Regularization and Simulation of Constrained Partial Differential Equations

vorgelegt von
Diplom-Mathematiker
Robert Altmann
aus Berlin

Von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
Dr. rer. nat

genehmigte Dissertation

Promotionsausschuss:

Vorsitzende:	Prof. Dr. Noemi Kurt
Berichter:	Prof. Dr. Volker Mehrmann
Berichterin:	Prof. Dr. Caren Tischendorf
Berichter:	Prof. Dr. Alexander Ostermann

Tag der wissenschaftlichen Aussprache: 29. 05. 2015

Berlin 2015

Contents

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Regularisierung von differentiell-algebraischen Gleichungen (DAEs) abstrakter Funktionen. Diese sogenannten *Operator DAEs* sind Operatorgleichungen, die die Struktur differentiell-algebraischer Gleichungen verallgemeinern. Sie bieten eine alternative Formulierung partieller Differentialgleichungen, die gewissen Nebenbedingungen genügen müssen. Die vorgestellte Regularisierung verbessert die Sensibilität der Operator DAEs gegenüber Störungen und resultiert in gut gestellten Systemen bei der Ortsdiskretisierung.

Operator DAEs sind hervorragend geeignet zur Modellierung physikalischer Systeme. Anwendungsbereiche findet man in der Strömungsmechanik, der Kontinuumsmechanik sowie im Bereich des Elektromagnetismus. Im Allgemeinen führen gekoppelte Systeme, die aus mehreren Subsystemen bestehen, oft auf diese Art von Gleichungen. Die Formulierung physikalischer Systeme als Operator DAE steht im direkten Zusammenhang zur schwachen Formulierung von partiellen Differentialgleichungen. Als Verallgemeinerung von DAEs kann auch die Nebenbedingung selbst einen Differentialoperator beinhalten wie zum Beispiel bei den Navier-Stokes Gleichungen, die durch die Divergenzfreiheit restringiert sind. Es handelt sich also um DAEs, die allgemein in einem Banachraum definiert sind. Eine weitere Charakterisierung ist gegeben durch die Eigenschaft, dass eine Semidiskretisierung im Ort auf eine DAE im ursprünglichen Sinne führt. Daraus resultieren auch die Stabilitätsprobleme wie die hohe Sensibilität gegenüber Störungen sowie die Notwendigkeit konsistenter Anfangsdaten.

Die Regularisierung folgt den Ideen der Indexreduktion für DAEs. Dabei sucht man eine äquivalente Operatorgleichung, die bessere numerische Eigenschaften aufweist. In dieser Arbeit wird speziell die Methode *minimal extension* betrachtet, die sich hervorragend für semi-expliziete Systeme eignet. Dies führt dann auf ein vergrößertes System, da im Regularisierungsprozess neue Variablen eingeführt werden. Dabei ist zu erkennen, dass sich der Index der semidiskreten Systeme verringert. In der Strömungsmechanik erhält man DAEs vom Index 1 statt Index 2 und im Bereich der Kontinuumsmechanik reduziert sich der Index sogar von 3 auf 1.

Diskretisiert man die Operatorgleichungen zuerst in der Zeit statt im Ort, so erhält man eine Folge von stationären partiellen Differentialgleichungen. Der letzte Teil der Arbeit analysiert die Konvergenz dieser Zeitdiskretisierung. Dabei ist zu beobachten, dass sich die einzelnen Variablen unterschiedlich verhalten. Der Lagrange Multiplikator, beziehungsweise der Druck im Bereich der Strömungsmechanik, benötigt stärkere Regularitätsannahmen, um die Konvergenz zu garantieren. Desweiteren wird der Einfluss von kleinen Störungen untersucht. Auch hierbei zeigt sich der Vorteil der präsentierten Regularisierung in Bezug auf die besseren Stabilitätseigenschaften verglichen mit dem ursprünglichen System.

Abstract

This thesis is devoted to the regularization of differential-algebraic equations in the abstract setting (operator DAEs) and the resulting positive impact on the corresponding semi-discrete systems and on the sensitivity to perturbations. The possibility of a modularized modeling and the maintenance of the physical structure of a dynamical system make operator DAEs convenient form the modeling point of view. They appear in all fields of applications such as fluid dynamics, elastodynamics, electromagnetics, as well as in multi-physics applications were different system types are coupled.

From a mathematical point of view, operator DAEs are constrained PDEs, written in the weak formulation. Therein, the constraint may itself be a differential equation such as in the Navier-Stokes equations where the velocity of a Newtonian fluid is constrained to be divergence-free. On the other hand, operator DAEs generalize the notion of DAEs to the infinite-dimensional setting, including abstract functions which map into a Banach space. Thus, a spatial discretization leads to a DAE in the classical sense. This also implies that typical stability issues known from the theory of DAEs such as the high sensitivity to perturbations also translate to the operator case.

The regularization of an operator DAE follows the concept of an index reduction for a DAE. Hence, an equivalent system is sought-after which has better properties from a numerical point of view. The presented regularization lifts the index reduction technique of *minimal extension* for semi-explicit DAEs to the abstract setting and leads to an extended operator DAE. A spatial discretization of the regularized system then leads to a DAE of lower index compared to the semi-discrete system arising from the original operator DAE. For flow equations we obtain a reduction from index 2 to index 1 whereas the applications from the field of elastodynamics yield a reduction from index 3 to index 1.

The last part of this thesis deals with the convergence of time discretization schemes applied to the regularized operator DAEs. Therein, we observe a qualitative difference for different variables. More precisely, we show that the Lagrange multiplier needs stronger regularity assumptions on the given data in order to guarantee the convergence to the exact solution of the operator DAE. Furthermore, the influence of perturbations in the right-hand sides of the system is analysed for the semi-discrete as well as for the continuous setting. This analysis shows the advantage of the presented regularization in terms of stability.

Published Papers

Several results of this theis were already published in preprints or journal publications. We give a short overview and indicate in which parts of the thesis the results have been used.

- [**Alt13a**] Index reduction for operator differential-algebraic equations in elastodynamics. *Z. Angew. Math. Mech. (ZAMM)*, 93(9):648–664, 2013.
  This paper introduces the idea of a regularization of operator DAEs as they appear in applications of elatodynamics. The procedure is based on the index reduction technique of minimal extension for semi-explicit DAEs. The results are used in Sections 7 and 9.

- [**Alt13b**] Modeling Flexible Multibody Systems by Moving Dirichlet Boundary Conditions. *Proceedings of Multibody Dynamics 2013 - ECCOMAS Thematic Conference (Zagreb, Croatia)*
  This proceedings contribution includes an example of an operator DAE including a coupling of an elastic body and a mass-spring-damper system. This example is only mentioned in the beginning of Section 7.

- [**Alt14**] Moving Dirichlet Boundary Conditions. *ESAIM Math. Model. Numer. Anal. (M2AN), 48(6):1859–1876, 2014*
  Problems were the part of the boundary, on which Dirichlet boundary conditions are prescribed, is time-dependent may be formulated as an operator DAE. This is mentioned in Section 5.1.

- [**AH13**] Finite element decomposition and minimal extension for flow equations. *Preprint 2013–11, Technische Universität Berlin*, 2013, *accepted for publication in ESAIM Math. Model. Numer. Anal. (M2AN)* (with Jan Heiland)
  This paper is devoted to the stable approximation of the pressure variable in fluid flow applications. For this, a regularization of the corresponding operator DAE is combined with a decomposition of finite element spaces. The results of this paper are mainly used within Section 8.

- [**AH14**] Regularization of constrained PDEs of semi-explicit structure. *Preprint 2014–05, Technische Universität Berlin*, 2014. (with Jan Heiland)
  This preprint provides a general framework for the regularization of semi-explicit operator DAEs of first order. In particular, this includes flow equations such as the (Navier-) Stokes equations. The results are used in Sections 6 and 8.

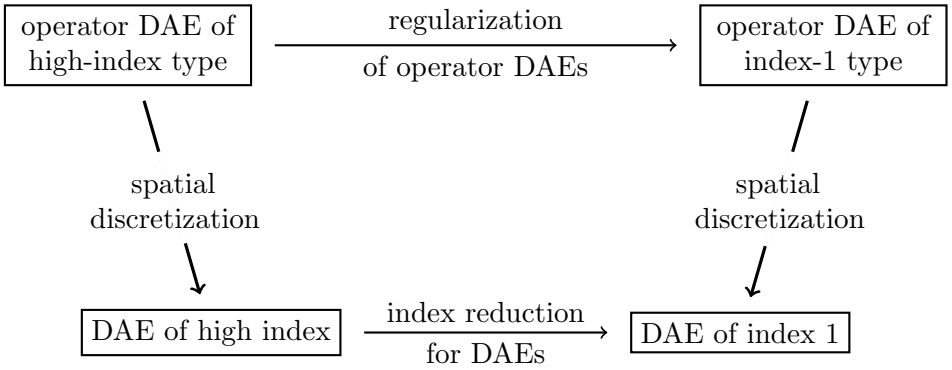
1. Introduction

With an increasing importance of automatic modeling, differential-algebraic equations (DAEs) register an increase in popularity. DAEs allow a quick and facile modeling procedure with a smaller need of system simplifications and exploit the system structure and sparsity [**CM99**]. In particular, they enable modularized modeling of several uni-physics components.

Nowadays, many models contain partial differential equations (PDEs). A coupling of systems then leads to a mixture of DAEs and PDEs which are called partial differential-algebraic equations (PDAEs) or, formulated in a weak functional analytic setting, *operator* or *abstract DAEs*. This approach follows the paradigm to include all available information to the system rather than implicitly eliminate variables. Thus, all variables remain a physically valid part of the system, also throughout the discretization process.

However, the simplicity of modeling shifts the difficulties to the mathematical part, i.e., to the analysis and simulation of such systems. DAEs suffer from instabilities, drift-off phenomena, and ill-posedness [**GM86, KM06, Ria08, LMT13**]. Note that the question of stability is of special importance if one considers applications including uncertain components, parameters, or inputs from other subsystems including themselves numerical errors.

The aim of this thesis is to regularize a specific class of PDAEs in the sense that these instabilities are reduced. Furthermore, the regularization increases the potential of adaptive methods for the simulation of constrained PDEs and is independent of the discretization scheme. The connection between the presented regularization on operator level and the index reduction for DAEs is illustrated in the following scheme.



Applications

Typical fields of applications which are modeled by DAEs are multibody systems such as problems in robotics [**ESF98**] as well as electrical circuit networks in which a (typically large) number of devices is coupled and constrained by Kirchhoff's laws [**Tis96**]. Another application comes from path control in which a specific part of a mechanical system is supposed to follow a prescribed trajectory. This typically leads to DAEs of very high index [**BK04**]. Note that the *index* measures, loosely speaking, the distance of a DAE from an ODE and thus, provides a measure of difficulty.

However, there are several applications for which the framework of DAEs is too restrictive. This is the case if the involved physical properties are described by PDEs. The Navier-Stokes equations illustrate a typical example as they consist, besides the dynamics, of an equation describing the conservation of mass, namely the incompressibility [**Tem77, EM13**]. For levitated droplet problems, in which one considers the effect of

the surface tension on a fluid interface, the coupling condition even contains the pressure variable [**BKZ92, EGR10**]. Multibody systems including flexible components are also described by PDAEs [**Sim00, Sim13**]. Further examples are the modeling of chemical kinetics [**KPSG85**] or the gas transfer in pipeline networks [**GJH$^+$13**]. Also applications in chemical engineering such as a non-reacting gas ignition or a superconductive coil often lead to PDAEs [**CM99**]. In general, one can say that multi-physics models which arise from a coupling of different components often lead to PDAEs [**EM13**].

