the commonalities and differences in the derivations of this part to the ones in Part I.A and introduce the crucial additional concepts needed in this part. In particular, the treatment of the network structure follows very similar for our kind of approximations. The major difference to the linear case is that additional tools from convex analysis are needed here, first and foremost the partial Legendre transform and its related variable transform.

1.1 Commonalities and differences to the linear case

Commonalities

Notation and framework is adopted from Part I-A where possible. In particular, the handling of network aspects completely relies on the framework of Part I.A. Let us therefore briefly recapitulate a few definitions of Part I.A-Section 1.3. A network $(\mathcal{E}, \mathcal{N}, l)$ is assumed to be given, fulfilling Assumption I.A-1.2. This means $\mathcal{N} = \mathcal{N}_0 \cup \mathcal{N}_\partial$, and every boundary node $\nu \in \mathcal{N}_\partial$ has exactly one incident edge. Recall that the spatial domain is described by the interconnection of edges and every edge can be identified with an interval $\omega \cong (0, l^\omega)$. Consequently, function spaces acting on the network can be introduced by the composition of edgewise function spaces, e.g., $\mathcal{L}^2(\mathcal{E})$, $\mathcal{H}^1_{pw}(\mathcal{E})$ and $\mathcal{H}^1_{div}(\mathcal{E})$. Recall that the related $\mathcal{L}^2$-product was also defined in terms of its edgewise contributions, i.e.,

$$\langle b, \tilde{b} \rangle_{\mathcal{E}} = \sum_{\omega \in \mathcal{E}} \langle b_{|\omega}, \tilde{b}_{|\omega} \rangle_\omega \qquad \text{and} \qquad ||b||_{\mathcal{E}} = \sqrt{\langle b, b \rangle_{\mathcal{E}}}$$

Subscripts are used to indicate the integration domain. When we integrate over the whole spatial domain, we also write $\langle \cdot, \cdot \rangle$ instead of $\langle \cdot, \cdot \rangle_{\mathcal{E}}$. The broken derivative-operator has been introduced as the edgewise weak derivative

$$\partial_x : \mathcal{H}^1_{pw}(\mathcal{E}) \to \mathcal{L}^2(\mathcal{E}), \qquad (\partial_x b)_{|\omega} = \partial_x b_{|\omega}, \qquad \text{for all } \omega \in \mathcal{E},$$

and the boundary operator $\mathcal{T} : \mathcal{H}^1_{pw}(\mathcal{E}) \to \mathbb{R}^{|\mathcal{N}_\partial|}$ by means of the trace theorem.

Remark 1.1. *In practice, the edges $\omega \in \mathcal{E}$ may have different weights A^ω associated to them. For ease of presentation, we assume $A^\omega = 1$ throughout the main part and postpone the generalization to $A^\omega \neq 1$ to Section 2.6.*

Remark 1.2. *When considering the equations on one edge only, the problem can be simplified as follows: The spatial domain $\Omega = (0, l^\omega)$ and the boundary $\mathcal{N}_\partial = \{0, l^\omega\}$ can be assumed. The set of inner nodes $\mathcal{N}_0$ is empty, and the coupling equations,* (2.1b) *below, do not appear. Also $\mathcal{E}$ can be replaced by $\Omega = (0, l^\omega)$ in all function spaces, and the scalar product on $\mathcal{L}^2(\mathcal{E})$ needs to be replaced by the standard scalar product on $\mathcal{L}^2((0, l^\omega))$. The boundary operator simplifies to*

$$\mathcal{T} b = \begin{bmatrix} -b[l^\omega] \\ b[0] \end{bmatrix}, \qquad \textit{for } b \in \mathcal{H}^1((0, l^\omega)).$$

With this simplification, most upcoming derivations can be followed without the framework of Part I.A.

Differences

According to the port-Hamiltonian wording, our weak formulations of Part I-A used a parametrization in the effort variable $\mathbf{e} = \nabla_{\mathbf{z}} H(\mathbf{z})$. Although a similar theory could be derived in effort variables for the nonlinear case, see [Egg19], this choice of parametrization is not viable for the application we have mainly in mind, which are the barotopic Euler equations. We instead rely on a mixed parametrization, composed of parts of the energy variable $\mathbf{z}$ and the effort variable $\mathbf{e}$, in this part. The resulting space discretization is similar to [Egg18] for the barotropic Euler equations, cf. Chapter 4, but we additionally preserve port-Hamiltonian structure. Moreover, the transformation between energy variables and effort variables is now generally nonlinear and needs more care in the examinations. To cope for that, we make heavy use of the so-called partial Legendre transform and the theory behind it, see Section 1.2.

Further, recall that we almost exclusively are concerned with Galerkin-type methods in space in Part I-A. Both, the space discretization by a finite element method and the projection-based model order reduction fit into the same Galerkin framework. To efficiently evaluate nonlinearities in projection-based reduced order models, it is necessary to add a so-called complexity reduction step in this part. We derive a structure-preserving complexity reduction by a quadrature-type approximations of integrals involving nonlinearities here, similar to [HCF17], [Jam08], [FACC14]. Finally, as existing structure-preserving time discretization schemes for port-Hamiltonian systems from literature, cf. [LM17], [Egg19], [MM19], cannot be applied to our formulations, due to our rather unconventional mixed parametrization in effort and energy variables, we deduce new structure-preserving time discretization schemes for our proposed parametrization. All in all, we go through each of the approximation steps illustrated in Fig. 1 in the introduction of Part I.

1.2 Tools from convex analysis

We present a concise summary of the results from convex analysis we draw from. For further reading, we refer to [RW98], [Roc70] and references therein.

Definition 1.3 (Convexity)**.** *Let $\mathbb{Z} \subset \mathbb{R}^m$ be a convex set.*

- *A function $\phi : \mathbb{Z} \to \mathbb{R}$ is called convex, if for every $\mathbf{z}, \bar{\mathbf{z}} \in \mathbb{Z}$ it holds*

$$\phi(\lambda \mathbf{z} + (1-\lambda)\bar{\mathbf{z}}) \leq \lambda \phi(\mathbf{z}) + (1-\lambda)\phi(\bar{\mathbf{z}}), \qquad \textit{for } \lambda \in (0,1).$$

- *A convex function is called strictly convex, when the upper inequality holds strictly for $\mathbf{z} \neq \bar{\mathbf{z}}$.*
- *A function $\Psi : \mathbb{Z} \to \mathbb{R}$ is called (strictly) concave, when $\mathbf{z} \mapsto -\Psi(\mathbf{z})$ is (strictly) convex.*

The following characterizations of strict convexity are used throughout the work, see [RW98, Theor. 2.14].

Theorem 1.4. *Let $\mathbb{Z} \subset \mathbb{R}^m$ be convex and $\phi : \mathbb{Z} \to \mathbb{R}$. Then it holds:*

1. *If ϕ is differentiable, it is strictly convex, if and only if,*

$$\phi(\bar{\mathbf{z}}) - \phi(\mathbf{z}) > \nabla_{\mathbf{z}} \phi(\mathbf{z}) \cdot (\bar{\mathbf{z}} - \mathbf{z}), \qquad \textit{for all } \mathbf{z}, \bar{\mathbf{z}} \in \mathbb{Z}.$$

2. *If ϕ is differentiable, it is strictly convex, if and only if,*

$$(\nabla_{\mathbf{z}} \phi(\bar{\mathbf{z}}) - \nabla_{\mathbf{z}} \phi(\mathbf{z})) \cdot (\bar{\mathbf{z}} - \mathbf{z}) > 0, \qquad \textit{for all } \mathbf{z}, \bar{\mathbf{z}} \in \mathbb{Z}.$$

3. *Let ϕ be twice differentiable. If its Hessian $\nabla_{\mathbf{zz}} \phi(\mathbf{z})$ is positive definite for all $\mathbf{z} \in \mathbb{Z}$, then it is also strictly convex. Moreover, it is convex, if and only if, the Hessian is positive semi-definite.*

Note that the characterization Theorem 1.4-(3) of strict convexity by the Hessian is only sufficient, but not necessary. An example for a smooth strict convex function with a Hessian not being positive definite is $x \mapsto x^4$. Clearly, its Hessian $x \mapsto 12x^2$ is zero at $x = 0$.

Legendre-transformation

Definition 1.5. *For a function $\phi : \mathbb{Z} \to \mathbb{R}$ with domain $\mathbb{Z} \subset \mathbb{R}^m$, the Legendre-transform ϕ^* is defined by*

$$\phi^*(\mathbf{y}) := \sup_{\mathbf{z} \in \mathbb{Z}} \mathbf{y} \cdot \mathbf{z} - \phi(\mathbf{z}), \qquad \mathbf{y} \in \mathbb{R}^m.$$

There exists a well-established theory around the duality of convexity- and differentiability-properties of a convex function and its Legendre transform. We give here a basic result, which is a special case of [RW98, Theorem 11.13].

Theorem 1.6. *Let $\mathbb{Z} \subset \mathbb{R}^m$ be an open convex set, and $\phi : \mathbb{Z} \to \mathbb{R}$ be differentiable and strictly convex. Then its Legendre transform ϕ^* is differentiable and strictly convex on every open convex subset of its domain. Moreover, the Legendre transform $(\phi^*)^*$ of ϕ^* is equal to ϕ.*

The Legendre transform ϕ^* of ϕ is constructed such that its gradient, if it exists, is the inverse function of the gradient of ϕ, as the following theorem states.

Theorem 1.7. *Let $\mathbb{Z} \subset \mathbb{R}^m$ be an open convex set, and $\phi : \mathbb{Z} \to \mathbb{R}$ be differentiable and strictly convex. Then the mapping $\mathbf{z} \mapsto \nabla_{\mathbf{z}}\phi(\mathbf{z})$ is injective, and its inverse with $\hat{\mathbf{z}}(\nabla_{\mathbf{z}}\phi(\mathbf{z})) = \mathbf{z}$ for all $\mathbf{z} \in \mathbb{Z}$ reads*

$$\hat{\mathbf{z}} : \nabla_{\mathbf{z}}\phi(\mathbb{Z}) \to \mathbb{Z}, \qquad \hat{\mathbf{z}}(\mathbf{y}) = \nabla_{\mathbf{y}}\phi^*(\mathbf{y}).$$

Moreover, the Legendre transform can equally be characterized as

$$\phi^*(\mathbf{y}) = \hat{\mathbf{z}}(\mathbf{y}) \cdot \mathbf{y} - \phi(\hat{\mathbf{z}}(\mathbf{y})), \qquad \textit{for } \mathbf{y} \in \nabla_{\mathbf{z}}\phi(\mathbb{Z}).$$

Proof. Let $\bar{\mathbf{z}}, \tilde{\mathbf{z}} \in \mathbb{Z}$ with $\bar{\mathbf{z}} \neq \tilde{\mathbf{z}}$. For injectivity of the map $\mathbf{z} \mapsto \nabla_{\mathbf{z}}\phi(\mathbf{z})$, we have to show that $\nabla_{\mathbf{z}}\phi(\bar{\mathbf{z}}) \neq \nabla_{\mathbf{z}}\phi(\tilde{\mathbf{z}})$. The latter is immediate by the strict convexity of ϕ and the second characterization of Theorem 1.4 of strict convexity. By that the map is injective, and the inverse $\hat{\mathbf{z}} : \nabla_{\mathbf{z}}\phi(\mathbb{Z}) \to \mathbb{Z}$ exists.

For the next part, recall that by Definition 1.5, the Legendre transform is given as the supremum over the map $\Psi : \mathbf{z} \mapsto \mathbf{y} \cdot \mathbf{z} - \phi(\mathbf{z})$. As Ψ is strictly concave and differentiable, its supremum is uniquely realized at $\bar{\mathbf{z}}$ with

$$\mathbf{0} \overset{!}{=} \nabla_{\mathbf{z}}\Psi\left(\bar{\mathbf{z}}\right) = \mathbf{y} - \nabla_{\mathbf{z}}\phi(\bar{\mathbf{z}}).$$

Clearly, $\bar{\mathbf{z}} = \hat{\mathbf{z}}(\mathbf{y})$, with $\hat{\mathbf{z}} : \nabla_{\mathbf{z}}\phi(\mathbb{Z}) \to \mathbb{Z}$ the inverse to $\mathbf{z} \mapsto \nabla_{\mathbf{z}}\phi(\mathbf{z})$, fulfills the latter equation. That verifies the claimed characterization of ϕ^*. Differentiating the just derived characterization of ϕ^* then also shows $\hat{\mathbf{z}}(\mathbf{y}) = \nabla_{\mathbf{y}}\phi^*(\mathbf{y})$. □

Remark 1.8. *Many authors also use the label 'convex conjugate' or 'Legendre-Fenchel transform' for ϕ^* defined in the general setting of Definition 1.5. We use the more classical label 'Legendre transform', which is typically used, when the additional assumptions of Theorem 1.6 hold.*

Partial Legendre-transformation

Instead of applying the Legendre transform onto all components, one can apply it only to parts of the components. This yields the partial Legendre transformation. Accordingly, we separate in the upcoming the vectors $\mathbf{a}, \mathbf{z} \in \mathbb{R}^n$ into sub-vectors

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}, \qquad \text{with } \mathbf{a}_i, \mathbf{z}_i \in \mathbb{R}^{n_i}, \quad n = n_1 + n_2.$$

Related state transformations in $\mathbb{R}^n$ are denoted by $\hat{\mathbf{a}}(\cdot), \hat{\mathbf{z}}(\cdot)$, for which also sub-components $\hat{\mathbf{a}}_i(\cdot), \hat{\mathbf{z}}_i(\cdot)$ with images in $\mathbb{R}^{n_i}$ are introduced. We consider the partial Legendre transform w.r.t. the second component throughout.

Definition 1.9. *Let $\mathbb{Z} \subset \mathbb{R}^n$ be a convex set. The partial Legendre transformation of $h : \mathbb{Z} \mapsto \mathbb{R}$ is defined as*

$$g(\mathbf{a}) = \sup_{\mathbf{z}_2 \in \{\bar{\mathbf{z}}_2 : [\mathbf{a}_1; \bar{\mathbf{z}}_2] \in \mathbb{Z}\}} \mathbf{a}_2 \cdot \mathbf{z}_2 - h([\mathbf{a}_1; \mathbf{z}_2]), \qquad \textit{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}.$$

Theorem 1.10. *Let $\mathbb{Z} \subset \mathbb{R}^n$ be a convex set. For $h : \mathbb{Z} \to \mathbb{R}$, let $\mathbf{z}_2 \mapsto h([\bar{\mathbf{z}}_1; \mathbf{z}_2])$ be strictly convex and differentiable for fixed $\bar{\mathbf{z}}_1$. Then the map*

$$\hat{\mathbf{a}} : \mathbb{Z} \to \mathbb{R}^n, \qquad \hat{\mathbf{a}}(\mathbf{z}) = \begin{bmatrix} \mathbf{z}_1 \\ \nabla_{\mathbf{z}_2} h(\mathbf{z}) \end{bmatrix}$$

is injective, and its inverse with $\hat{\mathbf{z}}(\hat{\mathbf{a}}(\mathbf{z})) = \mathbf{z}$ exists and is of the form

$$\hat{\mathbf{z}} : \mathbb{A} \to \mathbb{Z}, \quad \hat{\mathbf{z}}(\mathbf{a}) = \begin{bmatrix} \mathbf{a}_1 \\ \hat{\mathbf{z}}_2(\mathbf{a}) \end{bmatrix}, \qquad \text{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}$$

with $\mathbb{A} = \hat{\mathbf{a}}(\mathbb{Z}) \subset \mathbb{R}^n$.

Proof. Let $\bar{\mathbf{z}} = [\bar{\mathbf{z}}_1; \bar{\mathbf{z}}_2]$ and $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1; \tilde{\mathbf{z}}_2]$ be in $\mathbb{Z}$ with $\bar{\mathbf{z}} \neq \tilde{\mathbf{z}}$. Injectivity of $\hat{\mathbf{a}}(\cdot)$ is then equivalent to $\hat{\mathbf{a}}(\bar{\mathbf{z}}) \neq \hat{\mathbf{a}}(\tilde{\mathbf{z}})$. When $\bar{\mathbf{z}}_1 \neq \tilde{\mathbf{z}}_1$, this holds trivially. Let us therefore assume $\bar{\mathbf{z}}_1 = \tilde{\mathbf{z}}_1$, implying that $\bar{\mathbf{z}}_2 \neq \tilde{\mathbf{z}}_2$. The strict convexity of $\mathbf{z}_2 \mapsto h([\bar{\mathbf{z}}_1; \mathbf{z}_2])$ implies by Theorem 1.4 that

$$(\nabla_{\mathbf{z}_2} h([\bar{\mathbf{z}}_1; \bar{\mathbf{z}}_2]) - \nabla_{\mathbf{z}_2} h([\bar{\mathbf{z}}_1; \tilde{\mathbf{z}}_2])) \cdot (\bar{\mathbf{z}}_2 - \tilde{\mathbf{z}}_2) > 0,$$

i.e., $\nabla_{\mathbf{z}_2} h(\bar{\mathbf{z}}) \neq \nabla_{\mathbf{z}_2} h(\tilde{\mathbf{z}})$, which shows the injectivity. Clearly, its inverse $\hat{\mathbf{z}} : \mathbb{A} \to \mathbb{Z}$ then exists and is of the claimed form. □

Theorem 1.11. *Assume that the requirements of Theorem 1.10 hold true. Then the partial Legendre transformation g of h in Definition 1.9 can also be characterized by*

$$g : \mathbb{A} \to \mathbb{R}, \quad g(\mathbf{a}) = \hat{\mathbf{z}}_2(\mathbf{a}) \cdot \mathbf{a}_2 - h(\mathbf{a}_1, \hat{\mathbf{z}}_2(\mathbf{a})), \qquad \text{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} \in \mathbb{A}.$$

Moreover, it holds

$$\nabla_{\mathbf{a}_1} g(\mathbf{a}) = -\nabla_{\mathbf{z}_1} h(\mathbf{z})_{|\mathbf{z}=\hat{\mathbf{z}}(\mathbf{a})}, \qquad \nabla_{\mathbf{a}_2} g(\mathbf{a}) = \hat{\mathbf{z}}_2(\mathbf{a}),$$

and $\phi : \mathbf{a}_2 \mapsto g([\bar{\mathbf{a}}_1; \mathbf{a}_2])$ is strictly convex.

Proof. Let $\mathbf{a} = [\mathbf{a}_1; \mathbf{a}_2] \in \mathbb{A}$ be given. By construction, $\psi : \mathbf{z}_2 \mapsto \mathbf{a}_2 \cdot \mathbf{z}_2 - h(\mathbf{a}_1, \mathbf{z}_2)$ is strictly concave and differentiable. Its supremum is therefore uniquely realized for $\bar{\mathbf{z}}_2$ with

$$\mathbf{0} \overset{!}{=} \nabla_{\mathbf{z}_2} \psi(\bar{\mathbf{z}}_2) = \mathbf{a}_2 - \nabla_{\mathbf{z}_2} h(\mathbf{z})_{|\mathbf{z}=[\mathbf{a}_1; \bar{\mathbf{z}}_2]}.$$

Clearly, this holds for the choice $\bar{\mathbf{z}}_2 = \hat{\mathbf{z}}_2(\mathbf{a})$ with $\hat{\mathbf{z}}_2(\cdot)$ as in Theorem 1.10, which shows the claimed representation for the partial Legendre transformation. The equalities $\nabla_{\mathbf{a}_1} g(\mathbf{a}) = -\nabla_{\mathbf{z}_1} h(\mathbf{z})_{|\mathbf{z}=\hat{\mathbf{z}}(\mathbf{a})}$ and $\nabla_{\mathbf{a}_2} g(\mathbf{a}) = \hat{\mathbf{z}}_2(\mathbf{a})$ follow by straight forward calculations. As the partial Legendre transform can be considered as a Legendre transform of a function parametrized in $\mathbf{a}_1 = \mathbf{z}_1$, strict convexity of ϕ can be deduced from Theorem 1.6. □

Generalizations to other scalar products

Up to now, we have tacitly used the standard Euclidean scalar product, $(\mathbf{z}, \bar{\mathbf{z}}) \mapsto \mathbf{z} \cdot \bar{\mathbf{z}}$, in all our definitions. But the gradient and the Legendre transform are actually inner-product-dependent concepts, and other inner products as the Euclidean one will naturally appear later, when we consider our Galerkin approximations.

Definition 1.12. *Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\mathbf{M} \in \mathbb{R}^{n,n}$ a positive definite symmetric matrix inducing the inner product*

$$(\mathbf{z}, \bar{\mathbf{z}})_{\mathbf{M}} = \mathbf{z} \cdot (\mathbf{M}\bar{\mathbf{z}}), \qquad \textit{for } \mathbf{z}, \bar{\mathbf{z}} \in \mathbb{R}^N.$$

Then the gradient $\nabla_{\mathbf{M},\mathbf{z}}\phi$ with respect to the inner product $(\cdot,\cdot)_{\mathbf{M}}$ is defined by

$$(\nabla_{\mathbf{M},\mathbf{z}}\phi(\mathbf{z}), \bar{\mathbf{z}})_{\mathbf{M}} = \nabla_{\mathbf{z}}\phi(\mathbf{z}) \cdot \bar{\mathbf{z}}, \qquad \textit{for } \mathbf{z}, \bar{\mathbf{z}} \in \mathbb{R}^N.$$

It can be expressed by the Euclidean gradient as $\nabla_{\mathbf{M},\mathbf{z}}\phi(\mathbf{z}) = \mathbf{M}^{-1}\nabla_{\mathbf{z}}\phi(\mathbf{z})$.

Note that the criteria for strict convexity in terms of the gradient, Theorem 1.4, is invariant under the used scalar product. For the (partial) Legendre transform w.r.t. any other inner product, all results from before can be adapted. We only give the summarized results here, as the proofs are very similar to the Euclidean case.

Definition 1.13 (Generalization of Def. 1.9)**.** *Let $\mathbb{Z} = \mathbb{Z}_1 \times \mathbb{Z}_2 \subset \mathbb{R}^{n_1+n_2}$ be a convex set and $\mathbf{M} \in \mathbb{R}^{n_2,n_2}$ be symmetric positive definite. The partial Legendre transformation of $h : \mathbb{Z} \mapsto \mathbb{R}$ w.r.t. the inner product $(\cdot,\cdot)_{\mathbf{M}}$ is defined as*

$$g_{\mathbf{M}}(\mathbf{a}) = \sup_{\mathbf{z}_2 \in \{\bar{\mathbf{z}}_2 : [\mathbf{a}_1; \bar{\mathbf{z}}_2] \in \mathbb{Z}\}} (\mathbf{a}_2, \mathbf{z}_2)_{\mathbf{M}} - h([\mathbf{a}_1; \mathbf{z}_2]).$$

Lemma 1.14 (Generalization of Theorem 1.11)**.** *Assume that $h : \mathbb{Z} \mapsto \mathbb{R}$ in Definition 1.13 is such that $\phi : \mathbb{Z}_2 \to \mathbb{R}$, $\phi : \mathbf{z}_2 \mapsto h([\bar{\mathbf{z}}_1; \mathbf{z}_2])$ is strictly convex and differentiable for fixed $\bar{\mathbf{z}}_1$. Then*

$$\hat{\mathbf{a}} : \mathbb{Z} \to \mathbb{R}^{n_1+n_2}, \quad \hat{\mathbf{a}}(\mathbf{z}) = \begin{bmatrix} \mathbf{z}_1 \\ \nabla_{\mathbf{M},\mathbf{z}_2} h(\mathbf{z}) \end{bmatrix}, \qquad \textit{for } \mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}$$

is injective, and for $\mathbb{A} = \hat{\mathbf{a}}(\mathbb{Z})$ its inverse

$$\hat{\mathbf{z}} : \mathbb{A} \to \mathbb{Z}, \quad \hat{\mathbf{z}}(\mathbf{a}) = \begin{bmatrix} \mathbf{a}_1 \\ \hat{\mathbf{z}}_2(\mathbf{a}) \end{bmatrix}, \qquad \textit{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix},$$

exists. Moreover, the partial Legendre transform $g_{\mathbf{M}}(\cdot)$ can also be characterized as

$$g_{\mathbf{M}}(\mathbf{a}) = \hat{\mathbf{z}}_2(\mathbf{a}) \cdot \mathbf{a}_2 - h(\mathbf{a}_1, \hat{\mathbf{z}}_2(\mathbf{a})), \qquad \textit{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix} \in \mathbb{A},$$

and it holds

$$\nabla_{\mathbf{a}_1} g_{\mathbf{M}}(\mathbf{a}) = -\nabla_{\mathbf{z}_1} h(\mathbf{z})_{|\mathbf{z}=\hat{\mathbf{z}}(\mathbf{a})}, \qquad \nabla_{\mathbf{M},\mathbf{a}_2} g_{\mathbf{M}}(\mathbf{a}) = \hat{\mathbf{z}}_2(\mathbf{a}),$$

and $\mathbf{a}_2 \mapsto g_{\mathbf{M}}([\bar{\mathbf{a}}_1; \mathbf{a}_2])$ is strictly convex.

We also note that all parts of Theorem 1.7 generalize in the natural way to this case. This especially includes that we end up with the Hamiltonian itself, when the (partial) Legendre transform is applied twice w.r.t. the same inner product.

A comparison of Lemma 1.14 and Theorem 1.11 show that a scaling in the inner product leads to an inverse scaling of the gradient, as expressed by the following proposition.

Proposition 1.15. *Let the conditions of Lemma 1.14 hold. Let $g(\cdot)$ and $g_{\mathbf{M}}(\cdot)$ be the partial Legendre-transform w.r.t. the Euclidean scalar product and the scalar product $(\cdot,\cdot)_{\mathbf{M}}$, respectively. Then it holds*

$$\nabla_{\mathbf{M},\mathbf{a}_2} g_{\mathbf{M}}(\mathbf{a}) = \mathbf{M}^{-1} \nabla_{\mathbf{a}_2} g(\mathbf{a}), \qquad \textit{for } \mathbf{a} \in \mathbb{A}.$$

Remark 1.16. *Both, gradients and Legendre-transformations also take a prominent role in more general settings. While the generalization of a Riemannian gradient on finite dimensional nonlinear submanifolds can be calculated from the Euclidean gradient by suitable projection and rescaling, e.g. [AMS08, pp. 60], the situation is more delicate in the infinite dimensional case, where the existence of gradients is non-trivial and depends on the choice of inner product, see [Zei85], [Bre11].*

Chapter 2

Approximation procedure

This chapter presents the approximation procedure for our general nonlinear flow problem on networks. The problem is introduced in Section 2.1 and variable-transformed formulations as well as a variational principle are discussed. This lays the foundation for our Galerkin approximation and complexity reduction framework, which we present in Section 2.2 and Section 2.3, respectively. The structural properties of the approximations are further analyzed on the level of coordinate representations in Section 2.4. Afterwards, energy-stable time discretization schemes for our approximations are discussed in Section 2.5. Finally, Section 2.6 then addresses the adaptions to the case of networks with different edge-weights, which is needed for the the simulation of realistic network scenarios.

2.1 Network problem

Let a network be described by a directed graph $(\mathcal{N}, \mathcal{E}, l)$ and $\Omega = \{x : x \in \omega, \text{ for } \omega \in \mathcal{E}\}$ describe its spatial domain. We consider a class of prototypical partial differential equations on the network: The state $\underline{\mathbf{z}} = [z_1; z_2] : [0,T] \times \Omega \to \mathbb{R}^2$ is governed by

$$\partial_t \underline{\mathbf{z}}(t,x) = \begin{bmatrix} & -\partial_x \\ -\partial_x & -\tilde{r}(\underline{\mathbf{z}}(t,x)) \end{bmatrix} \nabla_{\mathbf{z}} h(\underline{\mathbf{z}}(t,x)), \qquad x \in \Omega, \quad t \in [0,T] \tag{2.1a}$$

with $\tilde{r} : \mathbb{R}^2 \to \mathbb{R}$ such that $\tilde{r}(\underline{\mathbf{z}}(t,x)) \geq 0$. The solution components are interconnected by the coupling conditions

$$\sum_{\omega \in \mathcal{E}(\nu)} n^{\omega}[\nu] \nabla_{z_2} h(\underline{\mathbf{z}}_{|\omega}(t,\nu)) = 0, \qquad \nabla_{z_1} h(\underline{\mathbf{z}}_{|\omega}(t,\nu)) = \nabla_{z_1} h(\underline{\mathbf{z}}_{|\tilde{\omega}}(t,\nu)) \quad \text{for } \omega, \tilde{\omega} \in \mathcal{E}(\nu). \tag{2.1b}$$

To close the system, data $\mathbf{z}_0 : \Omega \to \mathbb{R}^2$ is assumed to be given to prescribe initial conditions $\underline{\mathbf{z}}(0,x) = \mathbf{z}_0(x)$ for $x \in \Omega$, and one boundary data $u^{\nu} : [0,T] \to \mathbb{R}$ per boundary node $\nu \in \mathcal{N}_{\partial}$, each describing a boundary condition of one of the following types,

$$\textit{Type 1: } \nabla_{z_1} h(\mathbf{z}(t,\nu)) = u^{\nu}(t), \quad \textit{Type 2: } n^{\omega}[\nu] \nabla_{z_2} h(\underline{\mathbf{z}}(t,\nu)) = u^{\nu}(t) \quad \omega \in \mathcal{E}(\nu) \tag{2.1c}$$

for $t \in [0, T]$. We refer to $h(\cdot)$ as its Hamiltonian density and to

$$\mathcal{H}(\underline{\mathbf{z}}) = \langle h(\underline{\mathbf{z}}), 1\rangle = \sum_{\omega\in\mathcal{E}} \int_\omega h(\underline{\mathbf{z}})dx$$
$$\mathcal{M}(\underline{\mathbf{z}}) = \langle z_1, 1\rangle = \sum_{\omega\in\mathcal{E}} \int_\omega z_1 dx$$

as the Hamiltonian and the (total) mass, respectively. Given (2.1) has a strong solution, it can be shown that

$$\frac{d}{dt}\mathcal{M}(\underline{\mathbf{z}}) = \sum_{\nu\in\mathcal{N}_\partial,\, \omega\in\mathcal{E}(\nu)} n^\omega[\nu]\nabla_{z_2}h(\underline{\mathbf{z}}[\nu])$$
$$\frac{d}{dt}\mathcal{H}(\underline{\mathbf{z}}) \leq \sum_{\nu\in\mathcal{N}_\partial,\, \omega\in\mathcal{E}(\nu)} n^\omega[\nu]\nabla_{z_1}h(\underline{\mathbf{z}}[\nu])\nabla_{z_2}h(\underline{\mathbf{z}}[\nu]).$$

The first relation has in our application the interpretation of mass conservation, and the second one is referred to as energy dissipation from now on. Our approximation schemes is constructed such that they mimic these two relations on a discrete level, which greatly enhances their stability in comparison to standard methods. These properties are also used to derive, e.g., the existence of global solutions for our discretizations.

Assumption 2.1. *The domain $\mathbb{Z} \subset \mathbb{R}^2$ of the Hamiltonian density $h : \mathbb{Z} \to \mathbb{R}$ is an open convex set. Moreover, h is twice continuously differentiable with symmetric positive definite Hessian $\nabla_{\mathbf{zz}}h(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{Z}$.*

Lemma 2.2. *Under Assumption 2.1, it holds*

$$\nabla_{z_1z_1}h(\mathbf{z}) > 0, \qquad \nabla_{z_2z_2}h(\mathbf{z}) > 0,$$
$$\nabla_{z_1z_1}h(\mathbf{z})\,\nabla_{z_2z_2}h(\mathbf{z}) > (\nabla_{z_1z_2}h(\mathbf{z}))^2, \qquad \textit{for } \mathbf{z} \in \mathbb{Z}.$$

Proof. The eigenvalues of the Hessian are dependent on $\mathbf{z} \in \mathbb{Z}$ and read

$$\mu_{1/2}(\mathbf{z}) = \frac{1}{2}\left(\nabla_{z_1z_1}h(\mathbf{z}) + \nabla_{z_2z_2}h(\mathbf{z}) \pm \sqrt{(\nabla_{z_1z_1}h(\mathbf{z}) - \nabla_{z_2z_2}h(\mathbf{z}))^2 + 4\nabla_{z_1z_2}h(\mathbf{z})^2}\right).$$

By Assumption 2.1, $\mu_{1/2}(\mathbf{z}) > 0$ are positive. The assertions of the lemma can be derived by the latter by basic manipulations. □

Corollary 2.3. *Under Assumption 2.1, the map*

$$\hat{\mathbf{a}} : \mathbb{Z} \to \mathbb{A}, \quad \hat{\mathbf{a}}(\mathbf{z}) = \begin{bmatrix} z_1 \\ \nabla_{z_2}h(\mathbf{z}) \end{bmatrix}, \qquad \textit{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}$$

and its inverse $\hat{\mathbf{z}} : \mathbb{A} \to \mathbb{Z}$, defined in Theorem 1.10, are both continuously differentiable.

Proof. Theorem 1.10 is applicable, which shows that $\hat{\mathbf{a}}$ is injective, and its inverse $\hat{\mathbf{z}}$ is well-defined. The gradient $\nabla_{\mathbf{z}}\hat{\mathbf{a}}(\mathbf{z})$ of $\hat{\mathbf{a}}$ is continuous by Assumption 2.1, and a direct calculation shows that its eigenvalues read $\lambda_1 = 1$ and $\lambda_2 = \nabla_{z_2 z_2} h(\mathbf{z})$, which are strictly positive by Lemma 2.2. Therefore, the derivative of $\hat{\mathbf{z}} : \mathbb{A} \to \mathbb{Z}$ is a continuous function and is given as

$$\frac{d}{d\mathbf{a}}\hat{\mathbf{z}}(\mathbf{a})) = (\nabla_{\mathbf{a}}\hat{\mathbf{z}}(\mathbf{a}))^T = \left(\frac{d}{d\mathbf{z}}\hat{\mathbf{a}}(\mathbf{z})^{-1}\right)_{|\mathbf{z}=\hat{\mathbf{z}}(\mathbf{a})} = \begin{bmatrix} 1 & 0 \\ -\frac{\nabla_{z_1 z_2} h(\hat{\mathbf{z}}(\mathbf{a}))}{\nabla_{z_2 z_2} h(\hat{\mathbf{z}}(\mathbf{a}))} & \frac{1}{\nabla_{z_2 z_2} h(\hat{\mathbf{z}}(\mathbf{a}))} \end{bmatrix},$$

which can be deduced by applying the implicit function theorem locally for each $\mathbf{z} \in \mathbb{Z}$. □

Additionally, Assumption 2.1 determines our system (2.1) to be of strict hyperbolic type according to the following definition, see e.g, [LeV02].

Definition 2.4. *Let $\mathbb{Z} \subset \mathbb{R}^n$ and $\mathbf{F} : \mathbb{Z} \to \mathbb{R}^{n,n}$, $\mathbf{g} : \mathbb{Z} \to \mathbb{R}^n$ define the partial differential equation*

$$\partial_t \underline{\mathbf{z}}(t,x) = \mathbf{F}(\underline{\mathbf{z}}(t,x))\partial_x \underline{\mathbf{z}}(t,x) + \mathbf{g}(\underline{\mathbf{z}}(t,x)), \qquad x \in \Omega, \quad t \in [0,T],$$

for $\underline{\mathbf{z}} : [0,T] \times \Omega \to \mathbb{Z}$ on $\Omega \subset \mathbb{R}^d$ and $T > 0$. Assume further that $\mathbf{F}(\mathbf{z})$ has n real distinct eigenvalues $\mu_1, \dots, \mu_n$ for all $\mathbf{z} \in \mathbb{Z}$. Then we call this partial differential equation, and every partial differential equation formally transferable into the above form, strictly hyperbolic. Moreover, $\mu_1, \dots, \mu_n$ are called its characteristic speeds.

Proposition 2.5. *Under Assumption 2.1, the partial differential equation* (2.1a) *is strictly hyperbolic. The characteristic speeds read*

$$\lambda_{1/2}(\mathbf{z}) = \nabla_{z_1 z_2} h(\mathbf{z}) \pm \sqrt{\nabla_{z_1 z_1} h(\mathbf{z}) \nabla_{z_2 z_2} h(\mathbf{z})},$$

where $\lambda_1(\mathbf{z})$ is positive and $\lambda_2(\mathbf{z})$ is negative for $\mathbf{z} \in \mathbb{Z}$.

Proof. Equation (2.1a) can be formally rewritten as in Definition 2.4 with

$$\mathbf{F}(\mathbf{z}) = \begin{bmatrix} & -1 \\ -1 & \end{bmatrix} \nabla_{\mathbf{zz}} h(\mathbf{z}) \in \mathbb{R}^2, \qquad \text{for } \mathbf{z} \in \mathbb{Z}.$$

The eigenvalues of $\mathbf{F}(\mathbf{z})$ are given by $\lambda_{1/2}(\mathbf{z})$ as in the proposition. Using Lemma 2.2, it is readily seen that λ_1 is positive, and λ_2 is negative, which completes the proof. □

Change of variables have played a crucial role in the theoretical and numerical analysis of hyperbolic systems, e.g., [Moc80], [Har83], [CHS90], [Jam08]. Next, we discuss the state-transformed versions of the strong form (2.1a), our approximation schemes are based on. The partial Legendre transform plays a prominent role in the upcoming.

Corollary 2.6 (Partial Legendre-transformed strong form)**.** *Let $\hat{\mathbf{z}} : \mathbb{A} \to \mathbb{Z}$ and $g : \mathbb{A} \to \mathbb{R}$ be defined as in Theorem 1.10. Let further $\underline{\mathbf{a}} \in \mathcal{C}^1([0,T]; \mathcal{C}^1_{pw}(\mathcal{E}) \times \mathcal{C}^1_{pw}(\mathcal{E}))$ fulfill*

$$\partial_t \begin{bmatrix} a_1(t) \\ \nabla_{a_2} g(\underline{\mathbf{a}}(t)) \end{bmatrix} = \begin{bmatrix} & -\partial_x \\ -\partial_x & -r(\underline{\mathbf{a}}(t)) \end{bmatrix} \begin{bmatrix} -\nabla_{a_1} g(\underline{\mathbf{a}}(t)) \\ a_2(t) \end{bmatrix}, \qquad x \in \Omega, \quad t \in [0,T]$$

with $r(\underline{\mathbf{a}}(t)) := \tilde{r}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t)))$. Then $\hat{\mathbf{z}}(\underline{\mathbf{a}})$ fulfills equation (2.1a)*, and for the Hamiltonian $\mathcal{H}$ it holds*

$$\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) = \langle h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), 1\rangle = \langle \hat{z}_2(\underline{\mathbf{a}}(t)) a_2(t), 1\rangle - g(\underline{\mathbf{a}}(t)).$$

Remark 2.7. *Likewise as for the Hamiltonian density function h, also for $\mathcal{H}$, a partial Legendre transform $\mathcal{G}$ can be introduced. For sufficiently smooth $\underline{\mathbf{a}}(t)$, it reads*

$$\mathcal{G}(\underline{\mathbf{a}}(t)) = \langle g(\underline{\mathbf{a}}(t)), 1\rangle = \langle \hat{z}_2(\underline{\mathbf{a}}(t))a_2(t), 1\rangle - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))),$$

and it holds

$$\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) = \langle \hat{z}_2(\underline{\mathbf{a}}(t))a_2(t), 1\rangle - \mathcal{G}(\underline{\mathbf{a}}(t)),$$

which is in analogy to the last part of Theorem 1.7, stating that the Legendre transform of the Legendre transform again yields the Hamiltonian. For a related discussion of the dual theory in a function space setting, we, e.g., refer to [Zei85, Theorem 51.A] and [Bre11].

Both, the underlying partial Legendre transform, Corollary 2.6, and the interpretation of the system in terms of the energy variable $\mathbf{z}$ play a crucial role in our considerations. Referring to the latter interpretation, we construct a related state-transformed form of (2.1a) with the help of Theorem 1.11. It reads

$$\partial_t \begin{bmatrix} a_1(t) \\ \hat{z}_2(\underline{\mathbf{a}}(t)) \end{bmatrix} = \begin{bmatrix} & -\partial_x \\ -\partial_x & -r(\underline{\mathbf{a}}(t)) \end{bmatrix} \begin{bmatrix} \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) \\ a_2(t) \end{bmatrix} \qquad x \in \Omega, \quad t \in [0,T]. \tag{2.2}$$

Note also that by construction $a_2(t) = \nabla_{z_2} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t)))$. By testing the upper strong form and using partial integration once, the following variational principle can be deduced.

Theorem 2.8. *A strong solution $\underline{\mathbf{a}} \in \mathcal{C}^1([0,T]; \mathcal{C}^1_{pw}(\mathcal{E}) \times \mathcal{C}^1_{pw}(\mathcal{E}))$ of (2.2) fulfills the variational principle*

$$\begin{aligned} \langle \partial_t a_1(t), b_1\rangle &= -\langle \partial_x a_2(t), b_1\rangle \\ \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), b_2\rangle &= \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x b_2\rangle + \mathbf{e}_B(t) \cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}(t)) a_2(t), b_2\rangle \\ \mathbf{f}_B(t) &= \mathcal{T} a_2(t) \end{aligned}$$

for all $b_1 \in \mathcal{L}^2(\mathcal{E})$, $b_2 \in \mathcal{H}^1_{div}(\mathcal{E})$. Given $\mathcal{N}_\partial = \{\nu_1, \dots, \nu_p\}$, it holds for
$\mathbf{e}_B = [e_B[\nu_1]; \dots; e_B[\nu_p]]$, $\mathbf{f}_B = [f_B[\nu_1]; \dots; f_B[\nu_p]] \in \mathbb{R}^p$, and

$$\mathbf{e}_B[\nu_i] = \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}[\nu_i])), \qquad \mathbf{f}_B[\nu_i] = n^\omega[\nu_i] \nabla_{z_2} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}[\nu_i])), \quad \textit{for } \omega \in \mathcal{E}(\nu_i),$$

for $i = 1, \dots, p$.

Proof. The second equation in (2.2) reads

$$\partial_t \hat{z}_2(\underline{\mathbf{a}}(t)) = -\partial_x \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) - r(\underline{\mathbf{a}}(t)) a_2(t).$$

Testing this equation with $b_2 \in \mathcal{H}^1_{div}(\mathcal{E})$, integrating it over one edge $\omega = (\nu, \tilde{\nu}) \in \mathcal{E}$ and using integration by parts once, we obtain

$$\begin{aligned} \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), b_2\rangle_\omega = \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x b_2\rangle_\omega - \langle r(\underline{\mathbf{a}}(t)) a_2(t), b_2\rangle\omega \\ + \left[\nabla_{z_1} h(\hat{\mathbf{z}}(t,x))(-b_2[x])\right]_{x=\nu}^{\tilde{\nu}}. \end{aligned}$$

Repeating this for all edges and then summing over all equations, the interface terms at the inner nodes drop out, as $b_2 \in \mathcal{H}^1_{div}(\mathcal{E})$. All in all, this gives the second equation of the variational principle. The other equations follow similarly. □

Remark 2.9. *The variational principle may also hold for generalized solutions not being classical solutions. The needed regularities for all terms in Theorem 2.8 to be well-defined for every $t \in (0,T]$ depend, however, on the choice of the Hamiltonian density $h(\cdot)$ and the nonlinearity $r(\cdot)$. To the best of the author's knowledge, there only exist partial results on the well-posedness of weak solutions for this type of equations, e.g., [Che05], [LeV02].*

The variational principle can be used to show (local) mass-conservation and the energy-dissipation equality. The proof of this upcoming result is done in a way suitable for adaption to our Galerkin approximations later on.

Theorem 2.10. *Let $\underline{\mathbf{a}} \in \mathcal{C}^1([0,T];\mathcal{C}^1_{pw}(\mathcal{E}) \times \mathcal{C}^1_{pw}(\mathcal{E}))$ fulfill the variational principle of Theorem 2.8 for some $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{e}_B \in \mathcal{C}([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$. Then it holds for $[w_1,w_2] \subset \omega$, $\omega \in \mathcal{E}$*

$$\frac{d}{dt}\int_{[w_1,w_2]} a_1(t)dx = a_2(t)[w_1] - a_2(t)[w_2], \qquad \frac{d}{dt}\mathcal{M}(\underline{\mathbf{a}}(t)) = \sum_{\nu \in \mathcal{N}_\partial} f_B(t)[\nu]$$

$$\frac{d}{dt}\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}})(t)) = \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t))a_2(t), a_2(t)\rangle \leq \mathbf{e}_B(t) \cdot \mathbf{f}_B(t).$$

Proof. The first relation, the local mass conservation, follows by testing the first equation of the variational principle in Theorem 2.8 with $b_1 = \chi_{[w_1,w_2]}$, the indicator function of the domain $[w_1,w_2]$. By that we get

$$\frac{d}{dt}\int_{[w_1,w_2]} a_1 dx = \langle \partial_t a_1, \chi_{[w_1,w_2]}\rangle = -\langle \partial_x a_2, \chi_{[w_1,w_2]}\rangle = a_2[w_1] - a_2[w_2].$$

The second equation, the global mass conservation, follows similarly by summing up over the individual masses over the edges.

To derive the energy dissipation, we prove, as a preliminary step, the existence of $\xi(\underline{\mathbf{a}}(t)) \in \mathcal{L}^2(\mathcal{E})$ such that

$$\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1\rangle = \langle \xi(\underline{\mathbf{a}}), b_1\rangle \qquad \text{for all } b_1 \in \mathcal{L}^2(\mathcal{E}), \tag{2.3}$$

for solutions $\underline{\mathbf{a}}$ of the variational principle. By assuming $\underline{\mathbf{a}}$ fulfills the variational principle, the term $\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), \partial_x b_2\rangle$ has to be well-defined for all $b_2 \in \mathcal{H}^1_{div}(\mathcal{E})$ and $t \in [0,T]$, as the second equation of Theorem 2.8 implies. The compatibility

$$\mathcal{L}^2(\mathcal{E}) \subset \left\{\xi : \text{It exists } \zeta \in \mathcal{H}^1_{div}(\mathcal{E}) \text{ with } \partial_x \zeta = \xi\right\}$$

then implies that $b_1 \mapsto \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1\rangle$ is well-defined as an element of the dual space of $\mathcal{L}^2(\mathcal{E})$. The existence of $\xi(\underline{\mathbf{a}})$ fulfilling (2.3) can now be deduced from the Riesz representation theorem, as $\mathcal{L}^2(\mathcal{E})$ with $\langle \cdot,\cdot\rangle$ is a Hilbert space. Moreover, a formal application of the chain rule gives

$$\frac{d}{dt}\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}})) = \langle \partial_t a_1, \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}))\rangle + \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}), a_2\rangle.$$

As shown in the preliminary step, $\nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}))$ can be replaced by $\xi(\underline{\mathbf{a}})$ here. Applying the variational principle of Theorem 2.8 with $\underline{\mathbf{b}} = [\xi(\underline{\mathbf{a}}); a_2]$ then gives

$$\begin{aligned}\langle \partial_t a_1, \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}))\rangle + \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}), a_2\rangle &= -\langle \partial_x a_2, \xi(\underline{\mathbf{a}})\rangle + \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), \partial_x a_2\rangle \\ &\quad + \mathbf{e}_B \cdot \mathcal{T} a_2 - \langle r(\underline{\mathbf{a}}), a_2^2\rangle = \mathbf{e}_B \cdot \mathbf{f}_B - \langle r(\underline{\mathbf{a}}), a_2^2\rangle \leq \mathbf{e}_B \cdot \mathbf{f}_B,\end{aligned}$$

which finishes the proof. □

Remark 2.11. *Let $\{b_1^k\}_{k\in\mathbb{N}}$ be an orthogonal basis of $\mathcal{L}^2(\mathcal{E})$. The function $\xi : \mathcal{L}^2(\mathcal{E}) \times \mathcal{H}^1_{div}(\mathcal{E}) \to \mathcal{L}^2(\mathcal{E})$ fulfilling*

$$\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1 \rangle = \langle \xi(\underline{\mathbf{a}}), b_1 \rangle \qquad \textit{for all } b_1 \in \mathcal{L}^2(\mathcal{E}),$$

which we introduced in (2.3)*, then has the representation*

$$\xi(\underline{\mathbf{a}}) = \sum_{k\in\mathbb{N}} \frac{1}{||b_1^k||^2} \left\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1^k \right\rangle b_1^k.$$

It can be interpreted as the gradient w.r.t. the $\mathcal{L}^2$-scalar product of $a_1 \mapsto -\langle g([a_1; a_2]), 1\rangle$ for fixed a_2, see Remark 1.16.

The existence of gradient representations in the underlying variational principle is strongly related to the structure-preservation of our upcoming approximation schemes. They are a key tool for most analytical results on the Galerkin approximations.

2.2 Galerkin approximation

To transfer the structural properties on Galerkin approximations, we need the following structural assumption on the ansatz spaces.

Assumption 2.12 (Compatibility of spaces)**.** *Let $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, $\mathcal{V}_1 \subset \mathcal{L}^2(\mathcal{E})$ and $\mathcal{V}_2 \subset \mathcal{H}^1_{div}(\mathcal{E})$ be finite dimensional subspaces fulfilling the compatibility conditions*

A1) $\mathcal{V}_1 = \partial_x \mathcal{V}_2, \qquad$ *with* $\partial_x \mathcal{V}_2 = \{\xi : \textit{It exists } \zeta \in \mathcal{V}_2 \textit{ with } \partial_x \zeta = \xi\}$.

A2) $\{b_2 \in \mathcal{H}^1_{div}(\mathcal{E}) : \partial_x b_2 = 0\} \subset \mathcal{V}_2$.

For our analysis, we assume for each boundary node $\nu \in \mathcal{N}_\partial$ that one of the following boundary conditions is prescribed, as well as that initial conditions are given as follows:

$$\begin{aligned} &\textit{Type 1: } e_B[\nu] = u^\nu \in \mathcal{C}([0,\infty),\mathbb{R}), \qquad \textit{Type 2: } f_B[\nu] = u^\nu \in \mathcal{C}^1([0,\infty),\mathbb{R}) \\ &\underline{\mathbf{a}}(0) = \underline{\mathbf{a}}_0, \quad \text{for } \underline{\mathbf{a}}_0 \in \mathcal{V}, u^\nu : [0,T] \to \mathbb{R} \quad \text{given such that } \mathcal{T} a_2(0) = \mathbf{f}_B(0). \end{aligned} \tag{2.4}$$

Other types of boundary conditions could be handled similarly, given the boundary data is regular enough.

System 2.13. *Find $\underline{\mathbf{a}} \in \mathcal{C}^1([0,T]; \mathcal{V}_1 \times \mathcal{V}_2)$, $\mathbf{f}_B \in \mathcal{C}^1([0,T]; \mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{e}_B \in \mathcal{C}([0,T]; \mathbb{R}^{|\mathcal{N}_\partial|})$, solving*

$$\begin{aligned} \langle \partial_t a_1(t), b_1 \rangle &= -\langle \partial_x a_2(t), b_1 \rangle \\ \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), b_2 \rangle &= \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x b_2 \rangle + \mathbf{e}_B(t) \cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}(t)) a_2(t), b_2 \rangle \\ \mathbf{f}_B(t) &= \mathcal{T} a_2(t) \end{aligned}$$

for all $b_1 \in \mathcal{V}_1$, $b_2 \in \mathcal{V}_2$ and closing conditions (2.4)*. Assumption 2.12 is supposed to hold.*

The compatibility conditions yet imply two important properties on every solution of the Galerkin approximation, the local mass conservation and the energy dissipation equality, as we will derive next.

Lemma 2.14 (Local mass conservation)**.** *Let Assumption 2.12 hold, and let for $a_1 \in C^1([0,T];\mathcal{V}_1)$ and $a_2 \in C^1([0,T];\mathcal{V}_2)$ the variational principle*

$$\langle \partial_t a_1(t), b_1\rangle = -\langle \partial_x a_2(t), b_1\rangle, \qquad t \in [0,T]$$

hold for all $b_1 \in \mathcal{V}_1$. Then it follows $\partial_t a_1 \equiv -\partial_x a_2$.

Proof. By construction, $\partial_t a_1(t), \partial_x a_2(t) \in \mathcal{V}_1$ for $t \in [0,T]$, i.e., they lie in the test space. Therefore, the variational principle with $b_1 = \partial_t a_1 + \partial_x a_2$ gives

$$\langle \partial_t a_1(t), \partial_t a_1(t) + \partial_x a_2(t)\rangle = \langle -\partial_x a_2(t), \partial_t a_1(t) + \partial_x a_2(t)\rangle.$$

By subtracting the right-hand side from both sides of the latter equation, one can see that for all $t \in [0,T]$ the $\mathcal{L}^2$-norm of $\partial_t a_1(t) + \partial_x a_2(t)$ is zero, i.e., $\partial_t a_1 \equiv -\partial_x a_2$. □

Remark 2.15. *Lemma 2.14 shows that for any solution of System 2.13 mass conservation readily holds in a pointwise sense, i.e., local mass conservation. This is in contrast to the result for the Galerkin approximation of Part I.A, where we could only guarantee global mass conservation and needed the additional assumption that the global constant function was included in $\mathcal{V}_1$, cf. Assumption I.A-2.7.*

In the upcoming result, we state that all expressions involving the Hamiltonian density h can be re-expressed by representatives of the function spaces $\mathcal{V}_1, \mathcal{V}_2$. This yet reveals that the Galerkin approximations inherit the gradient structure of the underlying Hamiltonian density h and its Legendre transformed g.