Operator DAEs

The large range of applications calls for a good understanding of the resulting systems. However, the analysis of PDAEs in the general form is still far from complete [**EM13**]. This includes results on their well-posedness as well as a classification as it exists for DAEs in form of the index-concept [**Tis03, LMT13**].

An essential property used in this thesis is the possibility to formulate PDEs as ODEs in Banach spaces which we refer to as operator or abstract ODEs. Here we distinguish two approaches: the semigroup approach [**Paz83**] and the generalized formulation based on evolution triples which is used in this thesis. The use of generalized solutions provides the possibility to formulate problems from mathematical physics in an elegant functional analytic way [**Zei90a**, Ch. 19]. We consider here the generalized formulation since it appears naturally from the weak formulation of PDEs and allows for more general right-hand sides. The formulation as operator ODE is based on four principles [**Zei90a**, Ch. 23], namely

- to treat time and space variables in a different way,
- to use different spaces for the solution and its derivatives,
- to use generalized derivatives in time, and
- to search for solutions in appropriate Sobolev-Bochner spaces.

Following this framework, we may formulate constrained PDEs as operator DAEs. One advantage of the formulation is the resulting structure which retains the DAE structure although the problem is formulated in a Banach space. This facilitates the exploration of the interaction of DAE and operator theory.

As mentioned above, there exists no general theory for the existence and uniqueness of solutions. Since the systems of interest generalize DAEs as well as PDEs, we cannot expect a unified solution theory [**LMT13**]. For coupled systems, one difficulty may already arise in the modeling part when it is not specified which components require initial or boundary conditions. Note, however, that this problem does not appear within this thesis, as we only consider semi-explicit systems. The solvability of semi-linear PDAEs with nonlinear monotone operators, which are intended to study coupled systems as in circuit simulations, were analyzed in [**Mat12**], see also [**Gün01**].

Another open challenge is the missing index concept for general PDAEs, as this is much more complex than for DAEs. Recall that even in the DAE case there exist several nonequivalent notions of the index [**Meh13**]. However, first steps in this direction have been done as there exist classifications for particularly structured systems. An extension of the *tractability index* to a class of PDAEs, which is based on linearizations, was introduced in [**LMT01, Tis03**]. A generalization of the perturbation index for linear systems is given in [**CM99**]. Note that the involved perturbations are not restricted to the time variable, see also [**LSEL99, RA05**].

Contents

This thesis is divided into four parts. An introduction to several aspects from DAE theory and functional analysis is given in Part A. Since the thesis deals with constrained time-dependent PDEs and its formulation as operator DAEs, we have to analyse the interaction of these two topics. DAEs are characterized by its high sensitivity to perturbations and the resulting lack of robustness which carries over to the abstract setting. For the formulation of DAEs in an abstract setting, we need to introduce the concepts of Bochner spaces and Gelfand triples which provide the right spaces for the generalized formulation. Furthermore, we have to discuss the meaning of initial conditions and consistency conditions as they appear in the finite-dimensional setting. Finally, we recall basic discretization schemes in time and space which are needed for the simulation of time-dependent PDEs.

Part B is devoted to the regularization of constrained time-dependent PDEs of semi-explicit structure. This kind of remodeling approach goes along with the pattern of thought of maintaining all constraints within the system equations and even adds the so-called hidden constraints. Within the procedure, no variable transformation is needed such that the original physical meaning of the variables is preserved. The key for the presented regularization is the formulation of the system equations as operator DAE which allows to translate methods from the theory of DAEs to the abstract setting.

In Part C we analyse the positive effects of the regularization process in terms of the DAEs which result from a spatial discretization. We first consider the *method of lines* in which one discretizes in space first. A comparison of the DAEs arising from the original and regularized operator equation shows a decrease of the index and thus, an improvement in terms of stability. Because of this, we may consider the regularization procedure as an index reduction on operator level.

Discretizing the operator DAE first in time, i.e., following the *Rothe method*, we obtain a stationary PDE in every time step. Although this blurs the original DAE structure (because of the missing time dependence), the positive impact of the regularization from Part B is still apparent. These effects are analyzed in the final Part D. Furthermore, we prove the convergence of the Euler method for the semi-discretized operator DAE. Note that this corresponds to the limit case where the stationary PDEs of each time step are solved exactly. The analysis of the limiting case is important to anticipate problems which may appear for very small discretization parameters and helps to design discretization schemes which preserve properties from the continuous equations.

Acknowledgments

This thesis was developed in the scope of the ERC Advanced Grant "Modeling, Simulation and Control of Multi-Physics Systems" MODSIMCONMP. Additionally, I would like to thank the Berlin Mathematical School *BMS* for their support during the last years.

I would like to thank Prof. Volker Mehrmann for the possibility to work in his group, his supervision, and the provided freedom in the choice of the research directions. Furthermore, I want to thank Jan for his support and cooperation and Anne for the good atmosphere in the office.

Part

A

Preliminaries

The analysis, discretization, and simulation of constrained time-dependent PDEs requires the knowledge of different areas of numerical mathematics. For the formulation of such constrained PDEs in a weak sense, which is advantageous for the regularization as well as the simulation, we need several functional analytic concepts. This includes, in particular, the notion of Sobolev and Bochner spaces but also of Gelfand triples and Nemytskii mappings. On the other hand, semi-discretizations in space lead to DAEs. As a result, for an understanding of the occuring instabilities for such systems it is helpful to access the theory of DAEs. Well-known results such as the necessity of consistent initial values and the appearance of derivatives of the right-hand side in the solution also apply to the infinite-dimensional case. One may also observe the loss of convergence for low-order schemes applied to constrained PDEs.

Finally, speaking about discretizations, we have to deal with discretization schemes in time and space which lead to stable approximations. Here, we have to consider instabilities due to the differential-algebraic structure as well as instabilities due to the saddle point structure which occurs in the considered models. For this, we analyse stable mixed finite element schemes and time integration schemes which are suitable for DAEs.

This introductory part is organized as follows. In Section 2 we briefly review the characteristics of DAEs including the concept of the differentiation index and the corresponding stability problems. Driven by applications in fluid dynamics and structural mechanics, we do not consider the most general case and restrict ourselves to DAEs of semi-explicit structure. For such systems, special regularization techniques can be applied which form the basis of the reformulation of the operator DAEs in Part B.

For the analysis of time-dependent PDEs several functional analytic tools are needed. Weak formulations are stated in Sobolev spaces and the time-dependence additionally leads to so-called abstract functions. Within Section 3, we collect the necessary tools for

the analysis performed in this thesis including the integral notion for abstract functions. Using these functional analytic concepts, we can formulate time-dependent PDEs in the form of abstract differential equations in Section 4. Thus, we obtain ODEs and DAEs in an abstract setting of Banach spaces which are equivalent to time-dependent PDEs in the weak sense.

Section 5 closes the introductory part with a short overview of some discretization methods. This includes some basic finite element spaces, which we will use to discretize the problem in the space variables, as well as time integration schemes. Of special interest for the simulation of time-dependent PDEs is the interplay between spatial and temporal discretization. Here we distinguish between the method of lines and the Rothe method.

2. Differential-algebraic Equations (DAEs)

A convenient way of modeling, which allows modularized coupling, is provided by differential-algebraic equations (DAEs). For this, different systems may be coupled through the introduction of algebraic constraints which classify the type of connection. On the other side, this kind of modeling leads to systems which may cause difficulties for numerical simulation. DAEs are known for their stability issues since the solution needs besides integration also a numerical differentiation which may be an ill-posed problem.

In the first part of this section the notion of the index, which classifies DAEs, is introduced. Since we are interested in a special kind of systems, we focus on semi-explicit DAEs of first and second order. Typical examples with this structure are the semi-discretized Navier-Stokes equations and systems arising in (flexible) multibody dynamics. Second, problems arising from a high index are analyzed and index reduction methods are introduced. These methods provide a unified way to deal with DAEs of arbitrary high index - at least theoretically. Aim of an index reduction is to find an equivalent system with better stability properties which is beneficial for numerical simulations. Usually, the index-reduced systems can be treated similarly to stiff ODEs and thus, can be handled by well-analyzed methods. A summary of differences between DAEs and ODEs can be found in [**Pet82**].

2.1. Index Concepts. The index of a DAE is supposed to classify the difficulty of solving a given system. Nevertheless, there are several concepts which may lead to different indices. In the applications of interest we only have square systems such that we focus on the so-called *differentiation index*. Further notions are then shortly discussed afterwards. For a more detailed introduction, we refer to the monographs [**BCP96, KM06, Meh13**].

The most general form of a DAE of first order is given by

$$F(t, x, \dot{x}) = 0, \quad x(t_0) = x_0. \tag{2.1}$$

In particular, we will consider semi-explicit systems of the form

$$\dot{u} = f(u, p), \quad 0 = g(u) \tag{2.2}$$

or, of second order,

$$\ddot{u} = f(u, \dot{u}, \lambda), \quad 0 = g(u). \tag{2.3}$$

The best-known example of the form (2.2) is given by the semi-discrete Navier-Stokes equations. Examples of type (2.3) appear in mechanical systems such as multibody systems as well as semi-discretized elastodynamics.

2.1.1. *Differentiation Index.* As introductory example we consider the linear DAE

$$\begin{bmatrix} 0 & -1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & -1 \\ & & & 0 \end{bmatrix} \dot{x} + x = f$$

with a smooth right-hand side $f\colon [0,T] \to \mathbb{R}^n$ and some (consistent) initial condition for $x(0) \in \mathbb{R}^n$. This system is easily solvable and yields, via recursion, the solution formula

$$\begin{aligned} x_n &= f_n, \\ x_{n-1} &= f_{n-1} + \dot{x}_n = f_{n-1} + \dot{f}_n, \\ &\vdots \\ x_1 &= f_1 + \dot{x}_2 = \sum_{k=1}^{n} f_k^{(k-1)}. \end{aligned}$$

From this example we see that derivatives of the right-hand side may appear in the solution. Thus, small perturbations of f may lead to large errors, since the derivative of a small perturbation does not need to be small itself. Because of this instability, the number of required derivatives can be seen as a measure for the difficulty of solving the problem numerically. This motivates the definition of the *differentiation index* (d-index) [**HW96**, Chap. VII.1]. The d-index ν_d of a DAE (2.1) is the minimal number of analytical differentiations of the DAE,

$$F(t,x,\dot{x}) = 0, \quad \frac{\mathrm{d}F(t,x,\dot{x})}{\mathrm{d}t} = 0, \quad \ldots, \quad \frac{\mathrm{d}^{\nu_d}F(t,x,\dot{x})}{\mathrm{d}t^{\nu_d}} = 0, \tag{2.4}$$

which allows to extract algebraically from the equations in (2.4) an explicit ODE for $x(t)$. This resulting ODE is the so-called *underlying ODE*. For a precise definition of the differentiation index, we refer to [**BCP96**, Def. 2.2.2].

Remark 2.1. With additional information about the structure of the DAE, it is sufficient to consider differentiations of a part of (2.1). In particular, this is the case for systems of semi-explicit form.

For the semi-explicit DAEs (2.2) and (2.3) the d-index can be determined in an easy manner. We formulate these results in form of two lemmata [**HW96**, Ch. VII.1]. For the second-order case we consider a more specific case as it appears in the modeling of mechanical systems.

Lemma 2.2 (d-index for semi-explicit DAEs of first order). *The semi-explicit DAE* (2.2) *has d-index 2 if the matrix $g_u f_p$ with $g_u = \partial g/\partial u$ and $f_p = \partial f/\partial p$ is invertible.*

Lemma 2.3 (d-index for mechanical systems). *Consider the semi-explicit DAE of second order,*

$$\begin{aligned} M(q)\ddot{q} &= f(q,\dot{q}) - G^T(q)\lambda, \\ 0 &= g(q). \end{aligned}$$

If $M(q)$ is positive definite and the Jacobian $G(q) := \partial g/\partial q$ is of full row rank, then this system is of d-index 3.

Proof. The proof is based on the fact that the matrix $GM^{-1}G^T$ is non-singular. Details are given in [**HW96**, Ch. VII.1]. □

2.1.2. *Further Index Concepts.* Although the d-index is sufficient for the purpose of this thesis, we give a short overview of further notions. The *perturbation index* measures the effects of perturbations of the right-hand side [**HLR89**]. This approach is similar to the differentiation index but may lead to different indices in special cases. For an example we refer to page 460 in [**HW96**]. Note, however, that the perturbation index does not distinguish between the single components of the system. For semi-explicit systems one can observe that the differential variables are more robust to perturbations than the algebraic ones.

If the system is not square or underdetermined, one needs a more general concept. One possibility is given by the *strangeness index* [**KM06**, Ch. 3], which is closely related to the differentiation index. An analysis of the strangeness index for (2.2) in terms of the semi-discrete Navier-Stokes equations is given in [**Wei97**]. Therein the strangeness index was chosen because the differentiation index may not be defined if the divergence operator is discretized in such a way that g_u is not of full rank. This may happen if the non-uniqueness of the pressure variable is reflected within the discretization.

Further definitions of other index concepts such as the *tractability* or *structural index* can be found in [**LMT13, Meh13**]. For the applications considered in this thesis, all these concepts are essentially equivalent. In the following, we refer to the d-index simply as *index*.

2.2. High-index DAEs. The numerical integration of DAEs with index 1 works essentially as for stiff ODEs [**HW96**, Ch. VI.1]. Even for DAEs of index 2 the convergence of classical Runge-Kutta schemes is often preserved. However, the order of convergence may be limited by two [**Arn98a**].

For DAEs of higher index, i.e., with index $\nu_d \geq 2$, the situation turns out to be worse and may lead to numerical instabilities due to the occurrence of derivatives of the right-hand side. As a consequence, a direct treatment is not advisable as also the iteration matrix is very ill-conditioned [**BCP96**, Ch. 5.4]. In general, the application of standard numerical methods (for ODEs) to high-index problems may lead to a reduction of the convergence order or even a loss of convergence [**Meh13**]. For the implicit Euler scheme and two Runge-Kutta methods with 2 and 3 stages a survey of the convergence orders for DAEs is given in Table 2.1.

TABLE 2.1. Order of convergence of different time integration schemes for ODEs and DAEs, cf. [**KM06**, Ch. 5.2].

	ODE	index 1	index 2	index 3	index 4	index 5
Implicit Euler	order 1	order 1	order 1	-	-	-
Radau IIa (s=2)	order 3	order 3	order 2	order 1	-	-
Radau IIa (s=3)	order 5	order 5	order 3	order 2	order 1	-

Before we deal with methods to decrease the index of a DAE, we show how the concept of modularized coupling may lead to DAEs of arbitrary high index. We illustrate this by means of an example in which we couple two subsystems. Consider the two DAEs of index 2,

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ f \end{bmatrix}, \qquad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ g \end{bmatrix}.$$

Coupling the two systems via $g = -\dot{x}_2$, we obtain a DAE of index 4 since the solution involves the third derivative of the right-hand side f. This difficulty should be in mind

when using automatic modeling, in particular for multi-physics systems, where different types of models are coupled. Because of this it is advisable to couple systems which are itself at most of index 1. Note, however, that this may lead to high-index DAEs as well.

The numerical problems arising from high-index DAEs motivate the idea of an index reduction. For this, the given system is modified to a system of lower index which has the same solution set. Several strategies are introduced in the next subsection.

2.3. Index Reduction Techniques. A common approach for the reduction of the index of a general nonlinear DAE

$$F(t, x, \dot{x}) = 0, \quad x(t_0) = x_0$$

is given by the *derivative array* approach [**KM06**, Ch. 6.2]. Since this approach does not assume any structure of the given system, the method works for all DAEs, which satisfy a certain hypothesis, cf. [**KM06**, Hyp. 3.48]. Within this procedure, one has to differentiate all equations $(\nu_d - 1)$ times and to find suitable projections to extract the differential and algebraic equations. For large systems of high index the derivative array becomes very large and may cause memory problems. This holds especially for systems coming from the semi-discretization of PDEs such as for flexible multibody systems.

The complexity can be reduced if additional information about the structure of the system is available. This is the case for semi-explicit DAEs as systems (2.2) or (2.3). Then, it suffices to build up a reduced derivative array. In Section 2.3.2 we discuss a variant where no projection matrices are needed. Instead, so-called *dummy variables* are introduced which extend the system. Nevertheless, the systems dimension remains of moderate size for many applications. Such an approach was introduced in [**MS93**] and later extended in [**KM04**]. This method is of particular interest as it is the base of the regularization of the operator DAEs in Part B.

2.3.1. *Index Reduction by Differentiation.* Before introducing the method of minimal extension in the next subsection, we study the simplest index reduction technique of all. Consider a DAE of semi-explicit structure, e.g., a DAE of second order

$$M(q)\ddot{q} = f(q, \dot{q}) - G^T(q)\lambda, \tag{2.5a}$$

$$0 = g(q). \tag{2.5b}$$

We assume that the DAE fulfills all assumptions of Lemma 2.3 such that it is of index 3. If the constraint $0 = g(q)$ is replaced by its second derivative $0 = \frac{\mathrm{d}^2}{\mathrm{d}t^2} g(q)$, we obtain a DAE of index 1.

Although the DAE is now suitable for numerical integration, we observe a so-called *drift-off*. This means that the constraint $0 = g(q)$ is violated independent of the used step size. The magnitude of the drift-off is analyzed in [**HW96**, Th. VII.2.1]. A detailed illustration of this phenomenon by means of the mathematical pendulum for different formulations and solvers can be found in [**Ste06**, Ex. 5.3.1].

2.3.2. *Minimal Extension.* In this subsection we apply the index reduction technique of *minimal extension* [**KM04**] to a constrained multibody system, see also [**KM06**, Ch. 6.4]. Consider again the system (2.5) with symmetric and positive definite mass matrix $M(q) \in \mathbb{R}^{n,n}$ and the Jacobian of the constraint $G(q) = \partial g / \partial q \in \mathbb{R}^{m,n}$, which is assumed to be of full row rank with $m \leq n$. From Lemma 2.3 we know that this system represents a DAE of index 3. Since $G(q)$ is of full row rank, there exists an orthogonal matrix $Q \in \mathbb{R}^{n,n}$ such that $G(q)Q$ has the block structure

$$G(q)Q = \begin{bmatrix} G_1 & G_2 \end{bmatrix}$$

with an invertible matrix $G_2 \in \mathbb{R}^{m,m}$. Note that the choice of Q is not unique and that we assume Q to be independent of time which may restrict to length of the computational time interval. The matrix Q then allows to partition the position variable q into

$$\begin{bmatrix} q_1 \\ q_2 \end{bmatrix} := Q^T q.$$

Thereby, the new variables are of size $q_1 \in \mathbb{R}^{n-m}$ and $q_2 \in \mathbb{R}^m$, consistent with the splitting of $G(q)$. Since we can identify the equations which have to be differentiated, namely the algebraic constraint, we consider the reduced derivative array. For this, we add to the original system the two derivatives of the constraint, i.e.,

$$0 = G(q)\dot{q} + g_t(q) \quad \text{and} \quad 0 = G(q)\ddot{q} + z(q, \dot{q})$$

with $z(q, \dot{q}) = 2G_t(q)\dot{q} + g_{tt}(q) + \partial G(q)/\partial q(\dot{q}, \dot{q})$. These equations are called the *hidden constraints*. To avoid the expensive search for projectors, we introduce two dummy variables $p_2 := \dot{q}_2$ and $r_2 := \ddot{q}_2$. Thus, we apply an extension instead of projecting the system to its original size. With the variables q_1, q_2, p_2, r_2, and λ, the extended system is then square. Replacing every occurrence of $\dot{q}_2$ and $\ddot{q}_2$ by its corresponding dummy variable, we obtain the overall system

$$\begin{aligned} M(q)Q \begin{bmatrix} \ddot{q}_1 \\ r_2 \end{bmatrix} &= f(q_1, q_2, \dot{q}_1, p_2) - G^T(q_1, q_2)\lambda, \\ 0 &= g(q_1, q_2), \\ 0 &= \begin{bmatrix} G_1 & G_2 \end{bmatrix} \begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} + g_t(q_1, q_2), \\ 0 &= \begin{bmatrix} G_1 & G_2 \end{bmatrix} \begin{bmatrix} \ddot{q}_1 \\ r_2 \end{bmatrix} + z(q_1, q_2, \dot{q}_1, p_2). \end{aligned}$$

The proof that the resulting DAE is of index 1 is given in [**KM06**, Th. 6.12]. It is based on the implicit function theorem and the structure of $G(q)Q$ which allows to write q_2, p_2, and r_2 in terms of q_1 and its derivatives. Then, the DAE reduces to a quasi-linear ODE for q_1 and an algebraic equation for λ.

Note that the dimension of the overall system has been increased by twice the number of constraints. Thus, for most applications the system is still of moderate size. The difficulty of this method is to find a suitable transformation Q. For time-dependent constraints it may happen that the matrix Q has to be adapted over time in order to guarantee the full rank property of the block G_2.

On the other hand, there are several applications where Q can be chosen as the identity matrix if a suitable reordering of the variables is assumed. In this case, the needed variable transformation is just a permutation and thus, all variables keep their physical meaning.

3. Functional Analytic Tools

This section gives a summary of functional analytic tools which are needed to formulate constrained dynamical systems as operator DAEs, i.e., as DAEs on an abstract level. Starting from the definition of distributions, we introduce the concept of Sobolev spaces which is needed for the weak formulation of PDEs. In the analysis of PDEs, these spaces have proven to be more suitable than the classical C^k spaces of continuously differentiable functions.

For the notion of abstract differential equations, we consider so-called *abstract functions*, i.e., functions of the form

$$f\colon [0,T] \to X$$

with a real Banach space X and a bounded time interval $[0,T]$. We introduce Bochner integrals in Section 3.2, which allow to integrate abstract functions, and the corresponding function spaces which generalize the concept of Lebesgue spaces. A further important tool for abstract differential equations is the notion of Gelfand triples as well as the generalization of distributions. This then leads to Sobolev-Bochner spaces for which we summarize several properties in Section 3.3.

3.1. Fundamentals. Within this section, $\Omega \subset \mathbb{R}^d$ always denotes a *domain*, i.e., Ω is open, connected, and bounded. Furthermore, the domain is assumed to be non-empty. The boundary of Ω, namely $\partial\Omega$, can be classified in terms of smoothness.

Definition 3.1 (C^k-boundary [**RR04**, Def. 7.9]). A domain $\Omega \subset \mathbb{R}^d$ has a *C^k-boundary*, $k \geq 1$, if for every point $x \in \partial\Omega$ there exists a neighborhood N_x such that $N_x \cap \partial\Omega$ is a C^k-surface. Furthermore, $N_x \cap \Omega$ has to be 'on one side' of $N_x \cap \partial\Omega$.

Definition 3.2 (Lipschitz boundary [**RR04**, Def. 7.10]). The boundary of a domain $\Omega \subset \mathbb{R}^d$ is called *Lipschitz* if for every point $x \in \partial\Omega$ there exists a neighborhood N_x such that $N_x \cap \partial\Omega$ is the graph of a uniformly Lipschitz continuous function. Furthermore, $N_x \cap \Omega$ has to be 'on one side' of $N_x \cap \partial\Omega$.

Remark 3.3 (Polygonal domains). For simulations which rely on finite element discretizations and thus, triangulations of the domain Ω, polygonal domains play a special role. If two neighboring boundary edges touch each other only at nodes and each boundary node is the end of exactly two boundary edges, then the polygonal domain has a boundary of Lipschitz type. In particular, this excludes domains with crack.