Theorem 2.16. *Let $\{b_i^k\}_{k=1,\dots,N_i}$ be orthogonal bases of $\mathcal{V}_i$ for $i = 1,2$. Then the second equation of System 2.13 can be recast as*

$$\langle \tilde{\xi}(\underline{\mathbf{a}}(t)), b_2\rangle = \langle \xi(\underline{\mathbf{a}}(t)), \partial_x b_2\rangle + \mathbf{e}_B(t)\cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}(t)) a_2(t), b_2\rangle$$

with $\xi : \mathcal{V} \to \mathcal{V}_1$ and $\tilde{\xi} : \mathcal{V} \to \mathcal{V}_2$, defined by

$$\xi(\underline{\mathbf{a}}) = \sum_{k=1}^{N_1} \frac{1}{||b_1^k||^2} \left\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1^k \right\rangle b_1^k = -\sum_{k=1}^{N_1} \frac{1}{||b_1^k||^2} \left\langle \nabla_{a_1} g(\underline{\mathbf{a}}), b_1^k \right\rangle b_1^k$$

$$\tilde{\xi}(\underline{\mathbf{a}}) = \sum_{k=1}^{N_2} \frac{1}{||b_2^k||} \left\langle \partial_t \nabla_{a_2} g(\underline{\mathbf{a}}), b_2^k \right\rangle b_2^k.$$

The proof relies on the Riesz representation theorem, together with the compatibility condition. We refer to the more general result of Theorem 2.23.

Theorem 2.17 (Energy-dissipation equality)**.** *For any solution of System 2.13 it holds*

$$\frac{d}{dt}\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) = \mathbf{e}_B(t)\cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t)) a_2(t), a_2(t)\rangle \le \mathbf{e}_B(t)\cdot \mathbf{f}_B(t),$$

where the Hamiltonian for $\underline{\mathbf{a}} \in \mathcal{V}_1 \times \mathcal{V}_2$ is defined as $\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}})) = \langle h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), 1\rangle$.

As the proof is almost verbatim to the one of Theorem 2.10, up to replacing the spaces $\mathcal{L}^2(\mathcal{E})$ and $\mathcal{H}^1_{div}(\mathcal{E})$ by $\mathcal{V}_1$ and $\mathcal{V}_2$, it is omitted.

Remark 2.18. *Note that $\mathcal{H}^1_{div}(\mathcal{E})$, equipped with the $\mathcal{L}^2$-scalar product, is not a Hilbert space. Thus, the identification of $\partial_t \hat{z}_2(\underline{\mathbf{a}})$ with $\tilde{\xi}(\underline{\mathbf{a}}) \in \mathcal{H}^1_{div}(\mathcal{E})$, as we did it in Theorem 2.16, explicitly relies on the finite dimensional setting, whereas the proof of the energy dissipation does not, i.e., the existence of the representation $\tilde{\xi}(\underline{\mathbf{a}})$ from Theorem 2.16 is never explicitly used in our proofs.*

2.3 Complexity reduction

In order to formulate the complexity reduction, we have to specify the realization of the Galerkin approximation steps. In our approach, they consist of a space discretization by the finite element method and a subsequent model order reduction step, cf. Fig. 1 in the introduction of Part I. This yields the *full order model* and the *reduced order model*, respectively. Both models are of the form of System 2.13, but the ansatz space for the full order model is a typically large superset of the ansatz space for the reduced order model. In what follows, $\mathcal{V}$ refers to the ansatz space of the reduced order model. Moreover, the full order model is assumed to be associated to a partitioning of the spatial domain Ω described by finite elements $K_j \subset \Omega$, $j = 1, \ldots, J$. The complexity reduction aims for efficient approximation of nonlinear integral expressions in the reduced order model without evaluating them for each finite element. For that, we apply the following quadrature-type approximation.

Definition 2.19. *Let a partitioning of Ω be given by the full order model by finite elements $K_j \subset \Omega$, $j = 1, \ldots, J$, fulfilling*

$$\bigcup_{j \in J} K_j = \Omega, \qquad \int_{K_j \cap K_l} 1 dx = 0, \qquad \textit{for } j \neq l.$$

Let further an index-set $I \subset \{1, \ldots, J\}$ and weights $w_i \in \mathbb{R}$ for $i \in I$ be given. Then we define the complexity-reduced bilinear form $\langle \cdot, \cdot \rangle_c : \mathcal{L}_2(\Omega) \times \mathcal{L}_2(\Omega) \to \mathbb{R}$ and $|| \cdot ||_c$ by

$$\langle b, \bar{b} \rangle_c = \sum_{i \in I} w_i \int_{K_i} b(x) \bar{b}(x) dx, \qquad ||b||_c = \sqrt{\langle b, b \rangle_c}.$$

Note that $\langle \cdot, \cdot \rangle_c$ can be seen as an approximation of the $\mathcal{L}_2$-scalar product $\langle \cdot, \cdot \rangle$. For our analysis, we rely on the following assumptions.

Assumption 2.20. *The bilinear form $\langle \cdot, \cdot \rangle_c$ is given as in Definition 2.19 with $w_i > 0$ for $i \in I$. Moreover, there exists a constant $\tilde{C}$ such that it holds*

$$\frac{1}{\tilde{C}} ||b||_c \leq ||b|| \leq \tilde{C} ||b||_c, \qquad \textit{for } b \in \mathcal{V}_1 \cup \mathcal{V}_2.$$

Remark 2.21. *The last part of Assumption 2.20 is the same as requiring the $\mathcal{L}_2$-norm $|| \cdot ||$ and $|| \cdot ||_c$ to be equivalent norms on $\mathcal{V}_1 \cup \mathcal{V}_2$.*

The complexity-reduced approximation of System 2.13 we propose then reads as follows.

System 2.22. *Find $\underline{\mathbf{a}} \in \mathcal{C}^1([0, T]; \mathcal{V}_1 \times \mathcal{V}_2)$, $\mathbf{f}_B \in \mathcal{C}^1([0, T]; \mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{e}_B \in \mathcal{C}([0, T]; \mathbb{R}^{|\mathcal{N}_\partial|})$, solving*

$$\begin{aligned}
\langle \partial_t a_1(t), b_1 \rangle &= -\langle \partial_x a_2(t), b_1 \rangle \\
\langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), b_2 \rangle_c &= \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x b_2 \rangle_c + \mathbf{e}_B(t) \cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}(t)) a_2(t), b_2 \rangle_c \\
\mathbf{f}_B(t) &= \mathcal{T} a_2(t)
\end{aligned}$$

for all $b_1 \in \mathcal{V}_1$, $b_2 \in \mathcal{V}_2$. Closing conditions (2.4)*, Assumption 2.12 and Assumption 2.20 are presumed to hold.*

A crucial point is that we can recast the complexity reduction of the nonlinear terms as complexity reduction of the Hamiltonian, i.e., System 2.22 can be re-interpreted as a Galerkin-approximation with a modified Hamiltonian and friction term. The structural properties expressed in Theorem 2.16 carry over to the complexity-reduced system very naturally.

Theorem 2.23 (Counterpart of Theorem 2.16)**.** *Let $\{b_i^k\}_{k=1,\dots,N_i}$ be orthogonal bases of $\mathcal{V}_i$ for $i = 1, 2$. Then the second equation of System 2.22 can be recast as*

$$\langle \tilde{\xi}(\underline{\mathbf{a}}(t)), b_2 \rangle = \langle \xi(\underline{\mathbf{a}}(t)), \partial_x b_2 \rangle + \mathbf{e}_B(t) \cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}(t)) a_2(t), b_2 \rangle_c$$

with $\xi : \mathcal{V} \to \mathcal{V}_1$ and $\tilde{\xi} : \mathcal{V} \to \mathcal{V}_2$, defined by

$$\xi(\underline{\mathbf{a}}) = \sum_{k=1}^{N_1} \frac{1}{||b_1^k||^2} \left\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1^k \right\rangle_c b_1^k = - \sum_{k=1}^{N_1} \frac{1}{||b_1^k||^2} \left\langle \nabla_{a_1} g(\underline{\mathbf{a}}), b_1^k \right\rangle_c b_1^k$$

$$\tilde{\xi}(\underline{\mathbf{a}}) = \sum_{k=1}^{N_2} \frac{1}{||b_2^k||^2} \left\langle \partial_t \nabla_{a_2} g(\underline{\mathbf{a}}), b_2^k \right\rangle_c b_2^k.$$

Proof. When all expressions in System 2.22 are well-posed, this implies particularly

$$\langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x b_2 \rangle_c < \infty \qquad \text{and} \qquad \langle \partial_t \nabla_{a_2} g(\underline{\mathbf{a}}), b_2 \rangle_c = \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), b_2 \rangle_c < \infty.$$

By that and the compatibility condition $\mathcal{V}_1 \subset \partial_x \mathcal{V}_2$, the linear functionals

$$\Phi_1 : \mathcal{V}_1 \to \mathbb{R}, \quad b_1 \mapsto \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), b_1 \rangle_c, \qquad \Phi_2 : \mathcal{V}_2 \to \mathbb{R}, \quad b_2 \mapsto \langle \partial_t \nabla_{a_2} g(\underline{\mathbf{a}}), b_2 \rangle_c$$

are seen to be bounded. This means that Φ_i, for $i = 1, 2$, are elements of the dual spaces of $\mathcal{V}_i$. We consider the $\mathcal{L}^2$-scalar product here. As the latter spaces are Hilbert spaces, the Riesz representation theorem then shows the existence of $\xi(\underline{\mathbf{a}}(t)) \in \mathcal{V}_1$ and $\tilde{\xi}(\underline{\mathbf{a}}(t)) \in \mathcal{V}_2$ as claimed in the theorem. □

As a direct consequence, the energy-dissipation equality carries over from Theorem 2.17 with an almost verbatim proof.

Theorem 2.24 (Energy-dissipation equality)**.** *For any solution $\underline{\mathbf{a}}(t)$ of System 2.22 it holds*

$$\frac{d}{dt} \mathcal{H}_c(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) = \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t)) a_2(t), a_2(t) \rangle_c \leq \mathbf{e}_B(t) \cdot \mathbf{f}_B(t),$$

where the complexity-reduced Hamiltonian is given as

$$\mathcal{H}_c(\hat{\mathbf{z}}(\underline{\mathbf{a}})) = \langle h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), 1 \rangle_c \quad \textit{for } \underline{\mathbf{a}} \in \mathcal{V}_1 \times \mathcal{V}_2.$$

2.4 Coordinate representations

The identification of gradient structures in the function-space setting, which we did in Remark 2.11 and Theorem 2.16, strongly suggests that we should be able to find coordinate representations with gradient expressions of an underlying energy functional. We construct these representations in the upcoming. The notation follows the one from Part I.A-Section 2.5, where possible.

Definition 2.25. *Let $\mathcal{V} = \mathcal{V}_1 \times \mathcal{V}_2$, and let bases $b_i^1, \dots, b_i^{N_i}$ of $\mathcal{V}_i$ be fixed for $i = 1, 2$. For $\underline{a} = [a_1; a_2] \in \mathcal{V}$, we define the coordinate representation $\mathbf{a} = [\mathbf{a}_1; \mathbf{a}_2] \in \mathbb{R}^N$, with $N = N_1 + N_2$, by*

$$a_i(t) = \sum_{l=1}^{N_i} b_i^l \mathfrak{a}_i^l(t), \qquad \mathbf{a}_i = [\mathfrak{a}_i^1; \dots ; \mathfrak{a}_i^{N_i}] \in \mathbb{R}^{N_i}.$$

The transformation from the coordinate representation $\mathbf{a} \in \mathbb{R}^N$ to the function $\underline{a} = [a_1; a_2] \in \mathcal{V}$ is defined as

$$\Psi : \mathbb{R}^N \to \mathcal{V}, \qquad \Psi(\mathbf{a}) = \begin{bmatrix} \sum_{l=1}^{N_1} b_1^l \mathfrak{a}_1^l(t) \\ \sum_{l=1}^{N_2} b_2^l \mathfrak{a}_2^l(t) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}.$$

Galerkin approximation

Let us first consider the coordinate representation of System 2.13. Recall that in the function space setting, the partial Legendre transform $\mathcal{G} : \mathcal{V} \to \mathbb{R}$ and its density $g : \mathbb{R}^2 \to \mathbb{R}$ are given as

$$\mathcal{G}(\underline{a}) = \langle g(\underline{a}), 1 \rangle, \qquad g(\underline{a}) = \hat{z}_2(\underline{a}) a_2 - h(\hat{\mathbf{z}}(\underline{a})),$$

where h is the Hamiltonian density. The respective coordinate representation of $\mathcal{G}$ is then defined as

$$G : \mathbb{R}^N \to \mathbb{R}, \qquad G(\mathbf{a}) = \mathcal{G}\left(\Psi(\mathbf{a})\right).$$

By the chain rule, it follows for $l \in N_i$ that

$$\frac{\partial}{\partial_{\mathfrak{a}_i^l}} G(\mathbf{a}) = \left\langle \nabla_{a_i} g(\underline{a})_{|\underline{a} = \Psi(\mathbf{a})}, b_i^l \right\rangle.$$

Let for $i = 1, 2$, $\mathbf{M}_i \in \mathbb{R}^{N_i}$ be symmetric positive definite matrices. The partial gradients of G in the scalar product $(\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{M}_i \mathbf{x}) \cdot \mathbf{y}$ according to Definition 1.12 then read

$$\nabla_{\mathbf{M}_i, \mathbf{a}_i} G(\mathbf{a}) = \mathbf{M}_i^{-1} \nabla_{\mathbf{a}_i} G(\mathbf{a}) = \mathbf{M}_i^{-1} \left[\frac{\partial}{\partial_{\mathfrak{a}_i^1}} G(\mathbf{a}); \dots ; \frac{\partial}{\partial_{\mathfrak{a}_i^{N_i}}} G(\mathbf{a}) \right].$$

It readily can be seen, cf. Corollary 2.6, that System 2.13 naturally translates into a coordinate representation as follows.

Corollary 2.26. *Define the system matrices*

$$\mathbf{M}_1 = [\langle b_1^n, b_1^m \rangle]_{m,n=1,\dots,N_1}, \qquad \mathbf{M}_2 = [\langle b_2^n, b_2^m \rangle]_{m,n=1,\dots,N_2}$$

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \\ & \mathbf{M}_2 \end{bmatrix}, \qquad \mathbf{J} = [\langle -\partial_x b_2^n, b_1^m \rangle]_{m=1,\dots,N_1, n=1,\dots,N_2}$$

$$\mathbf{R}(\mathbf{a}) = [\langle r(\Psi(\mathbf{a})) b_2^n, b_2^m \rangle]_{m,n=1,\dots,N_2}$$

$$\mathbf{K}_2 = \left[\mathcal{T} b_2^1, \dots, \mathcal{T} b_2^{N_2}\right]^T, \qquad \mathbf{K} = \begin{bmatrix} \mathbf{0}_{N_1, |\mathcal{N}_\partial|} \\ \mathbf{K}_2 \end{bmatrix}.$$

Then System 2.13 can be equivalently described as:

Find $\mathbf{a} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$*,* $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ *and* $\mathbf{e}_B \in \mathcal{C}([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$*, solving*

$$\mathbf{M}\frac{d}{dt}\begin{bmatrix} \mathbf{a}_1(t) \\ \nabla_{\mathbf{M}_2,\mathbf{a}_2} G(\mathbf{a}(t)) \end{bmatrix} = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R}(\mathbf{a}(t)) \end{bmatrix} \begin{bmatrix} -\nabla_{\mathbf{M}_1,\mathbf{a}_1} G(\mathbf{a}(t)) \\ \mathbf{a}_2(t) \end{bmatrix} + \mathbf{K}\mathbf{e}_B(t)$$
$$\mathbf{f}_B(t) = \mathbf{K}^T\mathbf{a}(t).$$

Remark 2.27. *Note that only because of the compatibility of Assumption 2.12 on* $\mathcal{V}_1$ *and* $\mathcal{V}_2$*, we are able to recover the gradient structure fully in the coordinate representation, cf. Corollary 2.26. In particular, we utilize that for* $b_2 \in \mathcal{V}_2$*, it holds for an orthogonal basis* $\{b_1^l\}_{l=1,\dots,N_1}$*,*

$$\langle \nabla_{a_1} g(\underline{\mathbf{a}}), \partial_x b_2\rangle = \sum_{l=1}^{N_1} \langle b_1^l, \partial_x b_2\rangle \frac{1}{||b_1^l||^2}\langle \nabla_{a_1} g(\underline{\mathbf{a}}), b_1^l\rangle,$$

which relates to the term $\mathbf{J}^T\nabla_{\mathbf{M}_1,\mathbf{a}_1} G(\mathbf{a}(t))$ *in the coordinate representation, cf. Remark 2.11.*

In light of our discussion, the question arises, if we can recover a Hamiltonian and rewrite the system in Corollary 2.26 in terms of the energy variables in standard port-Hamiltonian form. This turns out to be the case, as the following result shows.

Theorem 2.28. *For the system in Corollary 2.26, the coordinate transform*

$$\hat{\mathfrak{z}} : \mathbb{A}_N \to \mathbb{R}^N, \qquad \hat{\mathfrak{z}}(\mathbf{a}) = \begin{bmatrix} \mathbf{a}_1 \\ \hat{\mathfrak{z}}_2(\mathbf{a}) \end{bmatrix}$$

is well-defined and injective on every open convex subset $\mathbb{A}_N \subset \mathbb{R}^N$ *of the domain of* G*. Moreover, for* $H : \hat{\mathfrak{z}}(\mathbb{A}_N) \to \mathbb{R}$ *defined as the partial Legendre transform of* G *by*

$$H(\mathbf{z}) = \sup_{\mathbf{a}_2 \in \{\bar{\mathbf{a}}_2 : [\mathbf{z}_1;\bar{\mathbf{a}}_2] \in \mathbb{A}_N\}} (\mathbf{z}_2, \mathbf{a}_2)_{\mathbf{M}_2} - G([\mathbf{z}_1;\mathbf{a}_2]),$$

it holds

$$\nabla_{\mathbf{M}_1,\mathbf{z}_1} H(\hat{\mathfrak{z}}(\mathbf{a})) = -\nabla_{\mathbf{M}_1,\mathbf{a}_1} G(\mathbf{a}), \qquad \nabla_{\mathbf{M}_2,\mathbf{z}_2} H(\hat{\mathfrak{z}}(\mathbf{a})) = \mathbf{a}_2,$$

and the coordinate representation can be equally rewritten as follows:

Find $\mathbf{z} \in \mathcal{C}^1([0,T];\hat{\mathfrak{z}}(\mathbb{A}_N))$*,* $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ *and* $\mathbf{e}_B \in \mathcal{C}([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ *solving*

$$\mathbf{M}\frac{d}{dt}\mathbf{z}(t) = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\tilde{\mathbf{R}}(\mathbf{z}(t)) \end{bmatrix} \nabla_{\mathbf{M}} H(\mathbf{z}(t)) + \mathbf{K}\mathbf{e}_B(t)$$
$$\mathbf{f}_B(t) = \mathbf{K}^T\nabla_{\mathbf{M},\mathbf{z}} H(\mathbf{z}(t)),$$

with $\tilde{\mathbf{R}}(\cdot)$ *defined such that* $\tilde{\mathbf{R}}(\hat{\mathfrak{z}}(\mathbf{a})) = \mathbf{R}(\mathbf{a})$ *for* $\mathbf{a} \in \mathbb{A}_N$*.*

Proof. The theorem can be derived by employing a partial Legendre transform on the functional G and the results from Lemma 1.14. □

Remark 2.29. *It can be shown that there exists a transformation* Ψ_z *from the coordinate representation* $\mathbf{z} \in \mathbb{R}^n$ *to the function* $\underline{\mathbf{z}} : \Omega \to \mathbb{R}^2$*, such that*

$$\hat{\mathbf{z}}(\Psi(\mathbf{a})) = \Psi_z(\hat{\mathfrak{z}}(\mathbf{a})), \qquad \text{for } \mathbf{a} \in \mathbb{A}_N \subset \mathbb{R}^n.$$

The Hamiltonian of Theorem 2.33 can then also be written as $H(\mathbf{z}) = \langle h(\Psi_z(\mathbf{z})), 1\rangle$*. This is in correspondence to the results in Theorem 1.6.*

Complexity-reduced system

All results we were able to derive for the purely projection-based approximation, System 2.13, can be carried over to the complexity-reduced System 2.22. While the first result, Corollary 2.30, holds independently of Assumption 2.20, the derivation of all other upcoming generalizations heavily rely on that assumption.

Corollary 2.30. *Let the system matrices be as in Corollary 2.26. Let, furthermore,*

$$G_c : \mathbb{A}_N \to \mathbb{R}^N, \quad G_c(\mathbf{a}) = \langle g(\Psi(\mathbf{a})), 1\rangle_c, \qquad \text{for } \mathbb{A}_N \subset \mathbb{R}^N$$
$$\mathbf{R}_c(\mathbf{a}) = [\langle r(\Psi(\mathbf{a}))b_2^n, b_2^m\rangle_c]_{m,n=1,\dots,N_2}\,.$$

Then System 2.22 can be equivalently described as:
Find $\mathbf{a} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$, $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ *and* $\mathbf{e}_B \in \mathcal{C}([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$*, solving*

$$\mathbf{M}\frac{d}{dt}\begin{bmatrix} \mathbf{a}_1(t) \\ \nabla_{\mathbf{M}_2,\mathbf{a}_2} G_c(\mathbf{a}(t)) \end{bmatrix} = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R}_c(\mathbf{a}(t)) \end{bmatrix} \begin{bmatrix} -\nabla_{\mathbf{M}_1,\mathbf{a}_1} G_c(\mathbf{a}(t)) \\ \mathbf{a}_2(t) \end{bmatrix} + \mathbf{K}\mathbf{e}_B(t)$$
$$\mathbf{f}_B(t) = \mathbf{K}^T\mathbf{a}(t).$$

Lemma 2.31. *For* G_c *as in Corollary 2.30, it holds for fixed* $\mathbf{a}_1$ *that*

$$\Phi : \mathbb{R}^{N_2} \to \mathbb{R}, \qquad \Phi : \mathbf{a}_2 \mapsto G_c([\mathbf{a}_1;\mathbf{a}_2])$$

is strictly convex on every convex open subset of its domain.

Proof. For fixed $\mathbf{a}_1 \in \mathbb{R}^{N_1}$, let a convex subset $\mathbb{S} \subset \mathbb{R}^{N_2}$ of the domain of Φ be given, and let $\mathbf{a}_2, \bar{\mathbf{a}}_2 \in \mathbb{S}$ with $\mathbf{a}_2 \neq \bar{\mathbf{a}}_2$. As Φ is differentiable, strict convexity can equivalently be shown by Theorem 1.4 by verifying the inequality

$$(\nabla_{\mathbf{a}_2}\Phi(\bar{\mathbf{a}}_2) - \nabla_{\mathbf{a}_2}\Phi(\mathbf{a}_2)) \cdot (\bar{\mathbf{a}}_2 - \mathbf{a}_2) \overset{!}{>} 0.$$

As a preparation, we introduce the following notation for the functions related to the coordinate representations,

$$a_2 = \sum_{l=1}^{N_1} b_2^l \mathfrak{a}_2^l \in \mathcal{V}_2, \quad \bar{a}_2 = \sum_{l=1}^{N_1} b_2^l \bar{\mathfrak{a}}_2^l \in \mathcal{V}_2, \quad \text{for } \mathbf{a}_2 = [\mathfrak{a}_2^1;\dots;\mathfrak{a}_2^{N_2}],\ \bar{\mathbf{a}}_2 = [\bar{\mathfrak{a}}_2^1;\dots;\bar{\mathfrak{a}}_2^{N_2}] \in \mathbb{R}^{N_2}.$$

Moreover, we define $a_1 \in \mathcal{V}_1$ as the function related to the coordinate representation $\mathbf{a}_1$. Then it holds

$$\begin{aligned}\nabla_{\mathbf{a}_2}\Phi(\mathbf{a}_2) \cdot \bar{\mathbf{a}}_2 &= \sum_{l=1}^{N_1} \frac{\partial}{\partial_{\mathfrak{a}_{2,l}}}\Phi(\mathbf{a}_2)\,\bar{\mathfrak{a}}_2^l = \sum_{l=1}^{N_1} \left\langle \nabla_{a_2} g([a_1;a_2]), b_2^l \right\rangle_c \bar{\mathfrak{a}}_2^l \\ &= \sum_{i\in I} w_i \int_{K_i} \nabla_{a_2} g([a_1;a_2]) \left(\sum_{l=1}^{N_1} \bar{\mathfrak{a}}_2^l b_2^l\right) dx = \sum_{i\in I} w_i \int_{K_i} \nabla_{a_2} g([a_1;a_2])\,\bar{a}_2 dx.\end{aligned}$$

And likewise, it holds

$$(\nabla_{\mathbf{a}_2}\Phi(\bar{\mathbf{a}}_2) - \nabla_{\mathbf{a}_2}\Phi(\mathbf{a}_2)) \cdot (\bar{\mathbf{a}}_2 - \mathbf{a}_2) = \sum_{i \in I} w_i \underbrace{\int_{K_i} \left(\nabla_{a_2} g([a_1;\bar{a}_2]) - \nabla_{a_2} g([a_1;a_2])\right) (\bar{a}_2 - a_2)\, dx}_{=:S_i}.$$

As $w_i > 0$ by definition, it only remains to show that $S_i \geq 0$ for $i \in I$, and for at least one $j \in I$ it is $S_j > 0$. Note that by construction, $\bar{a}_2 - a_2 \in \mathcal{V}_2$, and therefore we can follow from the equivalence of $||\cdot||_c$ and $||\cdot||$ on $\mathcal{V}_2$ that

$$\sum_{i \in I} w_i \int_{K_i} (\bar{a}_2 - a_2)^2 dx = ||\bar{a}_2 - a_2||_c^2 \geq \frac{1}{\tilde{C}^2} ||\bar{a}_2 - a_2||^2 > 0,$$

i.e., there exists a $j \in I$ with $(\bar{a}_2 - a_2)_{|K_j} \neq 0$. Further, the strict convexity of the density function $g : \mathbb{R}^2 \to \mathbb{R}$ with respect to its second argument, Theorem 1.11, then implies for this j

$$\int_{K_j} \left(\nabla_{a_2} g([a_1;\bar{a}_2]) - \nabla_{a_2} g([a_1;a_2])\right) (\bar{a}_2 - a_2)\, dx > 0, \qquad \text{and}$$
$$\int_{K_i} \left(\nabla_{a_2} g([a_1;\bar{a}_2]) - \nabla_{a_2} g([a_1;a_2])\right) (\bar{a}_2 - a_2)\, dx \geq 0, \qquad \text{for } i \in I.$$

Summarizing, it follows

$$(\nabla_{\mathbf{a}_2}\Phi(\bar{\mathbf{a}}_2) - \nabla_{\mathbf{a}_2}\Phi(\mathbf{a}_2)) \cdot (\bar{\mathbf{a}}_2 - \mathbf{a}_2) = \sum_{i \in I} w_i S_i \geq w_j S_j > 0.$$

□

From Lemma 2.31, the following two results can be derived in analogy to Lemma 1.14.

Lemma 2.32. *Let the domain of G_c, denoted by $\mathbb{A}_N$, be an open convex set. Then*

$$\hat{\mathfrak{z}} : \mathbb{A}_N \to \mathbb{R}^N, \qquad \hat{\mathfrak{z}}(\mathbf{a}) = \begin{bmatrix} \mathbf{a}_1 \\ \hat{\mathfrak{z}}_2(\underline{\mathbf{a}}) \end{bmatrix} := \begin{bmatrix} \mathbf{a}_1 \\ \nabla_{\mathbf{M}_2,\mathbf{a}_2} G_c(\mathbf{a}) \end{bmatrix}, \quad \textit{for } \mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}$$

is injective.

Theorem 2.33. *Let the system matrices be as in Corollary 2.30 and the assumptions of Lemma 2.32 hold. Let $H_c : \hat{\mathfrak{z}}(\mathbb{A}_N) \to \mathbb{R}$ be defined as partial Legendre transform of G_c by*

$$H_c(\mathbf{z}) = \sup_{\mathbf{a}_2 \in \{\bar{\mathbf{a}}_2 : [\mathbf{z}_1;\bar{\mathbf{a}}_2] \in \mathbb{A}_N\}} (\mathbf{z}_2, \mathbf{a}_2)_{\mathbf{M}_2} - G_c([\mathbf{z}_1;\mathbf{a}_2]), \qquad \textit{for } \mathbf{z} = [\mathbf{z}_1;\mathbf{z}_2].$$

Then it holds

$$\nabla_{\mathbf{M}_1,\mathbf{z}_1} H_c(\mathbf{z})_{|\mathbf{z}=\hat{\mathfrak{z}}(\mathbf{a})} = -\nabla_{\mathbf{M}_1,\mathbf{a}_1} G_c(\mathbf{a}), \qquad \nabla_{\mathbf{M}_2,\mathbf{z}_2} H_c(\mathbf{z})_{|\mathbf{z}=\hat{\mathfrak{z}}(\mathbf{a})} = \mathbf{a}_2.$$

Moreover, the coordinate representation can be equally rewritten as follows:

Find $\mathbf{z} \in \mathcal{C}^1([0,T];\hat{\mathfrak{z}}(\mathbb{A}_N))$, $\mathbf{f}_B \in \mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{e}_B \in \mathcal{C}([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$, solving

$$\begin{aligned} \mathbf{M}\frac{d}{dt}\mathbf{z}(t) &= \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\tilde{\mathbf{R}}(\mathbf{z}(t)) \end{bmatrix} \nabla_{\mathbf{M},\mathbf{z}} H_c(\mathbf{z}(t)) + \mathbf{K}\mathbf{e}_B(t) \\ \mathbf{f}_B(t) &= \mathbf{K}^T \nabla_{\mathbf{M},\mathbf{z}} H_c(\mathbf{z}(t)). \end{aligned}$$

with $\tilde{\mathbf{R}}$ defined such that $\tilde{\mathbf{R}}(\hat{\mathfrak{z}}(\mathbf{a})) = \mathbf{R}_c(\mathbf{a})$ for $\mathbf{a} \in \mathbb{A}_N$.

Note that $\hat{\mathfrak{z}}(\mathbb{A}_N)$ is a nonlinear N-dimensional manifold, which depends on the Hamiltonian density $h(\cdot)$ and the quadrature rule $\langle\cdot,\cdot\rangle_c$. The representation of Theorem 2.33 is therefore not suitable for a numerical realization. Also for the examination of well-posedness, the upcoming formulation in $\mathbf{a}$ is more convenient.

Lemma 2.34. *The solution $\underline{\mathbf{a}} \in \mathcal{C}^1([0,T];\mathcal{V})$ of System 2.22 can be equally characterized by an ordinary differential equation*

$$\frac{d}{dt}\mathbf{a}(t) = \mathbf{f}(t,\mathbf{a}(t)), \qquad\qquad \mathbf{a}(0) = \mathbf{a}_0 \in \mathbb{A}_N,$$

for $\mathbf{a} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$. In this formulation, $\mathbf{f} : [0,T] \times \mathbb{A}_N \to \mathbb{R}^N$ is continuous, and $\mathbb{A}_N$ denotes the open set of the domain of $\mathbf{f}$.

Proof. Firstly, the state transform $\hat{\mathfrak{z}}$ in Lemma 2.32 can be shown to be continuously differentiable with pointwise non-singular Jacobian. This readily follows from the non-singularity of the Jacobian of $\mathbf{a}_2 \mapsto \hat{\mathfrak{z}}(\bar{\mathbf{a}}_1,\mathbf{a}_2)$ for fixed $\bar{\mathbf{a}}_1$, which, in turn, can be derived with the help of Corollary 2.3 and a similar calculation as in the proof of Lemma 2.31.

In the second step, we consider the construction of the claimed standard form. Let $\mathbf{a} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$ be the coordinate representation of Corollary 2.30. As $\hat{\mathfrak{z}}$ is continuously differentiable, we may apply the chain-rule to get

$$\frac{d}{dt}\hat{\mathfrak{z}}(\mathbf{a}(t)) = \mathbf{Z}(\mathbf{a}(t))\frac{d}{dt}\mathbf{a}(t), \qquad \text{with } \mathbf{Z}(\mathbf{a}(t)) = \frac{d\hat{\mathfrak{z}}(\mathbf{a}(t))}{d\mathbf{a}}.$$

Using the representation of Corollary 2.30, we can write

$$\begin{aligned}(\mathbf{M}\,\mathbf{Z}(\mathbf{a}(t)))\,\frac{d}{dt}\mathbf{a}(t) &= \mathbf{M}\,\frac{d}{dt}\begin{bmatrix}\mathbf{a}_1(t)\\ \nabla_{\mathbf{M}_2,\mathbf{a}_2}G_c(\mathbf{a}(t))\end{bmatrix}\\ &= \begin{bmatrix} & \mathbf{J}\\ -\mathbf{J}^T & \mathbf{R}_c(\mathbf{a}(t))\end{bmatrix}\begin{bmatrix}-\nabla_{\mathbf{M}_1,\mathbf{a}_1}G_c(\mathbf{a}(t))\\ \mathbf{a}_2(t)\end{bmatrix} + \mathbf{K}\mathbf{e}_B(t).\end{aligned}$$

As $\mathbf{M}$ is non-singular, also $(\mathbf{M}\,\mathbf{Z}(\mathbf{a}(t)))$ is non-singular by the first part of the proof. Because of that and because the algebraic conditions $\mathbf{f}_B(t) = \mathbf{K}^T\mathbf{a}(t)$ present in the coordinate representation of Corollary 2.30 are linear in $\mathbf{a}(t)$, the differential equation in the above stated standard form can be derived by eliminating the boundary port variables $\mathbf{e}_B, \mathbf{f}_B$. The construction is in complete analogy to the one in Lemma I.A-2.25 for the linear case, which is the reason why we do not repeat it here. □

2.5 Time discretization

In the following, we construct time discretization schemes for System 2.13. The compatibility conditions for $\mathcal{V}_1$ and $\mathcal{V}_2$, Assumption 2.12, are supposed throughout. The schemes are constructed such that an analogous energy-bound as Theorem 2.17 can be shown for the fully discrete system. The first two ones, System 2.37 and System 2.40, feature numerical dissipation, though. Their energy bounds are therefore referred to as dissipation *inequalities*, instead of dissipation *equalities*. The third, System 2.45 is dissipation-free but more complex and restrictive in terms of implementation.

Remark 2.35. *The same kind of results hold, when the time discretization schemes are applied to the complexity-reduced System 2.22 instead. Also the proofs can be transferred almost verbatim, which is why we do not discuss them separately.*

In the following, let $0 = t_0 < t_1 < \ldots < t_K = T$ be a partition of $[0,T]$, $I_k = (t_{k-1}, t_k]$ and $\bar{I}_k = [t_{k-1}, t_k]$ for $k = 1, \ldots, K$. For ease of presentation, we assume constant time steps $\Delta_t = t_k - t_{k-1}$. The closing conditions (2.4) are realized for the time-discretized systems as follows: Given initial data $\underline{\mathbf{a}}_0$ and one boundary data $u^{\nu,k}$ per boundary node $\nu \in \mathcal{N}_\partial$ and time-step k, we set

$$\begin{aligned} &\textit{Type 1: } e_B^k[\nu] = u^{k,\nu}, \qquad \textit{Type 2: } f_B^k[\nu] = u^{k,\nu} \\ &\underline{\mathbf{a}}^0 \equiv \underline{\mathbf{a}}_0, \quad \text{for } \underline{\mathbf{a}}_0 \in \mathcal{V}. \end{aligned} \tag{2.5}$$

Remark 2.36. *Let us stress that the upcoming time discretization schemes employ different ansatz spaces for the solution. The realization of time-discrete boundary conditions $u^{k,\nu}$ has to be adapted accordingly: The first scheme, System 2.37, can be seen as a finite-difference-type scheme and thus $u^{k,\nu}$ is nothing but a point-evaluation in time then. For the Galerkin-in-time methods, System 2.40 and System 2.45, $u^{k,\nu}$ is a polynomial in time and is obtained by quadrature from the continuous boundary data in* (2.4)*.*

We start with the simplest and most dissipative discretization scheme.

System 2.37 (Implicit-Euler-type scheme)**.** *Let initial- and boundary-data be given by* (2.5)*. Then, find for $k \geq 1$, $\underline{\mathbf{a}}^k = [a_1^k; a_2^k] \in \mathcal{V}_1 \times \mathcal{V}_2$, $\mathbf{f}_B^k, \mathbf{e}_B^k \in \mathbb{R}^{|\mathcal{N}_\partial|}$ by solving*

$$\begin{aligned} \frac{1}{\Delta_t}\langle a_1^k - a_1^{k-1}, b_1\rangle &= -\langle \partial_x a_2^k, b_1\rangle \\ \frac{1}{\Delta_t}\langle \hat{z}_2(\underline{\mathbf{a}}^k) - \hat{z}_2(\underline{\mathbf{a}}^{k-1}), b_2\rangle &= \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)), \partial_x b_2\rangle + \mathbf{e}_B^k \cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}^k) a_2^k, b_2\rangle \\ \mathbf{f}_B^k &= \mathcal{T} a_2^k \end{aligned}$$

for all $b_1 \in \mathcal{V}_1$, $b_2 \in \mathcal{V}_2$.

Note that System 2.37 can be interpreted as an implicit-Euler-type discretization in the energy variable $\mathbf{z}$. This differs from the classical implicit Euler scheme, in which the second equation would read instead

$$\frac{1}{\Delta_t}\left\langle \left(\nabla_{\mathbf{a}} \hat{z}_2(\underline{\mathbf{a}}^k)\right) \cdot \left(\underline{\mathbf{a}}^k - \underline{\mathbf{a}}^{k-1}\right), b_2\right\rangle = \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)), \partial_x b_2\rangle + \mathbf{e}_B^k \cdot \mathcal{T} b_2 - \langle r(\underline{\mathbf{a}}^k) a_2^k, b_2\rangle.$$

The interpretation of the system in energy-variables plays a crucial role for all our schemes. A similar result as Theorem 2.17 for the time-continuous case can be derived.

Theorem 2.38 (Energy-dissipation inequality)**.** *For any solution of System 2.37 it holds for $0 \leq l < k \leq K$,*

$$\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^l)) < \Delta_t \left(\sum_{j=l+1}^{k} \mathbf{e}_B^j \cdot \mathbf{f}_B^j - \left\langle r(\underline{\mathbf{a}}^j), (a_2^j)^2 \right\rangle \right) \leq \Delta_t \sum_{j=l+1}^{k} \mathbf{e}_B^j \cdot \mathbf{f}_B^j,$$

where the Hamiltonian for $\underline{\mathbf{a}} \in \mathcal{V}_1 \times \mathcal{V}_2$ is defined as $\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}})) = \langle h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), 1\rangle$.

Proof. Clearly, it suffices to consider the case $l = k-1$, as $l < k-1$ can be directly derived from that case. Using the characterization of strict convexity of Theorem 1.4 onto $h(\cdot)$ yields

$$\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^{k-1})) = \left\langle h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) - h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^{k-1})), 1 \right\rangle < \left\langle \hat{\mathbf{z}}(\underline{\mathbf{a}}^k) - \hat{\mathbf{z}}(\underline{\mathbf{a}}^{k-1}), \nabla h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) \right\rangle.$$

As $\nabla_{z_2} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) = a_2^k \in \mathcal{V}_2$ and Lemma 2.14 holds, the time-discrete variational principle of System 2.37 with $[b_1; b_2] = \nabla_{\mathbf{z}} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k))$ can be used, leading to

$$\begin{aligned}\frac{1}{\Delta_t} \left\langle \hat{\mathbf{z}}(\underline{\mathbf{a}}^k) - \hat{\mathbf{z}}(\underline{\mathbf{a}}^{k-1}), \nabla h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) \right\rangle &= \mathbf{e}_B^k \cdot \mathbf{f}_B^k - \langle r(\underline{\mathbf{a}}^k) a_2^k, a_2^k \rangle \\ &\quad - \langle \partial_x a_2^k, \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) \rangle + \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)), \partial_x a_2^k \rangle \\ &= \mathbf{e}_B^k \cdot \mathbf{f}_B^k - \left\langle r(\underline{\mathbf{a}}^k), (a_2^k)^2 \right\rangle \leq \mathbf{e}_B^k \cdot \mathbf{f}_B^k.\end{aligned}$$

Setting together the two inequalities shows the assertion. □

The difference of the dissipation *inequality* of Theorem 2.38 to an *equality* is given by the numerical dissipation

$$D^{\mathrm{ImpEul}}(t_k) = \Delta_t \left(\sum_{j=1}^{k} \mathbf{e}_B^j \cdot \mathbf{f}_B^j - \left\langle r(\underline{\mathbf{a}}^j) a_2^j, a_2^j \right\rangle \right) - \left[\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^0)) \right]. \tag{2.6}$$

Next we construct more elaborate time discretization schemes by Galerkin methods in time, cf. [RSV13], [AM04], [LW16], [Egg19]. In the construction, the following well-known result is crucial, cf. [Hac12].

Lemma 2.39 (Tensor product Hilbert space). *Let $\mathcal{V}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and basis $(\mathbf{v}_i)_{i\in\mathbb{N}}$, and $\mathcal{Q}$ be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$ and basis $(\mathbf{q}_i)_{i\in\mathbb{N}}$. Then the tensor space $\mathcal{Q} \otimes \mathcal{V}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{QV}}$ defined by*

$$\langle \mathbf{q}_1 \otimes \mathbf{v}_1, \mathbf{q}_2 \otimes \mathbf{v}_2 \rangle_{\mathcal{QV}} := \langle \mathbf{q}_1, \mathbf{q}_2 \rangle_{\mathcal{Q}} \langle \mathbf{v}_1, \mathbf{v}_2 \rangle_{\mathcal{V}}, \qquad \textit{for } \mathbf{q}_1, \mathbf{q}_2 \in \mathcal{Q}, \quad \mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}.$$

Because of the latter lemma, the Hilbert-space structure obtained after space discretization can be transferred to a space-time Hilbert-space structure. For a given Hilbert-function space $\mathcal{W}$, we introduce the spaces of polynomials up to q-th order on $I_k = (t_{k-1}, t_k]$ with values in $\mathcal{W}$ as

$$\mathcal{Q}_q(I_k; \mathcal{W}) = \left\{ \psi : I_k \to \mathcal{W} : \psi(t) = \sum_{j=0}^{q} t^j w_j, \text{ with } w_j \in \mathcal{W} \right\}.$$

Clearly, these spaces are tensor Hilbert spaces as described in Lemma 2.39, when the $\mathcal{L}^2$-scalar product is used in time. Continuous and discontinuous compositions of the spaces are defined as

$$\begin{aligned}\mathcal{P}_q(\mathcal{W}) &= \left\{ \phi \in \mathcal{C}((0,T]; \mathcal{W}) : \ \phi_{|I_k} \in \mathcal{Q}_q(I_k; \mathcal{W}), \text{ for } k = 1, \dots, K \right\} \\ \mathcal{Q}_q(\mathcal{W}) &= \left\{ \phi : (0,T] \to \mathcal{W} : \quad \phi_{|I_k} \in \mathcal{Q}_q(I_k; \mathcal{W}), \text{ for } k = 1, \dots, K \right\}.\end{aligned}$$

A function $\phi \in \mathcal{Q}_q(\mathcal{W})$ may have jumps at the interval boundaries t_k, but for each $k \in \{1, \dots, K\}$, a continuous function can be defined on the closed interval $\bar{I}_k = I_k \cup \{t_{k-1}\}$ as

$$\phi^k \in \mathcal{C}(\bar{I}_k; \mathcal{W}), \qquad \phi^k(t) = \phi_{|I_k}(t) \quad \text{for } t \in I_k,$$

without ambiguities. When $\phi \in \mathcal{P}_q(\mathcal{W})$, it simply holds $\phi(t_k) = \phi^k(t_k) = \phi^{k+1}(t_k)$. In the upcoming, we consider the time discretization of System 2.13 by discontinuous Petrov-Galerkin methods.

System 2.40 (Time-discontinuous Galerkin)**.** *Find* $\underline{\mathbf{a}} = [a_1; a_2] \in \mathcal{P}_q(\mathcal{V}_1) \times \mathcal{Q}_{q-1}(\mathcal{V}_2)$ *and* $\mathbf{e}_B, \mathbf{f}_B \in \mathcal{Q}_{q-1}\left(\mathbb{R}^{|\mathcal{N}_\partial|}\right)$ *such that for* $k = 1, \ldots, K$ *and* $\underline{\mathbf{a}}^k \in \mathcal{C}(\bar{I}_k; \mathcal{V}_1 \times \mathcal{V}_2)$ *with* $\underline{\mathbf{a}}^k(t) = \underline{\mathbf{a}}_{|I_k}(t)$ *for* $t \in I_k$ *it holds*

$$\begin{aligned}
\int_{I_k} \langle \partial_t a_1^k(t), b_1 \rangle dt &= -\int_{I_k} \langle \partial_x a_2^k(t), b_1 \rangle dt \\
\int_{I_k} \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}^k(t)), b_2 \rangle dt &= \int_{I_k} \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t))), \partial_x b_2 \rangle - \langle r(\underline{\mathbf{a}}^k(t)) a_2^k(t), b_2 \rangle dt \\
&\quad + \int_{I_k} \mathbf{e}_B(t) \cdot \mathcal{T} b_2 dt - \left[\langle \hat{z}_2(\underline{\mathbf{a}}^k(t_{k-1})) - \hat{z}_2(\underline{\mathbf{a}}^{k-1}(t_{k-1})), b_2(t_{k-1}) \rangle \right] \\
\mathbf{f}_B(t) &= \mathcal{T} a_2(t).
\end{aligned}$$

for all $b_1 \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_1)$*,* $b_2 \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_2)$*, with closing conditions* (2.5)*.*

Lemma 2.41 (Space-time local mass conservation)**.** *Let Assumption 2.12 hold for* $\mathcal{V}_1$ *and* $\mathcal{V}_2$*, and let for* $[a_1; a_2] \in \mathcal{Q}_q(\bar{I}_k; \mathcal{V}_1) \times \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_2)$ *the variational principle*

$$\int_{I_k} \langle \partial_t a_1(t), b_1 \rangle dt = -\int_{I_k} \langle \partial_x a_2(t), b_1 \rangle dt$$

hold for all $b_1 \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_1)$*. Then* $\partial_t a_1 = -\partial_x a_2$*.*

Proof. By construction, $\partial_t a_1, \partial_x a_2 \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_1)$, i.e., they lie in the test space. Therefore, the variational principle with $b_1 = \partial_t a_1 + \partial_x a_2$ gives

$$\int_{I_k} \langle \partial_t a_1(t), \partial_t a_1(t) + \partial_x a_2(t) \rangle dt = \int_{I_k} \langle -\partial_x a_2(t), \partial_t a_1(t) + \partial_x a_2(t) \rangle dt.$$

The $\mathcal{L}^2$-product in time together with the inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{V}_1$ form an inner product on $\mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_1)$. By subtracting the right-hand side from both sides of the latter equation, one can see that the norm of $\partial_t a_1 + \partial_x a_2$ induced by the inner product on $\mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_1)$ is zero, i.e., $\partial_t a_1 = -\partial_x a_2$. □

Remark 2.42. *The variational principle in Lemma 2.41 then clearly holds for all test functions* b_1 *for which the expressions are well-defined. Mass conservation, locally in space and time, can be derived by testing the variational principle with local indicator functions.*

Theorem 2.43 (Energy-dissipation inequality)**.** *For any solution of System 2.40 it holds for* $0 \leq l < k \leq K$

$$\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_l))) < \int_{t_l}^{t_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t)) a_2(t), a_2(t) \rangle dt \leq \int_{t_l}^{t_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) dt,$$

where the Hamiltonian for $\underline{\mathbf{a}} \in \mathcal{V}_1 \times \mathcal{V}_2$ *is defined as* $\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}})) = \langle h(\hat{\mathbf{z}}(\underline{\mathbf{a}})), 1 \rangle$*.*

Proof. It suffices to consider the case $l = k - 1$, as $l < k - 1$ directly follows from that. For convenience, the abbreviation $\underline{\mathbf{c}}^j = [c_1^j; c_2^j] := \hat{\mathbf{z}}(\underline{\mathbf{a}}^j(t_{k-1}))$ for $j \in \{k-1, k\}$ is used throughout the proof. It then holds

$$\begin{aligned}
\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^{k-1}(t_{k-1}))) &= \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t_{k-1}))) + \mathcal{H}(\underline{\mathbf{c}}^k) - \mathcal{H}(\underline{\mathbf{c}}^{k-1}) \\
&= \int_{I_k} \left\langle \frac{d}{dt} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t))), 1 \right\rangle dt + \left\langle h(\underline{\mathbf{c}}^k) - h(\underline{\mathbf{c}}^{k-1}), 1 \right\rangle
\end{aligned}$$

By the strict convexity of $c \mapsto h([c_1^k; c])$ for fixed c_1^k and the fact that $c_1^k = a_1^k(t_{k-1}) = c_1^{k-1}$ by the choice of the ansatz spaces, it holds by Theorem 1.4 that

$$\left(h(\underline{\mathbf{c}}^k) - h(\underline{\mathbf{c}}^{k-1})\right) - \nabla_{z_2} h(\mathbf{z})_{|\mathbf{z}=\underline{\mathbf{c}}^k} \left(c_2^k - c_2^{k-1}\right) < 0.$$

This clearly implies

$$\left\langle h(\underline{\mathbf{c}}^k) - h(\underline{\mathbf{c}}^{k-1}), 1\right\rangle - \left\langle c_2^k - c_2^{k-1}, \nabla_{z_2} h(\mathbf{z})_{|\mathbf{z}=\underline{\mathbf{c}}^k}\right\rangle < 0. \tag{2.7}$$

By construction, $\underline{\mathbf{a}}^k \in \mathcal{Q}_q(I_k; \mathcal{V}_1) \times \mathcal{Q}_{q-1}(I_k; \mathcal{V}_2)$ is smooth in time on I_k, and the chain rule can be applied to get

$$\begin{aligned}\int_{I_k} \left\langle \frac{d}{dt} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t))), 1\right\rangle dt &= \int_{I_k} \langle \partial_t \hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t)), \nabla_{\mathbf{z}} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t)))\rangle dt \\ &= \int_{I_k} \langle \partial_t a_1^k(t), \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t)))\rangle + \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}^k(t)), a_2^k(t)\rangle dt.\end{aligned}$$

As $\nabla_{z_2} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k)) = a_2^k \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_2)$ and Lemma 2.41 holds, the variational principle with $[b_1; b_2] = \nabla_{\mathbf{z}} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k))$ can be used, leading to

$$\begin{aligned}\int_{I_k} \left\langle \frac{d}{dt} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t))), 1\right\rangle dt = &- \int_{I_k} \langle \partial_x a_2^k(t), \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t)))\rangle + \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t))), \partial_x a_2^k(t)\rangle dt \\ &- \langle r(\underline{\mathbf{a}}^k(t)) a_2^k(t), a_2^k(t)\rangle dt + \int_{I_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) dt - \left[\langle c_2^k - c_2^{k-1}, a_2^k(t_{k-1})\rangle\right] \\ = &- \langle r(\underline{\mathbf{a}}^k(t)) a_2^k(t), a_2^k(t)\rangle dt + \int_{I_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) dt - \left[\langle c_2^k - c_2^{k-1}, \nabla_{z_2} h(\underline{\mathbf{c}}^k)\rangle\right].\end{aligned}$$

Inserting the latter equality into the first equation of the proof and using estimation (2.7) finishes the proof. □

The difference of the dissipation *inequality* of Theorem 2.43 to an *equality* is given by the numerical dissipation

$$D^{\mathrm{DG}}(t_k) = \int_0^{t_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t)) a_2(t), a_2(t)\rangle dt - \left[\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(0)))\right]. \tag{2.8}$$

Towards energy-preserving time discretization schemes

At least theoretically, time discretization schemes fulfilling a dissipation-equality can also be constructed by Galerkin methods in time, which is demonstrated in the following.