3.1.1. *Dual Operators and Riesz Representation Theorem.* In this subsection we recall some basic properties of operators between Banach spaces such as the existence of a dual operator. For Hilbert spaces we obtain the so-called adjoint operator due to the representation theorem of Riesz. This subsection is based on the two chapters [**Yos80**, Ch. VII] and [**RR04**, Ch. 8.4].

Consider two real Banach spaces X and Y and a linear operator $A\colon \mathcal{D}(A) \subset X \to Y$, where $\mathcal{D}(A)$ denotes the *domain* of A. The *range* $\mathcal{R}(A)$ then denotes the subspace of Y, given by

$$\mathcal{R}(A) := \{y \in Y \mid \text{there exists an element } x \in \mathcal{D}(A) \text{ with } y = Ax\}.$$

The *null space* or *kernel* of the operator A is the subspace of X which is defined by $\ker(A) := \{x \in X \mid Ax = 0\}$. As in the finite-dimensional case, linear operators are invertible if and only if its kernel contains only the zero element. The inverse operator is then also linear [**RR04**, Th. 8.3].

For a given operator $A\colon \mathcal{D}(A) \subset X \to Y$ we want to define a mapping between the the *dual spaces* of Y and X, which generalizes the transpose of a matrix. The dual space, namely X^*, contains all linear functionals on X, i.e., linear bounded mappings from X to $\mathbb{R}$. Given such a functional $w \in X^*$, the action on an element $x \in X$ is defined by the *duality pairing*, $\langle w, x\rangle_{X^*,X} := w(x)$.

Definition 3.4 (Dual operator [**Yos80**, Def. VII.1.1]). Consider a linear operator $A\colon \mathcal{D}(A) \subset X \to Y$, where $\mathcal{D}(A)$ is dense in X. Let $\mathcal{D}(A^*)$ denote the following subset of Y^*: An element $v \in Y^*$ satisfies $v \in \mathcal{D}(A^*)$ if there exists an element $w \in X^*$ such that for all $x \in \mathcal{D}(A)$ it holds that

$$\langle v, Ax\rangle_{Y^*,Y} = \langle w, x\rangle_{X^*,X}.$$

This defines the mapping $A^*\colon \mathcal{D}(A^*) \subset Y^* \to X^*$ given by $A^*v := w$. The operator A^* is called the *dual operator* of A.

The dual of a linear operator is linear and satisfies for $x \in \mathcal{D}(A)$ and $v \in \mathcal{D}(A^*)$,

$$\langle v, Ax\rangle_{Y^*,Y} = \langle A^*v, x\rangle_{X^*,X}.$$

Furthermore, if A is linear and continuous, then $\mathcal{D}(A^*) = Y^*$ and A^* is linear and continuous as well [**Yos80**, Th. VII.1.2].

We now consider the situation for Hilbert spaces H which are isometric to their dual space. In particular, the following theorem provides a representation of functionals in H^* by elements in H.

Theorem 3.5 (Riesz representation theorem [**RR04**, Th. 6.52]). *Let H be a Hilbert space with inner product $(\cdot,\cdot)_H$. Then, there exists an invertible and isometric mapping $J\colon H^* \to H$ such that*

$$\langle h, x\rangle_{H^*,H} = (Jh, x)_H$$

for all $h \in H^*$ and $x \in H$. This operator is called the Riesz mapping.

Remark 3.6 (Embedding $H \hookrightarrow H^*$). The inverse of the Riesz mapping, $J^{-1}\colon H \to H^*$, which maps an element $x \in H$ to the functional $(x,\cdot)_H$, characterizes one possible continuous embedding $H \hookrightarrow H^*$. A second possibility will be introduced in Section 3.3.1 below by means of a Gelfand triple.

The combination of the dual operator and the Riesz mapping yields the so-called *adjoint operator* (or Hilbert adjoint) of A. For two Hilbert spaces H_1, H_2 and $A\colon H_1 \to H_2$, the adjoint operator $A^{\mathrm{ad}} := J_{H_1} A^* J_{H_2}^{-1}\colon H_2 \to H_1$ satisfies

$$(A^{\mathrm{ad}}y, x)_{H_1} = \langle A^* J_{H_2}^{-1} y, x\rangle_{H_1^*,H_1} = \langle J_{H_2}^{-1} y, Ax\rangle_{H_2^*,H_2} = (y, Ax)_{H_2}$$

for all $x \in H_1$ and $y \in H_2$.

3.1.2. *Test Functions and Distributions.* To generalize the concept of derivatives, which is necessary for the later analysis of differential equations, we have to introduce so-called *test functions*. For a domain $\Omega \subset \mathbb{R}^d$, these are smooth functions which have a compact support in Ω. The set of all test functions is denoted by $\mathcal{D}(\Omega) := C_0^\infty(\Omega)$. We say that a sequence of test functions Φ_n, $n \in \mathbb{N}$, *converges in* $\mathcal{D}(\Omega)$ to a function $\Phi \in \mathcal{D}(\Omega)$ if all derivatives of Φ_n converge uniformly to those of Φ. Several properties of test functions can be found in [**RR04**, Ch. 5.1]. The latter definition allows to introduce distributions as the generalization of a function.

Definition 3.7 (Distribution [**RR04**, Def. 5.8]). A linear mapping $\Phi \mapsto (f, \Phi)$, which maps from $\mathcal{D}(\Omega)$ to $\mathbb{R}$, is called a *distribution* if it is continuous, i.e., the convergence of a sequence $\Phi_n \to \Phi$ in $\mathcal{D}(\Omega)$ implies $(f, \Phi_n) \to (f, \Phi)$.

We remark that a continuous function f can be identified with a distribution due to

$$(f, \Phi) := \int_\Omega f(x)\Phi(x)\,\mathrm{d}x.$$

The set of distributions also includes the Dirac delta function which is no function in the classical sense. Since the definition of distributions is based on smooth functions, we can define derivatives of arbitrary order.

DEFINITION 3.8 (Generalized derivative [**RR04**, Ch. 5.2]). The derivative with respect to the multi-index α of a distribution f is defined by

$$\big(D^\alpha f, \Phi\big) := (-1)^{|\alpha|}\big(f, D^\alpha \Phi\big).$$

REMARK 3.9. The derivative of a distribution is again a distribution. Furthermore, the definition coincides with the classical derivative for functions $f \in C^1(\Omega)$ due to the integration by parts formula.

The generalization of the derivative permits to define weak solutions of differential equations. For this approach, mainly used for PDEs, the equation of interest is multiplied by a test function and then, the integration by parts formula is applied. Pushing some or all derivatives to the test function, we obtain the notion of weak solutions which may be of lower regularity than stated in the original formulation. In the following subsection, we use the notion of distributions to define Sobolev spaces.

3.1.3. *Sobolev Spaces.* This subsection is devoted to a short summary of Sobolev spaces and corresponding embedding results. These spaces are based on generalized derivatives from the previous subsection and the Lebesgue spaces $L^p(\Omega)$. The given definitions and results of this subsection can be found in standard text books on functional or numerical analysis, e.g., in [**AF03, Tar07**] or [**RR04**, Ch. 7].

DEFINITION 3.10 (Sobolev space $W^{k,p}(\Omega)$). Consider a domain $\Omega \subset \mathbb{R}^d$ and any integer $k \geq 0$ and $0 \leq p \leq \infty$. Then, the *Sobolev space* $W^{k,p}(\Omega)$ contains all distributions $u \in L^p(\Omega)$ which have (generalized) derivatives $D^\alpha u \in L^p(\Omega)$ for all multi-indices α with length $|\alpha| := \alpha_1 + \cdots + \alpha_d \leq k$.

Let $\|\cdot\|_p$ and $\|\cdot\|_\infty$ denote the norms of $L^p(\Omega)$ and $L^\infty(\Omega)$, respectively. Then, the space $W^{k,p}(\Omega)$ is a Banach space equipped with the norm

$$\|u\|_{k,p} := \Big(\sum_{|\alpha|\leq k} \|D^\alpha u\|_p^p\Big)^{1/p} \tag{3.1}$$

for $p < \infty$, and otherwise

$$\|u\|_{k,\infty} := \max_{|\alpha|\leq k} \|D^\alpha u\|_\infty.$$

In the special case $p = 2$, we obtain a Hilbert space and write $H^k(\Omega) := W^{k,2}(\Omega)$. For this, we equip the space with the inner product

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha|\leq k} \big(D^\alpha u, D^\alpha v\big)_{L^2(\Omega)} = \sum_{|\alpha|\leq k} \int_\Omega D^\alpha u D^\alpha v\,\mathrm{d}x.$$

Note that for $k = 0$ we obtain the Lebesgue space $H^0(\Omega) = L^2(\Omega)$. Since Sobolev spaces are based on Lebesgue spaces, we obtain the likewise result concerning separability and reflexivity of $W^{k,p}(\Omega)$.

THEOREM 3.11 (Separability and reflexivity). *For $p < \infty$ the spaces $W^{k,p}(\Omega)$ are separable. Furthermore, $W^{k,p}(\Omega)$ is reflexive if $1 < p < \infty$.*

Most of the proofs for elements of Sobolev spaces are based on density arguments. Here we only state that $C^\infty(\Omega) \cap H^k(\Omega)$ is dense in $H^k(\Omega)$, see [**RR04**, Lem. 7.48]. Furthermore, one is interested in embeddings of Sobolev spaces into each other as well as the question which Sobolev spaces are embedded in the space of continuous functions $C(\Omega)$ or even continuously differentiable functions (in the classical sense). Two negative examples are given by $H^1(\Omega) \not\subset L^p(\Omega)$ for $p > 6$, see [**Tar07**, Lem. 8.1], and $H^1(\Omega) \not\subset C(\Omega)$ for $d \geq 2$.

THEOREM 3.12 (Sobolev embedding I [**Ste08**, Th.2.5]). *Consider a domain $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary and $p > 1$. Then, for all parameters $s > d/p$ we obtain the continuous embedding $W^{s,p}(\Omega) \hookrightarrow C(\Omega)$.*

THEOREM 3.13 (Sobolev embedding II [**BS08**, Sect.1.4]). *Consider a domain $\Omega \subset \mathbb{R}^d$, non-negative integers $k \leq m$, and real numbers $1 \leq p \leq q \leq \infty$. Then, we obtain the continuous embedding $W^{m,q}(\Omega) \hookrightarrow W^{k,p}(\Omega)$.*

It is also possible to define Sobolev spaces $W^{s,p}(\Omega)$ with $s \in \mathbb{R}$, so-called *broken Sobolev spaces* [**AF03**]. We will only consider the special case of $s = 1/2$. For this exponent we obtain the space of traces as introduced in the following subsection.

3.1.4. *Traces.* As mentioned in the previous subsection, functions in Sobolev spaces are not necessarily continuous. This leads to the question whether Sobolev functions can be 'restricted' to surfaces of measure zero, in particular on the boundary of a domain Ω, the so-called *trace*. This property is crucial to enforce Dirichlet boundary conditions for PDEs. The presented results are taken from [**Tar07**, Ch. 13], [**BF91**, Ch. III.1], and [**Ste08**, Ch. 2].

For continuous functions in $C(\overline{\Omega})$ the restriction to the boundary $\partial\Omega$ is well-defined. This restriction defines a linear operator which can be continuously extended to functions in $H^1(\Omega)$. Note that the extension itself is not the restriction to the boundary, since this is not defined as $\partial\Omega$ is of measure zero [**Tar07**, Ch. 13]. The proof of the well-posedness of the extension is given in [**Ste08**, Th. 2.21] and motivates the following definition.

DEFINITION 3.14 (Trace operator). Consider a domain $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary. Then, the extension of the restriction operator on $\partial\Omega$ defines a linear and bounded operator $\gamma\colon H^1(\Omega) \to H^{1/2}(\partial\Omega)$, the so-called *trace operator*.

THEOREM 3.15 (Inverse trace theorem [**Ste08**, Th. 2.22]). *The trace operator from Definition 3.14 has a continuous right inverse, meaning that there exists a bounded operator $\mathcal{E}\colon H^{1/2}(\partial\Omega) \to H^1(\Omega)$ with $\gamma\mathcal{E}w = w$ for all $w \in H^{1/2}(\partial\Omega)$.*

REMARK 3.16. Justified by Theorem 3.15, Definition 3.14 also defines $H^{1/2}(\partial\Omega)$ as the trace space of $H^1(\Omega)$, i.e., the range of γ. Thus, a function defined on the boundary satisfies $w \in H^{1/2}(\partial\Omega)$ if and only if there exists a Sobolev function $v \in H^1(\Omega)$ with $\gamma v = w$. Note that the space $H^{1/2}(\partial\Omega)$ is a Hilbert space [**BF91**, Ch. III.1].

Remark 3.16 motivates the definition of a norm for the trace space with the help of the $H^1(\Omega)$-norm. For this, we may define

$$\|w\|_{H^{1/2}(\partial\Omega)} := \inf_{\substack{v \in H^1(\Omega),\\ \gamma v = w}} \|v\|_{1,2}.$$

An equivalent norm can be defined by the solution of a corresponding Dirichlet problem [**BF91**, Ch. III.1]. The norm then reads

$$\|w\|_{H^{1/2}(\partial\Omega)} := \|\bar{w}\|_{1,2}$$

where $\bar{w} \in H^1(\Omega)$ is the unique (weak) solution of

$$-\Delta \bar{w} + \bar{w} = 0 \qquad \text{in } \Omega, \tag{3.2a}$$

$$\bar{w} = w \qquad \text{on } \partial\Omega. \tag{3.2b}$$

We neglect the straightforward proof that this defines a norm on $H^{1/2}(\partial\Omega)$ but show that $\|w\|_{H^{1/2}(\partial\Omega)} = 0$ implies $w = 0$. For this, we deduce from $\|w\|_{H^{1/2}(\partial\Omega)} = 0$ that $\bar{w}$ has to vanish on Ω because $\|\cdot\|_{1,2}$ forms a norm. As solution of the corresponding Dirichlet problem, we finally get $0 = w$ on $\partial\Omega$ in the sense of traces.

Analogously, an inner product in $H^{1/2}(\partial\Omega)$ is given by

$$(v, w)_{H^{1/2}(\partial\Omega)} := (\bar{v}, \bar{w})_{H^1(\Omega)}.$$

Therein, $\bar{v}$ and $\bar{w}$ again denote the solution of the corresponding homogeneous Dirichlet problem (3.2) with boundary conditions v and w, respectively.

The space of traces is also defined for non-empty subsets (in the $(d-1)$-dimensional measure) of the boundary $\Gamma \subset \partial\Omega$, namely $H^{1/2}(\Gamma)$. It can be defined by the closure of all test functions $\mathcal{D}(\Gamma) \hookrightarrow \mathcal{D}(\partial\Omega)$ with respect to the norm $\|\cdot\|_{H^{1/2}(\partial\Omega)}$. A norm is given by

$$\|w\|_{H^{1/2}(\Gamma)} := \inf_{\substack{v \in H^1(\Omega), \\ \gamma v|_\Gamma = w}} \|v\|_{1,2}.$$

Remark 3.17. Clearly, test functions in $\mathcal{D}(\Gamma)$ can be extended by zero to the entire boundary $\partial\Omega$. However, one has to be aware of the fact that a function in $H^{1/2}(\Gamma)$ can not always be extended by zero to a function in $H^{1/2}(\partial\Omega)$, see [**BF91**, Ch. III.1].

Within this thesis, we often omit to write the trace operator explicitly, i.e., we write u instead of γu. Since we have defined the trace operator by a density argument, the operator γ is analogously defined for functions in $W^{1,p}(\Omega)$. Embedding theorems for Sobolev spaces then imply that the product of two traces is also well-defined [**Tar07**, Lem. 13.3]. An important subspace of $W^{1,p}(\Omega)$ is defined by the kernel of γ.

Definition 3.18 ($W_0^{1,p}(\Omega)$ and $H_0^1(\Omega)$). Let the boundary of the domain Ω be Lipschitz and $p > 1$. Then, the subspace $W_0^{1,p}(\Omega)$ is defined as the kernel of γ in $W^{1,p}(\Omega)$. In particular, $H_0^1(\Omega)$ denotes the subspace of $u \in H^1(\Omega)$ with $\gamma u = 0$.

Remark 3.19. An alternative to Definition 3.18 is given by the closure of $\mathcal{D}(\Omega)$ with respect to the norm $\|\cdot\|_{k,p}$ from (3.1), cf. [**Tar07**, Def. 6.6]. This then leads, more generally, to the subspaces $W_0^{k,p}(\Omega)$ and $H_0^k(\Omega)$ of $W^{k,p}(\Omega)$ and $H^k(\Omega)$, respectively.

Remark 3.20. The weak solution $\bar{w} \in H^1(\Omega)$ of the Dirichlet problem (3.2) is orthogonal to $H_0^1(\Omega)$ w.r.t. the inner product of $H^1(\Omega)$. Thus, $\bar{w}$ equals the unique element in $H_0^1(\Omega)^\perp$ which has the trace $\gamma\bar{w} = w$.

Remark 3.21. Similarly to Definition 3.18, $H_\Gamma^1(\Omega)$ denotes the subspace of $H^1(\Omega)$ with all functions that vanish along $\Gamma \subset \partial\Omega$ in the sense of traces. This definition requires that Γ is of positive surface measure.

3.1.5. *Poincaré Inequality and Negative Norms.* A peculiarity of the Sobolev norms (3.1) is the mixture of different units due to the involved derivatives. For some subspaces $V \subset W^{1,p}(\Omega)$ it is possible to avoid the $\|u\|_p$ term within the norm [**Tar07**, Ch. 10]. Within this subsection we assume Ω to be framed by a Lipschitz boundary.

We say that a subspace V of $W^{1,p}(\Omega)$ satisfies a *Poincaré inequality* if there exists a constant $c > 0$ such that

$$\|u\|_p \leq c \, \|\nabla u\|_p$$

for all $u \in V$. Such an inequality then implies that the norms $\|\cdot\|_{1,p}$ and $\|\nabla \cdot\|_p$ are equivalent on V. Obviously, the Poincaré inequality cannot hold for subspaces that contain the constant function 1.

LEMMA 3.22 (Poincaré inequality [**Tar07**, Lem. 10.2]). *Let $\Omega \subset \mathbb{R}^d$ be a domain with Lebesgue measure $|\Omega|$. Then, the space $W_0^{1,p}(\Omega)$ satisfies a Poincaré inequality of the form*

$$\|u\|_p \leq c(p)|\Omega|^{1/d}\|\nabla u\|_p$$

for all $u \in W_0^{1,p}(\Omega)$.

REMARK 3.23. This result can be generalized for functions which do not vanish along the entire boundary, i.e., the Poincaré inequality is also valid for functions in $H^1_\Gamma(\Omega)$ if $\Gamma \subset \partial\Omega$ has positive surface measure [**Rou05**, Ch. 1.4].

REMARK 3.24. Lemma 3.22 remains valid for Sobolev spaces of higher order [**RR04**, Rem. 7.33]. For $1 < p < \infty$ there exists a constant $c = c(k, p, d, \Omega) > 0$ such that

$$\|u\|_{k,p}^p \leq c \sum_{|\alpha|=k} \|D^\alpha u\|_p^p$$

for all $u \in W_0^{k,p}(\Omega)$.

Functions of subspaces which satisfy a Poincaré inequality do not necessarily vanish along the boundary. Consider the inequality

$$\|u\|_{1,p} \leq c \, \Big(\|\nabla u\|_p + \Big|\int_\Omega u \, \mathrm{d}x\Big|\Big). \tag{3.3}$$

which is valid for all $u \in W^{1,p}(\Omega)$ if Ω has a Lipschitz boundary [**Rou05**, Ch. 1.4]. This implies a Poincaré inequality also for the subspace of $W^{1,p}(\Omega)$ with vanishing mean value. Note that the integral term in (3.3) may be replaced by any other $W^{1,p}(\Omega)$-continuous seminorm (i.e., a seminorm $|\cdot|$ which satisfies $|\cdot| \leq c \, \|\cdot\|_{1,p}$) which does not vanish for the constant function 1. A particular result for convex Lipschitz domains is given by the following lemma.

LEMMA 3.25 (Payne–Weinberger [**PW60**]). *Let $\Omega \subset \mathbb{R}^2$ be a convex Lipschitz domain with diameter* $\operatorname{diam}(\Omega)$*. Then, every function $u \in H^1(\Omega)$ with integral mean $\bar{u} = \int_\Omega u \, dx$ satisfies*

$$\|u - \bar{u}\|_2 \leq \frac{\operatorname{diam}(\Omega)}{\pi}\|\nabla u\|_2. \tag{3.4}$$

We close this subsection with the introduction of *negative norms* and the corresponding Sobolev spaces of negative order.

DEFINITION 3.26 ($H^{-k}(\Omega)$). The space $H^{-k}(\Omega)$ is defined as the dual space of $H_0^k(\Omega)$.

Since the space $H^{-k}(\Omega)$ is defined by duality, the norm is given by

$$\|f\|_{-k,2} := \|f\|_{H^{-k}(\Omega)} := \sup_{v \in H_0^k(\Omega)} \frac{\langle f, v\rangle}{\|v\|_{k,2}}.$$

3.1.6. *Weak Convergence and Compactness.* A fundamental property of infinite-dimensional normed spaces is that the closed unit ball is not compact [**Ruž04**, Th. A.8.1]. As a consequence, bounded sequences do not need to have a convergent subsequence. Here, we mean *strong convergence* in X or convergence in norms, i.e., $x_n \to x$ if and only if

$$\|x_n - x\|_X \to 0 \quad \text{as } n \to \infty.$$

In order to retain this compactness property, we have to switch over to weaker topologies which leads to the notion of *weak convergence*. The results of this subsection are important for the convergence proofs in Part D of this thesis. All definitions and results from this subsection can be found in [**Alt92**, Ch. 5] and [**Ruž04**, App. A.8]. Furthermore, in what follows X always denotes a Banach space.

Definition 3.27 (Weak convergence). A sequence $(x_n) \subset X$ is *weakly convergent* to $x \in X$ if and only if for all functionals $f \in X^*$ is holds that

$$\langle f, x_n \rangle \to \langle f, x \rangle \quad \text{as } n \to \infty.$$

In this case, we write $x_n \rightharpoonup x$.

Because of the involved functionals, the definition can be seen as the generalization of the convergence in all coordinates in the finite-dimensional setting. At this point we note that the weak limit is unique and that a weakly convergent sequence is bounded. For sequences in a dual space, we define a second kind of weak convergence.

Definition 3.28 (Weak* convergence). A sequence $(f_n) \subset X^*$ is *weak* convergent* to $f \in X^*$ if and only if for all $x \in X$ is holds that

$$\langle f_n, x \rangle \to \langle f, x \rangle \quad \text{as } n \to \infty.$$

In this case, we write $f_n \overset{*}{\rightharpoonup} f$.

Remark 3.29. The two definitions above provide two different kinds of weak convergence for sequences in the dual space X^*. If X is a reflexive Banach space, then these two notions coincide.

In terms of the introduced weak topologies, we state the following compactness results. The first result is based on the theorem of Banach-Alaoglu [**Zei86**, App.] which states that the closed unit ball $B = \{f \in X^* \mid \|f\|_{X^*} \leq 1\} \subseteq X^*$ is compact with respect to the weak* topology.