Definition 2.44. *Let $\Pi_k : \mathcal{L}^2(\bar{I}_k; \mathbb{R}) \to \mathcal{Q}_{q-1}(\bar{I}_k; \mathbb{R})$ be the $\mathcal{L}^2$-orthogonal projection in time, defined for $f \in \mathcal{L}^2(\bar{I}_k; \mathbb{R})$ as*

$$\int_{I_k} (\Pi_k f(t)) g(t) dt = \int_{I_k} f(t) g(t) dt, \qquad \text{for all } g \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathbb{R}).$$

System 2.45 (Time-continuous Galerkin)**.** *Find $\underline{\mathbf{a}} = [a_1; a_2]$, $\underline{\mathbf{a}} : [0, T] \to \mathcal{V}_1 \times \mathcal{V}_2$ and $\mathbf{e}_B \in \mathcal{Q}_{q-1}\left(\mathbb{R}^{|\mathcal{N}_\partial|}\right)$, $\mathbf{f}_B : [0, T] \to \mathbb{R}^{|\mathcal{N}_\partial|}$ with $a_1 \in \mathcal{P}_q(\mathcal{V}_1)$ and $\hat{z}_2(\mathbf{a}[x]) \in \mathcal{P}_q(\mathbb{R})$ for $x \in \Omega$, such that for $k = 1, \ldots, K$ it holds*

$$\begin{aligned}
\int_{I_k} \langle \partial_t a_1(t), b_1 \rangle dt &= -\int_{I_k} \langle \partial_x a_2(t), b_1 \rangle dt \\
\int_{I_k} \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), b_2 \rangle dt &= \int_{I_k} \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x b_2 \rangle - \langle r(\underline{\mathbf{a}}(t))\, \Pi_k\, a_2(t), b_2 \rangle dt \\
&\quad + \int_{I_k} \mathbf{e}_B(t) \cdot \mathcal{T} b_2 dt \\
\mathbf{f}_B(t) &= \mathcal{T} a_2(t).
\end{aligned}$$

for all $b_1 \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_1)$, $b_2 \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathcal{V}_2)$, with closing conditions (2.5)*.*

Theorem 2.46 (Energy-dissipation equality)**.** *For any solution of System 2.45 it holds for $0 \leq l < k \leq K$*

$$\begin{aligned}
\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_l))) &= \sum_{j=l+1}^{k} \int_{I_j} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t)), (\Pi_j a_2(t))^2 \rangle dt \\
&\leq \int_{t_l}^{t_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) dt.
\end{aligned}$$

Proof. We consider only $l = k-1$, as $l < k-1$ can be directly derived from that case. By applying the chain rule we get

$$\begin{aligned}
\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_{k-1}))) &= \int_{I_k} \left\langle \frac{d}{dt} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), 1 \right\rangle dt \\
&= \int_{I_k} \langle \partial_t a_1(t), \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) \rangle + \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), a_2(t) \rangle dt \\
&= \int_{I_k} \langle \partial_t a_1(t), \Pi_k \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) \rangle + \langle \partial_t \hat{z}_2(\underline{\mathbf{a}}(t)), \Pi_k a_2(t) \rangle dt.
\end{aligned}$$

In the last step, we used that $\partial_t a_1[x], \partial_t \hat{z}_2(\mathbf{a}[x]) \in \mathcal{Q}_{q-1}(\bar{I}_k; \mathbb{R})$ for almost all $x \in \Omega$. Using now the variational principle of System (2.45), we get

$$\begin{aligned}
\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_{k-1}))) &= \int_{I_k} -\langle r(\underline{\mathbf{a}}(t))\, \Pi_k\, a_2(t), \Pi_k a_2(t) \rangle + \mathbf{e}_B(t) \cdot \mathcal{T}\left(\Pi_k a_2(t)\right) dt \\
+ \int_{I_k} -\langle \partial_x a_2(t), \Pi_k \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) \rangle &+ \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x (\Pi_k a_2(t)) \rangle dt \\
&= \int_{I_k} -\left\langle r(\underline{\mathbf{a}}(t)), (\Pi_k a_2(t))^2 \right\rangle + \mathbf{e}_B(t) \cdot \mathcal{T}\left(\Pi_k a_2(t)\right) dt.
\end{aligned}$$

The last equality holds, as

$$\begin{aligned}
\int_{I_k} \langle \partial_x a_2(t), \Pi_k \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) \rangle dt &= \int_{I_k} \langle \Pi_k \partial_x a_2(t), \Pi_k \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))) \rangle dt \\
&= \int_{I_k} \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t))), \partial_x (\Pi_k a_2(t)) \rangle dt.
\end{aligned}$$

Finally, the boundary-term can be rewritten as

$$\int_{I_k} \mathbf{e}_B(t) \cdot \mathcal{T}\left(\Pi_k a_2(t)\right) dt = \int_{I_k} \left(\Pi_k \mathbf{e}_B(t)\right) \cdot \mathcal{T} a_2(t) dt = \int_{I_k} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) dt,$$

which finishes the proof. □

Theorem 2.46 describes a dissipation *equality* in contrast to Theorem 2.43 and Theorem 2.38. Thus, the term

$$D^{\mathrm{CoG}}(t_k) = \sum_{l=1}^{k} \left[\int_{I_l} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) - \langle r(\underline{\mathbf{a}}(t)) \Pi_l a_2(t), a_2(t) \rangle dt \right] - \left[\mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(t_k))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(0))) \right] \quad (2.9)$$

is provable equal zero in exact arithmetic.

Remark 2.47. *The realization of System 2.45 is more delicate in comparison to the former methods due to the different parametrizations in space and time. Note furthermore that terms linear in the variable a_2, such as $\int_{I_k} \langle \partial_x a_2(t), b_1 \rangle dt$, lead to nonlinear expressions in the coefficients of the space-time discrete solution.*

For the special case of $\hat{z}_2 : \mathbb{R}^2 \to \mathbb{R}$ being a linear map, System 2.45 coincides with [LW16] for the conservation part. Only for this case, terms linear in the variable a_2, such as $\int_{I_k} \langle \partial_x a_2(t), b_1 \rangle dt$, lead to linear expressions in terms of the coefficient representations of the fully discrete system. The respective coordinate representation is stated in System 2.50.

Coordinate representations of discontinuous Galerkin in time

The detailed coordinate representation of the time-discrete System 2.40 for the lowest-order $q = 1$ is presented here. As by construction the test functions b_1, b_2 as well as the solution components a_2^k, $\partial_t a_1^k$, $\mathbf{f}_B^k$ and $\mathbf{e}_B^k$ are constant in time on I_k, most integrals in System 2.40 can be evaluated explicitly. Thus, for $k \geq 1$, the piecewise solution $\underline{\mathbf{a}}^k \in \mathcal{P}_1(\bar{I}_k; \mathcal{V}_1) \times \mathcal{Q}_0(\bar{I}_k; \mathcal{V}_2)$ and the constant values $\mathbf{e}_B^k, \mathbf{f}_B^k \in \mathbb{R}^{|\mathcal{N}_\partial|}$ fulfill

$$\begin{aligned}
\left\langle a_1^k(t_k) - a_1^k(t_{k-1}), b_1(t_{k-1}) \right\rangle &= -\Delta_t \left\langle \partial_x a_2^k(t_{k-1}), b_1(t_{k-1}) \right\rangle \\
\left\langle \hat{z}_2(\underline{\mathbf{a}}^k(t_k)) - \hat{z}_2(\underline{\mathbf{a}}^{k-1}(t_{k-1})), b_2(t_{k-1}) \right\rangle &= \int_{I_k} \langle \nabla_{z_1} h(\hat{\mathbf{z}}(\underline{\mathbf{a}}^k(t))), \partial_x b_2(t_{k-1}) \rangle \\
&\quad - \langle r(\underline{\mathbf{a}}^k(t)) a_2^k(t), b_2(t_{k-1}) \rangle dt + \Delta_t \, \mathbf{e}_B^k \cdot \mathcal{T} b_2(t_{k-1}) \\
\mathbf{f}_B^k &= \mathcal{T} a_2^k(t_{k-1}).
\end{aligned}$$

Given the transformation $\Psi : \mathbb{R}^N \to \mathcal{V}$ as in Definition 2.25, the related coordinate representation $\mathbf{a}^j = [\mathbf{a}_1^j; \mathbf{a}_2^j] \in \mathbb{R}^n$ is characterized by

$$\underline{\mathbf{a}}^k(t) = \Psi\left(\begin{bmatrix} \frac{t_k - t}{\Delta_t} \mathbf{a}_1^{k-1} + \frac{t - t_{k-1}}{\Delta_t} \mathbf{a}_1^k \\ \mathbf{a}_2^k \end{bmatrix} \right), \qquad \text{for } t \in [t_{k-1}, t_k].$$

For $\mathbf{a} = [\mathbf{a}_1; \mathbf{a}_2] \in \mathbb{R}^n$, we introduce the notation

$$\hat{\mathfrak{z}}(\mathbf{a}) = \begin{bmatrix} \mathbf{a}_1 \\ \nabla_{\mathbf{M}_2, \mathbf{a}_2} G(\mathbf{a}) \end{bmatrix}$$

for the energy variable. Finally, we employ a quadrature-rule for the remaining time-integrals. Given quadrature-weights ω_p and quadrature-points t_p for $p \in \{1, \dots, P\}$ the quadrature-rule for a vector- or matrix-valued function $\mathbf{f} : \mathbf{a} \to \mathbf{f}(\mathbf{a})$ is then defined as

$$\Xi^k\left[\mathbf{f}\right] := \sum_{p=1}^{P} \omega_p \mathbf{f}\left(\begin{bmatrix} \frac{t_k - t_p}{\Delta_t} \mathbf{a}_1^{k-1} + \frac{t_p - t_{k-1}}{\Delta_t} \mathbf{a}_1^k \\ \mathbf{a}_2^k \end{bmatrix} \right) \approx \frac{1}{\Delta_t} \int_{I_k} \mathbf{f}(\mathbf{a}^k(t)) dt. \tag{2.10}$$

The algebraic representation for the marching scheme reads then as follows:

System 2.48 (Coordinate representation of System 2.40)**.** *Given closing conditions* (2.5)*, find for* $k \geq 1$*,* $\mathbf{a}^k \in \mathbb{R}^N$*,* $\mathbf{e}_B^k, \mathbf{f}_B^k \in \mathbb{R}^{|\mathcal{N}_\partial|}$*, solving*

$$\frac{1}{\Delta_t} \mathbf{M} \left(\hat{\mathfrak{z}}\left(\mathbf{a}^k\right) - \hat{\mathfrak{z}}\left(\mathbf{a}^{k-1}\right) \right) = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\Xi^k\left[\mathbf{R}\right] \end{bmatrix} \begin{bmatrix} \Xi^k\left[-\nabla_{\mathbf{M}_1, \mathbf{a}_1} G\right] \\ \mathbf{a}_2^k \end{bmatrix} + \mathbf{K}\mathbf{e}_B^k$$
$$\mathbf{f}_B^k = \mathbf{K}^T \mathbf{a}^k.$$

Remark 2.49. *The quadrature in time* Ξ^k *makes the implementation much more convenient. We use high-order quadrature rules, which are expected to come with negligible small additional errors. The rigorous inclusion of quadrature rules in time into our analysis is left for future research here.*

Coordinate representations of continuous Galerkin in time

Here, the coordinate representation of the time-discrete System 2.45 for the lowest order $q = 1$ and the special case $\hat{z}_2([a_1; a_2]) = a_2$ is discussed. The latter means that $\underline{\mathbf{a}} = \hat{\mathbf{z}}(\underline{\mathbf{a}})$.

By construction it then holds $\underline{\mathbf{a}} = [a_1; a_2] \in \mathcal{P}_1([0, T]; \mathcal{V})$, $\mathbf{e}_B \in \mathcal{Q}_0([0, T]; \mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{f}_B \in \mathcal{P}_1([0, T]; \mathbb{R}^{|\mathcal{N}_\partial|})$. By basic manipulations, System 2.45 may also be simplified to

$$\begin{aligned} \langle a_1(t_k) - a_1(t_{k-1}), b_1(t_{k-1}) \rangle &= -\int_{I_k} \langle \partial_x a_2(t), b_1(t_{k-1}) \rangle \, dt \\ \langle a_2(t_k) - a_2(t_{k-1}), b_2(t_{k-1}) \rangle &= \int_{I_k} \langle \nabla_{a_1} h(\underline{a}(t)), \partial_x b_2(t_{k-1}) \rangle \\ &\qquad - \left\langle r(\underline{a}(t)) \left(\int_{I_k} a_2(\xi) d\xi \right), b_2(t_{k-1}) \right\rangle dt + \mathbf{e}_B^k \cdot \mathcal{T} b_2(t_{k-1}) \\ \mathbf{f}_B(t) &= \mathcal{T} a_2(t), \end{aligned}$$

where again $\mathbf{e}_B^k \in \mathbb{R}^{|\mathcal{N}_\partial|}$ is equal to the constant value $\mathbf{e}_B(t)$ for $t \in I_k$.

Given the transformation $\Psi : \mathbb{R}^N \to \mathcal{V}$ as in Definition 2.25, the related coordinate representation $\mathbf{a}^j = [\mathbf{a}_1^j; \mathbf{a}_2^j] \in \mathbb{R}^n$ and $\mathbf{f}_B^j \in \mathbb{R}^{|\mathcal{N}_\partial|}$ is characterized by

$$\begin{aligned} \underline{\mathbf{a}}(t) &= \Psi\left(\frac{t_k - t}{\Delta_t} \mathbf{a}^{k-1} + \frac{t - t_{k-1}}{\Delta_t} \mathbf{a}^k \right), \\ \mathbf{f}_B(t) &= \frac{t_k - t}{\Delta_t} \mathbf{f}_B^{k-1} + \frac{t - t_{k-1}}{\Delta_t} \mathbf{f}_B^k, \qquad \text{for } t \in [t_{k-1}, t_k]. \end{aligned}$$

The quadrature rule for a vector- or matrix-valued function $\mathbf{f} : \mathbf{a} \to \mathbf{f}(\mathbf{a})$ is then defined as

$$\Xi^k\left[\mathbf{f}\right] := \sum_{p=1}^{P} \omega_p \mathbf{f}\left(\frac{t_k - t_p}{\Delta_t} \mathbf{a}^{k-1} + \frac{t_p - t_{k-1}}{\Delta_t} \mathbf{a}^k \right) \approx \frac{1}{\Delta_t} \int_{I_k} \mathbf{f}(\mathbf{a}^k(t)) dt.$$

System 2.50 (Coordinate representation of System 2.45)**.** *Given closing conditions* (2.5)*, find for* $k \geq 1$, $\mathbf{a}^k \in \mathbb{R}^N$, $\mathbf{e}_B^k, \mathbf{f}_B^k \in \mathbb{R}^{|\mathcal{N}_\partial|}$*, solving*

$$\begin{aligned}\frac{1}{\Delta_t}\mathbf{M}\left(\hat{\mathfrak{z}}\left(\mathbf{a}^k\right)-\hat{\mathfrak{z}}\left(\mathbf{a}^{k-1}\right)\right) &= \begin{bmatrix} & \mathbf{J}\\ -\mathbf{J}^T & -\Xi^k\left[\mathbf{R}\right]\end{bmatrix}\begin{bmatrix}-\Xi^k\left[\nabla_{\mathbf{M}_1,\mathbf{a}_1}G\right]\\ \frac{1}{2}\left(\mathbf{a}_2^{k-1}+\mathbf{a}_2^k\right)\end{bmatrix}+\mathbf{K}\mathbf{e}_B^k\\ \mathbf{f}_B^k &= \mathbf{K}^T\mathbf{a}^k.\end{aligned}$$

2.6 Generalization for weighted edges

Throughout all our derivations, we assumed all edges $\omega \in \mathcal{E}$ to have a weight equal to one. If this is not the case, we define $\mathcal{A} = \{A^\omega,\ \omega \in \mathcal{E}\}$ as the set of edge-weights and extend the directed graph to $(\mathcal{N}, \mathcal{E}, l, \mathcal{A})$. To cope for the weights A^ω, the following adaptions are needed: In generalization to model problem (2.1), the state $\underline{\mathbf{z}} : [0,T] \times \Omega \to \mathbb{R}^2$ is then governed by

$$\partial_t\underline{\mathbf{z}}(t,x) = \begin{bmatrix} & -\partial_x\\ -\partial_x & -\tilde{r}(\underline{\mathbf{z}}(t,x))\end{bmatrix}\nabla_{\mathbf{z}}h(\underline{\mathbf{z}}(t,x)), \qquad\qquad x \in \Omega, \quad t \in [0,T],$$

together with coupling conditions at $\nu \in \mathcal{N}_0$ given as

$$\sum_{\omega\in\mathcal{E}(\nu)} n^\omega[\nu]A^\omega\nabla_{z_2}h(\underline{\mathbf{z}}_{|\omega}(t,\nu)) = 0, \qquad \nabla_{z_1}h(\underline{\mathbf{z}}_{|\omega}(t,\nu)) = \nabla_{z_1}h(\underline{\mathbf{z}}_{|\tilde{\omega}}(t,\nu)) \quad \text{for } \omega,\tilde{\omega} \in \mathcal{E}(\nu),$$

and one boundary condition per boundary node $\nu \in \mathcal{N}_\partial$, each of one of the following types,

$$\textit{Type 1: } \nabla_{z_1}h(\underline{\mathbf{z}}_{|\omega}(t,\nu)) = u^\nu(t), \qquad \textit{Type 2: } n^\omega[\nu]A^\omega\nabla_{z_2}h(\underline{\mathbf{z}}_{|\omega}(t,\nu)) = u^\nu(t), \quad \omega \in \mathcal{E}(\nu),$$

for $t \in [0,T]$ with prescribed boundary data $u^\nu : [0,T] \to \mathbb{R}$ and initial conditions. Note that scaling terms A^ω enter the coupling conditions and the boundary conditions of Type 2. Similarly, for the integration operator $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ over the network scaling terms A^ω have to be added,

$$\langle b, \tilde{b}\rangle_{\mathcal{E}} = \sum_{\omega\in\mathcal{E}} A^\omega \int_\omega b[x]\tilde{b}[x]dx. \tag{2.11}$$

The Hamiltonian is altered accordingly to

$$\mathcal{H}(\underline{\mathbf{z}}) = \langle h(\underline{\mathbf{z}}, 1\rangle_{\mathcal{E}} = \sum_{\omega\in\mathcal{E}} A^\omega \int_\omega h(\underline{\mathbf{z}})dx.$$

Finally, the ansatz space $\mathcal{H}^1_{div}(\mathcal{E})$, defined at the end of Part I.A-Section 1.3, is modified to

$$\mathcal{H}^1_{div}(\mathcal{E}) = \left\{ b \in \mathcal{H}^1_{pw}(\mathcal{E}) : \sum_{\omega\in\mathcal{E}(\nu)} n^\omega[\nu]A^\omega b_{|\omega}[\nu] = 0, \text{ for } \nu \in \mathcal{N}_0 \right\}.$$

With these adaptions, all our derived results generalize immediately to the case of a network with differently weighted edges.

Chapter 3

Realization of model reduction

In this chapter we discuss the implementation of our abstract model order reduction and complexity reduction approach from Chapter 2. The preceding space discretization step is assumed to be realized with the same finite element ansatz spaces as for the linear case in our studies, cf. Part I.A-Section 2.4. The finite element discretization serves as the *full order model* (FOM) in this chapter. Thus, our starting point is a coordinate representation similar to Corollary 2.26. Compared to the analysis in Chapter 2, we allow for more general boundary conditions here. They are realized by a generic nonlinear function $\mathbf{g} : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^p$.

System 3.1 (Full order model)**.** *Let input $\mathbf{u} : \mathbb{R}^p \to [0,T]$ be given and the system matrices be as in Corollary 2.26. Let further*

$$\hat{\mathbf{z}}_2(\bar{\mathbf{a}}) = \nabla_{\mathbf{a}_2} G(\bar{\mathbf{a}}), \quad \hat{\mathbf{e}}_1(\mathbf{a}) = -\nabla_{\mathbf{M}_1,\mathbf{a}_1} G(\bar{\mathbf{a}}), \quad \hat{\mathbf{d}}(\mathbf{a}) = \mathbf{R}(\mathbf{a})\mathbf{a}_2, \qquad \textit{for } \bar{\mathbf{a}} \in \mathbb{R}^N.$$

The full order model then reads: Find $\mathbf{a} \in \mathcal{C}^1([0,T];\mathbb{R}^N)$ and $\mathbf{e}_B \in \mathcal{C}([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$, solving

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} \mathbf{M}_1\mathbf{a}_1(t) \\ \hat{\mathbf{z}}_2(\mathbf{a}(t)) \end{bmatrix} &= \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & \end{bmatrix}\begin{bmatrix} \hat{\mathbf{e}}_1(\mathbf{a}(t)) \\ \mathbf{a}_2(t) \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{d}}(\mathbf{a}(t)) \end{bmatrix} + \mathbf{K}\mathbf{e}_B(t) \\ \mathbf{0}_{p,1} &= \mathbf{g}(\mathbf{e}_B(t), \mathbf{K}^T\mathbf{a}(t), \mathbf{u}(t)), \qquad\qquad \mathbf{a}(0) = \mathbf{a}_0. \end{aligned}$$

Let $\Psi : \mathbb{R}^N \to \mathcal{V}_f$ be the transformation from the coordinate representation to the related function space $\mathcal{V} = \mathcal{V}_f$ as in Definition 2.25. We suppose $\mathcal{V}_f = \Psi(\mathbb{R}^N)$ fulfills Assumption 2.12.

Note that we employed the equality $\nabla_{\mathbf{a}_2} G(\bar{\mathbf{a}}) = \mathbf{M}_2 \nabla_{\mathbf{M}_2,\mathbf{a}_2} G(\bar{\mathbf{a}})$ in the definition of $\hat{\mathbf{z}}_2(\bar{\mathbf{a}})$ in System 3.1 to get slightly shorter expressions compared to Corollary 2.26.

Remark 3.2. *The initial conditions $\mathbf{e}_B(0)$ are assumed to be prescribed consistent with the input function and its time-derivative at $t = 0$ and the initial state $\mathbf{a}_0$, cf. Part I.A-Lemma 2.25, and need therefore not to be prescribed separately.*

All upcoming methods will be snapshot-based. Consequently, we assume training data, the so-called snapshots $\mathbf{S}$, to be given as

$$\mathbf{S} = \left[\bar{\mathbf{a}}^1, \bar{\mathbf{a}}^2, \dots, \bar{\mathbf{a}}^L\right] \in \mathbb{R}^{N,L}, \qquad \bar{\mathbf{a}}^l = \begin{bmatrix} \bar{\mathbf{a}}_1^l \\ \bar{\mathbf{a}}_2^l \end{bmatrix}, \qquad \text{for } l \in \{1, \dots, L\}. \tag{3.1}$$

The training data typically consists of snapshots of one or several solution trajectories of System 3.1 for appropriate training setups.

3.1 Model order reduction

The model order reduction step by projection closely follows the ideas of Part I.A-Section 3.1. We thus are brief in the exposition here. The compatibility of the full order model, i.e., Assumption 2.12 on $\mathcal{V}_f = \Psi(\mathbb{R}^N)$ can be guaranteed for reduced order models under assumptions on the reduction basis only. In accordance to Part I.A-Assumption 3.5, they read as follows.

Assumption 3.3 (Compatibility of reduction basis)**.** *For System 3.1 given, let the reduction basis $\mathbf{V}_b$ have block structure*

$$\mathbf{V}_b = \begin{bmatrix} \mathbf{V}_1 & \\ & \mathbf{V}_2 \end{bmatrix}, \qquad \mathbf{V}_1 \in \mathbb{R}^{N_1,n_1}, \mathbf{V}_2 \in \mathbb{R}^{N_2,n_2}, \quad n = n_1 + n_2.$$

Then $\mathbf{V}_b$ is assumed to fulfill

A1) $im(\mathbf{M}_1\mathbf{V}_1) = im(\mathbf{J}\mathbf{V}_2)$

A2) $ker(\mathbf{J}) \subset im(\mathbf{V}_2)$.

Note that one compatibility condition less is needed as in Part I.A, as the proof for mass-conservation is slightly different, Remark 2.15. The ROM (without complexity reduction) is constructed by Galerkin projection.

System 3.4 (Order-reduced model)**.** *Given the full order model in System 3.1 and a reduction basis $\mathbf{V}_b$ fulfilling Assumption 3.3, the reduced order model is defined as follows: Find $\mathbf{a}_r = [\mathbf{a}_{r,1}; \mathbf{a}_{r,2}] \in \mathcal{C}^1([0,T]; \mathbb{R}^n)$ and $\mathbf{e}_B \in \mathcal{C}([0,T]; \mathbb{R}^{|\mathcal{N}_\partial|})$, solving*

$$\frac{d}{dt}\begin{bmatrix} \mathbf{M}_{r,1}\mathbf{a}_{r,1}(t) \\ \hat{\mathbf{z}}_{r,2}(\mathbf{a}_r(t)) \end{bmatrix} = \begin{bmatrix} & \mathbf{J}_r \\ -\mathbf{J}_r^T & \end{bmatrix} \begin{bmatrix} \hat{\mathbf{e}}_{r,1}(\mathbf{a}_r(t)) \\ \mathbf{a}_{r,2}(t) \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{d}}_r(\mathbf{a}_r(t)) \end{bmatrix} + \mathbf{K}_r\mathbf{e}_B(t)$$

$$\mathbf{0}_{p,1} = \mathbf{g}(\mathbf{e}_B(t), \mathbf{K}_r^T\mathbf{a}_r(t), \mathbf{u}(t)), \qquad \mathbf{a}_r(0) = \mathbf{M}_r^{-1}\mathbf{V}_b^T\mathbf{M}\mathbf{a}_0,$$

where $\mathbf{M}$ is as in Corollary 2.26, $\mathbf{M}_r = \mathbf{V}_b^T\mathbf{M}\mathbf{V}_b$ and

$$\begin{aligned} \mathbf{M}_{r,1} &= \mathbf{V}_1^T\mathbf{M}_1\mathbf{V}_1, & \mathbf{J}_r &= \mathbf{V}_1^T\mathbf{J}\mathbf{V}_2, & \hat{\mathbf{d}}_r(\mathbf{a}_r) &= \mathbf{V}_2^T\hat{\mathbf{d}}(\mathbf{V}_b\mathbf{a}_r) \\ \hat{\mathbf{e}}_{r,1}(\mathbf{a}_r) &= \mathbf{M}_{r,1}^{-1}\mathbf{V}_1^T\mathbf{M}_1\hat{\mathbf{e}}_1(\mathbf{V}_b\mathbf{a}_r), & \hat{\mathbf{z}}_{r,2}(\mathbf{a}_r) &= \mathbf{V}_2^T\hat{\mathbf{z}}_2(\mathbf{V}_b\mathbf{a}_r), & \mathbf{K}_r &= \mathbf{V}_b^T\mathbf{K}. \end{aligned}$$

Remark 3.5. *Both, the initial condition for $\mathbf{a}_r(0)$ and the term $\hat{\mathbf{e}}_{r,1}(\mathbf{a}_r)$ can be interpreted as orthogonal projections of their non-reduced counterparts w.r.t. the energy scalar products.*

To construct a reduction basis $\mathbf{V}_b$ fulfilling Assumption 3.3, we follow here Part I.A-Section 3.1, *Strategy 1.* Including the small adaptions needed for the slightly modified compatibility conditions in this part compared to Part-I.A, this leads us to Algorithm 3.6. The call 'ortho' in the algorithm refers to Part I.A-Algorithm 3.12, and the system matrices $\mathbf{M}_2$ and $\mathbf{K}_2$ are defined as in Corollary 2.26.

Algorithm 3.6 (Compatible basis for System 3.1)**.**

```
function [V1,V2]=compBasis(W1,M1,M2,J,K2,KernelJ,tolerance)
      V1 = ortho(W1,[],M1,tolerance);
      S = [J',K2];
      V2 = M2\(S*((S'*(M2\S))\[M1*V1;zeros(size(K2,2),size(V1,2))]));
      V2 = ortho([KernelJ,V2],[],M2,tolerance);
end
```

The high-fidelity basis $\mathbf{W}_1$ that our Algorithm 3.6 takes as input, is calculated from the snapshots $\mathbf{S}$ of (3.1). More precisely, we collect $\bar{\mathbf{a}}_1(t_l)$ and $\mathbf{M}_1^{-1}\mathbf{J}\bar{\mathbf{a}}_2(t_l)$ in one new snapshot matrix and apply proper orthogonal decomposition (POD) onto it. This is summarized in Algorithm 3.7. The call 'Podscp' in the algorithm refers to POD in a desired scalar product, see, e.g. [GH18], [KV01].

Algorithm 3.7 (POD basis W1)**.**

```
function [W1]=PodW1(S,M1,J,dimr)
  [N1,N2] = size(J);
  S1 = S(1:N1,:); S2 = S(N1+(1:N2),:);
  % construct POD-basis of dimension 'dimr' using scalar product induced by 'M1'
  W1 = Podscp([S1,M1\(J*S2)],M1,dimr);
end
```

Remark 3.8. *Given $\mathbf{a}(t) = [\mathbf{a}_1(t); \mathbf{a}_2(t)]$ is a snapshot of a solution to System 3.1 at time t, it holds*

$$\mathbf{M}_1^{-1}\mathbf{J}\mathbf{a}_2(t) = \partial_t\mathbf{a}_1(t).$$

Thus, our strategy for the snapshot collection can also be interpreted as collecting $\bar{\mathbf{a}}_1(t_l)$ and the time derivatives $\partial_t\bar{\mathbf{a}}_1(t_l)$ for $l = 1, \dots, L$ in a snapshot matrix and applying POD. Including time derivatives into the snapshots has been proposed, e.g., in [IW14], [KV01], also for problems without any special additional structure. The saddle-point-type structure and our aim to ensure Assumption 3.3-(A1) further motivates this for our problem.

3.2 Concept of complexity reduction

Although the reduced order model is typically of much lower dimension $n \ll N$ than the full order model, evaluations of nonlinearities may still scale with N. This is because there is in general no better way to evaluate them than performing a prolongation step $\bar{\mathbf{a}} = \mathbf{V}\mathbf{a}_r$ every time. To avoid this bottleneck, nonlinear terms are additionally approximated by so-called complexity reduction. Applying this approach to System 3.4, leads us to the complexity-reduced model.

System 3.9 (Complexity-reduced model)**.** *Under the definitions of System 3.4 and additional complexity-reduced approximations $\hat{\mathbf{z}}_{c,2}, \hat{\mathbf{d}}_c : \mathbb{R}^n \to \mathbb{R}^{n_2}$ and $\hat{\mathbf{e}}_{c,1} : \mathbb{R}^n \to \mathbb{R}^{n_1}$ of $\hat{\mathbf{z}}_{r,2}(\cdot)$, $\hat{\mathbf{d}}_r(\cdot)$ and $\hat{\mathbf{e}}_{r,1}(\cdot)$, the complexity-reduced model is defined as follows:*

Find $\mathbf{a}_r = [\mathbf{a}_{r,1}; \mathbf{a}_{r,2}] \in \mathcal{C}^1([0,T]; \mathbb{R}^n)$ *and* $\mathbf{e}_B \in \mathcal{C}([0,T]; \mathbb{R}^{|\mathcal{N}_\partial|})$ *solving*

$$\frac{d}{dt}\begin{bmatrix} \mathbf{M}_{r,1}\mathbf{a}_{r,1}(t) \\ \hat{\mathbf{z}}_{c,2}(\mathbf{a}_r(t)) \end{bmatrix} = \begin{bmatrix} & \mathbf{J}_r \\ -\mathbf{J}_r^T & \end{bmatrix} \begin{bmatrix} \hat{\mathbf{e}}_{c,1}(\mathbf{a}_r(t)) \\ \mathbf{a}_{r,2}(t) \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{d}}_c(\mathbf{a}_r(t)) \end{bmatrix} + \mathbf{K}_r \mathbf{e}_B(t)$$

$$\mathbf{0}_{p,1} = \mathbf{g}(\mathbf{e}_B(t), \mathbf{K}_r^T \mathbf{a}_r(t), \mathbf{u}(t)), \qquad \mathbf{a}_r(0) = \mathbf{M}_r^{-1}\mathbf{V}_b^T \mathbf{M}\mathbf{a}_0.$$

The rest of this section is concerned with the construction of the complexity-reduced function-approximations. For the exposition, we generically write $\mathbf{f} : \mathbb{R}^N \to \mathbb{R}^M$ for the nonlinearity of the full order model. Its counterpart in the order-reduced model reads

$$\mathbf{f}_r : \mathbb{R}^n \to \mathbb{R}^m, \qquad \mathbf{f}_r : \mathbf{a}_r \mapsto \hat{\mathbf{V}}^T \mathbf{f}(\mathbf{V}_b \mathbf{a}_r), \qquad \hat{\mathbf{V}} \in \mathbb{R}^{M,m},\ m \ll M.$$

Remark 3.10. *In our case it is* $M \in \{N_1, N_2, N\}$, $\hat{\mathbf{V}} \in \{\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_b\}$ *and, e.g.,* $\mathbf{f} = \hat{\mathbf{e}}_1$.

The complexity reduction now aims for a cheaper-to-evaluate $\mathbf{f}_c : \mathbb{R}^n \to \mathbb{R}^m$ with

$$\mathbf{f}_c(\bar{\mathbf{a}}_r^l) \approx \mathbf{f}_r(\bar{\mathbf{a}}_r^l), \qquad \text{for } \bar{\mathbf{a}}_r^l = \mathbf{V}_b^+ \bar{\mathbf{a}}^l, \quad l = 1, \ldots, L, \tag{3.2}$$

with $\bar{\mathbf{a}}^l$ the snapshots as in (3.1) and $\mathbf{V}_b^+$ the pseudo-inverse w.r.t. the given scalar product. The main ingredients in the construction of $\mathbf{f}_c$ are the formulation of an approximation ansatz and its solution by a snapshot-based heuristic. Similar to the model order reduction step, the implementation of compatibility conditions have to be explicitly implemented in the complexity reduction. The relevant condition here is Assumption 2.20.

3.3 Quadrature-type complexity reduction

Our complexity reduction framework from Section 2.3 employs one sparse quadrature-rule to approximate every nonlinear integral expression. Recall that the full order model infers a partitioning of the spatial domain Ω by the finite elements $K_j \subset \Omega$, $j = 1, \ldots, J$. Following Definition 2.19, the sparse quadrature-rule approximation $\langle \cdot, \cdot \rangle_c$, is defined by an index set I and quadrature weights w_i as

$$\langle \bar{b}, b \rangle_c = \sum_{i \in I} w_i \int_{K_i} \bar{b}[x] b[x] dx, \qquad \bar{b}, b \in \mathcal{L}^2(\mathcal{E}).$$

It is used to set up the complexity-reduced functions $\hat{\mathbf{z}}_{c,2}(\cdot), \hat{\mathbf{d}}_c(\cdot)$ and $\hat{\mathbf{e}}_{c,1}(\cdot)$ of System 3.9 as described in Corollary 2.30. A gain in efficiency is achieved, when only a few non-zero weights w_i are employed, i.e., $|I| \ll |\{1, \ldots, J\}|$. The construction of those weights is the topic of this section. For convenience, we gather them in a vector with one entry for each finite element,

$$\mathbf{w}^{qu} \in \mathbb{R}^J, \qquad [\mathbf{w}^{qu}]_i = \begin{cases} w_i, & i \in I \\ 0, & \text{else.} \end{cases}$$

The starting point is the representation of the $\mathcal{L}^2$-scalar product $\langle \cdot, \cdot \rangle$ as a sum of contributions from each finite element K_j,

$$\langle \bar{b}, b \rangle = \sum_{j=1}^{J} w_j^f \int_{K_j} \bar{b}[x] b[x] dx, \qquad \text{for } \bar{b}, b \in \mathcal{L}^2(\mathcal{E}). \tag{3.3}$$

The coefficients $w_j^f \in \mathbb{R}$ are nothing but the edge-weights from Section 2.6. For a network without edge-weighting it is $w_j^f = 1$, $j = 1, \dots, J$. For fixed functions $\bar{b}, b \in \mathcal{L}^2(\mathcal{E})$, (3.3) can be interpreted as a linear equality fulfilled by the weight-vector $\mathbf{w}^f = [w_1^f; \dots; w_J^f] \in \mathbb{R}^J$, i.e.,

$$\left[\int_{K_1} \bar{b}[x]b[x]dx \quad \int_{K_2} \bar{b}[x]b[x]dx \quad \dots \quad \int_{K_J} \bar{b}[x]b[x]dx\right] \mathbf{w}^f = \langle \bar{b}, b \rangle.$$

Similarly, one can reformulate each evaluation of the nonlinearity $\mathbf{f}_r$ at a snapshot of (3.1), i.e., $\mathbf{f}_r(\mathbf{V}^+\bar{\mathbf{a}}(t_l)) \in \mathbb{R}^{n_i}$ for $l \in \{1, \dots, L\}$, as m linear equalities fulfilled by $\mathbf{w}^f$. Collecting all of them in one large system, this then takes the form

$$\mathbf{A}\mathbf{w}^f = \mathbf{b}, \qquad \text{with } \mathbf{A} \in \mathbb{R}^{Lm,J},\ \mathbf{b} \in \mathbb{R}^{Lm} \tag{3.4}$$
$$\mathbf{A}, \mathbf{b} \text{ determined by } \mathbf{f}_r \text{ and } \mathbf{S} = \left[\bar{\mathbf{a}}^1, \bar{\mathbf{a}}^2, \dots, \bar{\mathbf{a}}^L\right] \text{ from (3.1).}$$

We refer to, e.g., [HCF17], [Jam08], [FACC14] for the detailed form of the entries in $\mathbf{A}$ and $\mathbf{b}$. The approximation ansatz for the quadrature-based approach then aims for a sparse vector of weights $\mathbf{w}^{qu} \in \mathbb{R}^J$ such that

$$\mathbf{A}\mathbf{w}^{qu} \approx \mathbf{b}, \qquad \text{with } \mathbf{A}, \mathbf{b} \text{ as in (3.4).}$$

The fundamental underlying heuristic we apply to find appropriate weights $\mathbf{w}^{qu}$ is the greedy search from Algorithm 3.12, similar to the ones used in [HCF17], [Jam08]. In contrast to the references, we additionally allow for initially fixed quadrature-points by the input $\mathbf{w}_0^{qu}$. These initial quadrature points are chosen towards the fulfillment of Assumption 2.20 independent of the snapshots, as explained below.

Remark 3.11. *The minimization problem in Algorithm 3.12-Step 2.d is an unconstrained least-squares problem, which can be easily solved exactly.*

Algorithm 3.12 (Greedy empirical quadrature weights)**.**

```
INPUT:
  • Data matrix $\mathbf{A}$ and vector $\mathbf{b}$ as in (3.4)
  • Initial sparse-quadrature vector $\mathbf{w}_0^{qu}$
  • Number of quadrature-points $n_c$
OUTPUT: Vector $\mathbf{w}^{qu}$ of quadrature-weights
  1. Set $I_0 = \{i \in \{1, \dots, J\} : [\mathbf{w}_0^{qu}]_i \neq 0\}$.
  2. for $k = 1 : n_c$
       a) Define set of candidates $I_c = \{1, \dots, J\} \setminus I_{k-1}$.
       b) Find $j_{max}$ solving: $\max_{j \in I_c} \left| \left[\mathbf{A}^T\left(\mathbf{A}\mathbf{w}^{k-1} - \mathbf{b}\right)\right]_j \right|$.
       c) Set $I_k = I_{k-1} \cup \{j_{max}\}$.
       d) Find $\mathbf{w}^k$ solving:
            $\min_{\substack{\mathbf{w}^k \in \mathbb{R}^J \\ [\mathbf{w}^k]_j = 0,\ for\ j \notin I_k}} \left|\left|\mathbf{A}\mathbf{w}^k - \mathbf{b}\right|\right|$
     endfor
  3. Set $\mathbf{w}^{qu} = \mathbf{w}^{n_c}$.
```

Towards compatibility

Regarding our analysis, we should aim for a quadrature-weight such that Assumption 2.20 is fulfilled, i.e.,

$$w_i \geq 0, \qquad\qquad i = 1, \dots, J \tag{3.5a}$$

$$\frac{1}{\tilde{C}}||b||_c \leq ||b|| \leq \tilde{C}||b||_c, \qquad\qquad b \in \mathcal{V}_i,\ i \in \{1, 2\}. \tag{3.5b}$$

Condition (3.5b) needs to be translated from function-space setting to an algebraic setting. For that, let bases $b_i^1, \dots, b_i^{n_i}$ of $\mathcal{V}_i$ for $i = 1, 2$ be given as in Definition 2.25. Then define mass matrices $\mathbf{M}_i, \mathbf{M}_{c,i} \in \mathbb{R}^{n_i, n_i}$ by

$$[\mathbf{M}_i]_{jl} = \langle b_i^l, b_i^j \rangle, \quad [\mathbf{M}_{c,i}]_{jl} = \langle b_i^j, b_i^l \rangle_c, \qquad j, l = 1, \dots, n_i. \tag{3.6}$$

Any $v \in \mathcal{V}_i$ can be written in terms of its coordinate representation $\mathbf{v} = [\mathfrak{v}^1; \dots; \mathfrak{v}^{n_i}] \in \mathbb{R}^{n_i}$ as $v = \sum_{l=1}^{N_i} \mathfrak{v}^l b_i^l$. It then holds

$$||v||_c^2 = \langle v, v \rangle_c = \sum_{j,l=1}^{n_i} \langle \mathfrak{v}^l b_i^l, \mathfrak{v}^j b_i^j \rangle_c = \sum_{j,l=1}^{n_i} \langle b_i^l, b_i^j \rangle_c \, \mathfrak{v}^l \mathfrak{v}^j = \mathbf{v}^T \mathbf{M}_{c,i} \mathbf{v}.$$

Similarly, the norm $||v||$ can be expressed in terms of $\mathbf{M}_i$. Condition (3.5b) thus holds for a bounded $\tilde{C}$, when $\mathbf{M}_{c,i}$ is non-singular for $i = 1, 2$. Moreover, $\tilde{C}$ can be seen as a stability-constant, and it is $\tilde{C} \approx 1$ for $\mathbf{M}_{c,i} \approx \mathbf{M}_i$. This clarifies our aim. Its strict fulfillment in a practical implementation is, however, a challenging task.

Remark 3.13. *The greedy procedures in [HCF17], [Jam08] enforce the positivity-constraint* (3.5a) *in contrast to our Algorithm 3.12. We omitted the constraint, as we only observed very minor violation in all our numerical experiments and the constraint* (3.5b)*, which showed to be more relevant in our tests, is not enforcible in such a straight forward manner in a greedy procedure.*

We therefore take a more pragmatic approach and include in a pre-selection step a few quadrature-points towards the fulfillment of the part of (3.5), which is the most troubled in practice. We observe that to be the following condition:

$$||b|| \leq \tilde{C}||b||_c, \qquad \text{for } b \in \mathcal{K}, \quad \mathcal{K} = \{w \in \mathcal{V}_2 : \partial_x w \equiv 0\}. \tag{3.7}$$

As we incorporate $\mathcal{K}$ in our reduced space $\mathcal{V}_{r,2}$ independently of the snapshots at hand, cf. Assumption 2.12, this is not surprising. With the same reasoning as above, a related algebraic characterization of (3.7) can be given in terms of a basis $w^1, \dots, w^k$ of $\mathcal{K}$ by the mass matrices $\tilde{\mathbf{M}}, \tilde{\mathbf{M}}_c \in \mathbb{R}^{k,k}$ defined by

$$\left[\tilde{\mathbf{M}}\right]_{ij} = \langle w^j, w^i \rangle, \quad \left[\tilde{\mathbf{M}}_c\right]_{ij} = \langle w^j, w^i \rangle_c, \qquad i, j = 1, \dots, k.$$

Our implementation of the pre-selection step in Algorithm 3.14 aims towards initial quadrature-weights $\tilde{\mathbf{w}}_0^{qu}$, for which $\tilde{\mathbf{M}}_c \approx \tilde{\mathbf{M}}$. Note that within the procedure it is employed that the entries of $\tilde{\mathbf{M}}$ can be expressed in terms of a basis $\mathbf{w}_1, \dots, \mathbf{w}_k$ of $ker(\mathbf{J}_r)$ and the function

$$\mathbf{f}_r : \mathbb{R}^{n_2} \to \mathbb{R}^k, \qquad \mathbf{f}_r : \mathbf{a}_{r,2} \mapsto \mathbf{W}^T \mathbf{a}_{r,2}, \quad \text{with } \mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_k].$$

The weights $\tilde{\mathbf{w}}_0^{qu}$ obtained from the pre-selection are taken as initialization in our greedy search. By this, respective quadrature points are forced to be included, independently of the snapshots.

Algorithm 3.14 (Pre-selection of weights for compatibility)**.**
INPUT: Matrix $\mathbf{J}_r$ *from System 3.4*
OUTPUT: Vector $\tilde{\mathbf{w}}_0^{qu}$ *of quadrature-weights*

1. *Form orthogonal basis* $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_k]$ *of* $ker(\mathbf{J}_r)$.

2. *Define* $\mathbf{f}_r : \mathbb{R}^{n_2} \to \mathbb{R}^k$, $\mathbf{f}_r : \mathbf{a}_{r,2} \mapsto \mathbf{W}^T \mathbf{a}_{r,2}$.

3. *Construct* $\mathbf{A}$ *and* $\mathbf{b}$ *according to* (3.4) *for* $\bar{\mathbf{a}}_r^l = \mathbf{w}_l$ *for* $l = 1, \dots, k$ *and* $\mathbf{f}_r$.

4. *Find* $\tilde{\mathbf{w}}_0^{qu}$ *by Algorithm 3.12 for* $\mathbf{A}$, $\mathbf{b}$ *from Step 3, and* $\mathbf{w}_0^{qu} = \mathbf{0}$, $n_c = k$.

Adaption to our model

For the greedy training phase in the construction of the quadrature-weights, we suggest to apply Algorithm 3.12 onto the snapshots (3.2) and the nonlinearity composed of $\hat{\mathbf{d}}_r(\cdot)$ and $\mathbf{a}_r \mapsto \mathbf{W}_J \mathbf{J}_r \hat{\mathbf{e}}_{r,1}(\mathbf{a}_r)$. The matrix $\mathbf{W}_J$ is constructed to be a basis matrix for $im(\mathbf{J}_r)$. Its inclusion only removes redundant data and enhances the performance of the training phase. Our full procedure consists of the pre-selection step and the subsequent greedy search. Its implementation is summarized in Algorithm 3.15.

Algorithm 3.15 (Training of sparse quadrature weights)**.**
INPUT:

- *Projected snapshots* $\bar{\mathbf{a}}_r^l$, $l = 1, \dots, L$ *from* (3.2)

- *Order-reduced model, System 3.4*

- *Number nonzero entries* n_c

OUTPUT: Vector $\mathbf{w}^{qu}$ *of quadrature-weights*

1. *Construct pre-selected quadrature-weight vector* $\mathbf{w}_0^{qu}$ *from Algorithm 3.14.*

2. *Construct orthogonal matrix* $\mathbf{W}_J$ *with* $im(\mathbf{W}_J) = im(\mathbf{J}_r)$.

3. *Define* $\mathbf{f}_r : \mathbf{a}_r \mapsto [\hat{\mathbf{d}}_r(\mathbf{a}_r); \mathbf{W}_J \mathbf{J}_r \hat{\mathbf{e}}_{r,1}(\mathbf{a}_r)]$.

4. *Construct* $\mathbf{A}$, $\mathbf{b}$ *depicted in* (3.4) *for* $\mathbf{f}_r$ *and snapshots* $\bar{\mathbf{a}}_r^l$, $l = 1, \dots, L$.

5. *Find* $\mathbf{w}^{qu}$ *by Algorithm 3.12 with* $\mathbf{A}$, $\mathbf{b}$, $\mathbf{w}_0^{qu}$ *and* n_c *as above.*

Remark 3.16. *While the pre-selection of* $\mathbf{w}_0^{qu}$ *in Algorithm 3.15 strongly improves the performance, cf. Section 5.5, it does not rigorously guarantee Assumption 2.20 to hold. The design of a more sophisticated heuristic strictly enforcing compatibility is a direction for future research.*

3.4 Complexity reduction by DEIM

Another popular complexity reduction method is the discrete empirical interpolation method (DEIM). This non-structure-preserving approach is instanced here for later numerical comparisons against our structure-preserving method from Section 3.3.

Let us explain the basic idea of DEIM with the example of the generic nonlinearity $\mathbf{f}_r$ from (3.2),

$$\mathbf{f}_r : \mathbb{R}^n \to \mathbb{R}^m, \qquad \mathbf{f}_r : \mathbf{a}_r \mapsto \hat{\mathbf{V}}^T \mathbf{f}(\mathbf{V}_b \mathbf{a}_r), \qquad \hat{\mathbf{V}} \in \mathbb{R}^{M,m},\ m \ll M.$$

Let $\mathbf{f}(\cdot)$ be the related full order model nonlinearity. The ansatz of DEIM aims for an interpolation of $\mathbf{f}(\cdot)$ by an appropriate sparse approximation. This is realized by setting up a quadratic matrix $\mathbf{D}_f$ with only a few non-zero columns, i.e., a small image, such that

$$\mathbf{f}(\bar{\mathbf{a}}^l) \approx \mathbf{D}_f \mathbf{f}(\bar{\mathbf{a}}^l)$$

for the trained snapshots $\bar{\mathbf{a}}^l$, (3.2). The DEIM-complexity-reduced function is then defined as

$$\mathbf{f}_c : \mathbf{a}_r \mapsto \hat{\mathbf{V}}^T \mathbf{D}_f \mathbf{f}(\mathbf{V}_b \mathbf{a}_r).$$

A gain in efficiency is achieved, when the sparsity of $\mathbf{D}_f$ is employed in the implementation. The construction of $\mathbf{D}_f$ follows a greedy procedure. We refer to [CS12], [CS10], [DS18], [PDG18] for details. Note that the references suggest slightly varying greedy methods, but the used approximation ansatz is the same for all of them.

Adaption to our model

In our implementation we employ the DEIM-procedure depicted in [PDG18, Algorithm 2] (without oversampling) to each of the nonlinearities $\hat{\mathbf{z}}_{r,2}(\cdot)$, $\hat{\mathbf{d}}_r(\cdot)$ and $\hat{\mathbf{e}}_{r,1}(\cdot)$ of System 3.4 separately. That is, sparse quadratic matrices $\mathbf{D}_z$, $\mathbf{D}_d$ and $\mathbf{D}_e$ are constructed in the training phase to set up the complexity-reduced nonlinearities for System 3.9 as

$$\begin{aligned} &\hat{\mathbf{z}}_{c,2} : \mathbf{a}_r \mapsto \mathbf{V}_2^T \mathbf{D}_z \hat{\mathbf{z}}_2(\mathbf{V}_b \mathbf{a}_r), &&\hat{\mathbf{d}}_c : \mathbf{a}_r \mapsto \mathbf{V}_2^T \mathbf{D}_d \hat{\mathbf{d}}(\mathbf{V}_b \mathbf{a}_r) \\ &\hat{\mathbf{e}}_{c,1} : \mathbf{a}_r \mapsto \mathbf{M}_{r,1}^{-1} \mathbf{V}_1^T \mathbf{M}_1 \mathbf{D}_e \hat{\mathbf{e}}_1(\mathbf{V}_b \mathbf{a}_r). \end{aligned}$$

Chapter 4

Application to Euler equations

Our approximation procedure is in particular applicable for the Euler equations with barotropic pressure models and for its simplifications sketched in Fig. 4.1. Besides the isothermal model, we are mainly interested in, e.g., the isentropic model is another instance for a barotropic pressure model. We refer to the introduction of Part I and [Che05], [BGH11], [Egg18] for details and alternative approximation methods. In this chapter we focus on the first two stages of the model hierarchy of Fig. 4.1, the barotropic Euler equations and what we refer to as the simplified barotropic Euler equations. The respective models for the network case are stated together with their approximation in our framework. We also derive a well-posedness result for the complexity-reduced system at the end of the chapter.

As our main motivation is the simulation of gas networks, we present the respective model assumptions here, cf. [SAB+17], [HGCATRM09], [BGH11], [DHLT17]. The typical choice for the pressure model in this application is the isothermal model. Further, edges $\omega \in \mathcal{E}$ represent (cross-sectionally averaged) pipes. The cross-sectional areas $\mathcal{A} = \{A^\omega,\, \omega \in \mathcal{E}\}$ are constant on a pipe but may be different for each one of them. They act as weights for the edges, see Section 2.6, and the

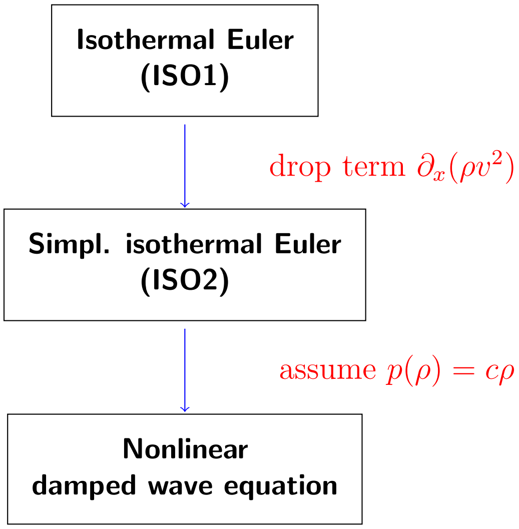

Figure 4.1: Model hierarchy for (isothermal) Euler-type equations. Equations of (ISO1) and (ISO2) are given by (4.1) and (4.2), respectively.

network description is thus given by a directed graph $(\mathcal{N}, \mathcal{E}, l, \mathcal{A})$. The pipes are assumed to be perfectly round. For convenience, we also introduce functions $A, D : \Omega \to \mathbb{R}$ for the cross-sectional area and the diameter. It then holds

$$A_{|\omega} = A^{\omega} \qquad \text{and} \qquad A_{|\omega} = \frac{\pi}{4} D_{|\omega}^2, \qquad \text{for } \omega \in \mathcal{E}.$$

In the upcoming, ρ and v denote density and velocity, respectively. The pressure p is assumed to be a function of density only, i.e., $p = p(\rho)$, which exactly is the characterization of a barotropic model. Further, λ describes a friction factor, which may be state-dependent.