Theorem 3.30 (Weak* compactness). *Let X be a separable Banach space. Then, every bounded sequence in X^* has a weak* convergent subsequence.*

For a reflexive Banach space this leads to the following theorem.

Theorem 3.31 (Weak compactness [**Alt92**, Th. 5.7]). *Let X be a reflexive Banach space. Then, every bounded sequence in X has a weakly convergent subsequence.*

3.2. Bochner Spaces. This subsection is devoted to the definition of an integral for abstract functions, i.e., for functions with values in a Banach space X, the so-called *Bochner integral*. The presented results are based on [**Emm04**, Ch. 7.1].

As in the theory of Lebesgue measures, we first consider simple functions, i.e., functions which take only a finite number of values $\{u_i\}_{i=1,\dots,n} \subset X$. Thus, for Lebesgue measurable sets $\{A_i\}_{i=1,\dots,n} \subseteq [0,T]$ with characteristic functions χ_{A_i}, a *simple function* $u\colon [0,T] \to X$ has the form $u(t) = \sum_{i=1}^n u_i\ \chi_{A_i}(t)$. The integral of a simple function is then defined as

$$\int_0^T u(t)\ \mathrm{d}t := \sum_{i=1}^n u_i \mu(A_i).$$

Therein, we have used the Lebesgue measure μ. Note that the integral is again an element of the Banach space X. Measurable functions are then defined as point-wise limits of simple functions.

Definition 3.32 (Bochner measurability [**Emm04**, Def. 7.1.9]). A function $u\colon [0,T] \to X$ is called *Bochner measurable* if there exists a sequence (u_n) of simple functions such that $u_n(t) \to u(t)$ for a.e. $t \in [0,T]$ as $n \to \infty$.

Remark 3.33. The convergence of the sequence (u_n) in Definition 3.32 is required to hold strongly in X. The concept of weak Bochner measurability is not considered here, since it coincides with the strong measurability for separable Banach spaces X. All applications throughout this thesis work on separable spaces.

As in the theory of Lebesgue, the next step is to introduce the notion of integrability.

Definition 3.34 (Bochner integrability [**Emm04**, Def. 7.1.14]). Consider a Bochner measurable function $u\colon [0,T] \to X$ and a sequence of simple functions (u_n) with $u_n(t) \to u(t)$. Then, u is called *Bochner integrable* if for every $\varepsilon > 0$ there exists a number $n_\varepsilon \in \mathbb{N}$ such that for all $n, m > n_\varepsilon$ it holds that

$$\int_0^T \|u_n - u_m\|_X \, \mathrm{d}t < \varepsilon.$$

For Bochner integrable functions, the integral over a Lebesgue measurable set $A \subseteq [0,T]$ with the corresponding characteristic function χ_A is defined via

$$\int_A u(t) \, \mathrm{d}t := \lim_{n\to\infty} \int_0^T u_n(t)\chi_A(t) \, \mathrm{d}t.$$

Note that the Bochner integral is a generalization of the Lebesgue integral since they coincide in the case $X = \mathbb{R}$. The strong connection between these two concepts is presented in the following result.

Proposition 3.35. *Let X be a separable Banach space. Then u is Bochner measurable if and only if $\langle f, u(\cdot)\rangle_{X^*,X}$ is Lebesgue measurable for every functional $f \in X^*$. Furthermore, a Bochner measurable function u is Bochner integrable if and only if $\|u(\cdot)\|_X$ is Lebesgue integrable.*

Proof. This result goes back to Pettis and can be found in [**Rou05**, Th. 1.34]. □

Given the relatedness of Bochner and Lebesgue integrability, it is no surprise that the Bochner integral adopts several properties from the theory of Lebesgue integrals. Some properties are summarized in the following proposition.

Proposition 3.36 (Properties of the Bochner integral [**Emm04**, Th. 7.1.15]). *Let X and Y be Banach spaces and let $u\colon [0,T] \to X$ be a Bochner integrable function. Then, for any Lebesgue measurable set $A \subseteq [0,T]$ and functional $f \in X^*$ it holds that*

$$\Big\| \int_A u(t) \; dt \; \Big\|_X \leq \int_A \|u(t)\|_X \; dt, \qquad \Big\langle f, \int_A u(t) \; dt \Big\rangle_{X^*,X} = \int_A \langle f, u(t)\rangle_{X^*,X} \; dt.$$

For a linear, continuous operator $\mathcal{K}\colon X \to Y$, the map $\mathcal{K}u(\cdot)$ is Bochner integrable and

$$\mathcal{K} \int_A u(t) \; dt = \int_A \mathcal{K}u(t) \; dt.$$

Remark 3.37. The latter proposition shows that Bochner integrals are fully defined by the action of linear functionals on the integrand.

In the sequel we utilize the notion $C([0,T];X)$ for abstract functions with values in X which are continuous in $[0,T]$. Accordingly, $AC([0,T];X)$ denotes the space of absolutely continuous functions with values in X. With this, the Bochner integral of abstract functions allows to introduce the concept of primitives. For a Bochner integrable function $u\colon [0,T] \to X$ we define the absolutely continuous function $\tilde{u} \in AC([0,T];X)$ by

$$\tilde{u}(t) := \int_0^t u(s)\,\mathrm{d}s.$$

The proof for $\tilde{u} \in AC([0,T];X)$ and the fact that $\tilde{u}$ is a.e. differentiable (in the classical sense) is shown in [**Emm04**, Th. 7.1.19]. The converse, i.e., the Bochner integrability of derivatives of absolutely continuous functions, only applies if X is reflexive, see [**Rou05**, Th. 1.39].

Collecting functions which coincide a.e. in equivalence classes, we obtain the notion of Bochner spaces.

Definition 3.38 (Bochner spaces $L^p(0,T;X)$). For $p \geq 1$ the linear space $L^p(0,T;X)$ is called *Bochner space* and contains the equivalence classes of Bochner integrable functions $u\colon [0,T] \to X$ which satisfy

$$\|u\|_{L^p(0,T;X)} := \Big(\int_0^T \|u(t)\|_X^p \,\mathrm{d}t \Big)^{1/p} < \infty$$

if $p < \infty$ and $\|u\|_{L^\infty(0,T;X)} := \operatorname{ess\,sup}_t \|u(t)\|_X < \infty$ in the case $p = \infty$.

As for Lebesgue integrable functions, we may also define the space $L^1_{\mathrm{loc}}(0,T;X)$ as the space of functions which are Bochner integrable on every compact subset of $]0,T[$. A number of properties of the Bochner spaces $L^p(0,T;X)$ are summarized in the following proposition.

Proposition 3.39 (Properties of Bochner spaces [**Emm04**, Th. 7.1.23]). *Let X and Y be Banach spaces with $X \hookrightarrow Y$, i.e., X is continuously embedded in Y, and H a Hilbert space. Then,*

(a) with $\|u\|_{L^p(0,T;X)}$ from Definition 3.38, $L^p(0,T;X)$ forms a Banach space,

(b) if X is separable, then so is $L^p(0,T;X)$ for all $1 \leq p < \infty$,

(c) if X is reflexive or X^* separable, then $L^p(0,T;X)$ is reflexive for all $1 < p < \infty$,

(d) $L^2(0,T;H)$ is a Hilbert space, and

(e) if $1 \leq q \leq p \leq \infty$, then $L^p(0,T;X) \hookrightarrow L^q(0,T;Y)$.

We close this subsection with a characterization of the dual space of $L^p(0,T;X)$.

Proposition 3.40 (Dual of Bochner spaces and Hölder inequality). *Consider $1 < p < \infty$ with conjugate exponent $p' = p/(p-1)$. If $L^p(0,T;X)$ is reflexive, then its dual space can be identified with the space $L^{p'}(0,T;X^*)$. The corresponding dual pairing is given by*

$$\langle f, x \rangle := \int_0^T \big\langle f(t), x(t) \big\rangle_{X^*,X} \, dt.$$

Furthermore, the Hölder inequality *holds, i.e., for $x \in L^p(0,T;X)$ and $f \in L^{p'}(0,T;X^*)$ we have*

$$\int_0^T \big\langle f(t), x(t) \big\rangle_{X^*,X} \, dt \leq \|f\|_{L^{p'}(0,T;X^*)} \|x\|_{L^p\,(0,T;X)}.$$

Proof. The first part of the claim is stated in [**Emm04**, Th. 7.1.23]. A proof of the Hölder inequality can be found in [**GGZ74**, Ch. IV.2]. □

3.3. Sobolev-Bochner Spaces. In this subsection, we discuss the interaction of Sobolev and Bochner spaces. This is of special relevance for the formulation of abstract differential equations which involves (generalized) derivatives of Bochner integrable functions. For this, we have to introduce Gelfand triples which then leads to certain embeddings for Sobolev-Bochner spaces.

3.3.1. *Gelfand Triples.* For the formulation of abstract ODEs in Section 4 it is beneficial to use different Sobolev spaces for the solution u and its derivative $\dot{u}$. In fact, a third space is needed to provide suitable initial conditions. A formalism, which has proven its worth, is the so-called Gelfand or evolution triple.

This subsection is based on the two chapters [**Emm04**, Ch. 8.1] and [**Wlo87**, Ch. 17.1].

Definition 3.41 (Gelfand triple [**Emm04**, Def. 8.1.7]). Consider a real, separable, and reflexive Banach space V and a real, separable Hilbert space H. If V is continuously and densely embedded in H, then the spaces V, H, V^* form a *Gelfand triple*. The space H is called the *pivot*.

A Gelfand triple is often written in the form $V \overset{d}{\hookrightarrow} H \cong H^* \overset{d}{\hookrightarrow} V^*$ which indicates the resulting continuous and dense embedding $H^* \hookrightarrow V^*$. This notion requires a justification which we provide in the following. The equivalence of the Hilbert spaces H and H^* is given by the Riesz representation theorem, see Theorem 3.5. Furthermore, the continuous embedding $V \hookrightarrow H$ implies the existence of a constant $c > 0$ with $\|v\|_H \leq c\|v\|_V$. Therein, $\|\cdot\|_V$ and $\|\cdot\|_H$ denote the norms in V and H, respectively. Consider a functional $f \in H^*$ which is, due to $V \hookrightarrow H$, also a linear functional on V, i.e., $f \in V^*$. We show that this embedding $H^* \hookrightarrow V^*$, characterized by the Gelfand triple, is continuous,

$$\|f\|_{V^*} = \sup_{v \in V} \frac{\langle f, v\rangle}{\|v\|_V} \leq c \cdot \sup_{v \in V} \frac{\langle f, v\rangle}{\|v\|_H} \leq c \cdot \sup_{v \in H} \frac{\langle f, v\rangle}{\|v\|_H} = c\, \|f\|_{H^*}.$$

Note that H^* is dense in V^* because $V \hookrightarrow H$ is assumed to be dense and V reflexive.

Another consequence of the Gelfand triple concerns the duality pairing $\langle\cdot,\cdot\rangle_{V^*,V}$. Because of $H \cong H^* \hookrightarrow V^*$, the duality pairing of V, V^* is the continuous extension of the inner product in H, namely $(\cdot,\cdot)_H$. Thus, for $h \in H$ and $v \in V$, we obtain

$$\langle h, v\rangle_{V^*,V} = (h, v)_H.$$

For a functional $f \in V^*$ there exists a sequence $(h_n) \subset H$ such that $J^* h_n \to f$ in V^* with the Riesz mapping J. Thus, for $v \in V$ is holds that

$$\langle f, v\rangle_{V^*,V} := \lim_{n\to\infty} (h_n, v)_H.$$

Remark 3.42. Consider the case where also V is a Hilbert space. We emphasize the fact that the embedding $V \hookrightarrow V^*$ from Theorem 3.5 does not coincide with the embedding given by the Gelfand triple V, H, V^*. For $u, v \in V$ we obtain the two different cases

$$\begin{aligned} &\text{Riesz:} && \langle v, u\rangle_{V^*,V} = (v, u)_V, \\ &\text{Gelfand:} && \langle v, u\rangle_{V^*,V} = (v, u)_H. \end{aligned}$$

Example 3.43. An example of a Gelfand triple which is used within this thesis is $H_0^1(\Omega)$, $L^2(\Omega)$, $H^{-1}(\Omega)$. But also the more general Sobolev spaces $W_0^{k,p}(\Omega)$ lead to Gelfand triples with the pivot space $L^2(\Omega)$, see [**Zei90a**, Ex. 23.12].

Remark 3.44 (Poincaré-Friedrich inequality). As mentioned before, the embedding $V \hookrightarrow H$ implies an inequality of the form $\|v\|_H \leq c\|v\|_V$. This inequality is called the *Poincaré-Friedrich inequality*, cf. Section 3.1.5 which includes the special case for the Gelfand triple of Example 3.43.

3.3.2. *Definition and Embeddings.* This subsection is devoted to a special class of Bochner spaces which occur in the analysis of abstract differential equations. For this, we have to combine the concept of Bochner spaces with Gelfand triples from the previous subsection. The results are taken from [**Rou05**, Ch. 7] and [**Emm04**, Ch. 8.1].

Similar to Definition 3.8, generalized derivatives can be defined for abstract functions by shifting the derivatives to the test function. This means that $\dot{u} \in L^1_{\mathrm{loc}}(0,T;X)$ is called the *generalized derivative* of $u \in L^1_{\mathrm{loc}}(0,T;X)$ if for all $\Phi \in C_0^\infty(0,T)$ it holds that

$$\int_0^T u(t)\dot{\Phi}(t)\ \mathrm{d}t = -\int_0^T \dot{u}(t)\Phi(t)\ \mathrm{d}t.$$

Consider two Sobolev spaces V_1 and V_2 with $V_1 \hookrightarrow V_2$. We define the *Sobolev-Bochner space*

$$W^{1;p,q}(0,T;V_1,V_2) := \big\{ v \in L^p(0,T;V_1) \;\big|\; \dot{v} \in L^q(0,T;V_2) \big\}.$$

Note that the occurring derivative should be understood in the generalized sense. Together with the norm

$$\|v\|_{W^{1;p,q}(0,T;V_1,V_2)} := \|v\|_{L^p(0,T;V_1)} + \|\dot{v}\|_{L^q(0,T;V_2)}$$

the space $W^{1;p,q}(0,T;V_1,V_2)$ is again a Banach space. For abstract differential equations of second order in time, we define in a similar manner

$$W^{2;p,q,r}(0,T;V_1,V_2,V_3) := \big\{ v \in L^p(0,T;V_1) \;\big|\; \dot{v} \in L^q(0,T;V_2),\ \ddot{v} \in L^r(0,T;V_3) \big\}.$$

Because of the assumed embedding $V_1 \hookrightarrow V_2$, we obtain the following result.

Lemma 3.45 (Embedding for general Sobolev-Bochner spaces). *Consider exponents $p,q \geq 1$ and continuously embedded Banach spaces $V_1 \hookrightarrow V_2$. Then, there exists a continuous embedding $W^{1;p,q}(0,T;V_1,V_2) \hookrightarrow C([0,T];V_2)$. Furthermore, $C^1([0,T];V_1)$ is dense in $W^{1;p,q}(0,T;V_1,V_2)$.*

Proof. The proof can be found in [**Rou05**, Lem. 7.1 and Lem. 7.2]. □

The application of Lemma 3.45 with $V = V_1 = V_2$ and $p = q$ yields the embedding

$$W^{1;p}(0,T;V) := W^{1;p,p}(0,T;V,V) \hookrightarrow C([0,T];V). \tag{3.5}$$

In particular, we have $H^1(0,T;V) := W^{1;2}(0,T;V) \hookrightarrow C([0,T];V)$. Yet another special case, which is important in the theory of abstract ODEs, is given by $V_2 = V_1^*$. If the embedding $V_1 \hookrightarrow V_1^*$ is given by a Gelfand triple with pivot space H, then we obtain a similar embedding result as in Lemma 3.45 but in a stronger topology.

Lemma 3.46 (Embedding with Gelfand triple [**Rou05**, Lem. 7.3]). *Consider a Gelfand triple V, H, V^* and conjugate exponents $p \geq p'$, i.e., $1/p + 1/p' = 1$. Then, the embedding $W^{1;p,p'}(0,T;V,V^*) \hookrightarrow C([0,T];H)$ is continuous. Furthermore, the integration by parts formula holds for all $u,v \in W^{1;p,p'}(0,T;V,V^*)$ and $0 \leq t_1 \leq t_2 \leq T$, i.e.,*

$$\big(u(t_2),v(t_2)\big)_H - \big(u(t_1),v(t_1)\big)_H = \int_{t_1}^{t_2} \big\langle \dot{u}(t),v(t)\big\rangle_{V^*,V} + \big\langle \dot{v}(t),u(t)\big\rangle_{V^*,V}\ dt.$$

The next result is concerned with derivatives of functions which lie in a certain subspace. Recall that a closed subspace W of a Banach space V does not necessarily have to possess a complement, i.e., a closed subspace Z with $V = W \oplus Z$ [**Mos06**]. Subspaces for which such a complement exists are called *complemented*.

Lemma 3.47. *Consider a complemented subspace W of a Banach space V and a Bochner integrable function $v \in L^p(0,T;W)$. Then, the existence of the time derivative $\dot{v} \in L^p(0,T;V)$ implies that $\dot{v} \in L^p(0,T;W)$.*

PROOF. Since W is complemented, there exists a projection $P\colon V \to W$, cf. [**Zei90a**, Ch. 21.12]. By assumption, it holds that $(\mathrm{id} - P)v(t) = 0$ for a.e. $t \in [0,T]$ with id denoting the identity. Since the time derivative of v exists - at least in a generalized sense - we may write $(\mathrm{id} - P)\dot{v}(t) = 0$, which finally implies for a.e. $t \in [0,T]$,

$$\dot{v}(t) = P\dot{v}(t) \in W. \qquad \square$$

Finally, we close this section with one existence result of a complemented subspace. This particular situation will be faced in Section 6.

LEMMA 3.48. *Let $A\colon V \to W$ denote a linear and continuous operator with real Banach spaces V and W. Assume there exists a closed subspace of V, namely V_2, such that $A_2 := A|_{V_2}\colon V_2 \to W$ is bijective. Then, the kernel of A satisfies $V = \ker A \oplus V_2$.*

PROOF. We show that $P := A_2^{-1}A\colon V \to V_2$ is a projection on V_2. For $v \in V$ we have $Pv \in V_2$. In addition, for $v \in V_2$ we know that $Av = A_2 v$ since A_2 is defined as the restriction of A to the subspace V_2. With this, we obtain

$$Pv = A_2^{-1}Av = A_2^{-1}A_2 v = v.$$

Thus, P defines a projection which also implies that $(\mathrm{id} - P)$ is a projection and

$$V = (\mathrm{id} - P)V \oplus V_2.$$

It remains to show that $\ker A = (\mathrm{id} - P)V$. The application of A to $v - Pv$ yields

$$A(v - Pv) = Av - A_2 A_2^{-1} Av = Av - Av = 0.$$

On the other hand, if $v \in \ker A$ and thus, $Pv = 0$, then its unique decomposition is given by

$$v = (v - Pv) + Pv = v + 0,$$

i.e., $\ker A \subset (\mathrm{id} - P)V$. Note that since $\ker A$ and V_2 are closed subspaces, the projection P is even continuous. $\square$

4. Abstract Differential Equations

With the functional analytic background of the previous section, we are able to formulate the generalization of classical differential equations in an abstract framework. Thus, we consider differential equations for abstract functions of the form

$$\dot{u} + \mathcal{K}u = \mathcal{F}, \qquad u(0) = g.$$

Instead of ODEs, where we search for a solution $u \in C^1([0,T], \mathbb{R}^n)$, we search here for a solution $u\colon [0,T] \to V$ with a separable and reflexive Banach space V. The restriction to separable and reflexive Banach spaces is reasonable in view of the considered applications within this thesis. More precisely, we search for solutions $u \in W^{1,p,q}(0,T;V,V^*)$ which corresponds to weak solutions in the context of PDEs. However, several notions and concepts of solutions exist as we will shortly discuss in the beginning of Section 4.2. Afterwards, we discuss precisely the meaning of initial conditions for such problems.

In Section 4.3 we then introduce abstract or operator DAEs, the corresponding generalization of DAEs to the abstract framework. In preparation for this, we introduce first the notion of Nemytskii mappings which deals with the extension of operators to Bochner spaces.

4.1. Nemytskii Mapping. To obtain well-defined operator differential equations, we need to extend possibly nonlinear operators $\mathcal{K}(t)\colon V \to V^*$ to operators defined for abstract functions of the Bochner space $L^p(0,T;V)$. More precisely, we are interested for which parameters $1 \le q,\ p < \infty$ such an operator $\mathcal{K}$ induces a bounded operator of the form

$$\mathcal{K}\colon L^p(0,T;V) \to L^q(0,T;V^*)$$

by $(\mathcal{K}u)(t) := \mathcal{K}(t,u(t))$. This extension is called a *Nemytskii map*, cf. [**Rou05**, Ch. 1.3]. The question of boundedness is answered in the following theorem.

Theorem 4.1 (Nemytskii map [**Rou05**, Th. 1.43]). *Consider an operator $\mathcal{K}\colon [0,T] \times V \to V^*$ which satisfies the properties*

(a) $\mathcal{K}(t,\cdot)\colon V \to V^*$ is continuous for a.e. $t \in [0,T]$,

(b) $\mathcal{K}(\cdot,v)\colon [0,T] \to V^*$ is measurable for all v, and

(c) $\|\mathcal{K}(t,v)\|_{V^*} \le \kappa(t) + c\|v\|_V^{p/q}$ for some $\kappa \in L^q(0,T)$.

Then, the mapping defined via $\mathcal{K}(v)(t) := \mathcal{K}(t,v(t))$ is continuous as a map from $L^p(0,T;V)$ to $L^q(0,T;V^*)$, where $1 \le p < \infty$ and $1 \le q \le \infty$.

In the remainder of this thesis, we do not distinguish between these two notions of an operator $\mathcal{K}$ and its corresponding Nemytskii map.

We give several examples which are of interest for miscellaneous applications. Of special interest is the case when the exponents $1 < p,q < \infty$ are conjugated, i.e, $1/p + 1/q = 1$. This is a basic assumption in the analysis of nonlinear evolution equations using monotonicity arguments [**Rou05**, Ch. 2 and Ch. 8]. The first example indicates that for nonlinear operators even the uniformly boundedness of $\mathcal{K}(t)\colon V \to V^*$ is not sufficient to obtain the conjugacy of the time exponents [**Emm04**, Ch. 8.2].

Example 4.2 (Navier-Stokes operator). Consider the nonlinear operator which arises in the weak formulation of the Navier-Stokes equations,

$$\mathcal{K}\colon V \to V^*, \quad \langle \mathcal{K}u, w\rangle_{V^*,V} := \int_\Omega \left(u \cdot \nabla\right) u \cdot w \,\mathrm{d}x.$$

Then, $\mathcal{K}\colon V \to V^*$ is bounded independently of t, cf. [**Tem77**, Lem. II.1.1], but, in the three-dimensional case, it is only bounded as an operator $\mathcal{K}\colon L^2(0,T;V) \cap L^\infty(0,T;\mathcal{H}) \to L^{4/3}(0,T;V^*)$, see e.g. [**Rou05**, Ch. 8.8.4].