4.1 Barotropic Euler equations

In the *full* barotropic Euler equations, which we also refer to as the ISO1 model for the isothermal case, the density and and velocity $\rho, v : [0, T] \times \Omega \to \mathbb{R}$ are governed by

$$\partial_t \rho + \partial_x (\rho v) = 0 \tag{4.1a}$$

$$\partial_t(\rho v) + \partial_x \left(\rho v^2 + p(\rho)\right) = -\frac{\lambda}{2D} \rho |v| v \tag{4.1b}$$

for $x \in \Omega$, $t \in [0, T]$. Given a potential energy function $P(\cdot)$ fulfilling

$$P''(\rho) = \frac{1}{\rho} p'(\rho),$$

the coupling conditions at inner nodes $\nu \in \mathcal{N}_0$ read

$$\sum_{\omega \in \mathcal{E}(\nu)} n^{\omega}[\nu] A^{\omega} \rho_{|\omega}[\nu] v_{|\omega}[\nu] = 0 \tag{4.1c}$$

$$P'(\rho_{|\omega})[\nu] + \frac{v_{|\omega}[\nu]^2}{2} = P'(\rho_{|\tilde{\omega}})[\nu] + \frac{v_{|\tilde{\omega}}[\nu]^2}{2} \qquad \text{for } \omega, \tilde{\omega} \in \mathcal{E}(\nu). \tag{4.1d}$$

The system has to be complemented with initial conditions and one boundary condition per boundary node. For the isothermal Euler equations, it holds in particular

$$\textit{ISO1:} \qquad p(\rho) = RT \frac{\rho}{1 - RT\alpha\rho}, \qquad P(\rho) = RT\rho \log\left(\rho_{sc} \frac{1 - RT\alpha\rho}{\rho}\right) \tag{4.1e}$$

for $R > 0$ the specific gas constant, T the temperature, a model parameter $\alpha < 0$ and a factor from the non-dimensionalization ρ_{sc}. For the latter, it holds $\rho_{sc} = 1$, when SI-units are used. Our abstract standard form (2.1) specializes to

$$\partial_t \begin{bmatrix} \rho \\ v \end{bmatrix} = \begin{bmatrix} & -\partial_x \\ -\partial_x & -r(\rho, \rho v) \end{bmatrix} \nabla h([\rho; v]), \qquad x \in \Omega, \quad t \in [0, T]$$

$$\text{with } h([\rho; v]) = \rho \frac{v^2}{2} + P(\rho), \quad r(\rho, \rho v) = \frac{\lambda}{2D} \frac{|v|}{\rho}, \qquad \underline{\mathbf{z}} = [\rho; v].$$

With mass flow m given as $m = \rho v$ it then holds

$$\nabla_{\mathbf{z}} h([\rho; v]) = \nabla_{\mathbf{z}} h\left(\left[\rho; \frac{m}{\rho}\right]\right) = \begin{bmatrix} \frac{m^2}{2\rho^2} + P'(\rho) \\ m \end{bmatrix}.$$

The complexity-reduced approximation, System 2.22, for this case thus reads as follows.

System 4.1. *Find* $[\rho; m] \in \mathcal{C}^1([0,T]; \mathcal{V}_1 \times \mathcal{V}_2)$*,* $\mathbf{f}_B \in \mathcal{C}^1([0,T]; \mathbb{R}^{|\mathcal{N}_\partial|})$ *and* $\mathbf{e}_B \in C^0([0,T]; \mathbb{R}^{|\mathcal{N}_\partial|})$ *such that it holds*

$$\begin{aligned}\langle \partial_t \rho, b_1 \rangle &= -\langle \partial_x m, b_1 \rangle \\ \left\langle \partial_t \frac{m}{\rho}, b_2 \right\rangle_c &= \left\langle P'(\rho) + \frac{m^2}{2\rho^2}, \partial_x b_2 \right\rangle_c + \mathbf{e}_B \cdot \mathcal{T} b_2 - \langle r(\rho, m) m, b_2 \rangle_c \\ \mathbf{f}_B &= \mathcal{T} m\end{aligned}$$

for all $b_1 \in \mathcal{V}_1$ *and* $b_2 \in \mathcal{V}_2$*, and* $[\rho(0); m(0)] = [\rho_0; m_0]$*. One boundary condition for each boundary node* $\nu \in \mathcal{N}_\partial$ *is prescribed, each of one of the following types for* $t \in [0,T]$*,*

$$\textit{Type 1: } e_B[\nu] = u^\nu \in \mathcal{C}([0,\infty), \mathbb{R}), \qquad \textit{Type 2: } f_B[\nu] = u^\nu \in \mathcal{C}^1([0,\infty), \mathbb{R})$$
$$\textit{and at least one } \nu \in \mathcal{V}_\partial \textit{ with boundary conditions of Type 1.}$$

The Hamiltonian of System 4.1 is then given for $\underline{\mathbf{a}} = [\rho; m]$ as

$$\mathcal{H}_c : \mathcal{V} \to \mathbb{R}, \qquad \mathcal{H}_c\left(\left[\rho; \frac{m}{\rho}\right]\right) = \left\langle P(\rho) + \frac{m^2}{2\rho}, 1 \right\rangle_c .$$

The specific choice of boundary conditions is chosen for the upcoming theoretical investigations in Section 4.3. For the application, more general boundary conditions can be easily included, cf. System 3.1.

4.2 Simplified barotropic Euler equations

In the *simplified* barotropic Euler equations, which we also refer to as the ISO2 model for the isothermal case, the density and and mass flow $\rho, m : [0,T] \times \Omega \to \mathbb{R}$ are governed by

$$\partial_t \rho + \partial_x m = 0 \tag{4.2a}$$

$$\partial_t m + \partial_x p(\rho) = -\frac{\lambda}{2D} \frac{1}{\rho} |m| m \tag{4.2b}$$

for $x \in \Omega$, $t \in [0,T]$. The coupling conditions at inner nodes $\nu \in \mathcal{N}_0$ then read

$$\sum_{\omega \in \mathcal{E}(\nu)} n^\omega[\nu] A^\omega m_{|\omega}[\nu] = 0 \tag{4.2c}$$

$$p(\rho_{|\omega}[\nu]) = p(\rho_{|\tilde{\omega}}[\nu]) \qquad \text{for } \omega, \tilde{\omega} \in \mathcal{E}(\nu). \tag{4.2d}$$

The system has to be complemented with initial conditions and one boundary condition per boundary node. Again, we can construct a Hamiltonian. For this, we assume a potential energy function $\tilde{P}(\cdot)$ to be given with

$$\tilde{P}''(\rho) = p'(\rho).$$

For the isothermal Euler equations, it in particular holds

$$\textit{ISO2:}\qquad p(\rho) = RT\frac{\rho}{1-RT\alpha\rho}, \qquad \tilde{P}(\rho) = \frac{-1}{RT\alpha^2}\left(\log\left(1-RT\alpha\rho\right)+RT\alpha\rho\right). \tag{4.2e}$$

Our abstract standard form (2.1) in this case specializes to

$$\partial_t \begin{bmatrix}\rho\\ m\end{bmatrix} = \begin{bmatrix} & -\partial_x\\ -\partial_x & -r(\rho,\rho v)\end{bmatrix}\nabla h([\rho;m]), \qquad x\in\Omega, \quad t\in[0,T]$$
$$\text{with } h([\rho;m]) = \frac{m^2}{2}+\tilde{P}(\rho), \quad r([\rho;m]) = \frac{\lambda}{2D}\frac{|m|}{\rho} \qquad \underline{\mathbf{z}} = [\rho;m].$$

The complexity-reduced approximation, System 2.22, for this case thus reads as follows.

System 4.2. *Find $[\rho;m]\in\mathcal{C}^1([0,T];\mathcal{V}_1\times\mathcal{V}_2)$, $\mathbf{f}_B\in\mathcal{C}^1([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{e}_B\in C^0([0,T];\mathbb{R}^{|\mathcal{N}_\partial|})$ such that it holds*

$$\begin{aligned}\langle\partial_t\rho,b_1\rangle &= -\langle\partial_x m,b_1\rangle\\ \langle\partial_t m,b_2\rangle_c &= \langle p(\rho),\partial_x b_2\rangle_c + \mathbf{e}_B\cdot\mathcal{T}b_2 - \langle r(\rho,m)m,b_2\rangle_c\\ \mathbf{f}_B &= \mathcal{T}m\end{aligned}$$

for all $b_1\in\mathcal{V}_1$ and $b_2\in\mathcal{V}_2$, and $[\rho(0);m(0)] = [\rho_0;m_0]$. One boundary condition for each boundary node $\nu\in\mathcal{N}_\partial$ has to be added.

The Hamiltonian of System 4.1 is then given for $\underline{\mathbf{a}} = [\rho;m]$ as

$$\mathcal{H}_c:\mathcal{V}\to\mathbb{R}, \qquad \mathcal{H}_c([\rho;m]) = \left\langle \tilde{P}(\rho)+\frac{m^2}{2},1\right\rangle_c.$$

Remark 4.3. *The nonlinear damped wave equation, Fig. 4.1, differs from the simplified barotropic Euler equations by the fact that the Hamiltonian density h is a quadratic function in the energy variable, i.e., there is constant $c>0$ such that*

$$h([\rho;m]) = \frac{1}{2}\left(m^2+c\rho^2\right).$$

As elaborated on in [Kug19], [EKLS20], most results from Part I.A can be generalized from the linear damped wave equation to the nonlinear damped wave equation for appropriate friction terms.

4.3 A well-posedness result

In this section we aim for a well-posedness result of System 4.1, which is the complexity-reduced approximation of the full barotropic Euler equations. We base our derivations on the following additional assumptions.

Assumption 4.4.

A1) Assumption 2.12 on $\mathcal{V}$ and Assumption 2.20 on $\langle\cdot,\cdot\rangle_c$ hold.

A2) The function $P : (0, M) \to \mathbb{R}$ for $M > 0$ is two times continuously differentiable. It holds $P''(y) > 0$, and $y \leq \max\{P(y), 1\}$ for $y \in (0, M)$, and $\lim_{y\to M} P(y) = \infty$.

A3) The initial conditions $[\rho_0; m_0] \in \mathcal{V}$ are chosen such that $\rho_0(x) \in (0, M)$ for $x \in K_i$, $i \in I$.

A4) The friction model takes the form $r(\rho, m) = 1/\rho^2$ (laminar friction model for $\lambda(\cdot)$).

A5) The underlying full order model uses the finite element spaces from Part I.A-Section 2.4, i.e., $\mathcal{V}_{f,1} = \mathcal{Q}_0(T_\mathcal{E})$ and $\mathcal{V}_{f,2} = \mathcal{P}_1(T_\mathcal{E})$, and the restriction of the density to one finite element K_i, $\rho(t)_{|K_i}$, is constant.

A6) It holds $A^\omega = 1$ for all $\omega \in \mathcal{E}$.

Assumption $(A1)$ ensures that $|| \cdot ||_c$ and $|| \cdot ||$ are equivalent norms on $\mathcal{V}_1$ and $\mathcal{V}_2$. Assumptions $(A2)$-$(A3)$ are used to show that the complexity-reduced Hamiltonian $\mathcal{H}_c$ can be evaluated and bounds the norm of the solution. Moreover, the underlying Hamiltonian density $h : (0, M) \times \mathbb{R} \to \mathbb{R}$ then fulfills Assumption 2.1. To establish a global existence-result, we follow the lead of [Egg18, Lemma 4.4] and derive uniform boundedness of the solution and strict positivity of the density. The proof of the latter relies on Assumptions $(A4)$-$(A5)$ and is needed to avoid zeros in the denominators of System 4.1. The last assumption is only made for ease of presentation.

Lemma 4.5. *For any solution of System 4.1, it holds that $\rho(t)_{|K_i} > 0$ for $t > 0$ and $i \in I$. Moreover, there exists a constant C, independent of the discretization parameters, such that*

$$\frac{1}{\rho(t)_{|K_i}} \leq \exp\left(C \frac{\sqrt{t}}{\Delta_{x_i}\sqrt{w_i}} \sqrt{R(t)} \right) \frac{1}{\rho(0)_{|K_i}}$$
$$\text{for } R(t) = \mathcal{H}_c\left(\left[\rho(0); \frac{m(0)}{\rho(0)} \right] \right) + \int_0^t \mathbf{e}_B(s) \cdot \mathbf{f}_B(s) ds, \qquad \text{for } i \in I,$$

where Δ_{x_i} is the grid size of the finite element K_i.

Proof. At several instances, we employ that $\rho(t)_{|K_i}$ and $\partial_x m_{|K_i}(t)$ are constant in space. As long as $1/\rho(t)_{|K_i}$ is well-defined, it therefore holds

$$\begin{aligned} \frac{d}{dt}\left(\frac{1}{\rho(t)_{|K_i}} \right) &= \left\langle \partial_t \frac{1}{\rho(t)}, 1 \right\rangle_{K_i} = - \left\langle \partial_t \rho(t), \frac{1}{\rho(t)^2} \right\rangle_{K_i} \\ &= \left\langle \partial_x m(t), \frac{1}{\rho(t)^2} \right\rangle_{K_i} \leq \left(\frac{1}{\rho(t)_{|K_i}} ||\partial_x m(t)||_{K_i,\infty} \right) \frac{1}{\rho(t)_{|K_i}}. \end{aligned}$$

By the inverse estimate, there exists a constant C with $||\partial_x m(t)||_{K_i} \leq C/\Delta_{x_i}||m(t)||_{K_i}$. Together with $1/\rho(t)$ and $\partial_x m(t)$ both being constant on K_i, this yields

$$\frac{1}{\rho(t)_{|K_i}} ||\partial_x m(t)||_{K_i,\infty} = \frac{1}{\rho(t)_{|K_i}} ||\partial_x m(t)||_{K_i} \leq c_i(t),$$
$$\text{with } c_i(s) = \frac{C}{\Delta_{x_i}} \left|\left| \frac{m(s)}{\rho(s)} \right|\right|_{K_i}.$$

Setting together the two estimates, we thus get by the Gronwall lemma, see [Chi06],

$$\frac{1}{\rho(t)_{|K_i}} \leq \exp\left(\int_0^t c_i(s)ds\right) \frac{1}{\rho(0)_{|K_i}}.$$

It remains to bound $\int_0^t c_i(s)ds$ for $i \in I$. Assuming $\rho(t)_{|K_i} > 0$ and $i \in I$, we can follow

$$\begin{aligned} w_i \int_0^T \left|\left|\frac{m(t)}{\rho(t)}\right|\right|_{K_i}^2 dt \leq \sum_{i\in I} w_i \int_0^T \left|\left|\frac{m(t)}{\rho(t)}\right|\right|_{K_i}^2 dt = \int_0^T \left\langle \left(\frac{m(t)}{\rho(t)}\right)^2, 1\right\rangle_c dt \\ \leq \int_0^T \left\langle \left(\frac{m(t)}{\rho(t)}\right)^2, 1\right\rangle_c dt + \mathcal{H}_c\left(\left[\rho(t); \frac{m(t)}{\rho(t)}\right]\right) = R(t). \end{aligned}$$

The latter equality corresponds to the energy-dissipation equality of Theorem 2.24, integrated in time. Now first applying the Jensen-inequality, see [RW98], and then inserting the former estimate yields

$$\int_0^t c_i(s)ds \leq \left(t\int_0^t c_i(s)^2 ds\right)^{1/2} = C\frac{\sqrt{t}}{\Delta_{x_i}}\left(\int_0^t \left|\left|\frac{m(t)}{\rho(t)}\right|\right|_{K_i}^2 ds\right)^{1/2} \leq C\frac{\sqrt{t}}{\Delta_{x_i}\sqrt{w_i}}\sqrt{R(t)}.$$

Inserting the latter bound on $\int_0^t c_i(s)ds$ into the estimate obtained by the Gronwall lemma finishes the proof. □

Remark 4.6. *In the special case of $\langle\cdot,\cdot\rangle_c$ chosen as the $\mathcal{L}^2$-scalar product, we recover the case of pure Galerkin-approximation without complexity reduction. Lemma 4.5 then shows strict positivity of $\rho(t)$ on all of the spatial domain. Such a result has been derived in [Egg18] in a similar setting.*

Note that positivity of $\rho(t)_{|K_i}$ can only be guaranteed for $i \in I$ by Lemma 4.5. This turns out to be sufficient, as the $\rho(t)$-terms in the denominator are only evaluated for $i \in I$. Next, we derive a boundedness result for the solution.

Theorem 4.7. *Let Assumptions 4.4 hold. Then there exist constants C_1, C_2, independent of the discretization parameters, such that for $[\rho; m] \in \mathcal{V}$ with $\rho_{|K_i} > 0$ for $i \in I$ it holds*

$$||\rho|| + ||m|| \leq \left(\max_{i\in I} w_i^{-\frac{1}{2}}\right)\left[C_1\mathcal{H}_c([\rho; m]) + C_2\right].$$

Proof. Let $k = \text{argmax}_{i\in I}\rho_{|K_i}$, and let $\bar{I} = \{i \in I : \rho_{|K_i} > 1]\}$ w.l.o.g. be non-empty. Otherwise we trivially can bound the terms $||\rho||_c, \rho_{|K_k}$ by a constant. From Assumption 4.4-(A1) we get

$$\begin{aligned} \frac{1}{\tilde{C}^2}||\rho||^2 \leq ||\rho||_c^2 = \langle\rho^2, 1\rangle_c = \sum_{i\in I} w_i\rho_{|K_i}^2 \leq \rho_{|K_k}\sum_{i\in I} w_i\rho_{|K_i} \\ = \rho_{|K_k}\left(\sum_{i\in\bar{I}} w_i\rho_{|K_i} + \sum_{j\in I\setminus\bar{I}} w_j\rho_{|K_j}\right) \leq P(\rho_{|K_k})\left(\sum_{i\in\bar{I}} w_i P(\rho_{|K_i}) + \sum_{j\in I\setminus\bar{I}} w_j\right). \end{aligned}$$

By Assumption 4.4-(A2) it follows that $P(\cdot)$ can be bounded from below (not necessarily by zero), which implies that there exists a constant $\hat{c}$ such that

$$w_k P(\rho_{|K_k}) \leq \sum_{i \in \bar{I}} w_i P(\rho_{|K_i}) \leq \langle P(\rho), 1 \rangle_c + \hat{c}.$$

Together, this shows that for some $\hat{c}_1, \hat{c}_2 > 0$ it holds

$$||\rho|| \leq \frac{1}{\sqrt{w_k}} \left(\hat{c}_1 \langle P(\rho), 1 \rangle_c + \hat{c}_2 \right).$$

Similarly, it follows

$$\begin{aligned} \frac{1}{\tilde{C}^2} ||m||^2 \leq ||m||_c^2 = \left\langle \frac{m^2}{\rho} \rho, 1 \right\rangle_c \leq \rho_{|K_k} \left\langle \frac{m^2}{\rho}, 1 \right\rangle_c \leq P(\rho_{|K_k}) \left\langle \frac{m^2}{\rho}, 1 \right\rangle_c \\ \leq \frac{1}{w_k} (\langle P(\rho), 1 \rangle_c + \hat{c}) \left\langle \frac{m^2}{\rho}, 1 \right\rangle_c, \end{aligned}$$

where in the last step the same estimate on $P(\rho[x_k])$ as before has been used. With the help of Young's inequality it follows

$$||m|| \leq \frac{\tilde{C}}{\sqrt{w_k}} \sqrt{(\langle P(\rho), 1 \rangle_c + \hat{c})} \sqrt{\left\langle \frac{m^2}{\rho}, 1 \right\rangle_c} \leq \frac{\tilde{C}}{2\sqrt{w_k}} \left(\langle P(\rho), 1 \rangle_c + \left\langle \frac{m^2}{\rho}, 1 \right\rangle_c + \hat{c} \right).$$

Setting together the estimates for $||\rho||$ and $||m||$ shows the assertion. □

Notably, the bound on the $\mathcal{L}^2$-norms in Theorem 4.7 is almost independent of the discretization parameters. Only the quadrature weights of the complexity reduction step enter. To make the bound uniform, one has to require the quadrature weights to be bounded from below by a positive constant. The concluding result now reads as follows.

Theorem 4.8. *Given $\int_{t=0}^{T} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t) dt$ is bounded for $T > 0$, System 4.1 has a unique solution $[\rho; m] \in \mathcal{C}^1([0,T); \mathcal{V}_1 \times \mathcal{V}_2)$, $\mathbf{f}_B \in \mathcal{C}^1([0,T); \mathbb{R}^{|\mathcal{N}_\partial|})$ and $\mathbf{e}_B \in \mathcal{C}([0,T); \mathbb{R}^{|\mathcal{N}_\partial|})$.*

Proof. Without loss of generality, we assume that the assumptions of the theorem hold for $T = \infty$ for the upcoming proof. Let the boundary conditions u^ν be grouped in two vectors $\mathbf{u}_1 \in \mathcal{C}([0,\infty), \mathbb{R}^{p_1})$ and $\mathbf{u}_2 \in \mathcal{C}^1([0,\infty), \mathbb{R}^{p_2})$ according to our distinction into boundary conditions of Type 1 and Type 2. In Corollary 2.30 and Lemma 2.32 characterizations of $\underline{\mathbf{a}} = [\rho; m] \in \mathcal{C}^1([0,T); \mathcal{V}_1 \times \mathcal{V}_2)$ in terms of a coordinate representation have been derived. This coordinate representation can be written as a system in $\mathbf{a} \in \mathcal{C}^1([0,T); \mathbb{R}^N)$ and $\boldsymbol{\mu} \in \mathcal{C}([0,\infty), \mathbb{R}^{p_2})$ given as

$$\begin{aligned} \mathbf{M} \frac{d}{dt} \begin{bmatrix} \mathbf{a}_1(t) \\ \hat{\mathfrak{z}}_2(\mathbf{a}(t)) \end{bmatrix} = \begin{bmatrix} & \mathbf{J} \\ -\mathbf{J}^T & -\mathbf{R}_c(\mathbf{a}(t)) \end{bmatrix} \begin{bmatrix} \nabla_{\mathbf{M}_1, \mathbf{a}_1} H_c(\mathbf{a}(t)) \\ \mathbf{a}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{P}_1 \end{bmatrix} \mathbf{u}_1(t) + \begin{bmatrix} \mathbf{0} \\ \mathbf{P}_2 \end{bmatrix} \boldsymbol{\mu}(t) \\ \mathbf{0} = \begin{bmatrix} \mathbf{0} & \mathbf{P}_2 \end{bmatrix}^T \mathbf{a}(t) - \mathbf{u}_2(t). \end{aligned}$$

The matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ describe sub-blocks of $\mathbf{K}_2$ in Corollary 2.30. To show global existence of solutions, we employ the underlying ordinary differential equation of the coordinated representation. It is obtained by eliminating $\boldsymbol{\mu}$. According to Lemma 2.34, this yields a system of the form

$$\frac{d}{dt} \mathbf{a}(t) = \mathbf{f}(t, \mathbf{a}(t)), \qquad \mathbf{a}(0) = \mathbf{a}_0 \in \mathbb{A}_N$$

with $\mathbf{f} : [0,T] \times \mathbb{A}_N \to \mathbb{R}^N$ continuous. Let $\Psi : \mathbb{R}^N \to \mathcal{V}$ be the transformation from the coordinate representation to the function in $\mathcal{V}$, Definition 2.25. The domain $\mathbb{A}_N$ can then be specified as

$$\mathbb{A}_N = \{\mathbf{a} \in \mathbb{R}^N : \text{ with } [\rho; m] = \Psi(\mathbf{a}) \in \mathcal{V}, \text{ and } 0 < \rho[x] < M \text{ for } x \in K_i,\, i \in I\}.$$

The local existence of a solution readily follows by the Peano existence theorem. The extension theorem [WP10, Theorem 6.2.1] then states that there exists a maximal continuation of this solution onto a maximal interval $[0, t^*)$. The theorem further states that either $t^* = \infty$ or one of the following scenarios holds for finite t^*:

- The solution diverges, i.e., $|| \lim_{t \uparrow t^*} \mathbf{a}(t)|| = \infty$.
- The solution tends to the boundary of $\mathbb{A}_N$, i.e., $\lim_{t \uparrow t^*} \mathbf{a}(t) \in \partial\mathbb{A}_N$.

The first scenario cannot be reached as the solution is bounded due to Theorem 2.24 and Theorem 4.7. The second scenario of $\mathbf{a}(t)$ leaving $\mathbb{A}_N$ is a contradiction to Lemma 4.5. Therefore, $t^* = \infty$, and the solution can be globally extended.

Let us now discuss the uniqueness. For that, we reinterpret the boundary conditions $\mathbf{u}_1(t)$, $\mathbf{u}_2(t)$ and the time-derivatives $\frac{d}{dt}\mathbf{a}_1(t)$, $\frac{d}{dt}\hat{\mathfrak{z}}_2(\mathbf{a}(t))$ as right hand sides for fixed $t \geq 0$ and gather them in a vector $\mathbf{g}(t)$. The coordinate representation then yields an algebraic system determining $[\mathbf{a}(t); \boldsymbol{\mu}(t)]$ in dependence of $\mathbf{g}(t)$, given as

$$\mathbf{F}\left(\begin{bmatrix}\mathbf{a}(t)\\ \boldsymbol{\mu}(t)\end{bmatrix}\right)\begin{bmatrix}\mathbf{a}(t)\\ \boldsymbol{\mu}(t)\end{bmatrix} = \mathbf{g}(t), \qquad \text{with} \quad \mathbf{F}\left(\begin{bmatrix}\mathbf{a}\\ \boldsymbol{\mu}\end{bmatrix}\right) = \begin{bmatrix} & \mathbf{J} & \\ -\mathbf{J}^T & -\mathbf{R}_c(\mathbf{a}(t)) & \mathbf{P}_2 \\ & \mathbf{P}_2^T & \end{bmatrix}$$

$$\mathbf{g}(t) = \begin{bmatrix}\mathbf{M} & \\ & \mathbf{I}\end{bmatrix} \frac{d}{dt}\begin{bmatrix}\mathbf{a}_1(t)\\ \hat{\mathfrak{z}}_2(\mathbf{a}(t))\\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ -\mathbf{P}_1\mathbf{u}_1(t)\\ \mathbf{u}_2(t)\end{bmatrix}.$$

Under the help of Part I.A-Lemma 2.18 and the inverse function theorem, unique solvability of the algebraic system can be derived. This, in turn, shows that the time derivatives $\frac{d}{dt}\mathbf{a}_1(t)$ and $\frac{d}{dt}\hat{\mathfrak{z}}_2(\mathbf{a}(t))$ are unique for each $t \geq 0$, i.e., the uniqueness of solutions of System 4.1. □

Remark 4.9. *The boundedness of the boundary exchange $\int_{t=0}^{T} \mathbf{e}_B(t) \cdot \mathbf{f}_B(t)dt$ was posed as an assumption in Theorem 4.8. Under certain circumstances on the boundary conditions or the form of $P(\cdot)$, one can derive an intrinsic bound. Sufficient is a bound depending linearly on the Hamiltonian and continuously on the boundary conditions u^ν. We obtain this, e.g., when only boundary conditions of Type 1 are posed, or for the isentropic model for $P(\cdot)$, cf. the introduction of Part I and [Egg18], [Che05].*

Let us briefly summarize the results of this subsection: We derived well-posedness, positivity of the densities, $\rho(t)_{|K_i} > 0$ for $i \in I$ and $t > 0$, and we were able to bound the norms of $\rho(t), m(t)$ in terms of $\mathcal{H}_c([\rho(t); m(t)])$ as described in Theorem 4.7 for $t > 0$, i.e., by the initial data and power exchange over the boundaries as described in Theorem 2.24.

Chapter 5

Numerical results

Different aspects of our theoretical findings are numerically illustrated in this chapter using the example of the isothermal Euler equations on networks. After our standard setup for the numerical experiments is settled in Section 5.1, the time discretization is examined in Section 5.2. All other sections are on the performance of our model reduction approach. Expectably, the latter depends on the parameter regimes and scales at hand. We briefly illustrate this for a small network example for varying friction terms in Section 5.3. More extensive studies are then carried out on a larger network example in Section 5.4. First this is done for purely projection-based reduced models and then for the additionally complexity-reduced ones. In particular, a comparison to the non-structure-preserving alternative for complexity reduction given by DEIM is made. Finally, the importance of the compatibility condition, Assumption 2.20, for the performance of our proposed complexity reduction is showcased in Section 5.5.

5.1 Setup for numerical results

For all calculations, equations and quantities are non-dimensionalized w.r.t. to SI-units. Where we assume it to be helpful, we state the respective SI-base in squared brackets once, e.g., $[kg]$ for kilogram. The constants

$$\bar{h} = 3600 \qquad \text{and} \qquad \overline{km} = 1000$$

are used to specify time- and space-variables in the illustrations. The isothermal Euler equations (ISO1) and its simplification (ISO2) from Chapter 4 are considered here in parameter regimes typical in the application for gas transport networks, [SAB$^+$17], [HGCATRM09]. For all test cases, the constitutive law for pressure p is set to

$$p(\rho) = RT\frac{\rho}{1 - RT\alpha\rho}, \qquad \text{with } R = 518\left[\frac{J}{kg\,K}\right], \quad T = 283\,[K], \quad \alpha = -3\cdot 10^{-8}.$$

If not stated differently, the ISO1 model is utilized. Although ISO1 and ISO2 lead to very similar solution trajectories for all our test scenarios, the ISO1 model clearly is the more challenging one in our framework due to the more complex Hamiltonian structure.

Errors are measured by the $\mathcal{L}^2$-norm in space and the supremum-norm in time. If not stated differently, the solution of the FOM acts as the reference solution. Thus given $\mathbf{a}$ as the FOM solution, and the approximation $\tilde{\mathbf{a}}$ at hand, the relative error E_T is defined as

$$E_T = \frac{\max_{t\in[0,T]} ||\mathbf{a}(t) - \tilde{\mathbf{a}}(t)||_{\mathcal{E}}}{\max_{t\in[0,T]} ||\mathbf{a}(t)||_{\mathcal{E}}}.$$

The FOM is obtained using a finite element discretization on an edge-wise uniform partitioning with the ansatz spaces from Part I.A-Section 2.4. Our standard discretization parameters are $\Delta_s \leq 500[m]$ chosen maximal as spatial step size, and a time discretization by the implicit Euler-type scheme, System 2.37, with $\Delta_t = 3[s]$ uniformly for FOM and all reduced models. Only deviations from these settings are reported in the scenarios. The resulting nonlinear algebraic equations after all discretization steps are solved by Netwton's method, which is ran until a relative tolerance of 10^{-8} is reached in our energy-norm. If the fixed-point iteration has not reached this tolerance after 20 iterations, we cancel the simulation and consider it as failed. For convenience of implementation, a few nonlinear integrals are approximated by the Gauss-quadrature with four quadrature-points [Bar16], [DH08]. These are the space integrals of the friction-term $\hat{\mathbf{d}}(\cdot)$ in System 3.1 and its reduced counterparts as well as the time-integrals in the Galerkin-in-time methods, Section 2.5, for which we already introduced the placeholder $\Xi^k[\cdot]$ for quadrature, see (2.10), and the boundary data for the Galerkin-in-time methods, cf. Remark 2.36.

The projection matrix in all our reduced models is obtained by Algorithm 3.7. This algorithm as well as the complexity reduction methods get the trajectory of the FOM solution sampled at 500 uniformly distributed time instances as snapshots $\mathbf{S}$ in (3.1), if not stated differently. The reduced dimension n and the number of degrees of freedom in the greedy search of the complexity reduction n_c then fully specify these models. We reserve the abbreviation *ROM* in this chapter for the reduced model gotten from Galerkin projection without complexity reduction, System 3.4. Additionally complexity-reduced models, System 3.9, are denoted as *ROMQ* and *ROMD*, depending on whether the quadrature-based method of Section 3.3 or the DEIM method of Section 3.4 is applied. Recall that only the complexity-reduced models *ROMQ* and *ROMD* are online-efficient in the sense that computational time is saved compared to FOM. The model *ROM*, in comparison, takes about the same simulation time as FOM. All implementations have been carried out in `MATLAB` Version 9.1.0.441655 (R2016b).

Remark 5.1 (Spatial representations for network solution)**.** *The spatial domain described by a single pipe can be identified with an interval. In our spatial representations for networks, we tacitly make these identifications for each pipe and string them together, see Fig. 5.4, Fig. 5.6 and Fig. 5.11. Note that discontinuities in the mass flows at junctions, marked by dotted vertical lines in the figures, are to be expected when more than two pipes meet due to the employed coupling conditions.*

5.2 Time discretization

In this section the performance of our time discretization schemes System 2.37, System 2.40 and System 2.45 are compared. We refer to them in the upcoming as *ImpEul*, *DG* and *CoG*, respectively. In contrast to all other sections, the space approximation is fixed and only the time discretization is varied. Consequently, the reference solution refers to a solution obtained from a highly resolved time discretization. In our case it is obtained by `MATLAB`'s solver '`ode15s`' with

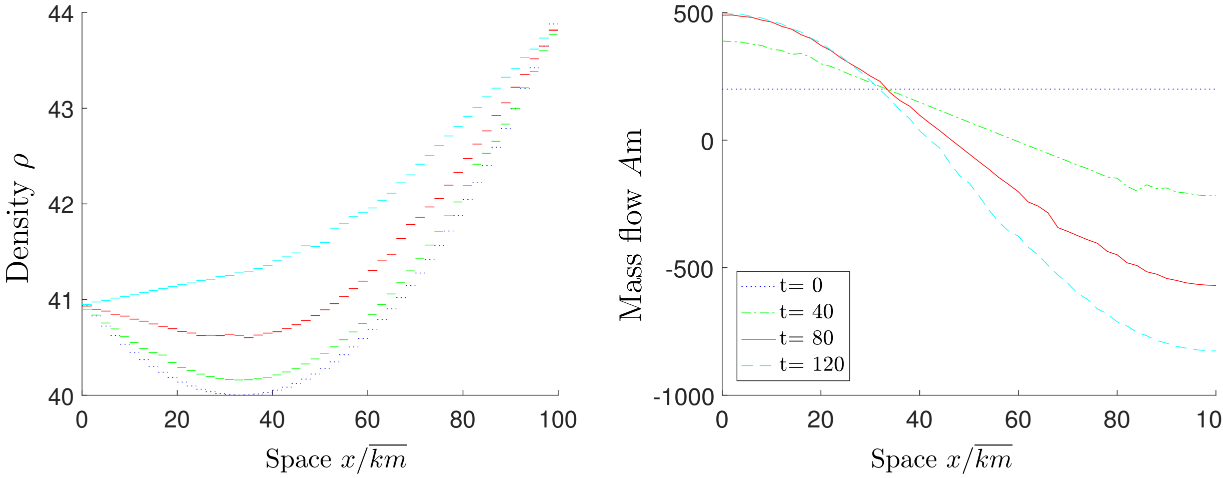

Figure 5.1: One pipe, Scenario 5.2. Spatial representation of reference solution.

the relative tolerance modified to 'RelTol $= 10^{-8}$' and exact Jacobian matrices forwarded to it. We start by considering a scenario with only one pipe.

Scenario 5.2 (One-pipe network)**.** *The network topology consists of one pipe* $\omega = (\nu_1, \nu_2)$ *with length* $l^\omega = 100\overline{km}$ *and diameter* $D = 2\sqrt{1/\pi}$ *(i.e.* $A = 1$*). Initial- and boundary conditions are chosen as*

$$\rho(0,x) = 40 + 9\left(\frac{x}{l^\omega} - \frac{1}{3}\right)^2, \qquad Am(0,x) = 200$$
$$\rho(t,\nu_1) = 41, \qquad \rho(t,\nu_2) = 44, \qquad t \in [0,T], \quad T = 120.$$

The ISO2 model is employed and the laminar friction model

$$\lambda(m) = 10^{-5}\frac{64}{mD}.$$

Moreover, step size $\Delta_s = 2\overline{km}$ *is chosen, resulting in a dimension* $N = 101$ *for FOM.*

The scenario involves initial conditions far afield from a stationary solution and a comparably small friction factor given by the laminar friction model [DHLT17].Consequently, relatively large temporal gradients are observed in the considered simulation period $t \in [0,120]$, see Fig. 5.1. The scenario is designed to reveal the different numerical dissipation rates of our methods distinctly.

Numerical dissipation

Recall the (absolute) numerical dissipation D^{ImpEul}, D^{DG} and D^{CoG} for the different methods from (2.6), (2.8) and (2.9). Defining the total dissipation for the final time T as

$$D_{oT} = \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(T))) - \mathcal{H}(\hat{\mathbf{z}}(\underline{\mathbf{a}}(0))),$$

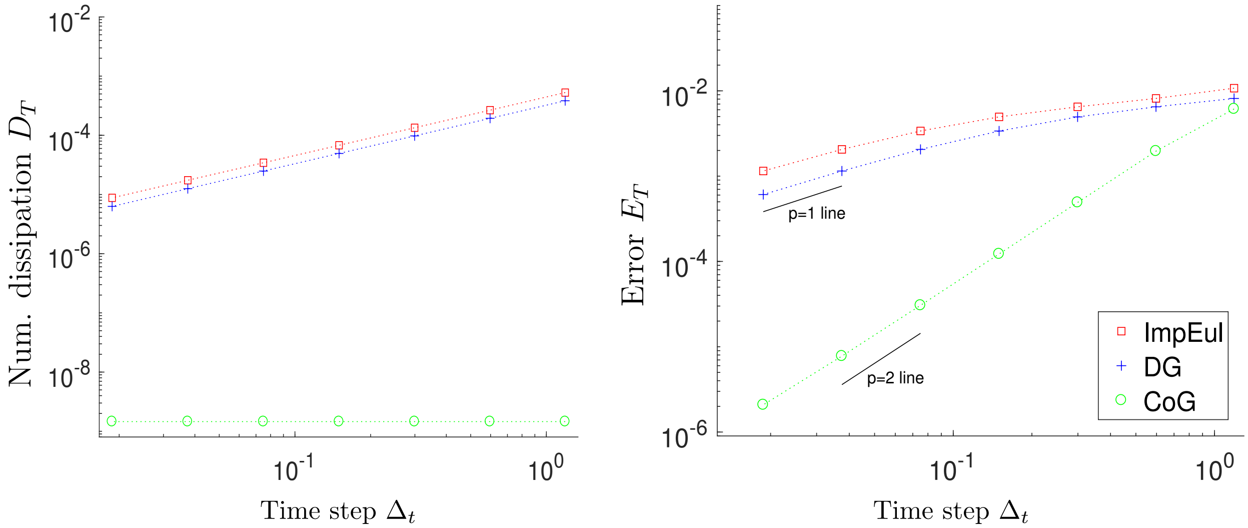

Figure 5.2: One pipe, Scenario 5.2. *Left:* Rel. numerical dissipation D_T; *Right:* Rel. errors E_T.

we can introduce a relative numerical dissipation rate D_T as the method-dependent quantity

$$D_T = \begin{cases} \frac{D^{\text{ImpEul}}(T)}{D_{oT}} & \text{for } \textit{ImpEul} \\ \frac{D^{\text{DG}}(T)}{D_{oT}} & \text{for } \textit{DG} \\ \frac{D^{\text{CoG}}(T)}{D_{oT}} & \text{for } \textit{CoG}. \end{cases} \tag{5.1}$$

This rate is depicted in Fig. 5.2-*left* for Scenario 5.2 for varying step size Δ_t. *ImpEul* and *DG* show a similar, rather high numerical dissipation, which scales approximately linear with the step size. *CoG*, on the other hand, inherits no step-size dependent numerical dissipation, but only a small step-size independent base-deviation in the Hamiltonian. With under 10^{-8} it is of an order comparable to the chosen tolerance in the underlying Newton's methods. This numerically confirms our energy dissipation inequalities and, in particular, the dissipation equality shown in Theorem 2.46.

Convergence

A convergence study in time for Scenario 5.2 is depicted in Fig. 5.2-*right*. It shows distinctly the convergence orders, which are first order for *ImpEuler* and *DG* and even second order for *CoG*. We note, however, that the handling of the friction term in *CoG* may lead to lower convergence orders for scenarios featuring other friction models.

One can observe this for the upcoming scenario, where the so-called Weymouth-model [DHLT17] is used for the friction factor.

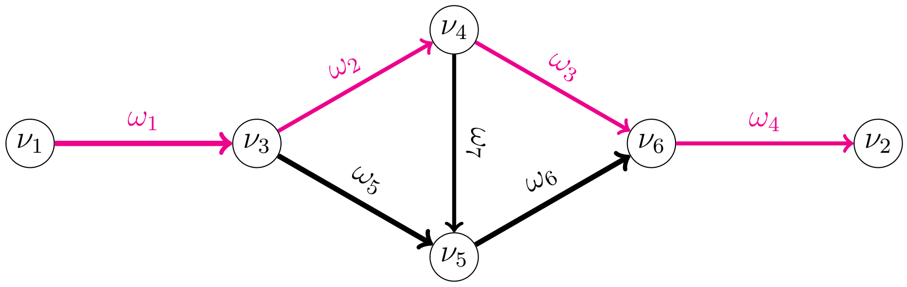

Figure 5.3: Network topology 'Diamond' used in Scenario 5.3. Lengths $\{l^{\omega_i}/\overline{km} : \omega_i \in \mathcal{E}\} = \{40, 38, 18, 15, 28, 27, 25\}$ and diameters $\{D^{\omega_i} : \omega_i \in \mathcal{E}\} = \{1.3, 1, 1, 1, 1.3, 1.3, 1\}$. The domain $\{\omega_i \in \mathcal{E} : i = 1, \dots, 4\}$ is marked in magenta.

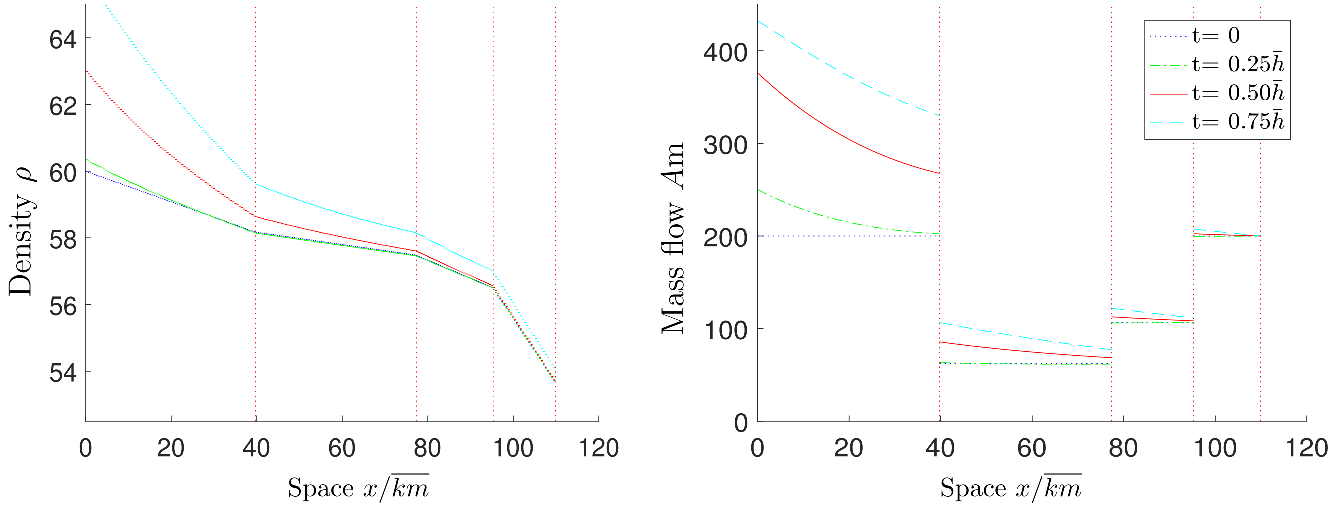

Figure 5.4: Diamond network, Scenario 5.3. Spatial representation of reference solution with the domain relating to the marked path in Fig. 5.3.

Scenario 5.3 (Diamond network)**.** *The network topology is given as in Fig. 5.3. The boundary conditions are chosen as*

$$\rho(t,\nu_1) = 60 + u_p(t), \qquad Am(t,\nu_2) = 200$$
$$\textit{with } u_p(t) = 5\left(\exp\left(-\frac{t}{\bar{h}}\right) - \cos\left(\frac{\pi t}{\bar{h}}\right)\right), \qquad t \in [0,T], \quad T = 2\bar{h},$$

and the solution is initialized with the respective stationary solution at $t = 0$. The employed friction model reads

$$\lambda = \frac{4}{(11.18\, D^{1/6})^2}.$$

The resulting dimension of FOM is $N = 771$.

The solution to Scenario 5.3 is visualized on the highlighted parts of spatial domain in Fig. 5.4 for different time instances. A convergence study can be found in Fig. 5.5, once for ISO1 and once

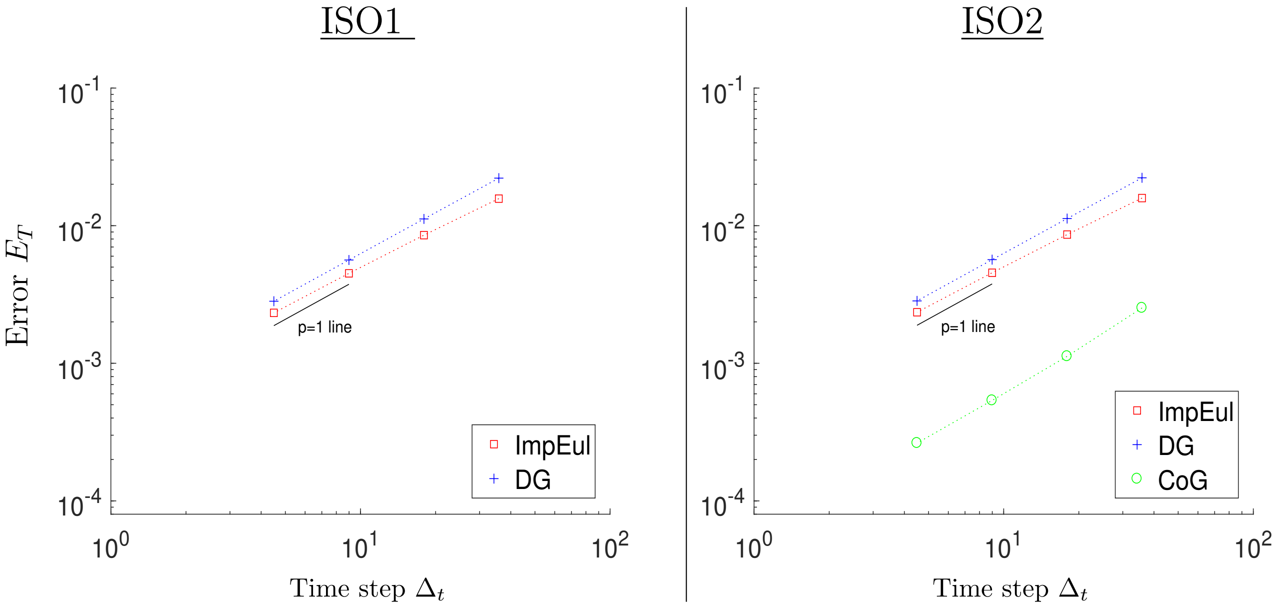

Figure 5.5: Diamond network, Scenario 5.3. time discretization errors, divided by employed models.

for ISO2. Convergence behavior is overall already seen for larger step sizes Δ_t compared to the former one-pipe scenario. All methods show here first order convergence, i.e., *CoG* is not of higher order. In terms of absolute error, it still shows the best performance. Further recall that *CoG* is only realizable for the ISO2 model and every single time step is more expensive, cf. Remark 2.47. Again, *DG* and *ImpEul* show very similar quantitative approximation behavior. As *ImpEul* shows to be slightly more efficient in our implementation, we employ this method here for all upcoming tests. Let us mention, though, that *DG* could be generalized to higher-order methods by choosing higher-order ansatz spaces, whereas a higher-order energy-stable generalization of *ImpEul* is not straight-forwardly obtained in our analysis.

5.3 Reducibility of gas networks

Of course not every model or every parameter regime is equally appropriate for the application of model reduction methods. One notoriously hard-to-reduce type of model consists of transport-dominated systems, which allow for sharp fonts, [CMS19], [RSSM18]. Luckily, in our application of gas networks, the friction term together with the high characteristic speeds insert considerable damping effects. How strong they are, depends on the chosen friction factor λ. To illustrate its influence, we repeat Scenario 5.3 with the modified input-profile

$$u_p(t) = 5u_{saw}\left(4\frac{t}{\bar{h}}\right), \qquad \text{with } u_{saw}(t) = \begin{cases} t, & 0 \le t < 1 \\ 2-t, & 1 \le t < 2 \\ 0, & t \ge 2, \end{cases} \tag{5.2}$$

and varying choices of friction factor λ. The choices are $\lambda = 0.01$ and $\lambda = 0.002$, both constant over the whole network. For both choices the respective solution on the spatial domain marked

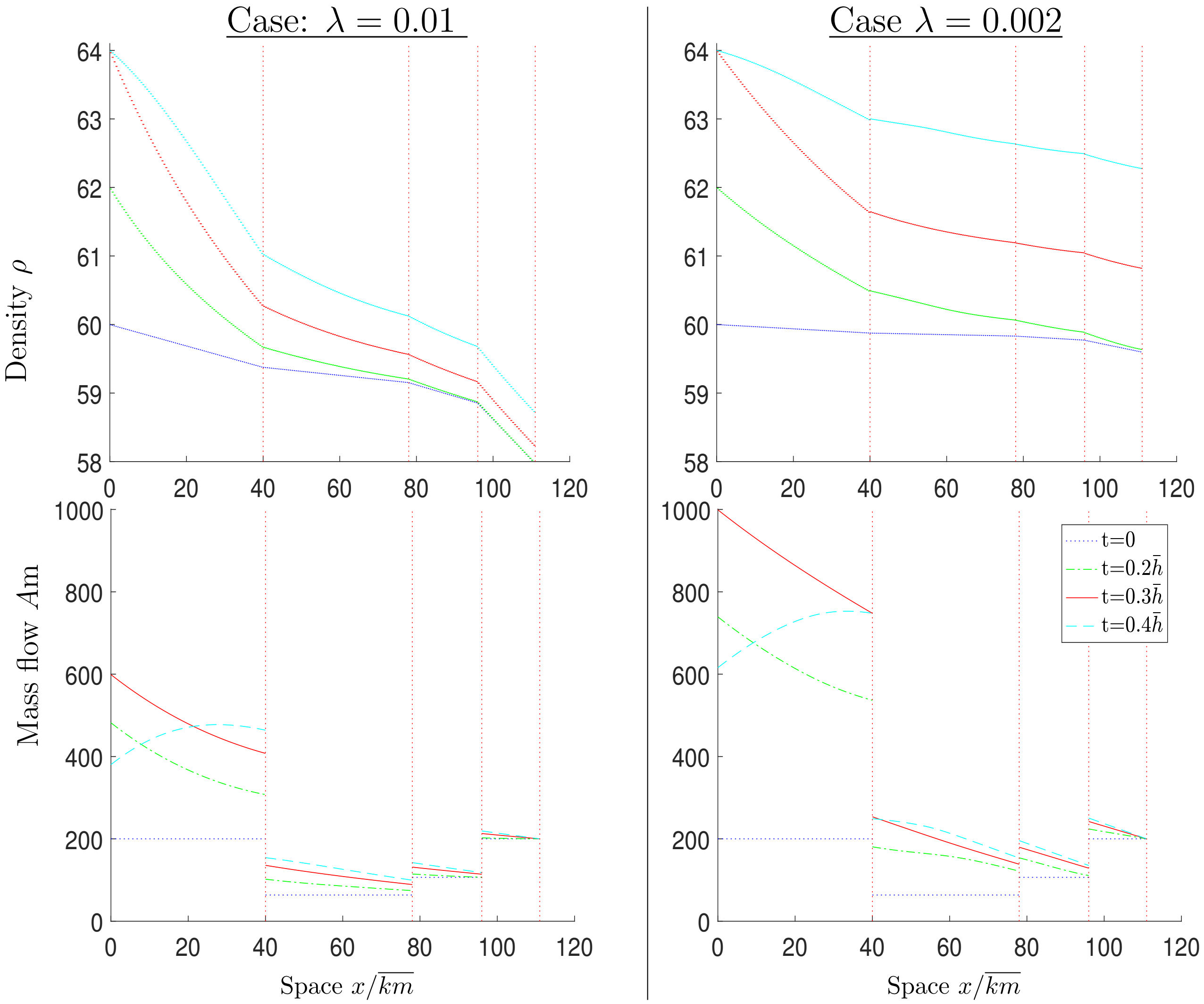

Figure 5.6: Diamond network, Scenario 5.3 with modified λ and input profile u_p from (5.2). *Left:* $\lambda = 0.01$; *Right:* $\lambda = 0.002$. Spatial representation of reference solution with domain relating to marked path in Fig. 5.3.

in Fig. 5.3 can be found in Fig. 5.6. The comparably high characteristic speeds inherent in the problem are apparent when considering the solution at $t = 0.2\bar{h}$. Already at this point in time, the influence of the varying boundary conditions at ν_1 have spread over the whole spatial domain. A visualization of the evolution over time for the density and mass flow at the pipe-end of ω_2 can be found in the top rows of Fig. 5.7 and Fig. 5.8.

The respective differences in time of FOM minus the reduced models are plotted for a selection of reduced models in these figures below. For all reduced models the dimension $n = 12$ is used. For *ROMQ*-12 and *ROMQ*-17 the complexity reduction is done with $n_c = 12$ and $n_c = 17$, respectively. While the (not online-efficient) *ROM* resolves the solution quite well for all scenarios, noticeable errors occur in the complexity-reduced models. Their approximation quality tends to be worse for the case with lower friction factor $\lambda = 0.002$, see Fig. 5.8-*bottom*. In particular, the *ROMQ* method with $n_c = 12$ seems under-resolved for this case. Notably also the under-resolved reduced models allow for a stable simulation. This behavior can not be taken for granted when applying

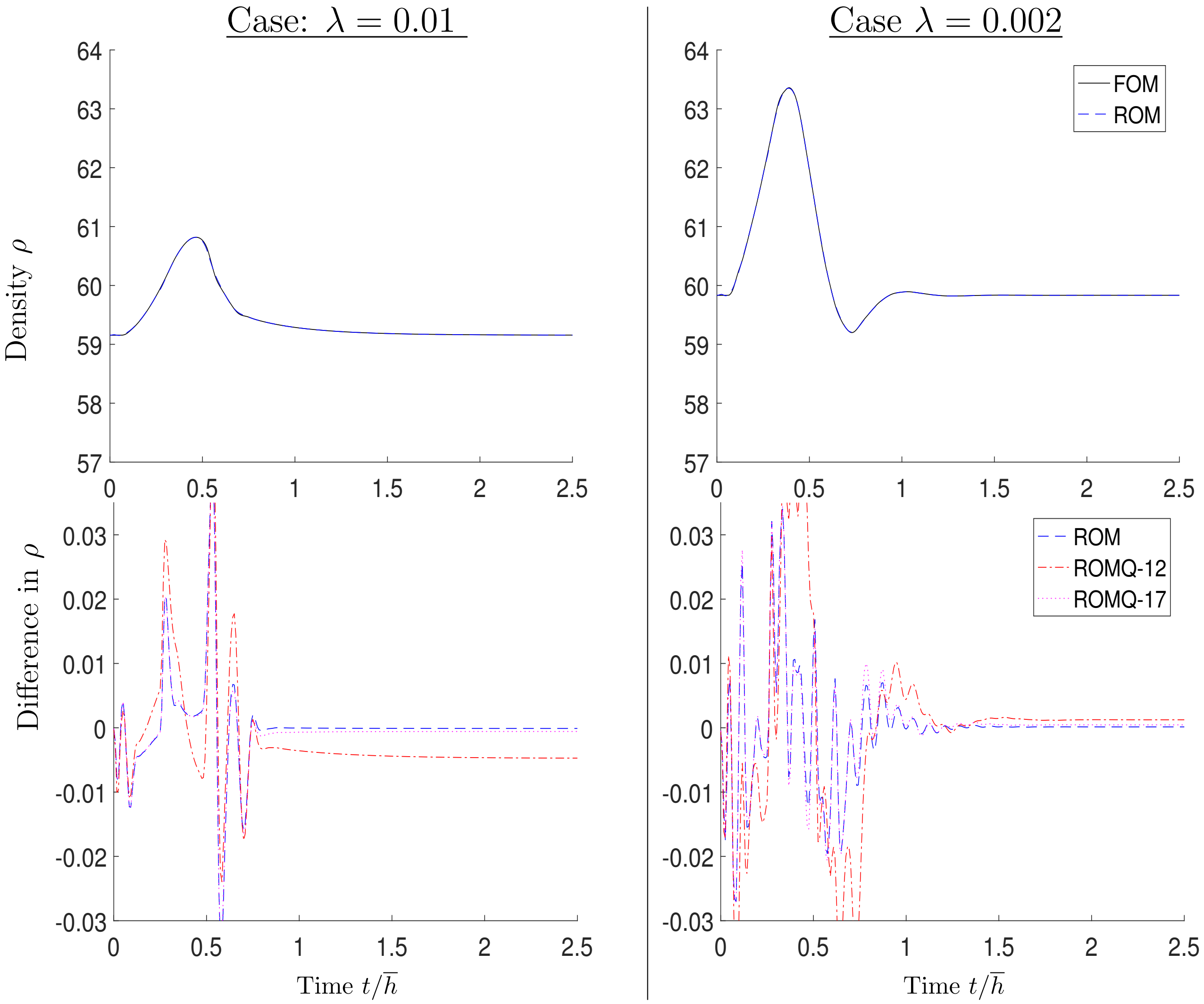

Figure 5.7: Diamond network, Scenario 5.3 with modified λ and input profile u_p from (5.2). *Left:* $\lambda = 0.01$; *Right:* $\lambda = 0.002$. Temporal representation of density ρ at pipe-end of ω_2. *Top:* FOM- and *ROM*-solution. *Bottom:* Difference of *ROM* and *ROMQ* to FOM.

non-structure-preserving model reduction methods to our kind of model, as seen in the second half of Section 5.4 and Section 5.5. Further, only scenarios with rather high friction factors seem appropriate for model reduction. We assume to be in such regimes for all upcoming tests.

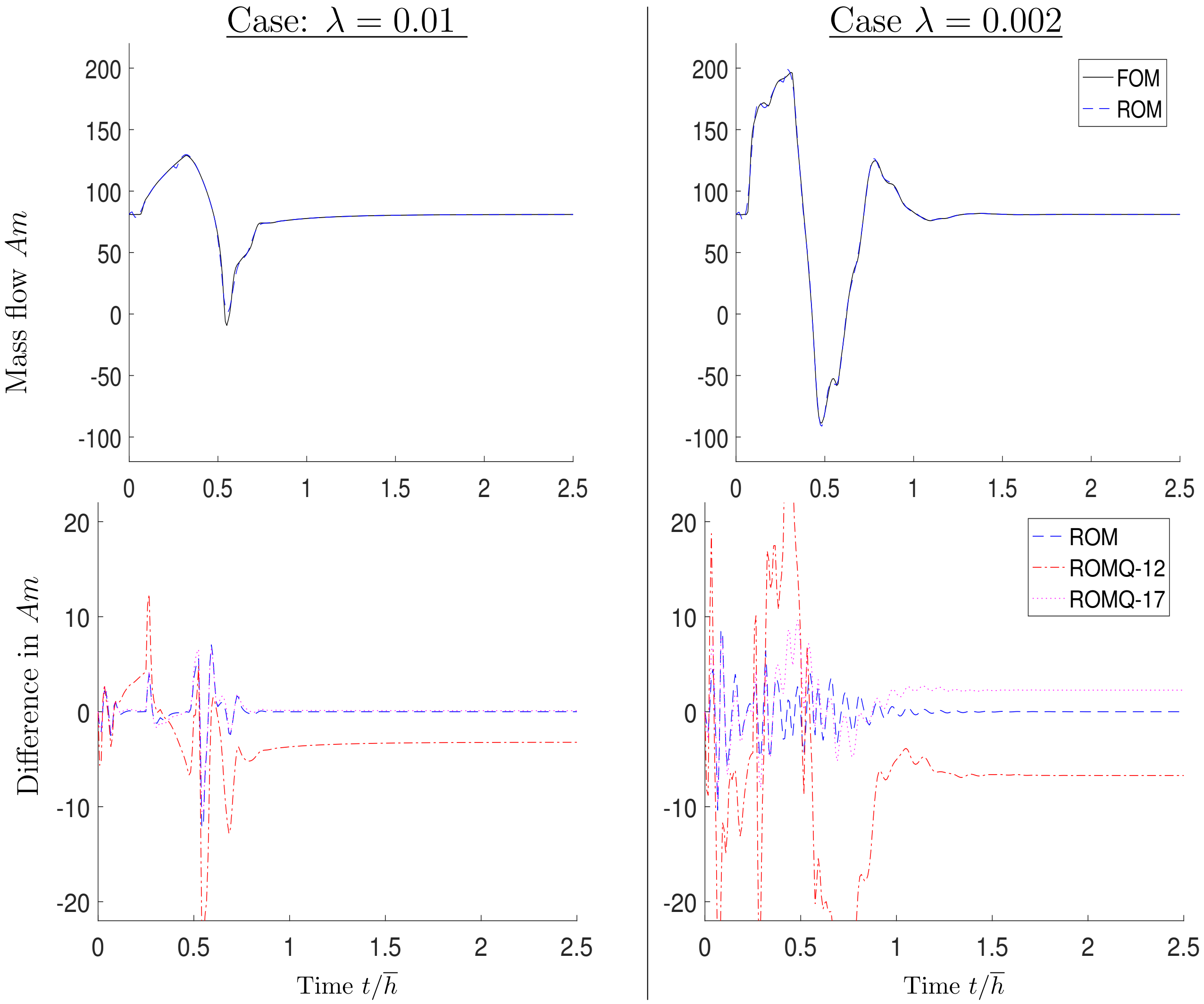

Figure 5.8: Diamond network, Scenario 5.3 with modified λ and input profile u_p from (5.2). *Left:* $\lambda = 0.01$; *Right:* $\lambda = 0.002$. Temporal representation of mass flow Am at pipe-end of ω_2. *Top:* FOM and *ROM*-solution. *Bottom:* Difference of *ROM* and *ROMQ* to FOM.

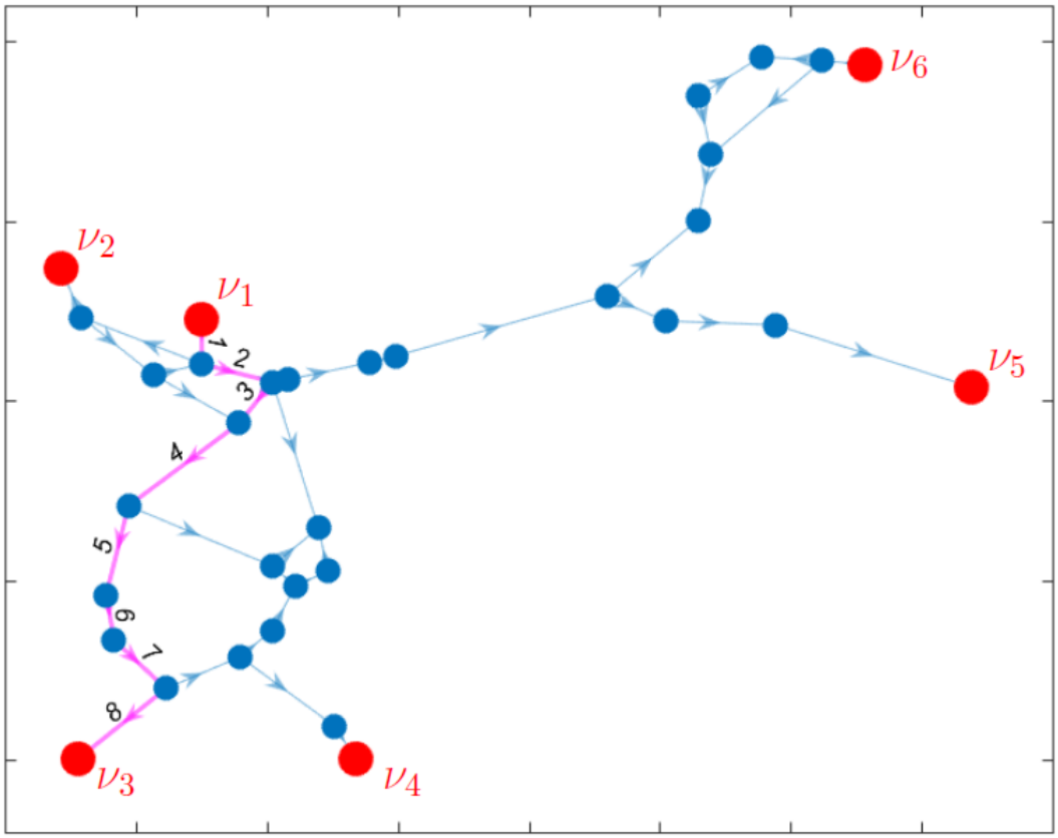

Figure 5.9: Large network topology used in Scenario 5.4. The network consits of 38 Pipes between $5\overline{km}$ to $74\overline{km}$ length, in total $1008\overline{km}$. Diameters are between 0.4 to 1. The spatial domain given by pipes 1-8 is marked in magenta.

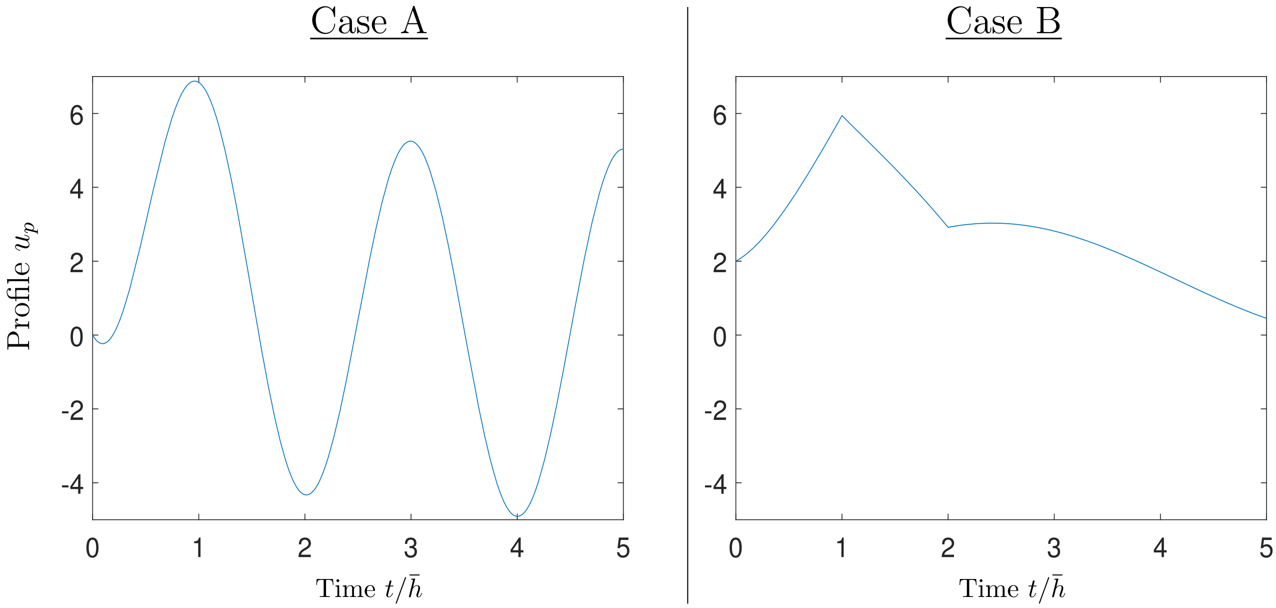

Figure 5.10: Input profiles for Scenario 5.4, divided by cases.

5.4 Model reduction for larger network

For all remaining numerical experiments in this part, we consider a larger pipe-network. The respective topology, shown in Fig. 5.9, is constructed from [SAB$^+$17, GasLib-40] by a few modifications, such as replacing compressors by pipes and rounding the pipe-lengths onto two leading digits.

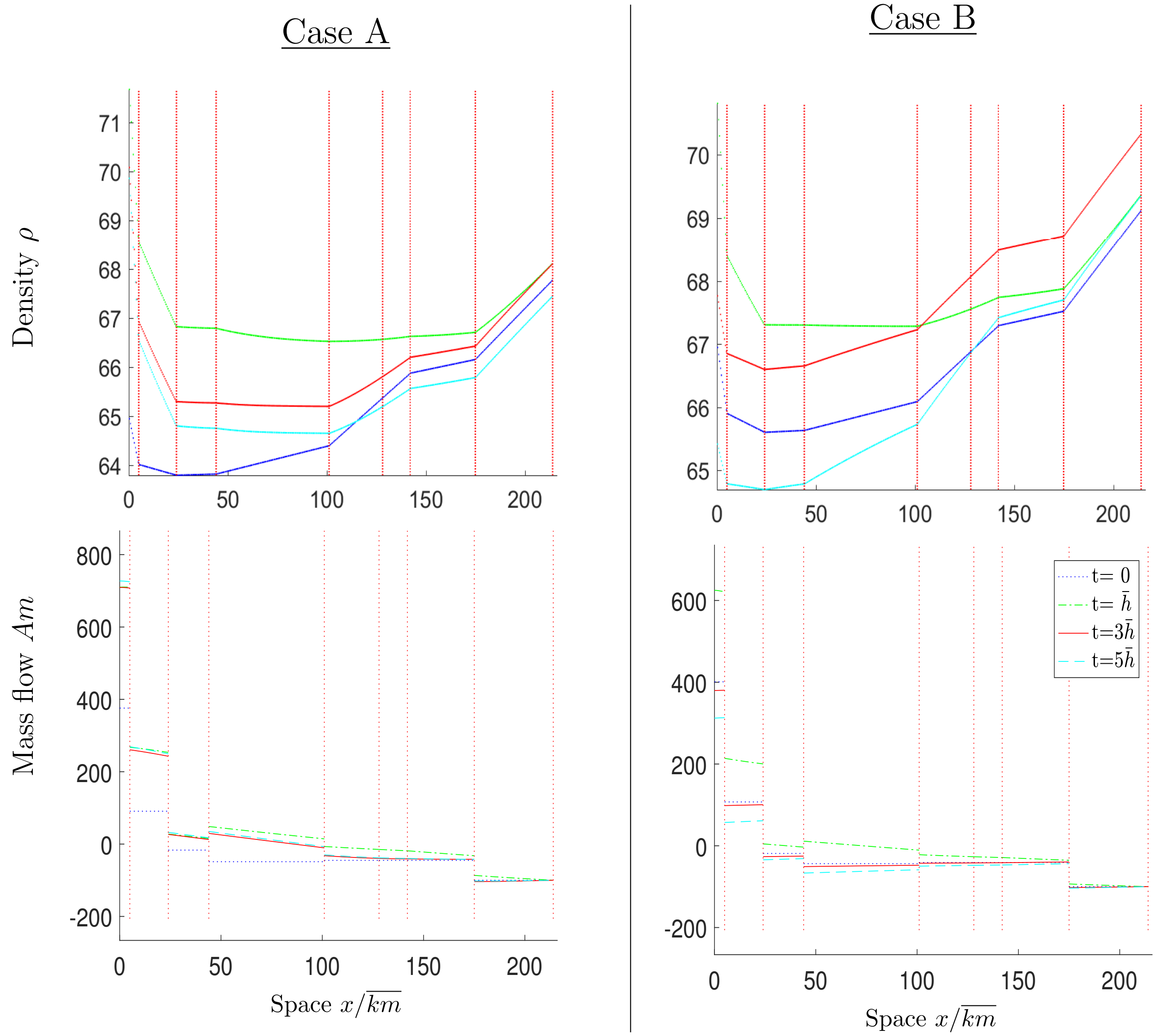

Figure 5.11: Large network, Scenario 5.4 divided by cases. Spatial representation of reference solution with domain representing pipes 1-8 (marked in magenta in Fig. 5.9).

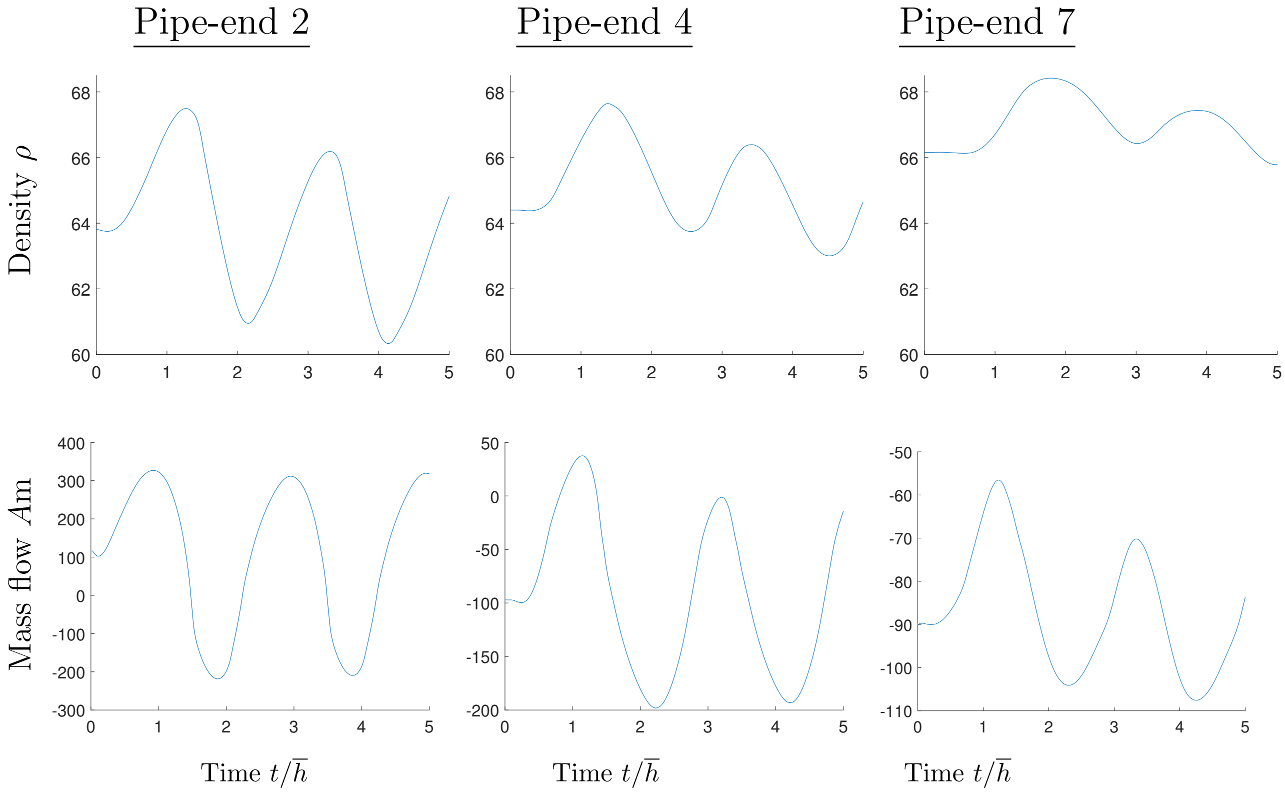

Figure 5.12: Large network, Scenario 5.4-Case A, temporal representation of reference solution at pipe-ends of pipes with number 2, 4 and 7 *(left to right)*. *Top:* Density ρ. *Bottom:* Mass flow Am.

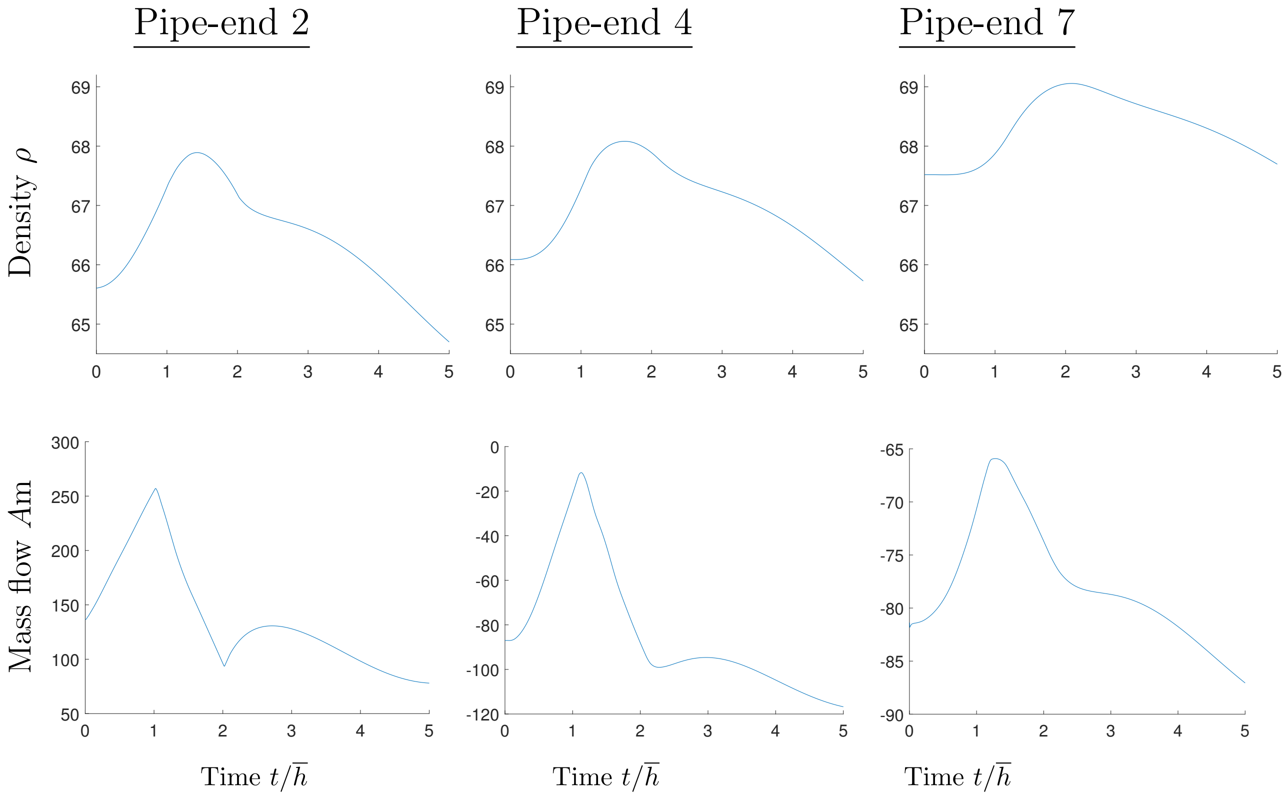

Figure 5.13: Large network, Scenario 5.4-Case B, temporal representation of reference solution at pipe-ends of pipes with number 2, 4 and 7 *(left to right)*. *Top:* Density ρ. *Bottom:* Mass flow Am.

Scenario 5.4 (Large network)**.** *The network topology is given as in Fig. 5.9. For given input-profile u_p, the boundary conditions are chosen as*

$$\begin{aligned} &\rho(t,\nu_1) = 65 + u_p(t), \qquad &&\rho(t,\nu_2) = 50 + u_p(t), \qquad &&\rho(t,\nu_4) = 60 - u_p(t) \\ &\rho(t,\nu_5) = 60, &&\rho(t,\nu_6) = 45, &&Am(t,\nu_3) = -100 \quad t \in [0,T], \end{aligned}$$

for $T = 5\bar{h}$. The solution is initialized with the respective stationary solution at $t = 0$. Constant friction factor $\lambda = 0.01$ is set. The input-profile is varied by cases:

- *Case A:* $u_p(t) = 5\left(\exp\left(-\frac{t}{\bar{h}}\right) - \cos\left(\pi\frac{t}{\bar{h}}\right)\right)$.
- *Case B:* $u_p(t) = 4u_{saw}\left(\frac{t}{\bar{h}}\right)) + 2\left(1 + sin\left(4\pi\frac{t}{\bar{h}}\right)\right) + -4\frac{t}{\bar{h}}\exp\left(\frac{t}{\bar{h}}\right)$,

$$\text{with } u_{saw}(t) = \begin{cases} t, & 0 \leq t < 1 \\ 2 - t, & 1 \leq t < 2 \\ 0, & t \geq 2. \end{cases}$$

The resulting dimension of FOM is $N = 4074$.

The input-profiles u_p for the cases of Scenario 5.4 are plotted in Fig. 5.10. Respective solutions are illustrated in Fig. 5.11 in a spatial representation. Temporal representations can be found in Fig. 5.12 for Case A and in Fig. 5.13 for Case B. Comparing these with the input-profiles from Fig, 5.10, the smoothing of the solution profiles by the damping is apparent.

For all upcoming tests, the solution trajectory of Scenario 5.4-Case A is used for the snapshots of the training phase. Consequently, larger errors for the reduced models can be expected in Case B.

Projection step

First, the error in the model reduction in dependence of the reduced dimension n is examined. For this, we consider in Fig. 5.14 the reduced models without complexity reduction, i.e., *ROM*. In the figure a direct comparison to *OrthP* is drawn, which is the orthogonal projection of the solution trajectory onto the respective reduced space. *OrthP* thus shows the pure projection error made in the model reduction and serves as a lower bound for the expected total reduction error.

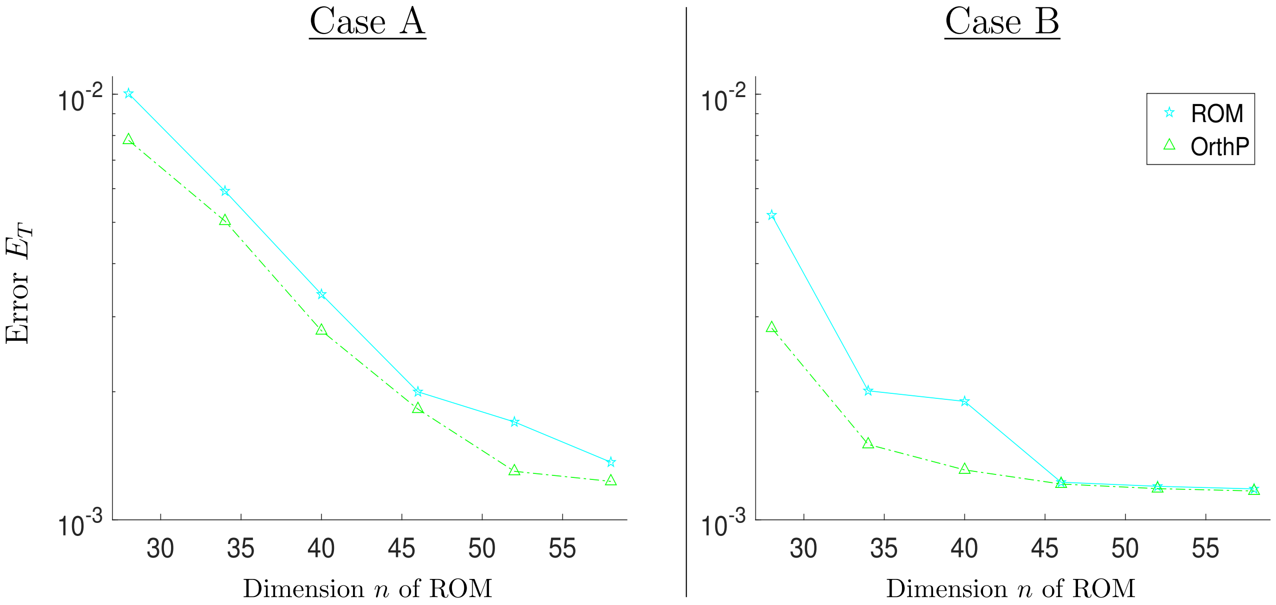

Figure 5.14: Large network, Scenario 5.4, relative errors E_T of *ROM* compared to orthogonal projection. *Left:* Case A (trained); *Right:* Case B (not trained).

We observe its error to monotonically decrease until an error of about 10^{-3} is reached. Further, the gap between *ROM* and *OrthP* is quite small and no significant drift-off effects can be observed in *ROM*. In other words, our reduced models show to be robust, both for the trained Case A and the not trained Case B. However, recall that *ROM* is not online-efficient and thus complexity reduction has to be taken into account to save computational time.

Complexity reduction step

Next, we examine the performance of the complexity reduction step for varying degrees of freedom n_c in the greedy search of the complexity reduction methods. For this, we fix the reduced dimension $n = 40$ in all models. The errors made for Scenario 5.4 by our proposed *ROMQ* and the non-structure-preserving *ROMD* are compared in Fig. 5.15. A few observations can be made. Firstly, the simulations of *ROMD* do not even run to completion for quite a few parameter settings, which are all points without a plot in Fig. 5.15. This in particular is the case for the not perfectly trained setting, Fig. 5.15-Case B, where only for $n_c = 75$ and $n_c = 85$ the simulation of *ROMD* does not break down. Our proposed *ROMQ*, on the other hand, does not suffer from these severe stability issues, and it runs to completion for all considered parameter settings. We also observe that for n_c sufficiently large, the complexity reduction errors can be neglected compared to the reduction errors. In the setting of Fig. 5.15, this holds for $n_c \approx 65$ and larger. Although *ROMQ* performs profoundly more robust than *ROMD*, it also shows some undesired effect. The error is not monotonously decreasing for increasing n_c for $n_c < 50$, though not as wavering as for *ROMD*. This points towards room for improvement for our heuristic, Algorithm 3.15, underlying the construction of *ROMQ*. The speed-up factor we observe for *ROMQ* in comparison to FOM is about a factor of six in our implementation (≈50 sec. vs. ≈300 sec.) for the highest tested order n_c. We want to emphasize, however, that this factor is strongly implementation-dependent and we assume bigger speed-up factors to be possible.

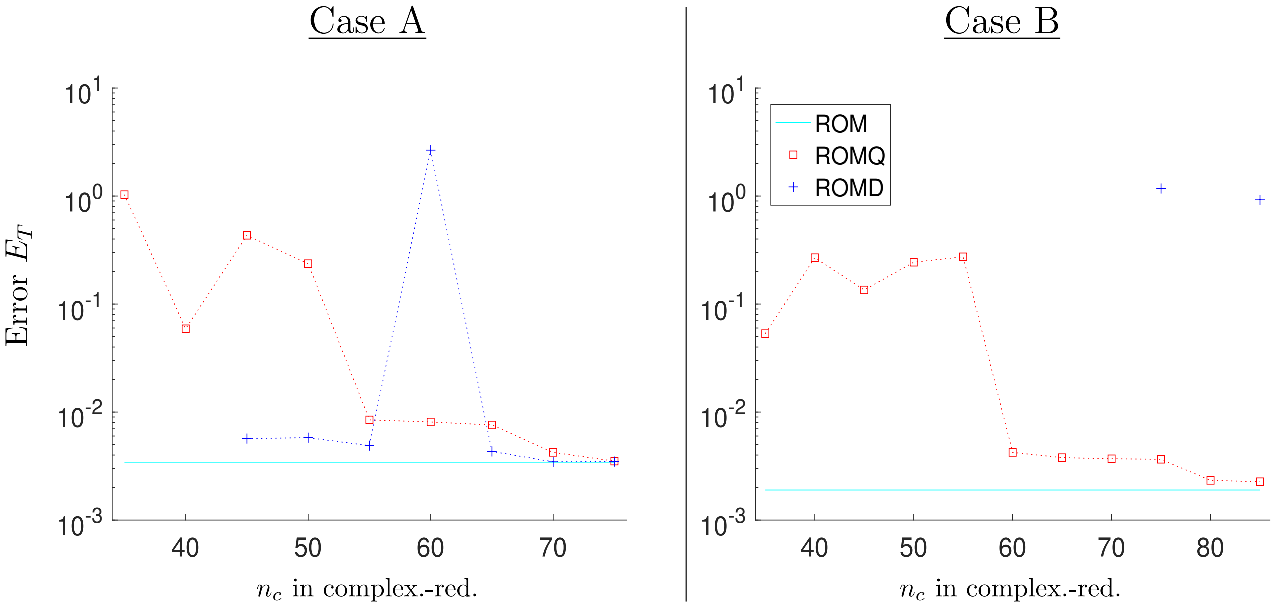

Figure 5.15: Large network, Scenario 5.4, relative errors E_T of complexity-reduced models. Underlying reduced dimension $n = 40$. *Left:* Case A (trained); *Right:* Case B (not trained).

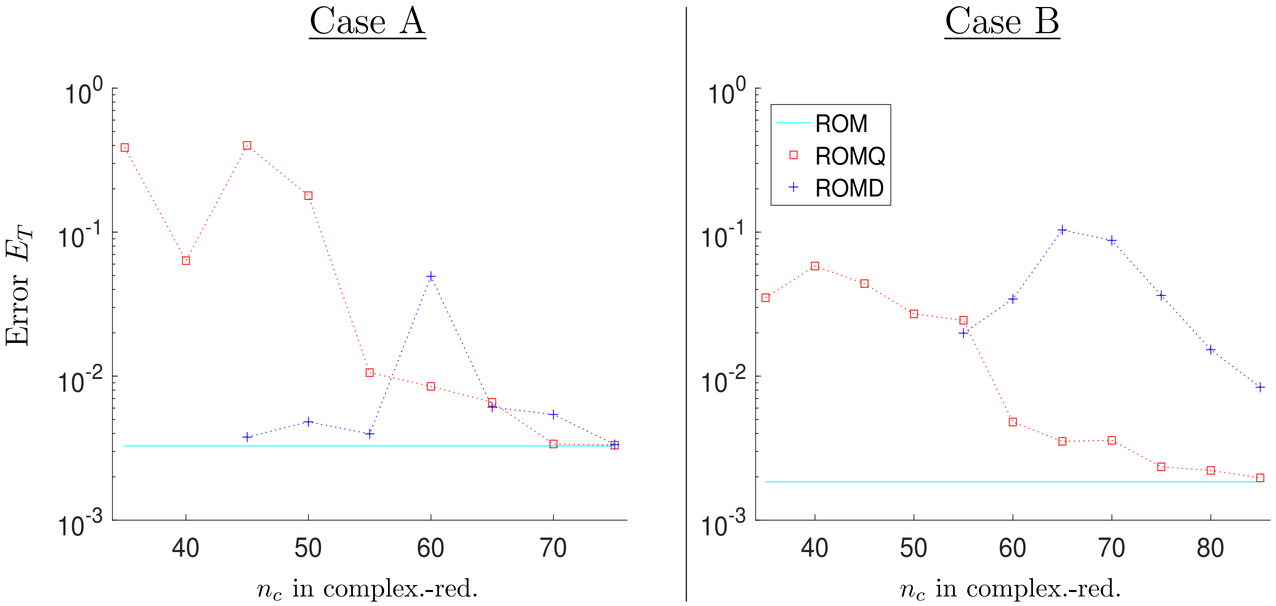

Figure 5.16: Large network, Scenario 5.4 with model modified to ISO2. Relative errors E_T of complexity-reduced models. Underlying reduced dimension $n = 40$. *Left:* Case A (trained); *Right:* Case B (not trained).

Complexity reduction for simpler model

Recall that our discretization of ISO1 involves nonlinearities in the time derivative as well, which represents a further challenge for the complexity reduction. By changing the model to ISO2, we have a setup without nonlinearities in the time derivatives, cf. Section 4.2. In *ROMD*, where all nonlinearities are complexity-reduced independently of each other, we thus can omit to approximate the now linear time-derivative-part. In the notation of Section 3.4, this relates to the choice $\mathbf{D}_z = \mathbf{I}$. We repeat Scenario 5.4 with this simplification and compare the performance of the complexity reduction methods in Fig. 5.16. Both *ROMQ* and *ROMD* perform better in terms of errors. In particular, the simulations in *ROMD* all run to completion for $n_c \geq 55$. Nonetheless, the same basic observations as in Fig. 5.16 can be made: The structure-preserving *ROMQ* model is more robust and has no parameter setting with simulation breakdowns. But its error is also not monotonically decreasing for increasing n_c on the whole parameter domain. The speed-up factor we observe in this setting for *ROMQ* in comparison to FOM is about a factor of four in our implementation (≈50 sec. vs. ≈200 sec.) for the highest tested order n_c.

5.5 Compatibility condition for complexity reduction

As discussed in Section 3.3, there is one part of our compatibility-conditions Assumption 2.20, which is notoriously violated for non-structure-preserving complexity reduction in our setting. That part is (3.7), i.e.,

$$||b|| \leq \tilde{C}||b||_c, \qquad \text{for } b \in \mathcal{K}, \quad \mathcal{K} = \{w \in \mathcal{V}_2 : \partial_x w \equiv 0\}.$$

Recall that the constant $\tilde{C}$ can be seen as a stability-constant and can be estimated from certain mass matrices, cf. the discussion in Section 3.3. In particular, it is bounded by the the condition number of the matrix $\mathbf{M}_{c,2}$ from (3.6),

$$\operatorname{cond}(\mathbf{M}_{c,2}) = ||\mathbf{M}_{c,2}||_2 \left|\left|\mathbf{M}_{c,2}^{-1}\right|\right|_2.$$

Recall further that a pre-selection step is assigned towards the fulfillment of (3.7) in Algorithm 3.15. In the upcoming, it is showcased that this pre-selection step does lower the condition number $\operatorname{cond}(\mathbf{M}_{c,2})$ and leads to overall more stable reduced results.

In what follows, a modified complexity-reduced model *ROMQ-nc* is compared to our proposed *ROMQ*. The modification *ROMQ-nc* is constructed to disregard for the compatibility condition, but to be otherwise similar to *ROMQ*. Concretely, we construct it by Algorithm 3.15 without its pre-selection step. In the notation of Algorithm 3.15, this relates to setting $\mathbf{w}_{qu}^0 = \mathbf{0}$.

We repeat our perfectly trained case, Scenario 5.4-Case A, for *ROMQ* and its non-compatible modification *ROMQ-nc*. From the plot Fig. 5.17-*left* of the resulting errors, we can see that *ROMQ-nc* even fails for a few parameter settings in contrast to *ROMQ*. Also, if it does not fail, it performs worse for almost all choices of n_c. Further, *ROMQ-nc* generally inherits significantly larger condition numbers $\operatorname{cond}(\mathbf{M}_{c,2})$ than our proposed method, Fig. 5.17-*right*, which is to be expected. Further, by comparing Fig. 5.17-*left* and -*right*, a coherence of $\operatorname{cond}(\mathbf{M}_{c,2})$ and the performance of the reduced models is apparent. Also the non-monotonicity of the errors in *ROMQ* w.r.t. n_c seem to correlate to a non-monotonicity in the condition number of $\operatorname{cond}(\mathbf{M}_{c,2})$. We want to stress that the error and $\operatorname{cond}(\mathbf{M}_{c,2})$ stabilize for *ROMQ* for large enough n_c. In our test case that is for $n_c \geq 55$. This stabilization can, on the other hand, not be observed for *ROMQ-nc*.

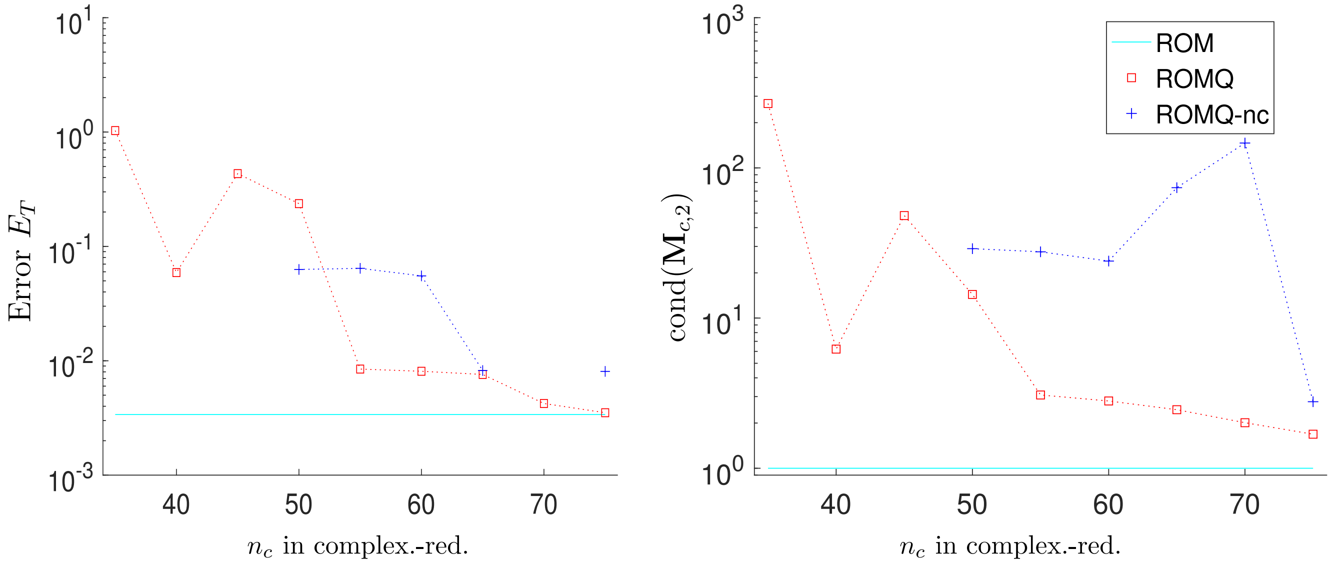

Figure 5.17: Large network, Scenario 5.4-Case A (trained). Comparing proposed *ROMQ* and non-compatible modification without pre-selection in Algorithm 3.14, *ROMQ-nc*. Underlying reduced dimension $n = 40$. *Left:* Rel. error E_T; *Right:* Condition number $cond(\mathbf{M}_{c,2})$ w.r.t. the Euclidean norm of complexity-reduced mass matrix.

Discussion and conclusion

Our aim in Part I was the systematic construction of online-efficient, structure-preserving reduced models for the approximation of a nonlinear flow problem on networks. In contrast to standard model reduction methods, our proposed procedure provides approximations that inherit the mass conservation and energy dissipation, which are fundamental features of the nonlinear flow problem. Furthermore, our reduced models can be shown to be of port-Hamiltonian structure. Most of our derivations relied on a careful energy-based modeling and appropriate variational principles. The discussion was divided into Part I.A and Part I.B.

In Part I.A we focused on the linear damped wave equation as a simple instance of our flow problem. The analysis for this special case relies on the Galerkin framework [EK18], [Kug19]. We studied the algebraic setting, which results after space discretization, in detail. This includes the differential-algebraic structure related to the network-character, and the interplay of the function-space setting and algebraic compatibility conditions, cf. Assumption 2.20.

We then embedded the results in a model reduction context, cf. Assumption 3.5, and proposed stable numerical procedures for the construction of compatible reduced models. In this regard, the problem of finding compatible bases has been analyzed, which showed some similarities to a problem occurring in so-called symplectic model reduction [PM16], [AH17]. We also showed in Theorem 3.16 that this problem significantly simplifies, when moment matching is chosen as the underlying model reduction procedure. With our subsequent numerical studies, we could verify the superior structural properties of our proposed structure-preserving approach compared to standard methods.

In Part I.B the approximation of the general nonlinear flow problem was then approached. In its treatment, we were faced with several additional challenges compared to the linear case: The energy-based modeling and the formulation of variational principles appropriate for structure-preserving approximations are more involved due to the nonlinearities. Our approach especially features a nonlinear parametrization of the energy variables of the solution, cf. Theorem 2.8, and makes heavy use of the theory around the partial Legendre transform. Notably, our variational characterization allows for a very convenient inclusion of the network-aspect, similar to the linear case.

Another new aspect was the need for complexity reduction for the treatment of nonlinearities. In this regard, we used a quadrature-type ansatz. The proposed complexity-reduced models were then shown to fulfill an energy-dissipation equality, Theorem 2.24. Furthermore, port-Hamiltonian structure could be verified, see Theorem 2.33, given appropriate compatibility conditions are posed on the ansatz spaces, Assumption 2.12, and the complexity reduction, Assumption 2.20. The former assumption is similar to the linear case, and the latter one implies that the complexity-reduced approximation of the $\mathcal{L}^2$-norm is a norm on the reduced ansatz space, cf. [EKLS20].

Moreover, we proposed adapted structure-preserving time discretization schemes for our space-discrete models, as existing structure-preserving methods like [LM17], [MM19],[CMKO11] could not be applied due to our nonlinear parametrization of the solution. Our schemes, which are an implicit-Euler-type scheme, a discontinuous Galerkin scheme and continuous Galerkin scheme in time, were all shown to either fulfill an energy-dissipation inequality, Theorem 2.38 and Theorem 2.43, or an energy dissipation equality in the latter case, Theorem 2.46.

In Chapter 3 a practical realization of the model order reduction and complexity reduction in our framework was discussed. The former consisted of an adaption of proper orthogonal decomposition and the latter of a greedy snapshot-based procedure for the training of quadrature weights. The emphasis was on the inclusion of the compatibility conditions in the implementation, which were exploited in our analysis.

The application of our approximation framework was then discussed in more detail using the barotropic Euler equations on networks as the underlying model. This discussion includes a well-posedness result of our proposed complexity-reduced models, Theorem 4.8.

To conclude Part I.B, we illustrated different aspects of our theoretical findings for the example of the isothermal Euler equations on networks. The parameter regimes and settings were oriented towards the simulation of gas network systems, cf. [SAB$^+$17], [DHLT17], [HGCATRM09]. Notably, we observed our proposed method to have good stability properties, particularly compared to popular standard methods like the discrete empirical interpolation method.

However, the numerical studies in Part I.B also indicated that the snapshot-based heuristics for the construction of the quadrature weights should be further studied, as some fluctuations in their fidelity w.r.t. the chosen degrees of freedom could be observed. Other possible questions for future research include the construction of compatible reduced spaces by a greedy method, cf. Part I.A-Remark 3.11, or the rigorous inclusion of quadrature-type approximation in time, Part I.B-Remark 2.49. Moreover, the extension of our arguments to more general situations, such as multi-dimensional systems or non-barotropic Euler equations, might be interesting and seems feasible.

Part II

Input-tailored system-theoretic model order reduction

Introduction

In this part we introduce a new system-theoretic model order reduction method for quadratic-bilinear dynamical systems of the form

$$\begin{aligned} \mathbf{E}\dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}\otimes\mathbf{x} + \mathbf{D}\mathbf{x}\otimes\mathbf{u} + \mathbf{B}\mathbf{u}, \quad t \geq 0 \\ \mathbf{y} &= \mathbf{C}\mathbf{x}, \qquad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^N, \qquad \mathbf{u}(t) \in \mathbb{R}^p, \quad t \geq 0 \end{aligned}$$

with nonsingular matrix $\mathbf{E}$ and Kronecker-tensor product $\otimes$, i.e., $\mathbf{x}\otimes\mathbf{x} \in \mathbb{R}^{N^2}$ and $\mathbf{x}\otimes\mathbf{u} \in \mathbb{R}^{Np}$. The system characterizes a map $\mathbf{u} \mapsto \mathbf{y}$ from typically low-dimensional input $\mathbf{u}$ to low-dimensional output $\mathbf{y}$ via a high-dimensional state $\mathbf{x}$. For a cheaper-to-evaluate reduced model, we seek for an appropriate basis matrix $\mathbf{V} \in \mathbb{R}^{N,n}$, $n \ll N$, and define the reduced model as

$$\begin{aligned} \mathbf{E}_r\dot{\mathbf{x}}_r &= \mathbf{A}_r\mathbf{x}_r + \mathbf{G}_r\mathbf{x}_r\otimes\mathbf{x}_r + \mathbf{D}_r\mathbf{x}_r\otimes\mathbf{u} + \mathbf{B}_r\mathbf{u} \\ \tilde{\mathbf{y}} &= \mathbf{C}_r\mathbf{x}_r, \qquad \mathbf{x}_r(0) = \mathbf{V}^T\mathbf{x}_0 \in \mathbb{R}^n \end{aligned}$$

with

$$\begin{aligned} &\mathbf{E}_r = \mathbf{V}^T\mathbf{E}\mathbf{V}, \qquad \mathbf{A}_r = \mathbf{V}^T\mathbf{A}\mathbf{V}, \qquad \mathbf{G}_r = \mathbf{V}^T\mathbf{G}\mathbf{V}\otimes\mathbf{V} \\ &\mathbf{B}_r = \mathbf{V}^T\mathbf{B}, \qquad \mathbf{D}_r = \mathbf{V}^T\mathbf{D}\mathbf{V}\otimes\mathbf{I}_p, \qquad \mathbf{C}_r = \mathbf{C}\mathbf{V}, \end{aligned}$$

and unit matrix $\mathbf{I}_p$ of dimension p. System-theoretic methods for linear systems are based on the frequency representation of the input-output map, which is a univariate algebraic mapping, called transfer function. For moment matching, the reduction basis $\mathbf{V}$ is developed such that the transfer function of the reduced model fulfills certain interpolation conditions. In the nonlinear case, the input-output map does in general not have a univariate frequency representation. Relaxations of the linear notions are needed to generalize it to the nonlinear case, see, e.g., recent multi-moment matching methods for multivariate frequency representations [ABJ16], [Gu12], [BB15], [ABJ16], [GAB15], [BB12b]. In our approach we pursue an other idea by using the following three relaxation steps:

1. Instead of considering the input-output map $\mathbf{u} \mapsto \mathbf{y}$ for arbitrary $\mathbf{u}$, we assume the input itself to be described by an autonomous quadratic differential system, the *signal generator*. The input-output system *driven by the signal generator* can then also be characterized by an enlarged autonomous output system, i.e., a system without any input, see Fig. 1 for an illustration.

2. We construct a *variational expansion* of the autonomous signal generator driven system w.r.t. its initial conditions. This results in an infinite series of linear systems.

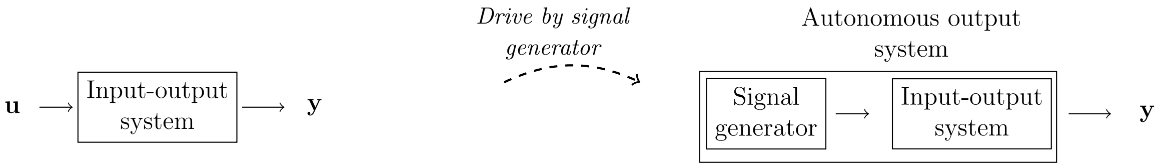

Figure 1: *Left:* Classical input-output modeling. *Right:* Modeling the same situation with an autonomous output system by replacing the external input with a respective signal generator.

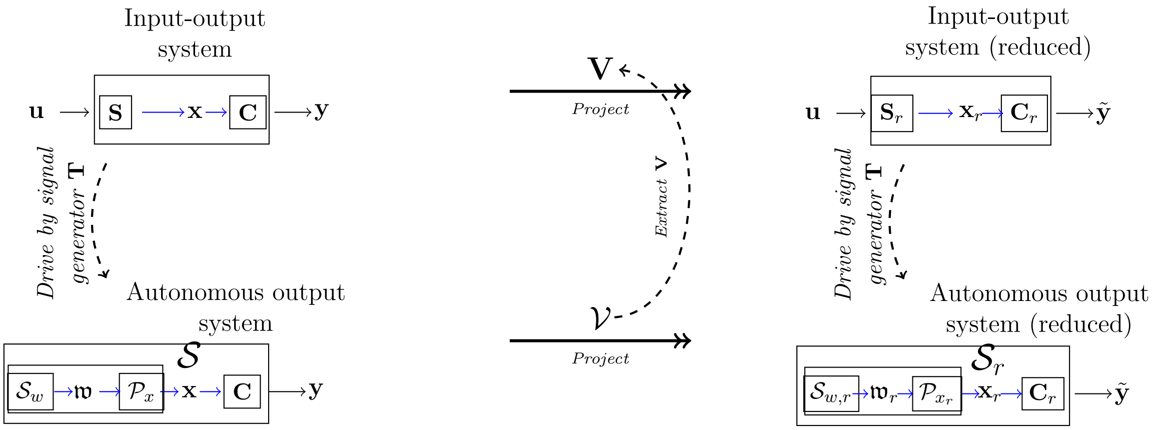

Figure 2: Sketch for input-tailored moment matching being based on the signal generator driven system and reduction via Galerkin projection.

3. For the first few terms of the variational expansion we construct univariate frequency representations and perform an *approximate moment matching.* This means the determination of the reduction basis $\mathbf{V}$ corresponds to approximating certain interpolation conditions for the univariate representations.

The idea of using signal generators for model reduction can also be found in [Ast10a], [Ast10b], [IA13]. But apart from that our approach is very different from theirs, as ours relies on variational expansions and by that considers families of solutions. In particular, our relaxation steps (1) and (2) induce a new input-tailored variational expansion of the state $\mathbf{x}$ of the high-dimensional dynamical system. The work that probably shares most similarities with ours, and which initially inspired us to look deeper into the subject, is [ZLW$^+$12], [ZW16]. The common feature is the univariate frequency representations derived for a variational expansion. Nonetheless, our approach exhibits profound differences to the former: Using the concept of signal generators we develop a framework that allows us to derive the variational expansion more rigorous and general. Our analysis suggests additional tensor-structured approximation conditions to be incorporated. Regarding the cascade- and low-rank tensor-structure present in the approximation problems yields a more efficient implementation. The latter point is crucial for practical usage, as the involved univariate frequency representations grow vastly in dimension when considered as unstructured linear ones. It turns out that the *exact* moment matching idea pursued classically in model reduction has to be relaxed to an *approximate* moment matching owed to the tensor structure of the problem. In this respect, our input-tailored moment matching is more involved as the multi-moment approaches [ABJ16], [BG17], [Gu12], [BB15], [BB12b]. However, our method corresponds to a one-dimensional interpolation problem unlike the multi-moment approaches corresponding to multi-dimensional interpolation problems. The latter consequently involve the choice of more expansion frequencies in multi-dimensional frequency space compared to ours, which involves fewer expansion frequencies to be chosen from a one-dimensional frequency space. A further difference to other system-theoretic reduction approaches is that ours extends very naturally to systems with more general input relations, such as, e.g., nonlinear functions and time derivatives. In this respect it is similarly flexible as the trajectory-based reduction methods like proper orthogonal decomposition [KV01], [AH14]. As a byproduct of the extension of our method to more general input relations, we also derive a respective extension for system-theoretic methods relying on multivariate frequency representations by incorporating input-weights. Although the use of input-weights in model reduction is not new [VA02], [BBG15], they have, to the best of the authors' knowledge, not been applied for this purpose before.

Outline

The outline of this part is as follows: The concept of a system to be driven by a signal generator as well as the proposed variational expansion and associated univariate frequency representation of the resulting autonomous system are presented in Chapter 1. We refer to the expansion and frequency representations as input-tailored, as they take into account the input described by the signal generator. The approximation conditions, which our reduction method aims for, resembles an approximate moment matching condition of the input-tailored frequency representations (Chapter 2). In this context the commuting diagram of Fig. 2 also takes a prominent role. In Chapter 3 our numerical realization is discussed. We particularly discuss the ability of handling non-standard input dependencies in our method in Section 4.1. In this context we also suggest an extension for other system-theoretic methods to handle non-standard input maps, which falls off as a byproduct of the discussion of our approach. The remainder of Chapter 4 provides expressions for higher-order univariate frequency representations as well as generalizations of the variational expansion and generalizations. The performance of our input-tailored moment matching method in comparison to the system-theoretic multi-moment matching and the trajectory-based proper orthogonal decomposition as well as the proposed handling of non-standard input maps are numerically studied in Chapter 5.

Chapter 1

Input-tailored expansion and frequency representation

In this chapter we develop our input-tailored variational expansion and frequency representation our reduction method is based on. Starting point is the concept of signal generator driven systems (Section 1.1). These signal generator driven systems are, by construction, autonomous. Variational expansions of autonomous systems and associated univariate frequency representations are the topic of Section 1.2. Then our input-tailored expansion and frequency representation are presented in Section 1.3. In Section 1.4, we embed our proposed expansion in the existing literature by relating it to the Volterra series and its frequency representations.

1.1 Signal generator driven system

In focus of this part are quadratic-bilinear dynamical systems of the form

$$\mathbf{S}: \quad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{\textcircled{2}} + \mathbf{D}\mathbf{x} \otimes \mathbf{u} + \mathbf{B}\mathbf{u}, \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^N \tag{1.1a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}, \quad \mathbf{u}(t) \in \mathbb{R}^p, \quad t \geq 0 \tag{1.1b}$$

with $\mathbf{E}$ nonsingular, all system matrices constant and $\mathbf{G} \in \mathbb{R}^{N,N^2}$, $\mathbf{D} \in \mathbb{R}^{N,Np}$. By slight abuse of notation, we identify throughout the part the realization of the state equation $\mathbf{S}$ with its input-to-state map $\mathbf{S} : \mathbf{u} \mapsto \mathbf{x}$.

Instead of considering $\mathbf{S}$ directly as abstract map, we use the concept of a system to be driven by a signal generator. A signal generator is an autonomous differential system describing the input $\mathbf{u}$. We employ here the class of signal generators with quadratic nonlinearities given as

$$\mathbf{T}: \qquad \mathbf{u} = \mathbf{C}_z\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_z\mathbf{z} + \mathbf{G}_z\mathbf{z}^{\textcircled{2}}, \qquad \mathbf{z}(0) = \mathbf{z}_0 \in \mathbb{R}^q. \tag{1.1c}$$

Remark 1.1 (Signal generators). *For example, an oscillation $u(t) = a\sin(\lambda t)$ for $t \geq 0$ and $a, \lambda \in \mathbb{R}$ is readily given by the signal generator*

$$u = [1\,|\,0]\mathbf{z}, \qquad \dot{\mathbf{z}} = \lambda \begin{bmatrix} & 1 \\ -1 & \end{bmatrix} \mathbf{z}, \qquad \mathbf{z}(0) = \begin{bmatrix} 0 \\ a \end{bmatrix}.$$

More generally, any linear combination of exponential pulses and sine- and cosine-oscillations can be described by a linear signal generator (as in (1.1c) *with* $\mathbf{G}_z = \mathbf{0}$*) by superposition of simple signal generators. Taking, e.g.,* $u(t) = a_1 \exp(\lambda_1 t) + a_2 \cos(\lambda_2 t)$*, the associated signal generator reads*

$$u = [1\,|\,0\,|1]\mathbf{z}, \qquad \dot{\mathbf{z}} = \begin{bmatrix} \lambda_1 & & \\ & & \lambda_2 \\ & -\lambda_2 & \end{bmatrix} \mathbf{z}, \qquad \mathbf{z}(0) = \begin{bmatrix} a_1 \\ 0 \\ a_2 \end{bmatrix}.$$

Arbitrary derivatives in frequency space of the above mentioned functions, such as e.g., $u(t) = t^k \exp(t)$ *for* $k \in \mathbb{N}$*, can also be described with linear signal generators. We refer to [ALM08], [Ast10a]. With nonlinear signal generators an even larger class of inputs can be described, see e.g., [ALM08], [Ast10a] for a broader discussion and some applications, or Section 5.2, Case 2 for an example of a quadratic signal generator.*

Similar to [Ast10a], [Ast10b], [IA13], the notion of a system to be driven by a signal generator is defined in the upcoming. It results from inserting a signal generator for the input $\mathbf{u}$ in system $\mathbf{S}$.

Definition 1.2 (Signal generator driven system)**.** *Let a quadratic-bilinear system* $\mathbf{S}$ *with an input* $\mathbf{u}$ *described by the signal generator* $\mathbf{T}$ *as in* (1.1) *be given,*

$$\begin{aligned} \mathbf{S}: &\quad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{\textcircled{2}} + \mathbf{D}\mathbf{x} \otimes \mathbf{u} + \mathbf{B}\mathbf{u}, && \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^N \\ \mathbf{T}: &\quad \mathbf{u} = \mathbf{C}_z\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_z\mathbf{z} + \mathbf{G}_z\mathbf{z}^{\textcircled{2}}, && \mathbf{z}(0) = \mathbf{z}_0 \in \mathbb{R}^q. \end{aligned}$$

Let $\mathbf{Q}$ *be the constant matrix such that*

$$\mathbf{Q}\begin{bmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{z}} \end{bmatrix}^{\textcircled{2}} = \begin{bmatrix} \bar{\mathbf{x}}^{\textcircled{2}} \\ \bar{\mathbf{x}} \otimes \bar{\mathbf{z}} \\ \bar{\mathbf{z}}^{\textcircled{2}} \end{bmatrix} \qquad \textit{for arbitrary } \bar{\mathbf{x}} \in \mathbb{R}^N,\ \bar{\mathbf{z}} \in \mathbb{R}^q.$$

Then we call the autonomous system

$$\mathcal{S}: \begin{aligned} &\mathcal{E}\dot{\mathfrak{w}} = \mathcal{A}\mathfrak{w} + \mathcal{G}\mathfrak{w}^{\textcircled{2}}, && \mathfrak{w}(0) = \mathfrak{b} \\ &\mathbf{x} = \mathcal{P}_x\,\mathfrak{w} \end{aligned}$$

with

$$\begin{aligned} &\mathcal{E} = \begin{bmatrix} \mathbf{E} & \\ & \mathbf{I}_q \end{bmatrix}, \qquad \mathcal{A} = \begin{bmatrix} \mathbf{A} & \mathbf{B}\mathbf{C}_z \\ & \mathbf{A}_z \end{bmatrix}, \qquad \mathcal{P}_x = [\mathbf{I}_N,\, \mathbf{0}], \qquad \mathfrak{b} = \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{z}_0 \end{bmatrix}, \\ &\mathcal{G} = \begin{bmatrix} \mathbf{G} & \mathbf{D}(\mathbf{I}_N \otimes \mathbf{C}_z) & \\ & & \mathbf{G}_z \end{bmatrix} \mathbf{Q} \end{aligned}$$

the signal generator driven system $\mathcal{S}$*.*

By definition, the solution $\mathbf{x}$ of system $\mathbf{S}$ for input $\mathbf{u}$ described by the signal generator $\mathbf{T}$ and the output $\mathbf{x}$ of the signal generator driven system $\mathcal{S}$ coincide. For an illustration, we refer to Fig. 2, left column. Note that the state equation of $\mathcal{S}$ (denoted by $\mathcal{S}_w$ in Fig. 2) is autonomous.

Remark 1.3. *Quantities associated to the original input-output system (e.g.* $\mathbf{S}$, $\mathbf{A}$, $\mathbf{x}$*) are written in another typesetting than the ones associated to the signal generator driven system (e.g.* $\mathcal{S}$, $\mathcal{A}$, $\mathfrak{w}$*). Moreover, frequency representations are written in a curved font (e.g.* $\mathscr{X}$, $\mathscr{W}$*), cf. Fig. 2.*

1.2 Variational expansion of autonomous systems

Our approach employs a variational expansion of the autonomous system $\mathcal{S}$ from Definition 1.2 and associated univariate frequency representations. The theoretical basis is given by the following theorem.

Theorem 1.4. *Let an α-dependent initial value problem of the autonomous quadratic differential equation*

$$\begin{aligned} \mathcal{E}\dot{\mathfrak{w}}(t;\alpha) &= \mathcal{A}\mathfrak{w}(t;\alpha) + \mathcal{G}\left(\mathfrak{w}(t;\alpha)\right)^{\textcircled{2}}, \qquad t \in (0,T) \\ \mathfrak{w}(0;\alpha) &= \alpha\mathfrak{b} \end{aligned}$$

be given for $T > 0$ and constant system matrices $\mathcal{E}, \mathcal{A} \in \mathbb{R}^{M,M}$, $\mathcal{G} \in \mathbb{R}^{M,M^2}$ and $\mathfrak{b} \in \mathbb{R}^M$ with $\mathcal{E}$ nonsingular. For parameter $\alpha \in I$, $0 \in I \subset \mathbb{R}$ bounded interval, the family of α-dependent solutions $\mathfrak{w}(\cdot,\alpha)$ can then be expanded as

$$\mathfrak{w}(t;\alpha) = \sum_{i=1}^{N} \alpha^i \mathfrak{w}_i(t) + O(\alpha^{N+1}), \qquad t \in [0,T), \quad \alpha \in I. \tag{1.2}$$

The univariate frequency representations $\breve{\mathscr{W}}_i$ of the first three functions $\mathfrak{w}_i$ for $s \in \mathbb{C}$ are

$$\breve{\mathscr{W}}_1(s) = (s\mathcal{E} - \mathcal{A})^{-1}\mathfrak{b} \tag{1.3a}$$

$$\breve{\mathscr{W}}_2(s) = (s\mathcal{E} - \mathcal{A})^{-1}\mathcal{G}\left(s\mathcal{E}^{\textcircled{2}} - \textcircled{2}_{\mathcal{E}}\mathcal{A}\right)^{-1}\mathfrak{b}^{\textcircled{2}} \tag{1.3b}$$

$$\breve{\mathscr{W}}_3(s) = 2(s\mathcal{E} - \mathcal{A})^{-1}\mathcal{G}\left(s\mathcal{E}^{\textcircled{2}} - \textcircled{2}_{\mathcal{E}}\mathcal{A}\right)^{-1}\mathcal{G} \otimes \mathcal{E}\left(s\mathcal{E}^{\textcircled{3}} - \textcircled{3}_{\mathcal{E}}\mathcal{A}\right)^{-1}\mathfrak{b}^{\textcircled{3}}. \tag{1.3c}$$

The proof of Theorem 1.4 relies on a variational expansion w.r.t. the initial conditions and on frequency space formulations using the so-called Associated Transform [Rug81]. Formal similarities to univariate frequency representations of [ZLW⁺12], [ZW16] are addressed and exploited within our proof. The variational expansion w.r.t. the initial conditions in Theorem 1.4 is particularly based on the following well-known result (Theorem 1.5) for which we state a proof for completeness.

Theorem 1.5. *Consider the α-dependent differential equation*

$$\begin{aligned} \dot{\mathfrak{w}}(t;\alpha) &= \mathfrak{f}(t,\mathfrak{w}(t;\alpha)) \qquad t \in (0,T) \\ \mathfrak{w}(0;\alpha) &= \hat{\mathfrak{b}} + \alpha\mathfrak{b}, \qquad \textit{with } \hat{\mathfrak{b}}, \mathfrak{b} \in \mathbb{R}^M \end{aligned}$$

for $T > 0$, $\mathfrak{f}$ being $N+1$ times continuously differentiable w.r.t. $\mathfrak{w}$ and continuous w.r.t. t. For $\alpha \in I$, $I \subset \mathbb{R}$ being a bounded interval containing zero, the family of α-dependent solutions $\mathfrak{w}(\cdot,\alpha)$ can be expanded as

$$\mathfrak{w}(t;\alpha) = \mathfrak{w}_0(t) + \sum_{i=1}^{N} \alpha^i \mathfrak{w}_i(t) + O(\alpha^{N+1}), \qquad t \in [0,T).$$

Proof. With the regularity assumptions on the right hand side $\mathfrak{f}$, unique solutions are given by the Picard-Lindelöf theorem for all $\alpha \in I$. Moreover, the solution $\mathfrak{w}$ is $N+1$ times continuously differentiable w.r.t. α. For both statements we refer to, e.g., [Har02, Sec. 5.4], [Chi06, Sec. 1]. Therefore, a Taylor series in α around $\alpha = 0$ gives

$$\mathfrak{w}(t;\alpha) = \mathfrak{w}_0(t) + \sum_{i=1}^{N} \alpha^i \mathfrak{w}_i(t) + \mathrm{O}(\alpha^{N+1})$$
$$\text{with } \mathfrak{w}_i(t) := \frac{1}{i!}\frac{\partial^i}{\partial \alpha^i}\mathfrak{w}(t;\alpha)_{|\alpha=0}.$$

$\square$

Furthermore, we use the following technical result from [BB12a], [Bre13].

Lemma 1.6. *Let* $\mathbf{P}, \mathbf{A} \in \mathbb{R}^{M,M}$, $\mathbf{B} \in \mathbb{R}^{M,K}$, $\mathbf{C} \in \mathbb{R}^{K,M}$, $\mathbf{D} \in \mathbb{R}^{K,K}$, *and let*

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_M \otimes \begin{bmatrix} \mathbf{I}_M \\ \mathbf{0}_{K,M} \end{bmatrix} & \mathbf{I}_M \otimes \begin{bmatrix} \mathbf{0}_{M,K} \\ \mathbf{I}_K \end{bmatrix} \end{bmatrix}.$$

Then it holds

$$\mathbf{M}^T \left(\mathbf{P} \otimes \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \right) \mathbf{M} = \begin{bmatrix} \mathbf{P}\otimes\mathbf{A} & \mathbf{P}\otimes\mathbf{B} \\ \mathbf{P}\otimes\mathbf{C} & \mathbf{P}\otimes\mathbf{D} \end{bmatrix}.$$

Moreover, $\mathbf{M}$ *is a permutation matrix and therefore orthogonal, i.e.,* $\mathbf{M}^{-1} = \mathbf{M}^T$.

Let us now turn to the proof of Theorem 1.4.

Proof. (Of Theorem 1.4) Theorem 1.5 with $\mathfrak{f}(t,\mathfrak{w}) = \mathcal{E}^{-1}(\mathcal{A}\mathfrak{w} + \mathcal{G}\mathfrak{w}^{\textcircled{2}})$ can be used to get

$$\mathfrak{w}(t;\alpha) = \sum_{i=1}^{N} \alpha^i \mathfrak{w}_i(t) + \mathrm{O}(\alpha^{N+1}).$$

The term $\mathfrak{w}_0$ scaling with α^0 drops out here as the solution for $\alpha = 0$ is $\mathfrak{w} \equiv \mathbf{0}$. Inserting this series representation into the differential equation and equating equal powers in α, we get

$$\begin{aligned} \mathcal{E}\dot{\mathfrak{w}}_1 &= \mathcal{A}\mathfrak{w}_1, & \mathfrak{w}_1(0) &= \mathfrak{b} \\ \mathcal{E}\dot{\mathfrak{w}}_2 &= \mathcal{A}\mathfrak{w}_2 + \mathcal{G}\mathfrak{w}_1^{\textcircled{2}}, & \mathfrak{w}_2(0) &= \mathbf{0} \\ \mathcal{E}\dot{\mathfrak{w}}_3 &= \mathcal{A}\mathfrak{w}_3 + \mathcal{G}\left(\mathfrak{w}_1 \otimes \mathfrak{w}_2 + \mathfrak{w}_2 \otimes \mathfrak{w}_1\right), & \mathfrak{w}_3(0) &= \mathbf{0}. \end{aligned}$$

On the upper equation for $\mathfrak{w}_1$, the standard Laplace-transform can be done, e.g., [Ant05], which gives the unique univariate frequency representation $\breve{\mathscr{W}}_1$ of $\mathfrak{w}_1$. Moreover, formally rewriting the equation for $\mathfrak{w}_1$ with the help of a Dirac impulse as

$$\mathcal{E}\dot{\mathfrak{w}}_1 = \mathcal{A}\mathfrak{w}_1 + \mathfrak{b}\delta(t), \qquad \lim_{\bar{t}\uparrow 0} \mathfrak{w}_1(\bar{t}) = \mathbf{0}$$

does not change its Laplace transform. Also multivariate frequency representations of $\mathfrak{w}_i$, $i = 2, 3$ can now be constructed following the standard procedure [Gu12], [Rug81], [ZW16]. To construct the desired univariate associated frequency representations $\breve{\mathscr{W}}_i$, the Associated Transform [Rug81] can be applied to the respective multivariate frequency representations of $\mathfrak{w}_i$. This step has already been performed for exactly our set of equations (using the Dirac impulse expression in the equation for $\mathfrak{w}_1$) in [ZW16], [ZLW$^+$12], see Remark 1.13. Therefore, our associated frequency representations coincide with their formally derived ones, and we can reuse their results. For $\breve{\mathscr{W}}_2$, the expression (1.3b) equals [ZW16, eq. (20)]. To derive expression (1.3c) for $\breve{\mathscr{W}}_3$, the following abbreviations are useful

$$\breve{\mathcal{E}}_2 = \begin{bmatrix}\mathcal{E} & \\ & \mathcal{E}^{\textcircled{2}}\end{bmatrix}, \quad \breve{\mathcal{A}}_2 = \begin{bmatrix}\mathcal{A} & \mathcal{G} \\ & \textcircled{2}_{\mathcal{E}}\mathcal{A}\end{bmatrix}, \quad \breve{\mathfrak{b}}_2 = \begin{bmatrix}\mathbf{0} \\ \mathfrak{b}^{\textcircled{2}}\end{bmatrix}, \quad \breve{\mathcal{C}}_2 = \begin{bmatrix}\mathbf{I}_M & \mathbf{0}\end{bmatrix},$$

cf. Lemma 1.8. Then expression [ZW16, eq. (23)] for $\breve{\mathscr{W}}_3$ is equivalent to

$$\begin{aligned}\breve{\mathscr{W}}_3(s) = {}& (s\mathcal{E} - \mathcal{A})^{-1}\mathcal{G} \\ & \Big[(\breve{\mathcal{C}}_2 \otimes \mathbf{I}_M)(s\breve{\mathcal{E}}_2 \otimes \mathcal{E} - (\breve{\mathcal{A}}_2 \otimes \mathcal{E} + \breve{\mathcal{E}}_2 \otimes \mathcal{A}))^{-1}(\breve{\mathfrak{b}}_2 \otimes \mathfrak{b}) \\ & + (\mathbf{I}_M \otimes \breve{\mathcal{C}}_2)(s\mathcal{E} \otimes \breve{\mathcal{E}}_2 - (\mathcal{E} \otimes \breve{\mathcal{A}}_2 + \mathcal{A} \otimes \breve{\mathcal{E}}_2))^{-1}(\mathfrak{b} \otimes \breve{\mathfrak{b}}_2)\Big].\end{aligned}$$

It remains to prove that this is equivalent to (1.3c). First we show that

$$\begin{aligned}&(\mathbf{I}_M \otimes \breve{\mathcal{C}}_2)(s\mathcal{E} \otimes \breve{\mathcal{E}}_2 - (\mathcal{E} \otimes \breve{\mathcal{A}}_2 + \mathcal{A} \otimes \breve{\mathcal{E}}_2))^{-1}(\mathfrak{b} \otimes \breve{\mathfrak{b}}_2) \\ &= (\breve{\mathcal{C}}_2 \otimes \mathbf{I}_M)(s\breve{\mathcal{E}}_2 \otimes \mathcal{E} - (\breve{\mathcal{A}}_2 \otimes \mathcal{E} + \breve{\mathcal{E}}_2 \otimes \mathcal{A}))^{-1}(\breve{\mathfrak{b}}_2 \otimes \mathfrak{b}).\end{aligned} \tag{1.4}$$

Using the respective orthogonal permutation matrix $\mathbf{M}$ from Lemma 1.6, we get

$$\begin{aligned}&(\mathbf{I}_M \otimes \breve{\mathcal{C}}_2)(s\mathcal{E} \otimes \breve{\mathcal{E}}_2 - (\mathcal{E} \otimes \breve{\mathcal{A}}_2 + \mathcal{A} \otimes \breve{\mathcal{E}}_2))^{-1}(\mathfrak{b} \otimes \breve{\mathfrak{b}}_2) \\ &= (\mathbf{I}_M \otimes \breve{\mathcal{C}}_2)\mathbf{M}(s\mathbf{M}^T\mathcal{E} \otimes \breve{\mathcal{E}}_2\mathbf{M} - \mathbf{M}^T(\mathcal{E} \otimes \breve{\mathcal{A}}_2 + \mathcal{A} \otimes \breve{\mathcal{E}}_2)\mathbf{M})^{-1}\mathbf{M}^T(\mathfrak{b} \otimes \breve{\mathfrak{b}}_2)\end{aligned}$$

Then by Lemma 1.6 we have

$$\begin{aligned}\mathbf{M}^T\mathcal{E} \otimes \breve{\mathcal{E}}_2\mathbf{M} &= \begin{bmatrix}\mathcal{E} \otimes \mathcal{E} & \\ & \mathcal{E} \otimes \mathcal{E}^{\textcircled{2}}\end{bmatrix} = \begin{bmatrix}\mathcal{E}^{\textcircled{2}} & \\ & \mathcal{E}^{\textcircled{3}}\end{bmatrix} = \breve{\mathcal{E}}_2 \otimes \mathcal{E} \\ \mathbf{M}^T(\mathcal{A} \otimes \breve{\mathcal{E}}_2 + \mathcal{E} \otimes \breve{\mathcal{A}}_2)\mathbf{M} &= \begin{bmatrix}\textcircled{2}_{\mathcal{E}}\mathcal{A} & \mathcal{G} \otimes \mathcal{E} \\ & \textcircled{3}_{\mathcal{E}}\mathcal{A}\end{bmatrix} = (\breve{\mathcal{A}}_2 \otimes \mathcal{E} + \breve{\mathcal{E}}_2 \otimes \mathcal{A}).\end{aligned}$$

A small calculation shows

$$\begin{aligned}\mathbf{M}^T\mathfrak{b} \otimes \breve{\mathfrak{b}}_2 &= \begin{bmatrix}\mathbf{0} \\ \mathfrak{b}^{\textcircled{3}}\end{bmatrix} = \breve{\mathfrak{b}}_2 \otimes \mathfrak{b} \\ \mathbf{I}_M \otimes \breve{\mathcal{C}}_2\mathbf{M} &= \breve{\mathcal{C}}_2 \otimes \mathbf{I}_M = [\mathbf{I}_{M^2}|\mathbf{0}],\end{aligned}$$

which together gives the equality (1.4). We therefore have

$$\begin{aligned}\breve{\mathscr{W}}_3(s) &= 2(s\mathcal{E} - \mathcal{A})^{-1}\mathcal{G}[\mathbf{I}_{M^2}|\mathbf{0}]\left[s\breve{\mathcal{E}}_2 \otimes \mathcal{E} - (\breve{\mathcal{A}}_2 \otimes \mathcal{E} + \breve{\mathcal{E}}_2 \otimes \mathcal{A})\right]^{-1}(\breve{\mathfrak{b}}_2 \otimes \mathfrak{b}) \\ &= 2(s\mathcal{E} - \mathcal{A})^{-1}\mathcal{G}\left(s\mathcal{E}^{\textcircled{2}} - \textcircled{2}_{\mathcal{E}}\mathcal{A}\right)^{-1}\mathcal{G} \otimes \mathcal{E}\left(s\mathcal{E}^{\textcircled{3}} - \textcircled{3}_{\mathcal{E}}\mathcal{A}\right)^{-1}\mathfrak{b}^{\textcircled{3}},\end{aligned}$$

i.e., representation (1.3c). In the last step we just used the upper-triangular structure of the matrix in the squared brackets to be inverted to factorize the term. □

Remark 1.7. *Certainly, the series in* (1.2) *can be formulated regarding terms of arbitrary high order in* α*. The tensor-structured explicit representations, however, get lengthy for high orders and the calculations more technical. In the main body of the part, we restrict ourselves from now on to terms up to order two to keep it more comprehensible. The tensor structure pattern that are observed and exploited for order two, are preserved for the expressions of higher order as well. For order three this can be seen in Theorem 1.4 and in respective generalizations of other important results provided in Section 4.3.*

Another point of view on the associated univariate frequency representation $\breve{\mathscr{W}}_2$ is highlighted in the following lemma that results from straight forward calculus (cf. Lemma 4.3 for $\breve{\mathscr{W}}_3$).

Lemma 1.8. *Assume that the requirements of Theorem 1.4 hold true. Then the associated frequency representation* $\breve{\mathscr{W}}_2$ *can be formulated with the linear representation*

$$\breve{\mathscr{W}}_2(s) = \breve{\mathcal{C}}_2 \left(s\breve{\mathcal{E}}_2 - \breve{\mathcal{A}}_2 \right)^{-1} \breve{\mathfrak{b}}_2,$$

$$\textit{with} \quad \breve{\mathcal{E}}_2 = \begin{bmatrix} \mathcal{E} & \\ & \mathcal{E}^{\textcircled{2}} \end{bmatrix}, \quad \breve{\mathcal{A}}_2 = \begin{bmatrix} \mathcal{A} & \mathcal{G} \\ & \textcircled{2}_{\mathcal{E}}\mathcal{A} \end{bmatrix}, \quad \breve{\mathfrak{b}}_2 = \begin{bmatrix} \mathbf{0} \\ \mathfrak{b}^{\textcircled{2}} \end{bmatrix}, \quad \breve{\mathcal{C}}_2 = \begin{bmatrix} \mathbf{I}_M & \mathbf{0} \end{bmatrix}.$$

Remark 1.9 (Cascade- and tensor-structure of associated frequency representations)**.** *The frequency representation* $\breve{\mathscr{W}}_1$ *associated to the first order term of the variational expansion is a usual linear input-to-state transfer function with dimension* M *equal to the dimension of the state* $\mathfrak{w}$*. According to Lemma 1.8 (and Lemma 4.3), also the higher-order terms possess linear state representations, which will strongly motivate our subsequently proposed procedure for setting up the approximation conditions in the approximate moment matching. However, since the frequency representations are of growing dimension,* $\mathbb{R}^{M+M^2}$ *for* $\breve{\mathscr{W}}_2$ *(*$\mathbb{R}^{M+M^2+M^3}$ *for* $\breve{\mathscr{W}}_3$*), operating directly on them – as done in [ZLW⁺12], [ZW16] – is unpractical for medium- to large-scale problems. For the development of a numerically tractable method, we instead exploit their special cascade- and tensor-structure that is revealed in Theorem 1.4. For example,* $\breve{\mathscr{W}}_2$ *can be interpreted as the cascade of the transfer functions* $\mathcal{G}\left(s\mathcal{E}^{\textcircled{2}} - \textcircled{2}_{\mathcal{E}}\mathcal{A}\right)^{-1}\mathfrak{b}^{\textcircled{2}}$ *and* $(s\mathcal{E} - \mathcal{A})^{-1}$*, where the former has low-rank tensor structure.*

1.3 Input-tailored variational expansion

Based on the notion of a system to be driven by a signal generator, we can now formulate our input-tailored expansion.

Definition 1.10 (Input-tailored variational expansion)**.** *Let the signal generator driven system* $\mathcal{S}$ *with enlarged state* $\mathfrak{w}$ *be as in Definition 1.2. Let*

$$\mathfrak{w}(t;\alpha) = \sum_{i=1}^{N} \alpha^i \mathfrak{w}_i(t) + O(\alpha^{N+1}), \qquad t \in [0,T)$$

be the variational expansion of $\mathfrak{w}$ *w.r.t. the initial conditions* $\mathfrak{w}(0;\alpha) = \alpha\mathfrak{b}$*. Let* $\breve{\mathscr{W}}_i$ *be the associated univariate frequency representations of* $\mathfrak{w}_i$ *as in Theorem 1.4.*

Then the input-tailored variational expansion of $\mathbf{x}$ *described by* $\mathcal{S}$ *(respectively by* $\mathbf{S}$ *and* $\mathbf{T}$*) is defined as*

$$\mathbf{x}(t;\alpha) = \sum_{i=1}^{N} \alpha^i \mathbf{x}_i(t) + O(\alpha^{N+1}), \qquad \mathbf{x}_i(t) = \mathcal{P}_x \, \mathfrak{w}_i(t)$$

with $\mathcal{P}_x = [\mathbf{I}_N, \mathbf{0}_{N,q}]$*. The input-tailored frequency representations* $\breve{\mathscr{X}}_i$ *are given as*

$$\breve{\mathscr{X}}_i(s) = \mathcal{P}_x \breve{\mathscr{W}}_i(s), \qquad s \in \mathbb{C}.$$

Let us emphasize that our input-tailored variational expansion is not tailored towards a single solution trajectory, but rather towards a family of solutions parametrized in the expansion parameter α. Given, e.g., the signal generator

$$u = [1\,|0]\mathbf{z}, \qquad \dot{\mathbf{z}} = \lambda \begin{bmatrix} & 1 \\ -1 & \end{bmatrix} \mathbf{z} \qquad \mathbf{z}(0) = \alpha \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \alpha \in \mathbb{R},$$

it relates to the inputs $u(t) = \alpha \sin(\lambda t)$, i.e., oscillations of varying amplitude.

Remark 1.11 (Possible generalizations)**.** *We point out that the definition of signal generator driven systems, Definition 1.2, and with that our whole approach can be generalized straightforwardly to systems with more sophisticated input maps, e.g., quadratic inputs, time derivatives, see Section 4.1.*

Moreover, the variational expansion from Theorem 1.4 itself can be generalized. Instead of considering families of solutions parametrized in initial conditions that dependent only on the single parameter α*, also families of solutions parametrized in a multidimensional parameter can be treated, see Section 4.4. This includes solutions parametrized in inputs* $u(t) = \sum_j \alpha_j u_j(t)$ *for varying* α_j*, where all* u_j *have a linear signal generator.*

1.4 Relation to Volterra series

In the following, we discuss the relation of our input-tailored variational expansion with the Volterra series, which is a variational expansion of the solution w.r.t. the input. The Volterra series has recently been extensively used as a basis for model reduction. For example, multi-moment matching has been discussed in [Gu12], [BB12b], [BB12c], hermite multi-moment matching in [BB15], [ABJ16], [BGG18], and balanced truncation in [BG17]. We recapitulate the variational ansatz from [Rug81], [LK78], [Gil77]. As the references are restricted to the scalar input case $u : \mathbb{R} \to \mathbb{R}$, we also use this restriction for convenience. Consider the state equation $\mathbf{S}$ with a scalar-valued input and trivial initial conditions, i.e.,

$$\mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{\textcircled{2}} + u\,\mathbf{D}\mathbf{x} + \mathbf{b}u, \quad \mathbf{x}(0) = \mathbf{0}, \quad u : \mathbb{R} \to \mathbb{R}$$

with $\mathbf{b} \in \mathbb{R}^N$. For appropriate input $u(t) = \alpha v(t)$ with $\alpha \in \mathbb{R}$ being sufficiently small and the system being uniquely solvable in an α-neighborhood containing zero, a variational expansion in the input holds, i.e., the solution can be expanded in α for $N > 0$ as

$$\mathbf{x}(t;\alpha) = \sum_{i=1}^{N} \alpha^i \mathbf{x}_i(t) + \mathrm{O}(\alpha^{N+1}) \qquad t \in [0, T) \tag{1.5}$$

for some $T > 0$. It can be shown, using the multivariate Laplace transform as in [BGG18], that the terms $\mathbf{x}_i$ have multivariate frequency representations $\mathscr{X}_i$ with

$$\begin{aligned} \mathscr{X}_1(s_1) &= \mathscr{G}_1(s_1)\mathscr{U}(s_1), \\ \mathscr{X}_2(s_1, s_2) &= \mathscr{G}_2(s_1, s_2)\mathscr{U}(s_1)\mathscr{U}(s_2) \\ \mathscr{X}_i(s_1, s_2, \ldots, s_i) &= \mathscr{G}_i(s_1, s_2, \ldots, s_i)\mathscr{U}(s_1)\mathscr{U}(s_2)\ldots\mathscr{U}(s_i), \qquad s_i \in \mathbb{C}, \quad i \leq N \end{aligned}$$

where $\mathscr{U}$ is the Laplace transform of the input u and $\mathscr{G}_i$ are the so-called symmetric transfer functions, see [Rug81], [LP06], [ZW16] for details on them. The model reduction methods relying on the Volterra series (1.5) typically formulate approximation conditions for the transfer functions $\mathscr{G}_i$.

At first glance there seems not to be a connection to our input-tailored variational expansion. The upcoming lemma, however, shows that for inputs described by linear signal generators, both expansions lead to the same result.

Lemma 1.12. *For a quadratic-bilinear differential system* $\mathbf{S}$ *with scalar input and trivial initial conditions, where the input is described by a signal generator* $\mathbf{T}$ *being linear, i.e.,*

$$\begin{aligned} \mathbf{S}: &\quad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{②} + u\,\mathbf{D}\mathbf{x} + \mathbf{b}u, && \mathbf{x}(0) = \mathbf{0}, \\ \mathbf{T}: &\quad u = \mathbf{C}_{\mathbf{z}}\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_{\mathbf{z}}\mathbf{z}, && \mathbf{z}(0) = \mathbf{z}_0, \end{aligned}$$

the same expansion of the solution

$$\mathbf{x}(t; \alpha) = \sum_{i=1}^{k} \alpha^i \mathbf{x}_i(t) + O(\alpha^{k+1})$$

can be obtained by the following two approaches:

a) By the input-tailored variational expansion of $\mathbf{x}$ *as in Definition 1.10.*

b) By the Volterra series: Expand the state $\mathbf{x}$ *for input* $u(t) = \alpha u_1(t)$ *in* α*, and then set the input* u *to be as in the linear signal generator.*

Proof. Proceeding from Approach a) we show the equality to Approach b). In Approach a) we assume for initial value $\mathfrak{w}_0 = \alpha[\mathbf{0}; \bar{\mathbf{z}}_0]$ that the extended state can be expanded as

$$\mathfrak{w}(t; \alpha) = \begin{bmatrix} \mathbf{x}(t; \alpha) \\ \mathbf{z}(t; \alpha) \end{bmatrix} = \sum_{i=1}^{k} \alpha^i \begin{bmatrix} \mathbf{x}_i(t) \\ \mathbf{z}_i(t) \end{bmatrix} + O(\alpha^{k+1}).$$

From the signal generator relation $\mathbf{T}$ it then follows

$$u(t; \alpha) = \sum_{i=1}^{k} \alpha^i u_i(t) + O(\alpha^{k+1}), \qquad u_i(t) = \mathbf{C}_{\mathbf{z}}\mathbf{z}_i(t).$$

As the signal generator is linear, it is easily seen that $\mathbf{z} \equiv \alpha\mathbf{z}_1$, thus also $u \equiv \alpha u_1$. Therefore, the expansion terms $\alpha^i\mathbf{x}_i$ scale with $u^i \equiv \alpha^i u_1^i$ as in Approach *b*), and hence the expansion terms $\mathbf{x}_i$ of both approaches coincide. □

Inputs described by linear signal generators are an important case. Alternatively to the derivation in [BGG18], the multivariate symmetric transfer functions $\mathscr{G}_i$ can already be derived by considering the response to sums of exponential functions

$$u(t) = \sum_{k=1}^{i} a_k \exp(\lambda_k t), \qquad \text{for arbitrary } a_k, \lambda_k \in \mathbb{R}$$

only, which is, e.g., used in the growing exponential approach, [Rug81], [Bre13]. Clearly, sums of exponential functions can be described by linear signal generators, cf. Remark 1.1. Therefore, loosely spoken, the associated univariate input-tailored frequency representation tailored towards the upper growing exponentials for different choices a_k, λ_k resemble the multivariate transfer functions $\mathscr{G}_i$. The works [LW13], [ZW16] indirectly heavily rely on the upper resemblance, but do not explicitly elaborate on it.

Finally, let us comment on the more formal approach by [ZW16], [ZLW$^+$12] that leads to similar univariate frequency representations as ours.

Remark 1.13. *In [ZW16], [ZLW$^+$12] the quadratic-bilinear equation of Theorem 1.4 with zero (pre-)initial conditions but an initial jump is considered, i.e.,*

$$\mathcal{E}\dot{\mathfrak{w}} = \mathcal{A}\mathfrak{w} + \mathcal{G}\mathfrak{w}^{②} + \mathfrak{b}u(t), \quad u(t) = \alpha\delta(t), \quad \lim_{\bar{t}\uparrow 0} \mathfrak{w}(\bar{t}) = \mathbf{0},$$

where $\delta(t)$ is the Dirac-impulse. There the solution $\mathfrak{w}$ is expanded formally as Volterra series with that distributional input $u(t) = \alpha\delta(t)$, yielding the same expansion terms as ours. However, the validity of the Volterra series when the input is a Dirac-impulse is not covered by the classical result on Volterra series expansions – as far as the authors know (cf., e.g., [Rug81], [LK78], [Gil77] or [Bor10]). This issue is also not further addressed or discussed in the respective works.

Chapter 2

Input-tailored system-theoretic model reduction framework

Aim of our method is to construct a reduced model such that for the input-tailored frequency representations $\breve{\mathscr{X}}_i$ the so-called moments

$$\frac{d^k}{ds^k}\breve{\mathscr{X}}_i(s)_{|s=s_0} \qquad \text{for } k,\ i,\ s_0 \text{ given}$$

of the full order model are *approximately* matched by their reduced counterparts. This is a relaxation of the linear moment matching idea, which we recapitulate in Section 2.1. Our input-tailored moment matching problem is formulated in Section 2.2. The notion of a signal generator driven system $\mathcal{S}$ and its reduced counterpart is herefore essential. The structure of the approximation problem is analyzed in Section 2.3. From a theoretical point of view, it can be characterized with linear theory. To do so, a change to high-dimensional state representations (cf. Lemma 1.8) is needed. Our projection ansatz, however, operates on the lower-dimensional original representation with tensor structure, which is why the relaxation from *exact* to *approximate* moment matching is needed. The proposed conditions aiming for approximate moment matching are presented in Section 2.4.

2.1 Moments and linear theory

The basic theory of linear moment matching is recalled here for convenience, for further reading we refer to, e.g., [Ant05], [Gri97], [Ast10a], and references therein.

Definition 2.1 (Moments)**.** *Given a univariate frequency representation $\mathscr{H}$ being k-times differentiable at $s_0 \in \mathbb{C}$, its k-th moment at s_0 is defined as*

$$\mathbf{m}_k = \frac{(-1)^k}{k!}\frac{d^k}{ds^k}\mathscr{H}(s)_{|s=s_0}.$$

Note that the moments $\mathbf{m}_k$ are dependent on the expansion frequency s_0 chosen, which we, however, suppress in our notation to keep it shorter.

Lemma 2.2. *Let a frequency representation $\mathscr{H}$ have the form $\mathscr{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$. Let, for given s_0, $\mathbf{A}_{s_0} = -s_0\mathbf{E} + \mathbf{A}$ be nonsingular. Then the k-th moment of $\mathscr{H}$ at s_0 reads*

$$\mathbf{m}_k = -\mathbf{C}\left[\mathbf{A}_{s_0}^{-1}\mathbf{E}\right]^k \mathbf{A}_{s_0}^{-1}\mathbf{B}, \quad \textit{for } k \geq 0.$$

The moments can be determined as follows: Calculate $\mathbf{k}_i$, the moments of $s \mapsto (s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ at s_0, by the recursion

$$\begin{aligned} i = 0: &\quad \mathbf{A}_{s_0}\mathbf{k}_0 = -\mathbf{B} \\ i > 0: &\quad \mathbf{A}_{s_0}\mathbf{k}_i = \mathbf{E}\mathbf{k}_{i-1}. \end{aligned}$$

Then set $\mathbf{m}_k = \mathbf{C}\mathbf{k}_k$.

For linear systems, reduced models fulfilling moment matching can be constructed by means of the following lemma.

Lemma 2.3. *Let $\mathscr{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ with $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{N,N}$, and let for given reduction basis $\mathbf{V} \in \mathbb{R}^{N,n}$ the reduced system be defined as*

$$\begin{aligned} &\mathscr{H}_r(s) = \mathbf{C}_r(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r \\ &\textit{with} \quad \mathbf{E}_r = \mathbf{V}^T\mathbf{E}\mathbf{V}, \quad \mathbf{A}_r = \mathbf{V}^T\mathbf{A}\mathbf{V}, \quad \mathbf{B}_r = \mathbf{V}^T\mathbf{B}, \quad \mathbf{C}_r = \mathbf{C}\mathbf{V}. \end{aligned}$$

If for prescribed s_0, it holds

$$\textit{span}\{\mathbf{k}_0, \mathbf{k}_1, \ldots, \mathbf{k}_k\} \subseteq \textit{im}\,(()\mathbf{V}) \tag{2.1}$$

for $\mathbf{k}_i$, $0 \leq i \leq k$, as defined in Lemma 2.2, then the (exact) moment matching condition

$$\frac{d^i}{ds^i}\mathscr{H}(s)_{|s=s_0} = \frac{d^i}{ds^i}\mathscr{H}_r(s)_{|s=s_0}, \qquad i \leq k \tag{2.2}$$

is satisfied. We say that the moments of the full and the reduced model match (up to k-th order at s_0). Moreover, it holds

$$\mathbf{k}_i = \mathbf{V}\mathbf{k}_{r,i}, \qquad i \leq k,$$

where $\mathbf{k}_{r,i}$ is recursively defined with $\mathbf{A}_{r,s_0} = -s_0\mathbf{E}_r + \mathbf{A}_r$ as

$$\begin{aligned} i = 0: &\quad \mathbf{A}_{r,s_0}\mathbf{k}_{r,0} = -\mathbf{B}_r \\ i > 0: &\quad \mathbf{A}_{r,s_0}\mathbf{k}_{r,i} = \mathbf{E}_r\mathbf{k}_{r,i-1}. \end{aligned}$$

Of course, projection errors play a crucial role in this kind of model reduction.

Lemma 2.4 (Error of projected solution)**.** *Let $\mathbf{V} \in \mathbb{R}^{N,n}$ be orthogonal. Let $\mathbf{b} \in \mathbb{R}^{N,p}$ and let $\mathbf{A} \in \mathbb{R}^{N,N}$, $\mathbf{A}_r = \mathbf{V}^T\mathbf{A}\mathbf{V}$ both be nonsingular, and*

$$\mathbf{X} := \mathbf{A}^{-1}\mathbf{b}, \qquad \mathbf{X}_r := \mathbf{A}_r^{-1}\mathbf{b}_r,$$

where $\mathbf{b}_r = \mathbf{V}^T\mathbf{b}$. Then the following approximation condition holds

$$\mathbf{X} - \mathbf{V}\mathbf{X}_r = \left[\mathbf{I} - \mathbf{V}\mathbf{A}_r^{-1}\mathbf{V}^T\mathbf{A}\right]\left[\mathbf{I} - \mathbf{V}\mathbf{V}^T\right]\mathbf{X},$$

where $\mathbf{I} - \mathbf{V}\mathbf{V}^T$ is the projector onto the orthogonal complement of the image of $\mathbf{V}$.

Lemma 2.4 can be shown by straight forward calculus. With the help of Lemma 2.4, and exploiting the recursive manner the moments can be defined, leads to an iterative proof of Lemma 2.3. It mainly relies on the fact that under condition (2.1) the projection error in the respective $\mathbf{k}_i$, i.e., $(\mathbf{I} - \mathbf{V}\mathbf{V}^T)\mathbf{k}_i$, is zero. It then follows iteratively that $\mathbf{V}\mathbf{k}_{r,i} = \mathbf{k}_i$, from which (2.2) can be deduced.

2.2 Reduced signal generator driven system

In this subsection we clarify our notion of a reduced signal generator driven system and its usage. We start by stating the basic result behind the commuting diagram sketched in Fig. 2.

Lemma 2.5. *Let a quadratic-bilinear system* $\mathbf{S}$*, a signal generator* $\mathbf{T}$*,*

$$\begin{aligned} \mathbf{S}: &\quad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{\textcircled{2}} + \mathbf{D}\mathbf{x} \otimes \mathbf{u} + \mathbf{B}\mathbf{u}, && \mathbf{x}(0) = \mathbf{x}_0, \\ \mathbf{T}: &\quad \mathbf{u} = \mathbf{C}_z\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_z\mathbf{z} + \mathbf{G}_z\mathbf{z}^{\textcircled{2}}, && \mathbf{z}(0) = \mathbf{z}_0, \end{aligned}$$

and the associated signal generator driven system $\mathcal{S}$*, as in Definition 1.2, be given. Let furthermore, for given reduction basis* $\mathbf{V} \in \mathbb{R}^{N,n}$*,* $n \ll N$*, the reduced state matrices be defined as*

$$\begin{aligned} &\mathbf{E}_r = \mathbf{V}^T\mathbf{E}\mathbf{V}, \qquad \mathbf{A}_r = \mathbf{V}^T\mathbf{A}\mathbf{V}, \qquad \mathbf{G}_r = \mathbf{V}^T\mathbf{G}\mathbf{V} \otimes \mathbf{V}, \\ &\mathbf{B}_r = \mathbf{V}^T\mathbf{B}, \qquad \mathbf{D}_r = \mathbf{V}^T\mathbf{D}\mathbf{V} \otimes \mathbf{I}_p. \end{aligned}$$

Let $\mathbf{Q}_r$ *be the constant matrix such that*

$$\mathbf{Q}_r \begin{bmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{z}} \end{bmatrix}^{\textcircled{2}} = \begin{bmatrix} \bar{\mathbf{x}}^{\textcircled{2}} \\ \bar{\mathbf{x}} \otimes \bar{\mathbf{z}} \\ \bar{\mathbf{z}}^{\textcircled{2}} \end{bmatrix} \qquad \textit{for arbitrary } \bar{\mathbf{x}} \in \mathbb{R}^n,\ \bar{\mathbf{z}} \in \mathbb{R}^q.$$

Introducing the reduced system as

$$\mathbf{S}_r: \qquad \mathbf{E}_r\dot{\mathbf{x}}_r = \mathbf{A}_r\mathbf{x}_r + \mathbf{G}_r\mathbf{x}_r \otimes \mathbf{x}_r + \mathbf{D}_r\mathbf{x}_r \otimes \mathbf{u} + \mathbf{B}_r\mathbf{u}, \qquad \mathbf{x}_r(0) = \mathbf{V}^T\mathbf{x}_0,$$

and setting up the signal generator driven system for $\mathbf{S}_r$ *and* $\mathbf{T}$ *gives*

$$\mathcal{S}_r: \quad \begin{aligned} \mathcal{E}_r\dot{\mathfrak{w}}_r &= \mathcal{A}_r\mathfrak{w}_r + \mathcal{G}_r\mathfrak{w}_r^{\textcircled{2}}, && \mathfrak{w}_r(0) = \mathfrak{b}_r, \\ \mathbf{x}_r &= \mathcal{P}_{x_r}\,\mathfrak{w}_r. \end{aligned}$$

with

$$\begin{aligned} &\mathcal{E}_r = \begin{bmatrix} \mathbf{E}_r & \\ & \mathbf{I}_q \end{bmatrix}, \quad \mathcal{A}_r = \begin{bmatrix} \mathbf{A}_r & \mathbf{B}_r\mathbf{C}_z \\ & \mathbf{A}_z \end{bmatrix}, \quad \mathcal{P}_{x_r} = [\mathbf{I}_n,\, \mathbf{0}], \quad \mathfrak{b}_r = \begin{bmatrix} \mathbf{V}^T\mathbf{x}_0 \\ \mathbf{z}_0 \end{bmatrix}, \\ &\mathcal{G}_r = \begin{bmatrix} \mathbf{G}_r & \mathbf{D}_r(\mathbf{I}_n \otimes \mathbf{C}_z) & \\ & & \mathbf{G}_z \end{bmatrix} \mathbf{Q}_r. \end{aligned}$$

Projecting the realization of $\mathcal{S}$ *as*

$$\begin{aligned} &\mathcal{E}_r = \mathcal{V}^T\mathcal{E}\mathcal{V}, \qquad \mathcal{A}_r = \mathcal{V}^T\mathcal{A}\mathcal{V}, \qquad \mathcal{G}_r = \mathcal{V}^T\mathcal{G}\mathcal{V} \otimes \mathcal{V} \qquad \mathfrak{b}_r = \mathcal{V}^T\mathfrak{b}, \\ &\textit{with reduction basis } \mathcal{V} = \begin{bmatrix} \mathbf{V} & \\ & \mathbf{I}_q \end{bmatrix}, \end{aligned}$$

and defining $\mathcal{P}_{x_r}$*, as above, leads to the same reduced signal generator driven system* $\mathcal{S}_r$*.*

The lemma is quite obvious, but nonetheless of high importance for us. The input-tailored frequency representations $\breve{\mathscr{X}}_{r,i}$ of $\mathcal{S}_r$ are accordingly obtained, as specified in Definition 1.10, by

$$\breve{\mathscr{X}}_{r,i}(s) = \mathcal{P}_{x_r}\breve{\mathscr{W}}_{r,i}(s), \qquad s \in \mathbb{C}$$

with $\breve{\mathscr{W}}_{r,i}$ being the frequency representation of the variational expansion terms $\mathfrak{w}_{r,i}$. The proposed *approximate* moment matching conditions we require on the reduced model $\mathbf{S}_r$ to be fulfilled are

$$\mathbf{V}\frac{d^k}{ds^k}\breve{\mathscr{X}}_{r,i}(s)_{|s=s_0} \overset{!}{\approx} \frac{d^k}{ds^k}\breve{\mathscr{X}}_i(s)_{|s=s_0} \qquad \text{for } k \leq L_i, \quad i \leq \bar{i}, \tag{2.3}$$

for L_i, $\bar{i}, s_0$ prescribed.

Remark 2.6 (Extracting reduction basis from extended problem)**.** *Note that the signal generator itself is not reduced in the construction of Lemma 2.5. This is also reflected in the block structure of $\mathcal{V}$ with a unit matrix block $\mathbf{I}_q$. Moreover, the lemma shows that projection and driving by a signal generator commute. Therefore, the input-tailored moment matching* (2.3) *can be approached in a two-step procedure:*

- *Find basis $\mathcal{V}$ such that*

$$\mathcal{V}\frac{d^k}{ds^k}\breve{\mathscr{W}}_{r,i}(s)_{|s=s_0} \overset{!}{\approx} \frac{d^k}{ds^k}\breve{\mathscr{W}}_i(s)_{|s=s_0} \qquad \textit{for } k \leq L_i, \quad i \leq \bar{i} \tag{2.4}$$

  holds, for L_i, $\bar{i}, s_0$ prescribed, where $\breve{\mathscr{W}}_i$, $\breve{\mathscr{W}}_{r,i}$ are as in Definition 1.10 given $\mathcal{S}$, $\mathcal{S}_r$.

- *Extract the basis $\mathbf{V}$ from $\mathcal{V}$.*

2.3 Input-tailored moments and projection

Up to now, the input-tailored moment matching problem has been tracked back to the extended problem (Remark 2.6), and it has been shown that the reduced signal generator driven system $\mathcal{S}_r$ can be seen as the projection of $\mathcal{S}$, Lemma 2.5. What remains to examine is the actual structure of the extended problem (2.4). It will be seen that, from a theoretical point of view, we can tackle the problem with linear theory by changing into high-dimensional linear representations.

Lemma 2.7 (Reduced associated frequency representation)**.** *Given the full order signal generator $\mathcal{S}$, and its reduced counterpart $\mathcal{S}_r$ as in Lemma 2.5, the reduced associated frequency representation $\breve{\mathscr{W}}_{r,2}$ is the Galerkin-projection of $\breve{\mathscr{W}}_2$ written in its high-dimensional linear representation of Lemma 1.8, i.e.,*

$$\begin{aligned}\breve{\mathscr{W}}_{r,2}(s) &= \breve{\mathcal{C}}_{r,2}\left(s\breve{\mathcal{E}}_{r,2} - \breve{\mathcal{A}}_{r,2}\right)^{-1}\breve{\mathfrak{b}}_{r,2} \\ &\textit{with} \quad \breve{\mathcal{E}}_{r,2} = \breve{\mathcal{V}}_2^T\breve{\mathcal{E}}_2\breve{\mathcal{V}}_2, \quad \breve{\mathcal{A}}_{r,2} = \breve{\mathcal{V}}_2^T\breve{\mathcal{A}}_2\breve{\mathcal{V}}_2, \quad \breve{\mathfrak{b}}_{r,2} = \breve{\mathcal{V}}_2^T\breve{\mathfrak{b}}_2, \quad \breve{\mathcal{C}}_{r,2} = \breve{\mathcal{C}}_2\breve{\mathcal{V}}_2, \\ &\textit{and} \quad \breve{\mathcal{V}}_2 = \begin{bmatrix}\mathcal{V}\\ \mathcal{V}^{\textcircled{2}}\end{bmatrix}.\end{aligned}$$

The proof is straight forward. Obviously, the inherent tensor-structure of the problem is handed over to the reduction basis $\breve{\mathcal{V}}_2$. Our method makes use of this special cascade- and tensor-structure that is also present in the moments, which we show in the upcoming.

Lemma 2.8. *For $s_0 \in \mathbb{C}$, $i > 0$ and quadratic matrices $\mathcal{E}, \mathcal{A}$ let $\mathcal{A}_{s_0} = -s_0\mathcal{E} + \mathcal{A}$. Then it holds*

$$\textcircled{i}_{\mathcal{E}}\mathcal{A}_{s_0/i} = -s_0\mathcal{E}^{\textcircled{i}} + \textcircled{i}_{\mathcal{E}}\mathcal{A}.$$

Proof. For $0 \leq k, m \leq i-1$, with $k+m+1 = i$ it holds

$$\mathcal{E}^{\textcircled{k}} \otimes \left(-\frac{s_0}{i}\mathcal{E} + \mathcal{A}\right) \otimes \mathcal{E}^{\textcircled{m}} = -\frac{s_0}{i}\mathcal{E}^{\textcircled{i}} + \mathcal{E}^{\textcircled{k}} \otimes \mathcal{A} \otimes \mathcal{E}^{\textcircled{m}}.$$

Since $\textcircled{i}_{\mathcal{E}}\mathcal{A}_{s_0/i}$ can be written as sum of i such expressions with $k = 0, \ldots, i-1$, and $m = i-k-1$, the lemma follows. □

A recursion formula for the moments $\mathbf{m}_i^{(2)}$ of $\breve{\mathscr{W}}_2$ can now be stated (cf. Theorem 4.4 for $\breve{\mathscr{W}}_3$). The super-index '$.^{(j)}$' in the moments is used throughout to indicate the correspondence to the j-th frequency representation $\breve{\mathscr{W}}_j$, $j = 2, 3$.

Theorem 2.9 (Extended input-tailored moments)**.** *Assume that the requirements of Theorem 1.4 and Lemma 1.8 hold, and let for given $s_0 \in \mathbb{C}$ the matrix $\mathcal{A}_{s_0} = -s_0\mathcal{E} + \mathcal{A}$ be nonsingular. Then the moments $\mathbf{m}_i^{(2)}$ of $\breve{\mathscr{W}}_2$ at s_0 are characterized by the recursion:*

$$\begin{aligned}
i = 0: \quad \textcircled{2}_{\mathcal{E}}\mathcal{A}_{s_0/2}\,\boldsymbol{\mu}_0^{(2)} &= -\mathfrak{b}^{\textcircled{2}} \\
\mathcal{A}_{s_0}\,\mathbf{m}_0^{(2)} &= -\mathcal{G}\,\boldsymbol{\mu}_0^{(2)} \\
i > 0: \quad \textcircled{2}_{\mathcal{E}}\mathcal{A}_{s_0/2}\,\boldsymbol{\mu}_i^{(2)} &= \mathcal{E}^{\textcircled{2}}\,\boldsymbol{\mu}_{i-1}^{(2)} \\
\mathcal{A}_{s_0}\,\mathbf{m}_i^{(2)} &= \mathcal{E}\,\mathbf{m}_{i-1}^{(2)} - \mathcal{G}\,\boldsymbol{\mu}_i^{(2)}.
\end{aligned}$$

Moreover, $\mathbf{k}_i^{(2)} = [\mathbf{m}_i^{(2)}; \boldsymbol{\mu}_i^{(2)}]$ are the moments of $s \mapsto \left(s\breve{\mathcal{E}}_2 - \breve{\mathcal{A}}_2\right)^{-1}\breve{\mathfrak{b}}_2$ at s_0.

Proof. The representation of Lemma 1.8 for $\breve{\mathscr{W}}_2$ is a linear state representation. Therefore, following Lemma 2.2, the factors $\mathbf{k}_i^{(2)}$ recursively defined by

$$\begin{aligned}
i = 0: \quad (-s_0\breve{\mathcal{E}}_2 + \breve{\mathcal{A}}_2)\mathbf{k}_0^{(2)} &= -\breve{\mathfrak{b}}_2 \\
i > 0: \quad (-s_0\breve{\mathcal{E}}_2 + \breve{\mathcal{A}}_2)\mathbf{k}_i^{(2)} &= \breve{\mathcal{E}}_2\,\mathbf{k}_{i-1}^{(2)},
\end{aligned}$$

are the moments of $s \mapsto \left(s\breve{\mathcal{E}}_2 - \breve{\mathcal{A}}_2\right)^{-1}\breve{\mathfrak{b}}_2$ at s_0. Let us introduce the following notation for the upper and lower blocks

$$\mathbf{k}_i^{(2)} = \begin{bmatrix} \mathbf{m}_i^{(2)} \\ \boldsymbol{\mu}_i^{(2)} \end{bmatrix}, \qquad \text{where } \mathbf{m}_i^{(2)} \in \mathbb{R}^M, \quad \boldsymbol{\mu}_i^{(2)} \in \mathbb{R}^{M^2}.$$

Then these blocks fulfill for $i > 0$

$$\begin{aligned}
(-s_0\mathcal{E} + \mathcal{A})\,\mathbf{m}_i^{(2)} + \mathcal{G}\,\boldsymbol{\mu}_i^{(2)} &= \mathcal{E}\,\mathbf{m}_{i-1}^{(2)} \\
(-s_0\mathcal{E}^{\textcircled{2}} + \textcircled{2}_{\mathcal{E}}\,\mathcal{A})\,\boldsymbol{\mu}_i^{(2)} &= \mathcal{E}^{\textcircled{2}}\,\boldsymbol{\mu}_{i-1}^{(2)}.
\end{aligned}$$

Using Lemma 2.8, we get the recursive expression for $\mathbf{m}_i^{(2)}$ for $i > 0$. The initial step $i = 0$ follows similarly. In fact, $\mathbf{m}_i^{(2)}$ is the i-th moment of $\breve{\mathscr{W}}_2$ at s_0, as it equals $\breve{\mathcal{C}}_2 \mathbf{k}_i^{(2)}$, which is the expression we get for the moment by applying the last part of Lemma 2.2. □

According to the linear theory, exact moment matching requires

$$\mathbf{k}_i^{(2)} \in \text{ image}(\breve{\mathcal{V}}_2). \tag{2.5}$$

This corresponds to a condition in a $(N+q)^2+(N+q)$-dimensional space. However, this condition cannot be fulfilled exactly because of the specific form our reduction basis has.

2.4 Approximation conditions

We propose an approximate moment matching that accounts for the special tensor structure of the problem.

Considering the reduction basis for $\breve{\mathscr{W}}_2$

$$\breve{\mathcal{V}}_2 = \begin{bmatrix} \mathcal{V}_2 \\ \mathcal{V}_2^{\textcircled{2}} \end{bmatrix},$$

we solve the following splitted problem: Find $\mathcal{V}_2$ such that it holds

$$||(\mathbf{I}_{N+q} - \mathcal{V}_2 \mathcal{V}_2^T)\, \mathbf{m}_i^{(2)}\, ||/||\, \mathbf{m}_i^{(2)}\, || \quad \textit{small} \text{ for } i = 0, 1, \ldots, L \tag{2.6a}$$

$$||(\mathbf{I}_{(N+q)^2} - \mathcal{V}_2^{\textcircled{2}} (\mathcal{V}_2^{\textcircled{2}})^T)\, \boldsymbol{\mu}_i^{(2)}\, ||/||\, \boldsymbol{\mu}_i^{(2)}\, || \quad \textit{small} \text{ for } i = 0, 1, \ldots, L \tag{2.6b}$$

for $\mathbf{m}_i^{(2)}$, $\boldsymbol{\mu}_i^{(2)}$ from Lemma 2.9. This aims for small projection errors

$$\left(\mathbf{I} - \breve{\mathcal{V}}_2 \breve{\mathcal{V}}_2^T\right) \mathbf{k}_i^{(2)}, \qquad \text{with } \mathbf{k}_i^{(2)} = [\mathbf{m}_i^{(2)}; \boldsymbol{\mu}_i^{(2)}],$$

which is a relaxation of the exact moment matching in (2.5).

In the assembly of the global reduction basis $\mathcal{V}$ that corresponds to all considered frequency representations $\breve{\mathscr{W}}_i$, $i \leq \bar{\imath}$, cf. (2.4), we provide a block structure of the form

$$\mathcal{V} = \begin{bmatrix} \mathbf{V} & \\ & \mathbf{I}_q \end{bmatrix}.$$

This reflects that the signal generator itself is not reduced and gives the desired reduction basis $\mathbf{V}$ of the original system.

Remark 2.10. *Let us stress the difference to former work on model reduction using univariate frequency representations for nonlinear systems. Comparing our approach with the one from [ZLW+12], [ZW16] there are, besides the more rigorous treatment of the variational expansion (cf. Remark 1.13), three major differences: The first and most important one is that our analysis reveals an additional tensor-structured approximation condition* (2.6b) *to naturally appear when aiming for approximate moment matching. Such a condition is not present in the former approach. Second, our framework using the concept of signal generator driven systems enables us to* *consider a larger class of input scenarios within the process. And finally, the inherent cascade- and sparse-tensor-structure has not been exploited in the former algorithmic implementation. It will be seen in Chapter 3 that the appearing tensor-structured problems can be formulated as Lyapunov-type equations with 'sparse right hand sides'. We deal with them using recently proposed low-rank solvers from literature, which is known to save memory- and time-effort by orders of magnitude, cf. [SKB16], [Sim07], [KT10].*

Chapter 3

Numerical realization

In this chapter we present and discuss the algorithms for the numerical realization of our input-tailored moment matching method.

3.1 Low-rank calculations of input tailored moments

The main part of the numerics consists in constructing the subspace for basis $\mathcal{V}$ such that (2.6) hold. Clearly, it is easy to construct a basis matrix $\mathcal{V}$ fulfilling (2.6a) exactly, namely just use the matrix composed of the moments $\mathbf{m}_i^{(2)}$ itself. The question remains, why a low-rank basis fulfilling (2.6b) should exist. Let us herefore look at the zeroth auxiliary moment $\boldsymbol{\mu}_0^{(2)}$ around s_0. It fulfills the equation

$$\left[\mathcal{E} \otimes \mathcal{A}_{s_0/2} + \mathcal{A}_{s_0/2} \otimes \mathcal{E}\right] \boldsymbol{\mu}_0^{(2)} + \mathfrak{b}^{②} = \mathbf{0},$$

which is the well-known Lyapunov equation, written in tensor notation, with a sparse 'right hand side' $\mathfrak{b}^{②}$. Low-rank solutions for these kind of equations exist under reasonable conditions [KT10], [Sim07], [BB12a], and take the form

$$\sum_{k=1}^{n_i} \mathbf{z}_i^k \otimes \mathbf{z}_i^k \approx \boldsymbol{\mu}_i^{(2)} \quad \text{for small } n_i. \tag{3.1}$$

For the higher order terms, e.g., $\boldsymbol{\mu}_1^{(2)}$, we suggest to follow up the iteration with the new sparse 'right hand side' $\mathcal{E}^{②}\,\boldsymbol{\mu}_0^{(2)}$, i.e., the low-rank approximation from the former step, and so on. By that, we do not only have a strategy to efficiently approximate $\boldsymbol{\mu}_i^{(2)}$ and $\mathbf{m}_i^{(2)}$ up to a certain extend, but also a candidate for a low-rank basis, namely the span over all $\mathbf{z}_i^k$. The upcoming Algorithm 3.1 summarizes our approach aiming towards (2.6).

Note that the moments involved are the ones for the signal generator driven system $\mathcal{S}$. Albeit the reduction basis $\mathbf{V}$ is constructed for the original system $\mathbf{S}$. Thus, the selection matrix $\mathcal{P}_x : \mathfrak{w} \mapsto \mathbf{x}$ appears here.

Algorithm 3.1 (Moment-matching-bases for $\breve{\mathscr{X}}_2$).
INPUT:

- *Realization matrices of signal generator driven system* $\mathcal{S}$ *(cf. Definition 1.2):* $\mathcal{E}$, $\mathcal{A}$, $\mathcal{G}$, $\mathfrak{b}$
- *Dimension of state variable* N*; Dimension of signal generator:* q
- *Expansion frequencies:* $(s_1, s_2, \ldots, s_\mu)$*; Number of moments:* $(L_1, L_2, \ldots, L_\mu)$
- *Tolerance for low-rank approximations: tol*
- *Basis for space not considered in low-rank approximation:* $\mathbf{V}_\perp$

OUTPUT: Reduction bases: $\mathbf{V}_a$, $\mathbf{V}_b$.

```
1. Set $\mathcal{P}_x = [\mathbf{I}_N,\ \mathbf{0}_{N,q}]$.
2. for $j = 1, \ldots \mu$
    a) Set $s_0 := s_j$ and $L := L_j$.
    b) Calculate low-rank factors $\mathbf{z}_i^k$ for $k = 1, \ldots, n_i$, $i = 0, \ldots, L-1$, see (3.1), i.e.,
       $$\mathbf{z}_i^k \text{ with: } \sum_{k=1}^{n_i} \mathbf{z}_i^k \otimes \mathbf{z}_i^k \approx \left( \left( \textcircled{2}_{\mathcal{E}} \mathcal{A}_{s_0/2} \right)^{-1} \mathcal{E}^{\textcircled{2}} \right)^i \left( \textcircled{2}_{\mathcal{E}} \mathcal{A}_{s_0/2} \right)^{-1} \mathfrak{b}^{\textcircled{2}}.$$
    c) Gather all $(\mathcal{P}_x \mathbf{z}_i^k)$ in $\mathbf{Z}_{s_j}$, i.e.,
       $$\mathbf{Z}_{s_j} := \mathcal{P}_x[\mathbf{z}_0^1, \mathbf{z}_0^2, \ldots, \mathbf{z}_0^{n_0}, \mathbf{z}_1^1, \ldots, \mathbf{z}_1^{n_1}, \ldots, \mathbf{z}_{L-1}^{n_{L-1}}]$$
    endfor
3. Gather all $\mathbf{Z}_{s_j}$ in $\mathbf{Z}$, i.e.,
       $$\mathbf{Z} := [\mathbf{Z}_{s_1}, \mathbf{Z}_{s_2}, \ldots, \mathbf{Z}_{s_\mu}]$$
4. for $j = 1, \ldots \mu$
    a) Set $s_0 := s_j$ and $L := L_j$.
    b) Calculate $\mathbf{m}_i^{(2)}$ for $s_0$ from Lemma 2.9 (using the low-rank approximations on $\boldsymbol{\mu}_i^{(2)}$ from
       Step (2b))
    c) Gather all $(\mathcal{P}_x \, \mathbf{m}_i^{(2)})$ in $\mathbf{M}_{s_j}$, i.e.,
       $$\mathbf{M}_{s_j} := \mathcal{P}_x[\mathbf{m}_0^{(2)}, \mathbf{m}_1^{(2)}, \ldots, \mathbf{m}_{L-1}^{(2)}]$$
    endfor
5. Construct $\mathbf{V}_a$ as orthogonal basis of $[\mathbf{M}_{s_1}, \ldots, \mathbf{M}_{s_j}]$.
```

6. *Define* $\mathbf{P}_\perp$ *as orthogonal projection onto the orthogonal complement of span of* $[\mathbf{V}_a, \mathbf{V}_\perp]$*. Then set* $\mathbf{V}_b$ *to consist of all left-singular vectors of* $(\mathbf{P}_\perp \mathbf{Z})$ *with singular value bigger than tol.*

In terms of numerical calculation, the most delicate step is the construction of the low-rank factors $\mathbf{z}_i^k$. Note that for each i in Step (2b) we actually need to construct a low-rank solution on a Lyapunov equation. The projection step with $\mathbf{P}_\perp$ removes components of the dominant space already present in the former constructed bases, and therefore allows for lower-order truncation in step (6).

3.2 Constructing the reduction basis

In this subsection we conclude our approach for the construction of a reduced model, which aims at approximate moment matching of the input-tailored frequency representations $\breve{\mathscr{X}}_1$, $\breve{\mathscr{X}}_2$ from Definition 1.10.

For the basis construction with regards to $\breve{\mathscr{X}}_1$, the signal generator does not need to be considered. This is because $\breve{\mathscr{W}}_1$ can be factored as

$$\breve{\mathscr{W}}_1(s) = \left[(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\right] \mathbf{C}_z (s\mathbf{I}_q - \mathbf{A}_z)^{-1}\mathbf{z}_0,$$

i.e., into the standard linear transfer function $(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ and the signal generator. As discussed in Chapter 2, the signal generator is not reduced, and therefore moment matching of the linear transfer function automatically imposes moment matching on $\breve{\mathscr{W}}_1$. Concluding, the following algorithm for the construction of a reduced model is proposed.

Algorithm 3.2 (Input-tailored approximate moment matching)**.** *INPUT:*

- *Realization matrices of the quadratic-bilinear dynamical system* $\mathbf{S}$ *to reduce:* $\mathbf{E}$*,* $\mathbf{A}$*,* $\mathbf{G}$*,* $\mathbf{D}$*,* $\mathbf{B}$*,* $\mathbf{C}$
- *Realization matrices of the signal generator* $\mathbf{T}$*:* $\mathbf{A}_z$*,* $\mathbf{G}_z$*,* $\mathbf{C}_z$
- *Initial value vectors:* $\mathbf{x}_0, \mathbf{z}_0$
- *Concerning* $\breve{\mathscr{X}}_2$*: Expansion frequencies:* $(s_1, s_2, \dots, s_\mu)$*; Number of moments to match:* $(L_1, L_2, \dots, L_\mu)$*; Tolerance for low-rank approximations in Algorithm 3.1: tol*
- *Concerning* $\breve{\mathscr{X}}_1$*: Expansion frequencies:* $(\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_\nu)$*; Number of moments to match:* $(\tilde{L}_1, \tilde{L}_2, \dots, \tilde{L}_\nu)$

OUTPUT: Reduced realization: $\mathbf{E}_r$*,* $\mathbf{A}_r$*,* $\mathbf{G}_r$*,* $\mathbf{D}_r$*,* $\mathbf{B}_r$*,* $\mathbf{C}_r$*.*

1. *Construct reduction basis* $\mathbf{V}_1$ *for* $\breve{\mathscr{X}}_1$ *as orthonormal basis*
   for the union of the Krylov spaces $\mathcal{K}_{\tilde{L}_j}(\mathbf{A}_{\tilde{s}_j}^{-1}\mathbf{E}, \mathbf{A}_{\tilde{s}_j}^{-1}\mathbf{b})$ *for* $j = 1, \dots \nu$*.*
2. *Construct realization for signal generator driven system* $\mathcal{S}$ *(Definition 1.2):* $\mathcal{E}$*,* $\mathcal{A}$*,* $\mathcal{G}$*,* $\mathfrak{b}$
3. *Construct reduction bases* $\mathbf{V}_a$*,* $\mathbf{V}_b$ *for* $\breve{\mathscr{X}}_2$ *by Algorithm 3.1 for*
   frequencies $(s_1, s_2, \dots, s_\mu)$*, number of moments* $(L_1, L_2, \dots, L_\mu)$*, tolerance tol and* $\mathbf{V}_\perp = \mathbf{V}_1$*.*

4. *Construct* $\mathbf{V}$ *as orthogonal basis of span of* $[\mathbf{V}_a, \mathbf{V}_b, \mathbf{V}_1]$.

5. *Calculate reduced state representation as* $\mathbf{E}_r = \mathbf{V}^T\mathbf{E}\mathbf{V}$, $\mathbf{A}_r = \mathbf{V}^T\mathbf{A}\mathbf{V}$, $\mathbf{G}_r = \mathbf{V}^T\mathbf{G}\mathbf{V}\otimes\mathbf{V}$, $\mathbf{D}_r = \mathbf{V}^T\mathbf{D}\mathbf{V}\otimes\mathbf{I}_p$, $\mathbf{B}_r = \mathbf{V}^T\mathbf{B}$, $\mathbf{C}_r = \mathbf{C}\mathbf{V}$.

For Step (1) in Algorithm 3.2 we just use the standard Krylov method as in [Gri97], [Ant05]. Note furthermore that in the calculation of $\mathbf{G}_r$ it is advisable to avoid the memory-demanding explicit calculation of $\mathbf{V}\otimes\mathbf{V}$, see [Bre13], which we also do.

Remark 3.3. *Algorithm 3.1 is only assumed to be stable if the order of moments matched* L_j *are all chosen moderate. This is, because we actually seek for a special so-called Krylov space. For matrices* $\mathbf{M}, \mathbf{L}$ *of appropriate dimension and* $L \in \mathbb{N}$ *the Krylov space is defined as*

$$\mathcal{K}_L(\mathbf{M}, \mathbf{L}) := \mathit{span}\left\{\left[\mathbf{L},\, \mathbf{M}\mathbf{L}, \ldots,\, \mathbf{M}^{L-1}\mathbf{L}\right]\right\}.$$

Then Step (2b)*, thought of in* $\mathbb{R}^{N^2}$*, consists in constructing the Krylov space*

$$\mathcal{K}_L\left(\left(\textcircled{2}_{\mathcal{E}}\mathcal{A}_{s_0/2}\right)^{-1}\mathcal{E}^{\textcircled{2}},\, \left(\textcircled{2}_{\mathcal{E}}\mathcal{A}_{s_0/2}\right)^{-1}\mathfrak{b}^{\textcircled{2}}\right)$$

without any orthogonalization between the iteration. This is known to be unstable for high orders, see, e.g., [Gri97], [Ant05]. However, orthogonalization in $\mathbb{R}^{N^2}$ *destroys our tensor structure. It is possible to recover a low-rank tensor structure by additional truncation, but this goes with further approximation errors [KK18]. Therefore, we recommend to match the moments at several frequencies* s_i *rather than at high-order moments as it is also usual practice for linear moment matching.*

Chapter 4

Generalizations

Our input-tailored method can be generalized in several directions. In particular, the handling of non-standard input dependencies is possible as shown in Section 4.1. To the best of the author's knowledge, this has not been discussed before for system-theoretic methods relying on multivariate frequency representations. We therefore also address a respective generalization for these methods in Section 4.2. It is based on the notion of input-weighting, which shows some formal similarities to our input-tailoring. Section 4.3 provides expressions for higher-order univariate frequency representations and Section 4.4 generalizations of the variational expansion.

4.1 Handling of non-standard input-dependencies

In practical applications the state equation $\mathbf{S}$ to reduce may take a more general form as in (1.1a), e.g.,

$$\mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{②} + \mathbf{D}\mathbf{x} \otimes \mathbf{u} + \mathbf{B}\mathbf{u} + \mathbf{K}(\mathbf{u})$$

with $\mathbf{K}(\mathbf{u})$ describing input dependencies not affine-linear in $\mathbf{u}$. For example, quadratic terms in the inputs can come from boundary control terms, when systems with quadratic nonlinearities are discretized, as shown for the Burgers' equation in Section 5.2. Also time derivatives in the input can appear, when the state equation $\mathbf{S}$ originates from an index-reduced differential-algebraic equation [KM06], [LMT13]. The usual work-around in system-theoretic model reduction is to introduce artificial augmented inputs for all non-standard terms. Obviously, this enlarges the input and ignores known input-structure, which leads to worse results in model reduction.

Our input-tailored approach can incorporate a large class of input-relations directly, as we discuss for some cases in the following.

Input map with quadratic term and/or time derivative For $\mathbf{K}(\mathbf{u}) = \mathbf{G}_u\mathbf{u}^{②} + \mathbf{B}_p\dot{\mathbf{u}}$ our signal generator driven system, and with that the core of our approach generalizes as follows.

Definition 4.1 (Generalization of Definition 1.2, Signal generator driven system)**.** *Let a system* $\mathbf{S}$ *of the following form with an input* $\mathbf{u}$ *described by the signal generator* $\mathbf{T}$ *(as in* (1.1c)*) be given*

$$\begin{aligned}
\mathbf{S}: &\quad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{②} + \mathbf{D}\mathbf{x} \otimes \mathbf{u} + \mathbf{B}\mathbf{u} + \mathbf{G}_u\mathbf{u}^{②} + \mathbf{B}_p\dot{\mathbf{u}}, && \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^N\\
\mathbf{T}: &\quad \mathbf{u} = \mathbf{C}_z\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_z\mathbf{z} + \mathbf{G}_z\mathbf{z}^{②}, && \mathbf{z}(0) = \mathbf{z}_0 \in \mathbb{R}^q.
\end{aligned}$$

Let $\mathbf{Q}$ *be the constant matrix such that*

$$\mathbf{Q}\begin{bmatrix}\bar{\mathbf{x}}\\ \bar{\mathbf{z}}\end{bmatrix}^{\textcircled{2}} = \begin{bmatrix}\bar{\mathbf{x}}^{\textcircled{2}}\\ \bar{\mathbf{x}}\otimes\bar{\mathbf{z}}\\ \bar{\mathbf{z}}^{\textcircled{2}}\end{bmatrix} \qquad \textit{for arbitrary } \bar{\mathbf{x}}\in\mathbb{R}^N,\ \bar{\mathbf{z}}\in\mathbb{R}^q.$$

Then we call the autonomous system

$$\mathcal{S}:\ \begin{aligned} \mathcal{E}\dot{\mathfrak{w}} &= \mathcal{A}\mathfrak{w} + \mathcal{G}\mathfrak{w}^{\textcircled{2}}, \qquad\qquad \mathfrak{w}(0) = \mathfrak{b}\\ \mathbf{x} &= \mathcal{P}_x\,\mathfrak{w} \end{aligned}$$

with

$$\mathcal{E} = \begin{bmatrix}\mathbf{E} & \\ & \mathbf{I}_q\end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix}\mathbf{A} & \mathbf{B}\mathbf{C}_z + \mathbf{B}_p\mathbf{C}_z\mathbf{A}_z\\ & \mathbf{A}_z\end{bmatrix}, \quad \mathcal{P}_x = [\mathbf{I}_N,\, \mathbf{0}], \quad \mathfrak{b} = \begin{bmatrix}\mathbf{x}_0\\ \mathbf{z}_0\end{bmatrix},$$
$$\mathcal{G} = \begin{bmatrix}\mathbf{G} & \mathbf{D}(\mathbf{I}_N\otimes\mathbf{C}_z) & \mathbf{G}_u\mathbf{C}_z\otimes\mathbf{C}_z + \mathbf{B}_p\mathbf{C}_z\mathbf{G}_z\\ & & \mathbf{G}_z\end{bmatrix}\mathbf{Q},$$

the signal generator driven system $\mathcal{S}$*.*

Note that the solution $\mathbf{x}$ of system $\mathbf{S}$ for input $\mathbf{u}$ described by the signal generator $\mathbf{T}$ and the output $\mathbf{x}$ of the signal generator-driven system $\mathcal{S}$ from the definition coincide.

Input map with higher-order time derivatives When higher-order time derivatives occur in the input map, the further procedure depends on the signal generator. If the signal generator is linear, we can use that for

$$\mathbf{u} = \mathbf{C}_z\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_z\mathbf{z} \qquad \mathbf{z}(0) = \mathbf{z}_0, \quad \text{it holds } \frac{d^i}{dt^i}\mathbf{u} = \mathbf{C}_z\mathbf{A}_z^i\mathbf{z}.$$

Thus, a signal generator driven system, which is quadratic in the extended state $[\mathbf{x};\mathbf{z}]$, can be directly constructed. Only the system matrices $\mathcal{A}$, $\mathcal{G}$ have to be slightly adjusted.

If the signal generator is nonlinear, we suggest to further extend the signal generator driven system. We exemplarily discuss this for the case of second order derivatives $\ddot{\mathbf{u}}$: Introduce $\mathbf{z}_1 = \dot{\mathbf{z}}$ as a dependent variable and extend the signal generator driven state to $\mathfrak{w} = [\mathbf{x};\mathbf{z};\mathbf{z}_1]$. Add the additional equation

$$\dot{\mathbf{z}}_1 = \mathbf{A}_z\mathbf{z}_1 + \mathbf{G}_z(\mathbf{z}_1\otimes\mathbf{z} + \mathbf{z}\otimes\mathbf{z}_1), \qquad \mathbf{z}_1(0) = \mathbf{z}_{10}$$

with $\mathbf{z}_{10}$ chosen consistently to $\mathbf{z}$ to the signal generator driven system. Then proceed as in Definition 4.1 to construct the quadratic signal generator driven system with extended state $\mathfrak{w} = [\mathbf{x};\mathbf{z};\mathbf{z}_1]$.

4.2 Input-weighted concept for input-output systems

At least formally, our input-tailoring shows some similarities to the concept of input-weighting. The latter has been used in system-theoretic model reduction of linear systems to get reduced models with enhanced fidelity in certain frequency ranges. We refer to [VA02], [BBG15] and references therein for details.

Motivated by our approach, we propose the usage of input-weights to incorporate non-standard input maps in the system-theoretic methods like multi-moment matching or balanced truncation [BG17] based on multivariate frequency representations. To stress the formal similarities to our input-tailored approach, we use a similar notation.

Definition 4.2 (Input-weighted system). *Let a system* $\mathbf{S}$ *and an input-weight* $\mathbf{F}$ *be given as*

$$\begin{aligned} \mathbf{S}:&\quad \mathbf{E}\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{G}\mathbf{x}^{②} + \mathbf{D}\mathbf{x}\otimes\mathbf{u} + \mathbf{B}\mathbf{u} + \mathbf{G}_u\mathbf{u}^{②} + \mathbf{B}_p\dot{\mathbf{u}}, && \mathbf{x}(0) = \mathbf{0}\in\mathbb{R}^N \\ \mathbf{F}:&\quad \mathbf{u} = \mathbf{C}_z\mathbf{z}, \qquad \dot{\mathbf{z}} = \mathbf{A}_z\mathbf{z} + \mathbf{G}_z\mathbf{z}^{②} + \mathbf{B}_z\mathbf{u}_F, && \mathbf{z}(0) = \mathbf{0}\in\mathbb{R}^q. \end{aligned}$$

Then we call $\mathbf{S}_F : \mathbf{u}_F \mapsto \mathbf{x}$

$$\mathbf{S}_F: \quad \begin{aligned} &\mathcal{E}\dot{\mathfrak{w}} = \mathcal{A}\mathfrak{w} + \mathcal{G}\mathfrak{w}^{②} + \mathcal{B}\mathbf{u}_F, && \mathfrak{w}(0) = \mathbf{0} \\ &\mathbf{x} = \mathcal{P}_x\,\mathfrak{w} \end{aligned}$$

with

$$\mathcal{B} = \begin{bmatrix} \mathbf{B}_p\mathbf{C}_z\mathbf{B}_z \\ \mathbf{B}_z \end{bmatrix} \qquad \text{and } \mathcal{E}, \mathcal{A}, \mathcal{G} \text{ as in Definition 4.1,}$$

the input-weighted system.

The upper input-weighted system $\mathbf{S}_F$ results from the assumption that the inputs of interest $\mathbf{u}$ can be constructed from input-weight $\mathbf{F}$ and some auxiliary input $\mathbf{u}_F$, and then incorporating the input-weight into the input-output description. By construction, $\mathbf{S}_F$ has a linear input map. Therefore, any standard system-theoretic model reduction method based on the input-independent multivariate frequency representations can be used on it to construct an extended reduction basis $\mathcal{V}$. The reduction basis $\mathbf{V}$ for the original system $\mathbf{S}$ can then be extracted from the extended basis $\mathcal{V}$ in the same fashion as we do it in our input-tailored approach, cf. Remark 2.6. Of course, the choice of input-weight $\mathbf{F}$ and its influence on the reduction method is an important issue in this approach, but beyond the scope of this work.

4.3 Univariate frequency representation of third order

Our approach uses univariate frequency representations tailored towards user-pre-defined families of inputs. For convenience, most of our discussion is restricted to the first two expansion terms. Let us stress that higher order terms can be similarly considered. The cascade- and tensor-structured pattern, which the second order terms and their moments evidently have, is preserved. This section provides the expressions and results associated to the third order term $\mathscr{W}_3$. In particular, we state the respective extensions of Lemma 1.8 and Theorem 2.9.

Lemma 4.3 (Counterpart of Lemma 1.8). *Assume that the requirements of Theorem 1.4 hold true. Then the associated frequency representation* $\breve{\mathscr{W}}_3$ *can also be formulated with the linear* *representation*

$$\breve{\mathscr{W}}_3(s) = \breve{\mathcal{C}}_3 \left(s\breve{\mathcal{E}}_3 - \breve{\mathcal{A}}_3\right)^{-1} \breve{\mathfrak{b}}_3,$$

$$\textit{with} \quad \breve{\mathcal{E}}_3 = \begin{bmatrix} \mathcal{E} & & \\ & \mathcal{E}^{\textcircled{2}} & \\ & & \mathcal{E}^{\textcircled{3}} \end{bmatrix}, \quad \breve{\mathcal{A}}_3 = \begin{bmatrix} \mathcal{A} & 2\mathcal{G} & \\ & \textcircled{2}_{\mathcal{E}}\mathcal{A} & \mathcal{G} \otimes \mathcal{E} \\ & & \textcircled{3}_{\mathcal{E}}\mathcal{A} \end{bmatrix}$$

$$\breve{\mathfrak{b}}_3 = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathfrak{b}^{\textcircled{3}} \end{bmatrix}, \quad \breve{\mathcal{C}}_3 = \begin{bmatrix} \mathbf{I}_M & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

The linear state representation follows by straight forward calculus from Theorem 1.4.

Theorem 4.4 (Counterpart of Theorem 2.9)**.** *Assume the requirements of Theorem 1.4 and Lemma 4.3 hold, and let for given $s_0 \in \mathbb{C}$ the matrix $\mathcal{A}_{s_0} = -s_0\mathcal{E} + \mathcal{A}$ be nonsingular. Then the moments $\mathbf{m}_i^{(3)}$ of $\breve{\mathscr{W}}_3$ at s_0 are characterized by the recursion formula:*

$$\begin{aligned}
i = 0: \quad & \textcircled{3}_{\mathcal{E}}\mathcal{A}_{s_0/3}\, \boldsymbol{\eta}_0^{(3)} &&= -\mathfrak{b}^{\textcircled{3}} \\
& \textcircled{2}_{\mathcal{E}}\mathcal{A}_{s_0/2}\, \boldsymbol{\mu}_0^{(3)} &&= -\mathcal{G} \otimes \mathcal{E}\, \boldsymbol{\eta}_0^{(3)} \\
& \mathcal{A}_{s_0}\, \mathbf{m}_0^{(3)} &&= -2\mathcal{G}\, \boldsymbol{\mu}_0^{(3)} \\
i > 0: \quad & \textcircled{3}_{\mathcal{E}}\mathcal{A}_{s_0/3}\, \boldsymbol{\eta}_i^{(3)} &&= \mathcal{E}^{\textcircled{3}}\, \boldsymbol{\eta}_{i-1}^{(3)} \\
& \textcircled{2}_{\mathcal{E}}\mathcal{A}_{s_0/2}\, \boldsymbol{\mu}_i^{(3)} &&= \mathcal{E}^{\textcircled{2}}\, \boldsymbol{\mu}_{i-1}^{(3)} - \mathcal{G} \otimes \mathcal{E}\, \boldsymbol{\eta}_i^{(3)} \\
& \mathcal{A}_{s_0}\, \mathbf{m}_i^{(3)} &&= \mathcal{E}\, \mathbf{m}_{i-1}^{(3)} - 2\mathcal{G}\, \boldsymbol{\mu}_i^{(3)}.
\end{aligned}$$

Moreover, $\mathbf{k}_i^{(3)} = [\mathbf{m}_i^{(3)}; \boldsymbol{\mu}_i^{(3)}; \boldsymbol{\eta}_i^{(3)}]$ are the moments of $s \mapsto \left(s\breve{\mathcal{E}}_3 - \breve{\mathcal{A}}_3\right)^{-1} \breve{\mathfrak{b}}_3$ at s_0.

The proof follows similarly as the one of Theorem 2.9.

4.4 Variational expansion for parametrized initial conditions

This section deals with the generalization of the variational expansion in Theorem 1.4. Instead of just dealing with α-dependent initial conditions $\mathfrak{w}(0; \alpha) = \alpha\mathfrak{b}$ as in the main part of this part, initial conditions parametrized in a multidimensional linear space spanned by the column span of a matrix $\mathcal{B}_0 \in \mathbb{R}^{M,K}$ can be regarded.

We consider the $\mathbf{r}$-dependent dynamical system

$$\begin{aligned}
\mathcal{E}\dot{\mathfrak{w}}(t; \mathbf{r}) &= \mathcal{A}\mathfrak{w}(t; \mathbf{r}) + \mathcal{G}\left(\mathfrak{w}(t; \mathbf{r})\right)^{\textcircled{2}}, \qquad t \in (0, T) \\
\mathfrak{w}(0; \mathbf{r}) &= \mathcal{B}_0\mathbf{r}, \quad \text{ for } \mathbf{r} \in \mathbb{R}^K
\end{aligned}$$

with $T > 0$ and system matrices $\mathcal{E}, \mathcal{G}$ as in Theorem 1.4. The respective generalization of Theorem 1.4 then leads to the expansion

$$\mathfrak{w}(t; \mathbf{r}) = \sum_{i=1}^{N} \mathfrak{w}_i(t)\mathbf{r}^{\textcircled{i}} + \textit{higher order terms.}$$

Again, the $\mathfrak{w}_i$ have Laplace transforms $\breve{\mathscr{W}}_i$ analogously as in Theorem 1.4, where only $\mathfrak{b}$ is replaced by $\mathcal{B}_0$ at all instances.

To see that this holds true, we note that for each concrete choice of $\mathbf{r}$ one can define $\tilde{\mathfrak{b}}$ such that $\tilde{\mathfrak{b}} = \mathcal{B}_0\mathbf{r}$. Then the state equation for $\mathfrak{w}$ also can be written as

$$\mathcal{E}\dot{\mathfrak{w}}(t) = \mathcal{A}\mathfrak{w}(t) + \mathcal{G}(\mathfrak{w}(t))^{\textcircled{2}}, \qquad \mathfrak{w}(0) = \tilde{\mathfrak{b}}.$$

Now use Theorem 1.4 and expand for $\tilde{\mathfrak{b}} = \alpha\mathfrak{b}$ the solution $\mathfrak{w}$ in α. Afterwards re-substitute the '$\tilde{\mathfrak{b}}^{\textcircled{i}}$'-terms with the relation $\tilde{\mathfrak{b}}^{\textcircled{i}} = (\mathcal{B}_0\mathbf{r})^{\textcircled{i}} = \mathcal{B}_0^{\textcircled{i}}\mathbf{r}^{\textcircled{i}}$, which gives the expressions we claimed.

Chapter 5

Numerical results

In this chapter we investigate the numerical performance of our new input-tailored approximate moment matching in comparison to the system-theoretic multi-moment matching, the trajectory-based proper orthogonal decomposition, and the method [ZLW^{+}12], [ZW16] based on univariate frequency representations. We consider three benchmarks, which have been used in literature to test especially, but not exclusively, nonlinear system-theoretic model reduction methods, e.g., [ABJ16], [BG17], [Gu12], [BB12c], [BB15], [Gu11]. Apart from a general performance comparison, certain aspects are further highlighted in the different benchmark tests: The viscous Burgers' equation (Section 5.2) is used to demonstrate the applicability of the extensions from Section 4.1 to handle non-standard input maps. The difference of input-tailoring in our method against the use of training trajectories in proper orthogonal decomposition is illustrated on the other two benchmarks. On the one hand, different input-scenarios may lead to the same input-tailored expansion, although the solution trajectories differ nonlinearly. Such an example is discussed for the Chafee-Infante equation (Section 5.3). On the other hand, our method is overall less dependent on the input-scenario than the proper orthogonal decomposition, which is showcased for the nonlinear RC-ladder (Section 5.4). A discussion on the difference and computational advantage of our approach to [ZLW^{+}12], [ZW16] concludes Section 5.4 and the numerical section.

5.1 Setup for numerical results

We have implemented our approach, which we refer to as *AssM*, in `MATLAB`. For an efficient realization of Step (2b) in Algorithm 3.1 the routine '`mess_lyap`' from `M.E.S.S.` Toolbox [SKB16] with its default settings is used. The full order model simulations, referred to as *FOM* as well as the reduced simulation are done using `MATLAB`'s solver '`ode15s`', where the tolerances are modified to 'AbsTol $= 10^{-8}$' and 'RelTol $= 10^{-6}$' and the exact Jacobian matrices are forwarded to the solver. For the intended comparison, the one-sided multi-moment matching approach from [BB12b, Alg. 2], [BB12c], which we refer to as *MultM*, has been implemented. It aims at matching the moments of the symmetric transfer functions $\mathscr{G}_1(s)$, at given frequencies $(\sigma_1, \sigma_2, \ldots, \sigma_\mu)$ up to order q_1 as well as the multi-moments of $\mathscr{G}_2(s_1, s_2)$ at the diagonal frequency pairs $((\sigma_1, \sigma_1), (\sigma_2, \sigma_2) \ldots (\sigma_\mu, \sigma_\mu))$ up to order q_2, where $q_1 \geq q_2$ has to be chosen. We refer to [BB12c], [Bre13], [Gu12] for details. Just for convenience we choose the same expansion frequencies for *AssM*, both for the first and the second associated transfer function and equal moment orders $L_1 = \cdots = L_\mu$ and $\tilde{L}_1 = \cdots = \tilde{L}_\mu$

for all expansion frequencies, which we denote by L and $\tilde{L}$. As a heuristic to construct expansion frequencies, we apply IRKA method to the linear transfer function $\mathscr{G}_1(s)$, and select the first few real calculated values, similarly to [BB12b], [BB15], [ABJ16]. Reduction results are also compared to those gotten from the proper orthogonal decomposition method [KV01], [AH14], referred to as *POD*. For the construction of *POD* we use time snapshots of the training trajectory, which, if not indicated differently, is chosen as the solution trajectory. We use 300 uniformly distributed time snapshots in all benchmark test cases, as we experienced no improvements in the results when increasing the number of snapshots. The results have been generated on an Intel Core i7-8700 CPU, with 3.20GB RAM, and `MATLAB` Version 9.3.0.713579 (R2017b).

5.2 Burgers' equation

On the spatial domain $\Omega = (0,1)$ we consider the nonlinear viscous Burgers' equation given by

$$\begin{aligned}
\partial_t v(\xi,t) &= -v(\xi,t)\,\partial_\xi v(\xi,t) + \nu\,\partial_{\xi\xi} v(\xi,t) && \text{in } (0,1)\times(0,T)\\
v(0,t) &= u(t), \qquad\qquad \partial_\xi v(1,t) = 0 && \text{in } (0,T)\\
v(\xi,0) &= 0 && \text{on } [0,1]
\end{aligned}$$

with $\nu = 0.01$. The input u particularly prescribes a Dirichlet boundary condition on the left boundary ($\xi = 0$). We choose the output to be the boundary value on the right, $y(t) = v(1,t)$. The two input-scenario cases we present relate to one linear and one nonlinear signal generator:

Case 1 Linear signal generator.

$$u(t) = 0.5\left(\cos\left(1.3\pi t\right) - \cos\left(5.4\pi t\right) - \sin\left(0.6\pi t\right) + 1.2\sin\left(3.1\pi t\right)\right)$$

The input u is a sum of sine- and cosine-functions. Every summand can be described by a dynamic system, e.g., the last summand $\tilde{u}(t) = 1.2\sin\left(3.1\pi t\right)$ has the linear signal generator

$$\tilde{u} = [1\,|\,0]\mathbf{z}, \qquad \dot{\mathbf{z}} = 3.1\pi\begin{bmatrix} & 1\\ -1 & \end{bmatrix}\mathbf{z} \qquad \mathbf{z}(0) = 1.2\begin{bmatrix}0\\1\end{bmatrix},$$

and analogously for the others. Superposing these single generators gives the linear signal generator for u.

Case 2 Nonlinear signal generator.

$$u(t) = \frac{1}{0.5 - \exp\left(2t\right)} + 2\exp\left(-t\right)$$

The respective signal generator is nonlinear and reads

$$u = [-0.5\,|\,2]\mathbf{z}, \quad \dot{\mathbf{z}} = \begin{bmatrix}-2 & \\ & -1\end{bmatrix}\mathbf{z} + \begin{bmatrix}-0.5 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\mathbf{z}^{②}, \quad \mathbf{z}(0) = \begin{bmatrix}4\\1\end{bmatrix}.$$

Case 1 is particularly similar to a test case considered in [Bre13], [BB12b], [BB15] for multi-moment matching.

In the upcoming, we employ two different discretization schemes for the upper partial differential equations to construct one quadratic-bilinear system with linear input-dependency (Section 5.2), and one with nonlinear input-dependency (Section 5.2). Note that both discretized systems describe up to a small discretization error the same dynamics, and should therefore serve as an equally valid basis for model reduction. The nonlinear input-dependency, however, cannot directly be incorporated into the input-output based system-theoretic methods. As will be demonstrated, our input-tailored and input-weighted extensions, respectively, can be used for both discretizations and show to be almost independent of the underlying discretization.

Full order model formulation with linear input map

The Burgers' equation is discretized in space with standard central finite differences and uniform mesh size h implicitly defined by $h = 1/(N+2)$ with N inner grid points. The equations for the inner node values $v_i(t) \approx v(\xi_i, t)$ with $\xi_i = ih$ read

$$\dot{v}_i = -v_i \frac{v_{i+1} - v_{i-1}}{2h} + \nu \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}, \quad 1 \leq i \leq N.$$

The discretized boundary conditions give $v_0 = u$ and $(v_{N+1} - v_N)/h = 0$, which we use to eliminate v_0 and v_{N+1}. This leaves us with a quadratic-bilinear full order model of the form (1.1) with state $\mathbf{x}(t) = [v_1(t); v_2(t); \ldots; v_N(t)]$ and $\mathbf{E} = \mathbf{I}_N$. The output matrix becomes $\mathbf{C} = [0, \ldots, 0, 1]$, as $y = v_{N+1} = v_N$ due to the boundary conditions.

The parameters used in the model reduction for *AssM* and *MultM* are summarized in Table 5.1. Proceeding from the *FOM* with $N = 4000$, this leads to reduced models of dimension $n = 16$, which is also the dimension we choose for the reduced model of *POD*. The respective results concerning output behavior and absolute error over time are illustrated in Fig. 5.1. As can be observed, all methods (*AssM*, *MultM* and *POD*) perform comparably well, showing a similar error behavior with moderate numerical oscillations near steep gradients of the solution output in both cases. Notably, also the *POD* trained with the solution trajectory itself does not lead to significantly better results, which indicates that this benchmark example is rather hard to reduce for any kind of model reduction method.

Our main motivation to introduce input-weighted multi-moment matching, i.e., the extension of Section 4.2, is the incorporation of non-standard input-dependencies. The latter is showcased

Expansion frequencies	*AssM* & *MultM*	0.03, 0.22
Order moments	*AssM*	$\tilde{L} = 3, \quad L = 2$
	MultM	$q_1 = 3, \quad q_2 = 2$
Tolerance	*AssM*: *tol*	0.001 (*Case 1*)
		0.0001 (*Case 2*)
Resulting dimensions	*AssM* & *MultM*	16

Table 5.1: Reduction parameters for Burgers' equation Case 1 & Case 2 (*FOM* with $N = 4000$).

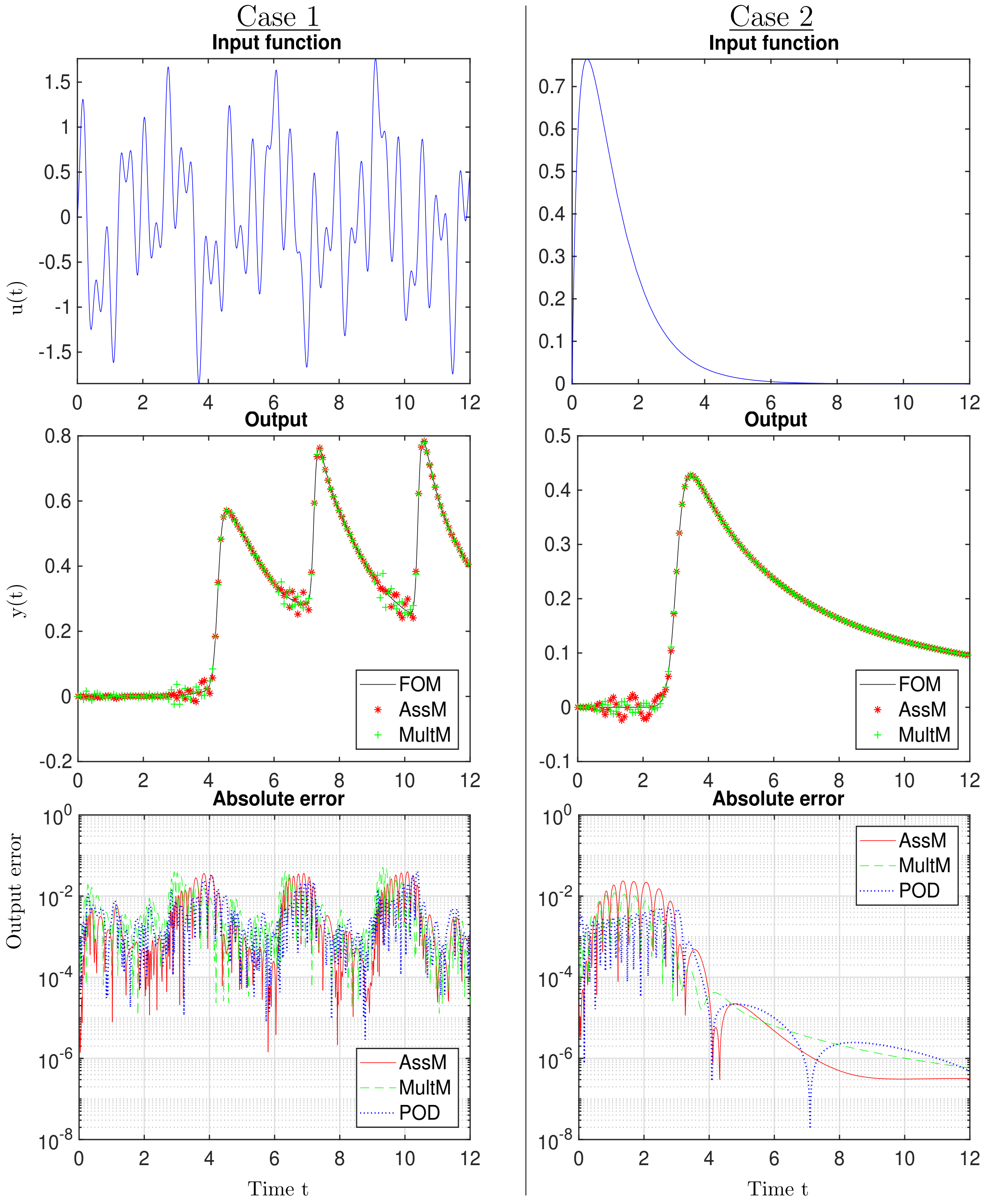

Figure 5.1: Reduction results for Burgers' equation with linear input map. *Top to bottom:* Input u, output y, output errors. Dimensions: *FOM*: $N = 4000$, *Reduced models*: $n = 16$ (cf. Table 5.1).

in the next subsection. But we want to mention that incorporation of an input-weight influences the reduction method itself. To illustrate this, a preliminary comparison of the input-weighted multi-moment matching, which we refer to as *MultM-iw*, and the unweighted *MultM* is done. We construct reduced models with *MultM-iw* using the exemplary choice of input-weight

$$u = z, \qquad \dot{z} = -z + u_F, \qquad\qquad z(0) = 0,$$

(i.e., $\mathbf{C}_z = 1$, $\mathbf{G}_z = 0$, $\mathbf{A}_z = -1$ and $\mathbf{B}_z = 1$ in Definition 4.2) and otherwise the same parameters as for *MultM* in Table 5.1.

The *MultM-iw* leads to the smaller dimension $n = 12$ compared to $n = 16$ for *MultM*, which is due to the input-weighted system, cf. Definition 4.2, not having any bilinear parts to be considered in the multi-moment matching. In this example the smaller dimension goes hand in hand with a slightly larger output error, cf. Fig. 5.2. But we note that we also tested both methods with altered reduction parameters (including different choices of input-weights), and observed comparable results when the reduced models are constructed to be of equal dimension, e.g., by incorporating an additional expansion frequency for *MultM-iw*.

Full order model formulation with nonlinear input map

The *FOM* with quadratic input map for the Burger's equation is constructed as follows: Instead of using the advective form $v\,\partial_\xi v$ for the nonlinearity as in Section 5.2, we rewrite it in conservative form $0.5\,\partial_\xi(v^2)$ and then apply the central finite difference scheme. Then the equations for the inner node values v_i read

$$\dot{v}_i = -\frac{v_{i+1}^2 - v_{i-1}^2}{4h} + \nu\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}, \quad 1 \le i \le N.$$

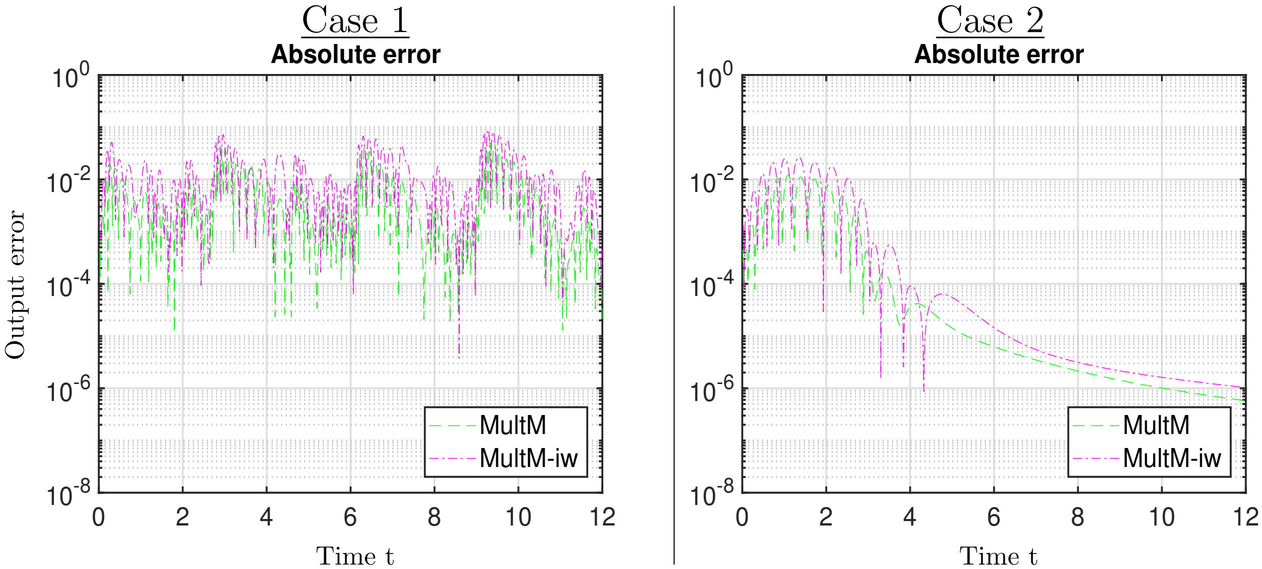

Figure 5.2: Reduction results for Burgers' equation with linear input map. Comparing reduction error for unweighted *MultM* with $n = 16$ and input-weighted *MultM-iw* with $n = 12$. (cf. Table 5.1 and Fig. 5.1).

Incorporating the boundary conditions in this new discretization gives a quadratic term in the input u, resulting in a *FOM* with nonlinear input map. We use this new *FOM* and apply, with the extensions proposed in Section 4.1, the input-tailored method and the input-weighted multi-moment matching with the reduction parameters as before, cf. Table 5.1. The resulting reduced models are referred to as *AssM-q* for the input-tailored, and *MultM-q-iw* for the input-weighted multi-moment matching method, respectively.

Independence of the underlying discretization in dimensions of the reduced models is observed, i.e., $n = 16$ for *AssM-q* and *AssM*, and $n = 12$ for *MultM-q-iw* and *MultM-iw*. Also the output response does not change beyond a negligible order much smaller than the reduction errors, which can be seen comparing the output differences in Fig. 5.3 with the output errors in Fig. 5.1 and Fig. 5.2.

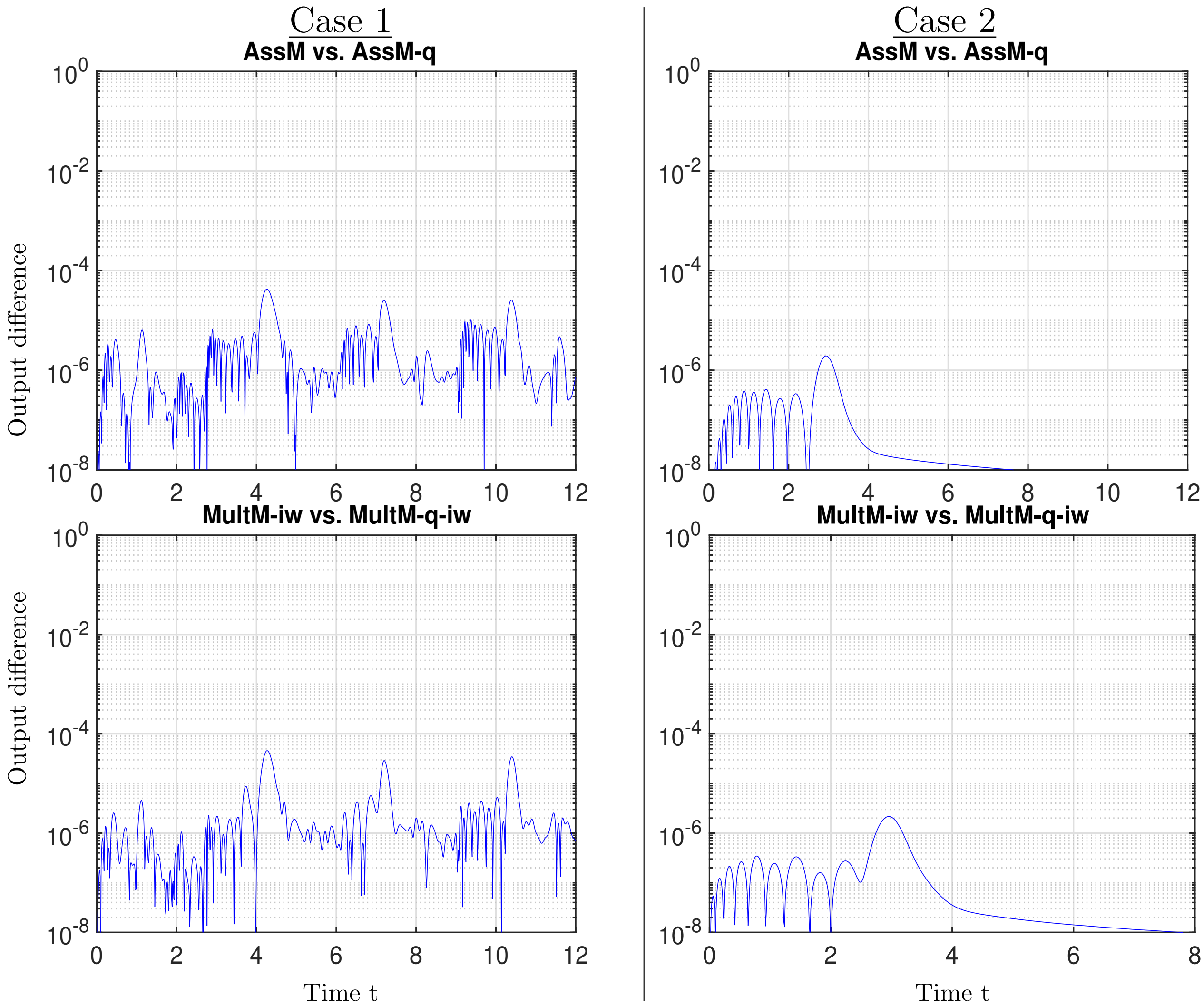

Figure 5.3: Difference of reduced outputs for Burgers' equation under different underlying discretizations. *AssM* and *MultM-iw* (advective discretization, input-linear) versus *AssM-q* and *MultM-q-iw* (conservative discretization, input-nonlinear) (cf. Table 5.1, Fig. 5.1 and Fig. 5.2).

5.3 Chafee-Infante equation

The Chafee-Infante equation is a one-dimensional convection-diffusion equation for $v = v(\xi, t)$ with a cubic nonlinearity in v. Following [BB15], we introduce the augmented function w by $w = v^2$, and consider an artificial differential equation describing w by differentiating the algebraic relation to get $\partial_t w = 2v\,\partial_t v$. By that a representation with only quadratic nonlinearities results, reading

$$\begin{aligned}
\partial_t v(\xi,t) &= -v(\xi,t)\,w(\xi,t) + \partial_{\xi\xi} v(\xi,t) + v(\xi,t) && \text{in } (0,1)\times(0,T)\\
\partial_t w(\xi,t) &= -2w(\xi,t)^2 + 2v(\xi,t)\,\partial_{\xi\xi} v(\xi,t) + 2v(\xi,t)^2 && \text{in } (0,1)\times(0,T)\\
v(0,t) &= u(t), \qquad \partial_\xi v(1,t) = 0 && \text{in } (0,T)\\
w(0,t) &= v(0,t)^2, \qquad w(1,t) = v(1,t)^2 && \text{in } (0,T)\\
v(\xi,0) &= v^0(\xi), \qquad w(\xi,0) = v^0(\xi)^2 && \text{on } [0,1].
\end{aligned}$$

The equations for $w(0,t)$ and $w(1,t)$ should be read as consistency conditions. Trivial initial conditions for v^0 are employed here, and the input u, prescribing a Dirichlet boundary condition on the left boundary ($\xi = 0$), is varied over the test cases. In particular, two test cases distinguishing by a linear scaling α in the input are set up:

$$\begin{aligned}
&u(t) = \alpha\left[\cos(1.3\pi t) - \cos(5.4\pi t) - \sin(0.6\pi t) + 1.2\sin(3.1\pi t)\right]\\
&\textit{Case 1}: \qquad \alpha = 1.\\
&\textit{Case 2}: \qquad \alpha = 0.125.
\end{aligned}$$

Therefore, the corresponding signal generators of both cases coincide up to a scaling. (They are a scaling of the signal generator in Case 1 of Section 5.2.) Similarly as for the Burgers' equation, we discretize the system in space using central finite differences with a uniform mesh with $\tilde{N}$ inner grid points, and eliminate the boundary node values by means of the boundary conditions. This leads to a quadratic-bilinear system of the form (1.1) with state $\mathbf{x}(t) = [v_1(t); \ldots; v_{\tilde{N}}(t); w_1(t); \ldots; w_{\tilde{N}}(t)]$ and $\mathbf{E} = \mathbf{I}_N$, $N = 2\tilde{N}$. As output we consider $y(t) = v(1,t)$, implying the output matrix $\mathbf{C} = [\mathbf{0}_{1,\tilde{N}-1}, 1, \mathbf{0}_{1,\tilde{N}}]$, since $v_{\tilde{N}} = v_{\tilde{N}+1}$ due to the boundary conditions.

Expansion frequencies	*AssM* & *MultM*	1.5, 21.5, 48.3
Order moments	*AssM*	$\tilde{L} = 1$, $L = 2$
	MultM	$q_1 = 2$, $q_2 = 2$ $(\sigma \in \{1.5, 21.5\})$
		$q_1 = 2$, $q_2 = 1$ $(\sigma = 48.3)$
Tolerance	*AssM*: *tol*	0.001 (*Case 1*)
		0.0001 (*Case 2*)
Resulting dimensions	*AssM* & *MultM*	12

Table 5.2: Reduction parameters for Chafee-Infante equation Case 1 & Case 2 (*FOM* with $N = 1500$).

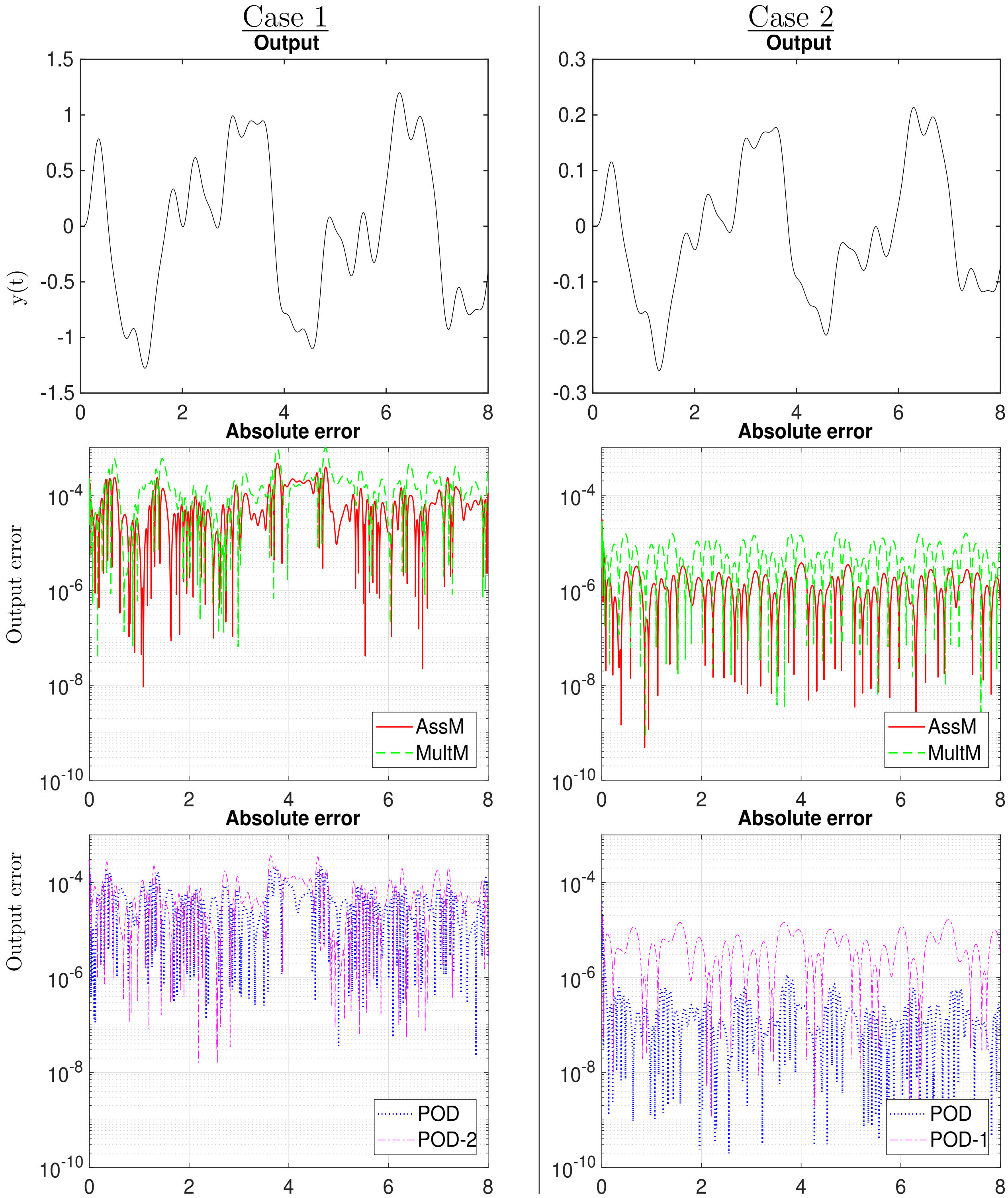

Figure 5.4: Reduction results for Chafee-Infante equation. *Top to bottom:* Output y (with *FOM*), output errors for case-independent *AssM* and *MultM*, output errors for case-dependent *POD* models. Dimensions: *FOM*: $N = 1500$, *Reduced models*: $n = 12$ (cf. Table 5.2).

Reduced models of dimension $n = 12$ are constructed for *MultM* and *AssM* with the parameters of Table 5.2. Reduced models of the same size are also constructed by *POD*, where we deviated from the standard procedure and constructed separately bases of dimension 6 each for the physical variables $[v_1(t); \ldots; v_{\tilde{N}}(t)]$ and $[w_1(t); \ldots; w_{\tilde{N}}(t)]$, which were then combined to a full block basis of dimension 12. The direct application of *POD* onto the full state leads to significantly worse results, which is known to possibly happen when different physical variables are mixed [AH14]. As *POD* depends on the chosen training trajectory, the two cases lead to two distinct *POD* reduced models. When we test *POD* models, where training trajectory and solution trajectory of the test case differ, we indicate by a suffix '*-1*' or '*-2*' the case used in the training phase. In contrast, *AssM* leads to the same reduced model in both cases, because by construction the input-tailored variational expansions coincide. (The scaling in *tol*, cf. Table 5.2, is only needed to compensate for the scaling in the input.)

As seen in Fig. 5.4, the case-independent *AssM* performs well for both cases, especially also better than *MultM*. The results of proper orthogonal decomposition, in contrast, depend on the training trajectory. Worse results are seen for *POD-1* (trained with Case 1) in Case 2, and *POD-2* (trained with Case 2) in Case 1 than for the perfectly trained models. In particular, if not trained perfectly, the proper orthogonal decomposition performs worse than our *AssM*.

5.4 Nonlinear RC-ladder

This benchmark describes a nonlinear RC-ladder with $\tilde{N}$ capacitors and I-V diodes. The nonlinearity is due to the diode I-V characteristics, given by $g(v) = \exp(40v - 1)$ for voltages v. We use the same setup as in [ABJ16], [BG17], [BB15], but also in [ZLW$^+$12], [ZW16] a similar example has been studied. The node voltages v_i ($2 \leq i \leq \tilde{N} - 1$, and $\tilde{N} = 500$) are described by

$$\begin{aligned}
\dot{v}_1(t) &= -2v_1(t) + v_2(t) - g(v_1(t)) - g(v_1(t) - v_2(t)) + u(t) \\
\dot{v}_i(t) &= -2v_i(t) + v_{i-1}(t) + v_{i+1}(t) + g(v_{i-1}(t) - v_i(t)) - g(v_i(t) - v_{i+1}(t)) \\
\dot{v}_{\tilde{N}}(t) &= -v_{\tilde{N}}(t) + v_{\tilde{N}-1}(t) + g(v_{\tilde{N}-1}(t) - v_{\tilde{N}}(t)).
\end{aligned}$$

The input u corresponds to a current source. As detailed, e.g., in [Gu11], [SLSM19], the system can be recast as a quadratic-bilinear system of size $N = 2\tilde{N} = 1000$ in the new variables $x_1 = v_1$, and $x_i = v_{i-1} - v_i$ for $2 \leq i \leq \tilde{N}$, and $x_i = \exp(40x_{i-\tilde{N}} - 1)$ for $\tilde{N} + 1 \leq i \leq 2\tilde{N}$. The output is chosen as $y = x_1$, and the benchmark is treated with trivial initial conditions and two different cases of inputs:

Case 1 Exponential pulse. $u(t) = \exp(-t)$ with corresponding signal generator

$$u = z, \qquad \dot{z} = -z, \qquad z(0) = 1.$$

Case 2 Oscillation. $u(t) = 1 + \cos(10\pi t)$ with corresponding signal generator

$$u = [1 \,|\, 0 \,|1]\mathbf{z}, \qquad \dot{\mathbf{z}} = \begin{bmatrix} 0 & & \\ & & 10\pi \\ & -10\pi & \end{bmatrix} \mathbf{z}, \qquad \mathbf{z}(0) = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Case 1 and the reduction parameters for *MultM* are directly taken from [Bre13], [BB12c], whereas Case 2 is modified from the reference to have a higher amplitude and frequency.

The reduction parameters for *AssM* and *MultM* are summarized in Table 5.3. Additionally, standard *POD* is used to construct reduced models of size $n = 11$ equal to the size of the other reduced models. As the inputs in both cases differ nonlinearly from each other, both, *POD* and *AssM* lead to different reduced models depending on the case used in the reduction step. We indicate the models where the training scenario and the test case do not coincide by a suffix ’*-1*’ or ’*-2*’ for the case used in the training trajectory or input-tailoring, respectively.

As seen in Fig. 5.5, *AssM* outperforms *MultM* here by up to two orders. Perfectly trained *POD* is yet superior to both system-theoretic methods, but falls off strongly when the training scenario differs from the test case. In contrast, our *AssM* method shows to be much less sensitive to the training scenario. We observe reduction errors of up to two orders smaller than for *POD*, when training and test case disagree (cf. last row of Fig. 5.5).

Remark 5.1 (Choice of signal generator)**.** *There is no necessity to choose the signal generator in the reduction phase of AssM such that it generates the signal of the test case, as we did mostly throughout this part. Robust (input-independent) other choices for signal generators are yet an open issue to us. Let us, however, note that in our experience the impact of the chosen signal generator in AssM is not comparably strong as the impact of training trajectories in POD.*

Expansion frequencies	*AssM* & *MultM*	1
Order moments	*AssM*	$\tilde{L} = 3,\ L = 2$
	MultM	$q_1 = 5, \quad q_2 = 2$
Tolerance	*AssM*: *tol*	0.0006
Resulting dimensions	*AssM* & *MultM*	11

Table 5.3: Reduction parameters for nonlinear RC-ladder (*FOM* with $N = 1000$).

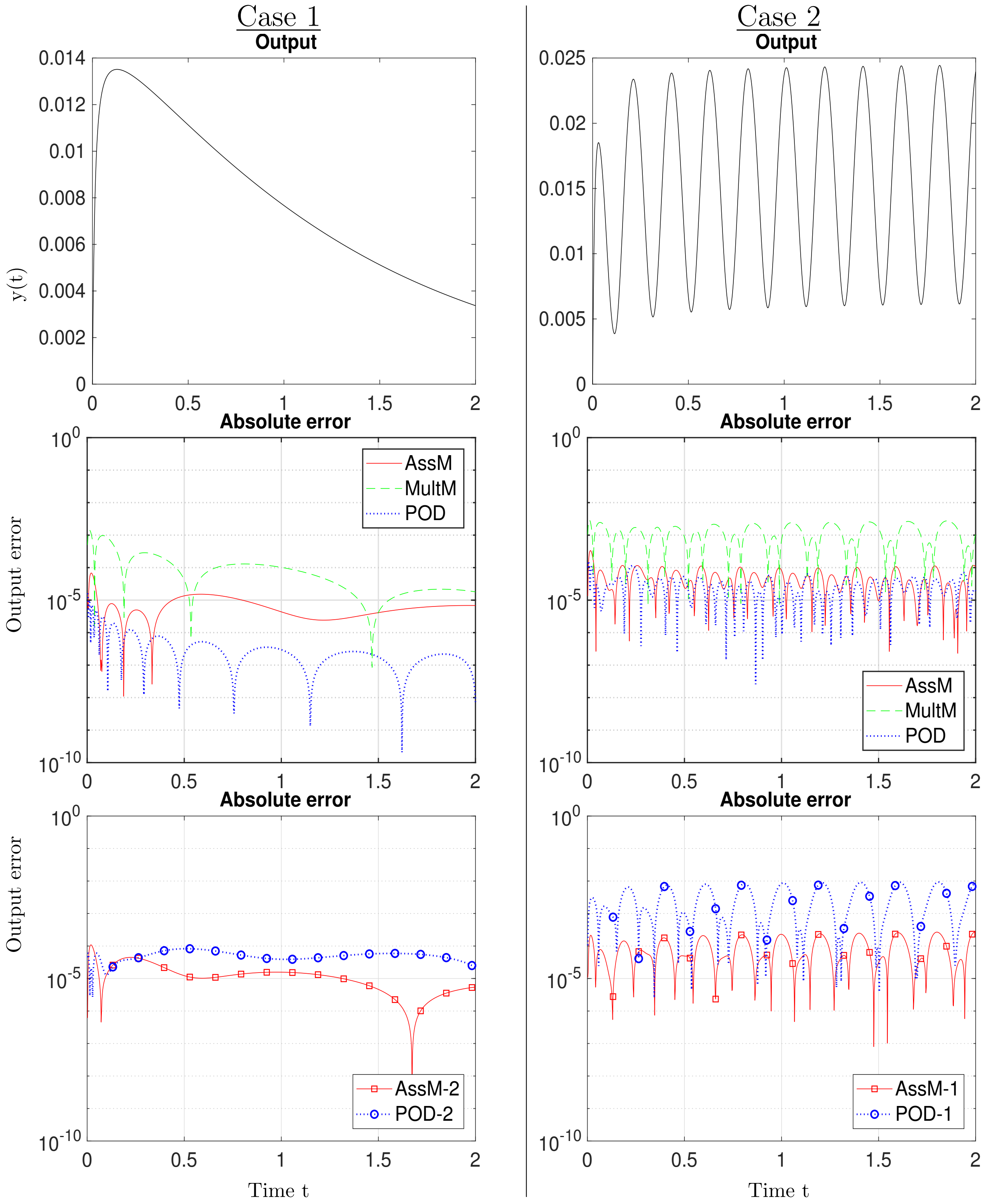

Figure 5.5: Reduction results for nonlinear RC-ladder. *Top to bottom:* Output y (with *FOM*), output errors of methods with training case and test case coinciding, output errors of *AssM* and *POD* with training case and test case disagreeing. Dimensions: *FOM*: $N = 1000$, *Reduced models*: $n = 11$ (cf. Table 5.3).

Expansion frequencies	*AssM-mu*	1.2, 8.8, 37.7, 108.2
	AssM-inf	0.2, 1.3, 5.9, 20.0, 56.1, 121.3
Order moments	*AssM-mu* & *AssM-inf*	$\tilde{L} = 1, \quad L = 1$
Tolerance	*AssM-mu*: *tol*	0.0005
	AssM-inf: *tol*	∞
Resulting dimensions	*AssM-mu* & *AssM-inf*	12

Table 5.4: Reduction parameters using multiple expansion frequencies for nonlinear RC-ladder (*FOM* with $N = 1000$).

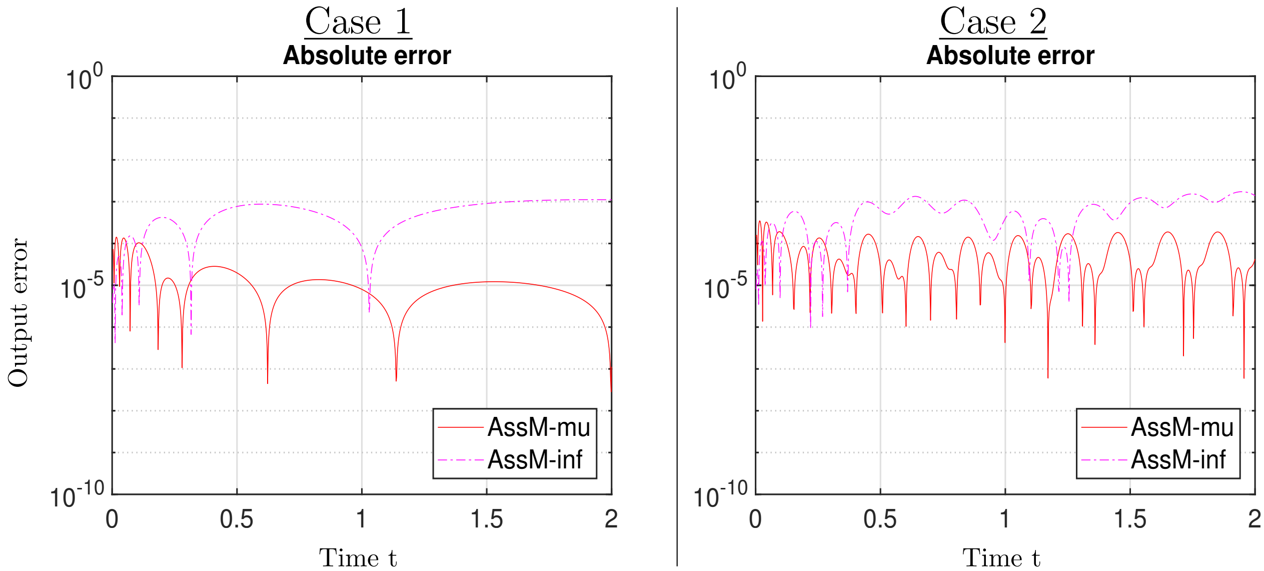

Figure 5.6: Reduction errors for nonlinear RC-ladder with multiple expansion frequencies. Dimensions: *FOM*: $N = 1000$, *Reduced models*: $n = 12$ (cf. Table 5.4).

In a last test, we exemplarily showcase the importance of the newly proposed approximation condition (2.6b). This condition has no analog in the former method [ZLW$^+$12], [ZW16]. We repeat the test cases for the RC-ladder using *AssM* with altered parameters involving multiple expansion frequencies as described in Table 5.4. The expansion frequencies are, as mentioned, found by applying IRKA onto the first transfer function of the Volterra series. The model *AssM-mu* aims for the approximation condition (2.6b) up to a small tolerance, whereas in *AssM-inf* the approximation condition is ignored and instead more expansion frequencies are used. The latter therefore relates to [ZLW$^+$12], [ZW16]. Although both models are of equal size $n = 12$, *AssM-mu* leads to profoundly better results, as seen in Fig. 5.6.

Remark 5.2 (Performance)**.** *Our method AssM yields low order high fidelity models that are competitive and overall similar to other system-theoretic model reduction methods as the multi-moment matching. It naturally extends to systems with non-standard input maps, cf. Section 4.1 and comparisons in Section 5.2, which makes it in this respect similarly flexible as the trajectory-*

based methods like proper orthogonal decomposition. In contrast to trajectory-based methods, AssM does not rely on pre-calculated full order model simulations. The most crucial part of its offline phase consists in the solution of Lyapunov-type equations, which makes it more costly than the offline phase of simple multi-moment matching. Nonetheless, our implementation is a profound enhancement over the approach in [ZLW$^+$12], [ZW16] in terms of offline times. We refer to [SLSM19], where we showcased the latter. The time savings stem from the exploitation of the appearing low-rank tensor structures, cf. Section 3.1.

Discussion and conclusion

In this part we suggested a new system-theoretic model reduction approach for quadratic-bilinear dynamical systems, which is based on a different perspective as compared to the multivariate frequency-based ones. Instead of relying on input-output modeling, we used the notion of signal generator driven systems. By that, input-tailored variational expansions were constructed for a large class of inputs. We compared our approach to the *system-theoretic* multi-moment matching and the *trajectory-based* proper orthogonal decomposition, and observed rather similar performance to the former. Compared to the method [ZLW$^+$12], [ZW16], which also utilizes univariate frequency representations, our method shows profound enhancements regarding analytical results and numerical performance. We stress that in contrast to existing system-theoretic reduction methods, our method naturally extends to systems with non-standard input dependencies, such as, e.g., quadratic terms, time derivatives. As a byproduct of the latter, we also suggested a modification of input-output based system-theoretic methods able to handle non-standard input dependencies.

We restricted the discussion in the main part to variational expansion terms up to order two. Nonetheless, the results are presented in a tensor notation allowing for convenient generalizations to higher order, as provided in Section 4.3 for the third order terms. Regarding higher order terms in the numerical implementation, of course, the typical adaptions for the handling of tensors with order higher than two, have to be integrated, cf. [KT10], [KK18]. Other possible extensions of our approach could include more sophisticated automated choices of expansion frequencies or generic (input-independent) signal generators as well as the handling of systems with more general nonlinearities.

Bibliography

[ABJ16] M. I. Ahmad, P. Benner, and I. M. Jaimoukha. Krylov subspace methods for model reduction of quadratic-bilinear systems. *IET Control Theory Appl.*, 10:2010–2018(8), 2016.

[AH14] D. Amsallem and U. Hetmaniuk. Error estimates for Galerkin reduced-order models of the semi-discrete wave equation. *ESAIM Math. Model. Numer. Anal*, 48(1):135–163, 2014.

[AH17] B. Afkham and J. Hesthaven. Structure preserving model reduction of parametric Hamiltonian systems. *SIAM J. Sci. Comput.*, 39(6):A2616–A2644, 2017.

[AH19] B. Afkham and J. Hesthaven. Structure-preserving model-reduction of dissipative Hamiltonian systems. *SIAM J. Sci. Comput.*, 81(1):3–21, 2019.

[ALM08] A. Astolfi and A. Lorenzo Marconi, eds. *Analysis and Design of Nonlinear Control Systems.* Springer, 2008.

[AM04] G. Akrivis and C. Makridakis. Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM-Math. Model. Num.*, 38(2):261–289, 2004.

[AMS08] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds.* Princeton University Press, 2008.

[Ant05] A. Antoulas. *Approximation of Large-Scale Dynamical Systems.* Society for Industrial and Applied Mathematics, 2005.

[Ast10a] A. Astolfi. Model reduction by moment matching for linear and nonlinear systems. *IEEE Trans. Autom. Control*, 55(10):2321–2336, 2010.

[Ast10b] A. Astolfi. Model reduction by moment matching, steady-state response and projections. In *49th IEEE Conference on Decision and Control (CDC)*, pages 5344–5349, 2010.

[Bar16] S. Bartels. *Numerik 3x9.* Springer, 1 edition, 2016.

[BB12a] P. Benner and T. Breiten. Interpolation-based H_2-model reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.*, 33(3):859–885, 2012.

[BB12b] P. Benner and T. Breiten. Krylov-subspace based model reduction of nonlinear circuit models using bilinear and quadratic-linear approximations. In *Progress in Industrial Mathematics at ECMI 2010*, volume 17, pages 153–159. Springer, 2012.

[BB12c] P. Benner and T. Breiten. Two-sided moment matching methods for nonlinear model reduction. Preprint MPIMD/12-12, Max Planck Institute Magdeburg, 2012.

[BB15] P. Benner and T. Breiten. Two-sided projection methods for nonlinear model order reduction. *SIAM J. Sci. Comput.*, 37(2):B239–B260, 2015.

[BBF14] U. Baur, P. Benner, and L. Feng. Model order reduction for linear and nonlinear systems: A system-theoretic perspective. *Arch. Comput. Methods Eng.*, 21(4):331–358, 2014.

[BBG15] T. Breiten, C. Beattie, and S. Gugercin. Near-optimal frequency-weighted interpolatory model reduction. *Systems & Control Letters*, 78:8–18, 2015.

[BG17] P. Benner and P. Goyal. Balanced Truncation Model Order Reduction For Quadratic-Bilinear Control Systems. arXiv e-prints 1705.00160, 2017.

[BGG18] P. Benner, P. Goyal, and S. Gugercin. $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems. *SIAM J. Matrix Anal. Appl.*, 39(2):983–1032, 2018.

[BGH11] J. Brouwer, I. Gasser, and M. Herty. Gas pipeline models revisited: Model hierarchies, nonisothermal models, and simulations of networks. *Multiscale Model. Simul.*, 9(2):601–623, 2011.

[BH15] P. Benner and J. Heiland. Time-dependent Dirichlet conditions in finite element discretizations. *ScienceOpen Research*, pages 1–18, 2015.

[BMBM18] A. Moses Badlyan, B. Maschke, C. A. Beattie, and V. Mehrmann. Open physical systems: from GENERIC to port-Hamiltonian systems. arXiv e-prints 1804.04064, 2018.

[BMXZ17] C. A. Beattie, V. Mehrmann, H. Xu, and H. Zwart. Port-Hamiltonian descriptor systems. arXiv e-prints 1705.09081, 2017.

[Bor10] A. Borys. Consideration of Volterra series with excitation and/or impulse responses in the form of Dirac impulses. *IEEE Trans. Circuits Syst., II, Exp. Briefs*, 57(6):466–470, 2010.

[Bra07] D. Braess. *Finite Elements. Theory, Fast Solvers and Applications in Elasticity Theory.* Springer, 4 edition, 2007.

[Bre11] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer, 2011.

[Bre13] T. Breiten. *Interpolatory Methods for Model Reduction of Large-Scale Dynamical Systems.* Ph.D. Thesis, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany, 2013.

[CBG16] S. Chaturantabut, C. Beattie, and S. Gugercin. Structure-preserving model reduction for nonlinear port-Hamiltonian systems. *SIAM J. Sci. Comput.*, 38(5):B837–B865, 2016.

[CCS18] K. Carlberg, Y. Choi, and S. Sargsyan. Conservative model reduction for finite-volume models. *J. Comput. Phys.*, 371:280 – 314, 2018.

[CGM$^+$12] E. Celledoni, V. Grimm, R. I. McLachlan, D. I. McLaren, D. O'Neale, B. Owren, and G. R. W. Quispel. Preserving energy resp. dissipation in numerical PDEs using the Average Vector Field method. *J. Comput. Phys.*, 231(20):6770 – 6789, 2012.

[Che05] G.-Q. Chen. Euler equations and related hyperbolic conservation laws. In C. M. Dafermos and E. Feireisl, eds., *Handbook of Differential Equations Evolutionary Equations*, volume 2, pages 1–104. Elsevier, 2005.

[Chi06] C. Chicone. *Ordinary Differential Equations with Applications.* Texts in Applied Mathematics. Springer, 2006.

[CHS90] F. Chalot, T. J. Hughes, and F. Shakib. Symmetrization of conservation laws with entropy for high-temperature hypersonic computations. *Comput. Syst. Eng.*, 1(2-4):495–521, 1990.

[CMKO11] S. H. Christiansen, H. Z. Munthe-Kaas, and B. Owren. Topics in structure-preserving discretization. *Acta Numerica*, 20:1–119, 2011.

[CML19] F. L. Cardoso-Ribeiro, D. Matignon, and L. Lefevre. A partitioned finite element method for power-preserving discretization of open systems of conservation laws. arXiv e-prints 1906.05965, 2019.

[CMS19] N. Cagniart, Y. Maday, and B. Stamm. *Model Order Reduction for Problems with Large Convection Effects*, volume 47 of *Computational Methods in Applied Sciences*, pages 131–150. Springer, 2019.

[CS10] S. Chaturantabut and D. C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, 2010.

[CS12] S. Chaturantabut and D. C. Sorensen. A state space error estimate for POD-DEIM nonlinear model reduction. *SIAM J. Numer. Anal.*, 50(1):46–63, 2012.

[DH08] P. Deuflhard and A. Hohmann. *Numerische Mathematik 1: Eine algorithmisch orientierte Einführung*, volume 1. De Gruyter, 2008.

[DHLT17] P. Domschke, B. Hiller, J. Lang, and C. Tischendorf. Modellierung von Gasnetzwerken: Eine Übersicht. Technical Report 2717, Technische Universität Darmstadt, 2017.

[Dom11] P. Domschke. *Adjoint-Based Control of Model and Discretization Errors for Gas Transport in Networked Pipelines.* PhD thesis, TU Darmstadt, Verlag Dr. Hut, 2011.

[DS18] Z. Drmac and A. K. Saibaba. The discrete empirical interpolation method: Canonical structure and formulation in weighted inner product spaces. *SIAM J. Matrix Anal. Appl.*, 39(3):1152–1180, 2018.

[Egg18] H. Egger. A robust conservative mixed finite element method for compressible flow on pipe networks. *SIAM J. Sci. Comput*, 40(1):A108–A129, 2018.

[Egg19] H. Egger. Structure preserving approximation of dissipative evolution problems. *Numer. Math.*, 143(1):85–106, 2019.

[EK18] H. Egger and T. Kugler. Damped wave systems on networks: Exponential stability and uniform approximations. *Numer. Math.*, 138(4):839 – 867, 2018.

[EKLS$^+$18] H. Egger, T. Kugler, B. Liljegren-Sailer, N. Marheineke, and V. Mehrmann. On structure-preserving model reduction for damped wave propagation in transport networks. *SIAM J. Sci. Comput.*, 40(1):A331–A365, 2018.

[EKLS20] H. Egger, T. Kugler, and B. Liljegren-Sailer. Stability preserving approximations of a semilinear hyperbolic gas transport model. In *Hyperbolic Problems: Theory, Numerics, Applications*, volume 10, pages 427–433. AIMS Series on Appl. Math, 2020.

[FACC14] C. Farhat, P. Avery, T. Chapman, and J. Cortial. Dimensional reduction of nonlinear finite element dynamic models with finite rotations and energy-based mesh sampling and weighting for computational efficiency. *Int. J. Numer. Meth. Eng.*, 98(9):625–662, 2014.

[FKJ$^+$13] O. Farle, D. Klis, M. Jochum, O. Floch, and R. Dyczij-Edlinger. A port-Hamiltonian Finite-Element formulation for the Maxwell equations. In *2013 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, pages 324–327, 2013.

[Fre08] R. W. Freund. On Pade-type model order reduction of J-Hermitian linear dynamical systems. *Linear Algebra Appl.*, 429(10):2451–2464, 2008.

[GAB15] P. Goyal, M. I. Ahmad, and P. Benner. Model reduction of quadratic-bilinear descriptor systems via Carleman bilinearization. In *European Control Conference, ECC 2015, Linz, Austria, July 15-17, 2015*, pages 1177–1182, 2015.

[GH18] C. Gräßle and M. Hinze. POD reduced-order modeling for evolution equations utilizing arbitrary finite element discretizations. *Adv. Comput. Math.*, 44(6):1941–1978, 2018.

[Gil77] E. Gilbert. Functional expansions for the response of nonlinear differential systems. *IEEE Trans. Autom. Control*, 22(6):909–921, 1977.

[GPBvdS09] S. Gugercin, R. V. Polyuga, C. A. Beattie, and A. van der Schaft. Interpolation-based H_2 model reduction for port-Hamiltonian systems. In *Proceedings of the 48th IEEE Conference on Decision and Control, and the 28th Chinese Control Conference, Shanghai*, pages 5362—-5369, 2009.

[Gri97] E. J. Grimme. *Krylov projection methods for model reduction.* Ph.D. Thesis, Univ. of Illinois at Urbana-Champaign, USA, 1997.

[GSW13] S. Gugercin, T. Stykel, and S. Wyatt. Model reduction of descriptor systems by interpolatory projection methods. *SIAM J. Sci. Comput.*, 35(5):B1010–B1033, 2013.

[GTvdSM04] G. Golo, V. Talasila, A. van der Schaft, and B. Maschke. Hamiltonian discretization of boundary control systems. *Autom.*, 40(5):757–771, 2004.

[Gu11] C. Gu. QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 30(9):1307–1320, 2011.

[Gu12] C. Gu. *Model Order Reduction of Nonlinear Dynamical Systems.* PhD thesis, EECS Department, University of California, Berkeley, 2012.

[GVL96] G. H. Golub and C. F. Van Loan. *Matrix Computations.* The John Hopkins University Press, 3 edition, 1996.

[Hac12] W. Hackbusch. *Tensor Spaces and Numerical Tensor Calculus.* Springer, 2012.

[Har83] A. Harten. On the symmetric form of systems of conservation laws with entropy. *J. Comput. Phys.*, 49(1):151–164, 1983.

[Har02] P. Hartman. *Ordinary Differential Equations.* Society for Industrial and Applied Mathematics, second edition, 2002.

[HCF17] J. A. Hernandez, M. A. Caicedo, and A. Ferrer. Dimensional hyper-reduction of nonlinear finite element models via empirical cubature. *Comput. Methods Appl. Mech. Engrg.*, 313:687–722, 2017.

[HGCATRM09] A. Herran-Gonzalez, J. M. De La Cruz, B. De Andres-Toro, and J. L. Risco-Martin. Modeling and simulation of a gas distribution pipeline network. *Appl. Math. Model.*, 33(3):1584–1600, 2009.

[HSS08] M. Heinkenschloss, D. C. Sorensen, and K. Sun. Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations. *SIAM J. Sci. Comput.*, 30(2):1038–1063, 2008.

[IA13] T. C. Ionescu and A. Astolfi. Families of reduced order models that achieve nonlinear moment matching. In *2013 American Control Conference*, pages 5518–5523, 2013.

[IW14] T. Iliescu and Z. Wang. Are the snapshot difference quotients needed in the proper orthogonal decomposition? *SIAM J. Sci. Comput.*, 36(3):A1221–A1250, 2014.

[Jam08] A. Jameson. The construction of discretely conservative finite volume schemes that also globally conserve energy or entropy. *J. Sci. Comput.*, 34(2):152–187, 2008.

[KK18] V. Khoromskaia and B. N. Khoromskij. *Tensor Numerical Methods in Quantum Chemistry*. De Gruyter, 2018.

[KM06] P. Kunkel and V. Mehrmann. *Differential Algebraic Equations*. European Mathematical Society, 2006.

[Kot13] P. Kotyczka. Discretized models for networks of distributed parameter port-Hamiltonian systems. In *nDS'13; Proceedings of the 8th International Workshop on Multidimensional Systems*, pages 1–5, 2013.

[KT10] D. Kressner and C. Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.*, 31(4):1688–1714, 2010.

[Kug19] T. Kugler. *Galerkin methods for simulation of wave propagation on a network of pipes*. PhD thesis, TU Darmstadt, Verlag Dr. Hut, 2019.

[KV01] K. Kunisch and S. Volkwein. Galerkin proper orthogonal decomposition methods for parabolic systems. *Numer. Math.*, 90:117–148, 2001.

[LeV90] R. J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser, 1990.

[LeV02] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, 2002.

[LK78] C. Lesiak and A. Krener. The existence and uniqueness of Volterra series for nonlinear systems. *IEEE Trans. Autom. Control*, 23(6):1090–1095, 1978.

[LM17] L. Lefevre and S. Medianu. Symplectic discretization of port controlled Hamiltonian systems. *IFAC-PapersOnLine*, 50:3629–3634, 2017.

[LM18] J. Lang and P Mindt. Entropy-preserving coupling conditions for one-dimensional Euler systems at junctions. *Netw. Heterog. Media*, 13(1):177, 2018.

[LMT13] R. Lamour, R. März, and C. Tischendorf. *Differential-Algebraic Equations: A Projector Based Analysis*. Springer, 2013.

[LP06] P. Li and L. T. Pileggi. Compact reduced-order modeling of weakly nonlinear analog and RF circuits. *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, 24(2):184–203, 2006.

[LSM17] B. Liljegren-Sailer and N. Marheineke. A structure-preserving model order reduction approach for space-discrete gas networks with active elements. In *Progress in Industrial Mathematics at ECMI 2016*, pages 439–446. Springer, 2017.

[LSM18] B. Liljegren-Sailer and N. Marheineke. Input-tailored system-theoretic model order reduction for quadratic-bilinear systems. arXiv e-prints 1809.08979, 2018.

[LSM19] B. Liljegren-Sailer and N. Marheineke. Structure-preserving Galerkin approximation for a class of nonlinear port-Hamiltonian partial differential equations on networks. In *Proc. Appl. Math. Mech.*, volume 19, page e201900399, 2019.

[LW13] H. Liu and N. Wong. Autonomous Volterra algorithm for steady-state analysis of nonlinear circuits. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, 32(6):858–868, 2013.

[LW16] Y.-W. Li and X. Wu. Functionally fitted energy-preserving methods for solving oscillatory nonlinear Hamiltonian systems. *SIAM J. Numer. Anal*, 54(4):2036–2059, 2016.

[MLD19] P. Mindt, J. Lang, and P. Domschke. Entropy-preserving coupling of hierarchical gas models. *SIAM J. Math. Anal.*, 51(6):4754–4775, 2019.

[MM14] Y. Miyatake and T. Matsuo. A general framework for finding energy dissipative/conservative H_1-Galerkin schemes and their underlying H_1-weak forms for nonlinear evolution equations. *BIT Numerical Mathematics*, 54(4):1119–1154, 2014.

[MM19] V. Mehrmann and R. Morandin. Structure-preserving discretization for port-Hamiltonian descriptor systems. In *58th IEEE Conference on Decision and Control, CDC 2019, Nice, France, December 11-13, 2019*, pages 6863–6868. IEEE, 2019.

[Moc80] M. S. Mock. Systems of conservation laws of mixed type. *J. Differ. Equations*, 37(1):70–88, 1980.

[MS05] V. Mehrmann and T. Stykel. *Dimension Reduction of Large-Scale Systems: Proceedings of a Workshop held in Oberwolfach, Germany, October 19–25, 2003*, chapter Balanced Truncation Model Reduction for Large-Scale Systems in Descriptor Form, pages 83–115. Springer, 2005.

[MvdS92] B. M. Maschke and A. van der Schaft. Port-controlled Hamiltonian systems: Modelling origins and systemtheoretic properties. *IFAC Proceedings Volumes*, 25(13):359 – 365, 1992. 2nd IFAC Symposium on Nonlinear Control Systems Design 1992, Bordeaux, France, 24-26 June.

[MvdS01] B. M. Maschke and A. van der Schaft. *Hamiltonian representation of distributed parameter systems with boundary energy flow*, volume 2, pages 137–142. Springer, 2001.

[PAvdS12] R. Pasumarthy, V. Ambati, and A. van der Schaft. Port-Hamiltonian discretization for open channel flows. *Syst. Control. Lett.*, 61(9):950–958, 2012.

[PDG18] B. Peherstorfer, Z. Drmac, and S. Gugercin. Stabilizing discrete empirical interpolation via randomized and deterministic oversampling. arXiv e-prints 1808.10473, 2018.

[PM16] L. Peng and K. Mohseni. Symplectic model reduction of Hamiltonian systems. *SIAM J. Sci. Comput.*, 38(1):A1–A27, 2016.

[Roc70] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Mathematical Series. Princeton University Press, 1970.

[RSSM18] J. Reiss, P. Schulze, J. Sesterhenn, and V. Mehrmann. The shifted proper orthogonal decomposition: A mode decomposition for multiple transport phenomena. *SIAM J. Sci. Comput.*, 40(3):A1322–A1344, 2018.

[RSV13] T. Richter, A. Springer, and B. Vexler. Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems. *Numer. Math.*, 124(1):151–182, 2013.

[Rug81] W. J. Rugh. *Nonlinear system theory: the Volterra/Wiener approach*. Johns Hopkins series in information sciences and systems. Johns Hopkins University Press, 1981.

[RW98] R. Rockafellar and R. Wets. *Variational Analysis*. Springer, 1998.

[SAB$^+$17] M. Schmidt, D. Aßmann, R. Burlacu, J. Humpola, I. Joormann, N. Kanelakis, T. Koch, D. Oucherif, M. Pfetsch, L. Schewe, R. Schwarz, and M. Sirvent. GasLib – A Library of Gas Network Instances. *Data*, 2(4), 2017.

[Sim07] V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM J. Sci. Comput*, 29:1268–1288, 2007.

[SKB16] J. Saak, M. Köhler, and P. Benner. M-M.E.S.S.-1.0.1 – the matrix equations sparse solvers library. DOI:10.5281/zenodo.50575, 2016.

[SLSM19] N. Stahl, B. Liljegren-Sailer, and N. Marheineke. Moment matching based model order reduction for quadratic-bilinear systems. In *Progress in Industrial Mathematics at ECMI 2018*, pages 551–557. Springer, 2019.

[SMH19] A. Serhani, D. Matignon, and G. Haine. Partitioned finite element method for port-Hamiltonian systems with boundary damping: Anisotropic heterogeneous 2D wave equations. *IFAC-PapersOnLine*, 52(2):96 – 101, 2019.

[VA02] A. Varga and B. D. O. Anderson. Frequency-weighted balancing related controller reduction. *IFAC Proceedings Volumes*, 35(1):113–118, 2002.

[vdSJ14] A. van der Schaft and D. Jeltsema. Port-Hamiltonian systems theory: An introductory overview. *Foundations and Trends in Systems and Control*, 1:173–378, 2014.

[vdSM13] A. van der Schaft and B. M. Maschke. Port-Hamiltonian systems on graphs. *SIAM J. Control Optim.*, 51:906–937, 2013.

[VL85] C. Van Loan. Computing the CS and the generalized singular value decompositions. *Numer. Math.*, 46:479–491, 1985.

[WLEK10] T. Wolf, B. Lohmann, R. Eid, and P. Kotyczka. Passivity and structure preserving order reduction of linear port-Hamiltonian systems using Krylov subspaces. *Eur. J. Control*, 16(4):401–406, 2010.

[WP10] M. Wilke and J. Prüss. *Gewöhnliche Differentialgleichungen und dynamische Systeme*. Springer, 1 edition, 2010.

[Zei85] E. Zeidler. *Nonlinear Functional Analysis and its Applications*. Springer, 1985.

[ZLW$^+$12] Y. Zhang, H. Liu, Q. Wang, N. Fong, and N. Wong. Fast nonlinear model order reduction via Associated Transforms of high-order Volterra transfer functions. In *Proceedings of the 49th Annual Design Automation Conference*, DAC'12, pages 289–294. ACM, 2012.

[ZW16] Y. Zhang and N. Wong. Compact model order reduction of weakly nonlinear systems by Associated Transform. *Int. J. Circ. Theor. App.*, 44(7):1367–1384, 2016.

Scientific Career

Björn Liljegren-Sailer,
born on October, 7th 1989 in Malmö, Sweden

Education

01/2015 – 07/2020	PhD student in mathematics, FAU Erlangen-Nürnberg / Universität Trier, doctoral advisor Prof. Dr. Nicole Marheineke
10/2012 – 12/2014	Master of Science Technomathematics, FAU Erlangen-Nürnberg Master-thesis: *Analysis and numerics for a PDAE of hyperbolic type describing viscoelastic curved jets*
08/2013 – 01/2014	Study abroad, Lund University, Schweden
10/2009 – 10/2012	Bachelor of Science in Technomathematics, FAU Erlangen-Nürnberg, Bachelor-thesis: *Matched Asymptotic Expansions am Beispiel des elastischen Balkens*
09/2000 – 06/2009	Abitur, Gymnasium Höchstadt