Second, we give a positive example which leads to a bounded Nemytskii mapping with conjugate exponents.

EXAMPLE 4.3 (p-Laplacian). For the p-Laplacian, i.e.,

$$\mathcal{K}\colon V \to V^*, \quad \langle \mathcal{K}u, v\rangle_{V^*,V} := \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \,\mathrm{d}x,$$

we take the Sobolev space $V = W_0^{1,p}(\Omega)$. This then induces an operator $\mathcal{K}\colon L^p\big(0,T;V) \to L^{p'}(0,T;V^*)$ with $1/p + 1/p' = 1$, see [**Rož04**, Ch. 3.3.6].

Finally, we give a corollary of Theorem 4.1 which applies for instance to linear operators that are uniformly bounded with respect to time.

COROLLARY 4.4. *Consider any $1 \le p < \infty$ and an operator $\mathcal{K}\colon [0,T] \times V \to V^*$ which is measurable for fixed $v \in V$ and uniformly bounded in the sense that there exists a constant $C_\mathcal{K}$ such that $\|\mathcal{K}(t)v\|_{V^*} \le C_\mathcal{K}\|v\|_V$ for all $v \in V$ and a.e. $t \in [0,T]$. Then, $(\mathcal{K}v)(t) := \mathcal{K}(v(t))$ defines a continuous operator from $L^p(0,T;V)$ to $L^p(0,T;V^*)$.*

PROOF. The application of Theorem 4.1 with $p = q$ and $\gamma = 0$ yields the result. □

EXAMPLE 4.5 (Linear elasticity). In the case of linear isotropic material laws, i.e.,

$$\mathcal{K}\colon V \to V^*, \quad \langle \mathcal{K}u, v\rangle_{V^*,V} := \int_\Omega \big(2\mu\varepsilon(u) + \lambda \operatorname{trace}\varepsilon(u) I_{2\times 2}\big)\colon \varepsilon(v)\,\mathrm{d}x$$

with $\varepsilon(u)$ denoting the symmetric gradient, μ, λ the Lamé constants [**BS08**, Ch. 11], and $A : B := \sum_{i,j} A_{ij}B_{ij}$ the inner product for matrices considered as vectors, we use as ansatz space $V = H^1(\Omega)$. This setting then induces the bounded operator $\mathcal{K}\colon L^2(0,T;V) \to L^2(0,T;V^*)$.

4.2. Operator ODEs. The generalization of an ODE, which allows solutions in function spaces, is called abstract ODE, abstract Cauchy problem, or evolution equation. However, not all of these notions are equivalent since they consider the differential equation in different function spaces with different regularity assumptions. Consistent with classical ODEs, the *abstract Cauchy problem* considers the equation

$$\dot{u} + \mathcal{K}u = \mathcal{F}$$

in a Banach space V. This means that the operator $\mathcal{K}$ maps from its domain $\mathcal{D}(\mathcal{K}) \subset V$ to V and that the right-hand side satisfies $\mathcal{F}\colon [0,T] \to V$. Then, a *classical solution* satisfies $u \in C^1([0,T];V)$ and the corresponding initial condition reads $u(0) = g \in V$. For this approach to the problem, there exists a generalization of the theorem of Picard-Lindelöf [**Emm04**, Th. 7.2.3] for the local existence of solutions. This approach is closely related to semigroups and often deals with unbounded operators $\mathcal{K}$, see [**Paz83**, Ch. 4]. Also in this framework weaker notions of solutions are used such as *mild solutions*.

Following the concept of weak formulations in the theory of PDEs, it seems more natural to consider operators of the form $\mathcal{K}\colon V \to V^*$ and right-hand sides $\mathcal{F}\colon [0,T] \to V^*$. This leads to the theory of *weak solutions* which we use within this thesis. For an introduction we refer to the book chapters [**Wlo87**, Ch. 26], [**Emm04**, Ch. 8], or [**Rou05**, Ch. 8.1]. The interrelation between the classical and weak solution concept is discussed in [**Zei90a**, Ch. 23]. Because of the weakened regularity assumptions, one important issue is

the well-posedness of the initial condition and the question in which space this condition has to be posed.

REMARK 4.6. One has to be careful with the different terminology in PDE and operator theory. A strong solution of an operator ODE corresponds to a weak solution of the corresponding time-dependent PDE. Going further, weak solutions of operator equations correspond to *very weak solutions* of the equivalent PDEs [**Rou05**, Ch. 8.1].

4.2.1. *First-order Equations.* This subsection is devoted to the formulation of semi-linear parabolic PDEs as operator equations. The operator equation corresponds to the weak formulation of the PDE in time and space and has the form

$$\dot{u} + \mathcal{K}u = \mathcal{F}, \qquad u(0) = g. \tag{4.1}$$

Note that this formulation equals an ODE in an abstract setting, since we assume the equation to hold in a Banach space. Thus, equation (4.1) is called an *abstract* or *operator ODE*. In addition, a discretization of the Banach space by finite elements would lead to an ODE in the common sense. In order to make this formulation reasonable, one has to specify the search space for the solution u and in which space the system should be understood. Considering also the weak form in time, the meaning of the initial condition has to be clarified as well.

The solution should satisfy $u(t) \in V$ for a.e. $t \in [0,T]$ for some separable and reflexive Banach space V. Thus, we consider u to be an element of the Bochner space $L^p(0,T;V)$ with $1 \leq p$. Further, we assume the operator $\mathcal{K}$ to satisfy $\mathcal{K}\colon L^p(0,T;V) \to L^q(0,T;V^*)$, cf. the previous subsection on Nemytskii maps. For the right-hand side we assume $\mathcal{F} \in L^q(0,T;V^*)$ such that it is sufficient for the (weak) time-derivative of u to take values in the dual space V^*. Thus, in the given model it is natural to search for a solution in the space

$$u \in W^{1;p,q}(0,T;V,V^*).$$

It remains to find a reasonable interpretation of the initial condition. For this, we assume a Gelfand triple V, H, V^*. Lemma 3.46 then implies for $q \geq p/(p-1)$ that u is embedded in the space $C([0,T],H)$ such that the initial condition is well-posed for $g \in H$.

REMARK 4.7 (Regularity of initial data). If the prescribed initial data satisfies $g \in V$, then we obtain $\|u(0)-g\|_H = 0$. Because of the embedding $V \hookrightarrow H$ as part of the Gelfand triple V, H, V^*, this implies $\|u(0)-g\|_V = 0$. Thus, the triangle inequality yields

$$\|u(0)\|_V \leq \|u(0)-g\|_V + \|g\|_V = \|g\|_V < \infty.$$

As a result, the additional regularity of the initial data translates to $u(0) \in V$.

As mentioned above, the operator ODE corresponds to a PDE in weak form. For this, we need to consider the PDE multiplied by test functions in V. The corresponding composition in the operator form is to state equation (4.1) in the dual space V^*. In summary, the abstract ODE has the form:

For given data $\mathcal{F} \in L^q(0,T;V^*)$ and $g \in H$ find $u \in W^{1;p,q}(0,T;V,V^*)$ such that for a.e. $t \in [0,T]$ it holds that

$$\dot{u}(t) + \mathcal{K}u(t) = \mathcal{F}(t) \quad \text{in } V^* \tag{4.2}$$

with initial condition $u(0) = g \in H$.

Note that equation (4.2) should be understood pointwise in L^1, i.e., equation (4.2) means that for all $v \in V$ and $\phi \in C_0^\infty(0,T)$ it holds that

$$\int_0^T \left\langle \dot{u}(t), v \right\rangle_{V^*,V} \phi(t) + \langle \mathcal{K}u(t), v \rangle_{V^*,V} \phi(t) \, \mathrm{d}t = \int_0^T \left\langle \mathcal{F}(t), v \right\rangle_{V^*,V} \phi(t) \, \mathrm{d}t.$$

As an example, we formulate the heat equation in this abstract notion.

EXAMPLE 4.8 (Heat equation). The heat equation on a domain Ω is given by

$$\dot{u} - \Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \quad u(0) = g.$$

For the weak formulation we consider the spaces $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ and define the operator $\mathcal{K}\colon V \to V^*$, which arises from the integration by parts formula, by $\langle \mathcal{K}u, v\rangle := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x$. If we understand this system also weakly in time, then we obtain the operator ODE (4.2) with the initial condition stated in the space H.

Similarly, the concepts of this subsection can be applied to second-order PDEs as they appear e.g. in the dynamics of elastic media. Nevertheless, the spaces have to be adapted in this case.

4.2.2. *Second-order Equations.* We close the discussion on operator ODEs with the formulation of hyperbolic PDEs in operator form. This includes applications such as the wave equation as well as elastodynamics which are considered in detail in Section 7.1. As in the previous subsection, the operator formulation is based on a Gelfand triple V, H, V^*, see also [**Wlo87**, Ch. 29] and [**LM72**, Ch. 3.8].

In this subsection we consider $\mathcal{K}$ to be an operator of the form $\mathcal{K}\colon L^2(0,T;V) \to L^2(0,T;V^*)$. With a right-hand side $\mathcal{F} \in L^2(0,T;V^*)$ and initial data g and h, we may consider the operator ODE of second order. Without damping term, this system has the form

$$\ddot{u}(t) + \mathcal{K}u(t) = \mathcal{F}(t) \quad \text{in } V^* \tag{4.3a}$$

with initial conditions

$$u(0) = g, \qquad \dot{u}(0) = h. \tag{4.3b}$$

A suitable ansatz space for u is given by $W^{2;2,2,2}(0,T;V,H,V^*)$, i.e., we search for $u \in L^2(0,T;V)$ with derivatives in H and second derivatives in V^*. As for first-order systems in Section 4.2.1, we have to discuss reasonable spaces for the initial conditions. With the embedding result from Lemma 3.46 we can only assure

$$u \in C([0,T],H), \quad \dot{u} \in C([0,T],V^*).$$

This would call for initial conditions in H and V^*, respectively. However, with more regular data $\mathcal{F} \in L^2(0,T;H)$ we even obtain continuity of u in V and of $\dot{u}$ in H, see [**LM72**, Ch. 3, Th. 8.1]. In this case, it is reasonable to state initial conditions $g \in V$ and $h \in H$.

Several existence and uniqueness results are known for operator ODEs of second order for different assumptions on the included operators and right-hand sides. These also include an additional viscous damping term of the form $\mathcal{D}\dot{u}(t)$. Note that a damping term may need an adjustment of the ansatz space as shown in Section 7. Existence results can be found, e.g., in the monographs [**LS65**], [**GGZ74**, Ch. 7], [**Zei90b**, Ch. 33], or [**Rou05**, Part II]. More recent results can be found in [**ET10a**] and the references therein. For linear equations of second order, the analysis can be found in [**Fat85**, Ch. 2]. An example of a linear system is given by the wave equation.

EXAMPLE 4.9 (Wave equation). The wave equation with homogeneous boundary conditions in a domain Ω is given by

$$\ddot{u} - \Delta u = f \quad \text{in } \Omega, \quad u(0) = g, \quad \dot{u}(0) = h.$$

As for the heat equation in Example 4.8, we consider the Gelfand triple with $V = H_0^1(\Omega)$ and pivot space $H = L^2(\Omega)$. The weak form of the wave equation is then given by equation (4.3) if we define $\mathcal{K}$ as for the heat equation by $\langle \mathcal{K}u, v\rangle := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x$.

Equation 4.3 is just one prototype of a hyperbolic PDE which shows that the concept of Gelfand triples is also valuable for second-order operator equations. The inclusion of a damping term $\mathcal{D}\dot{u}$ requires small adjustments as shown in the case of elastodynamics in Section 7.1. This is the case if the damping operator is of the form $\mathcal{D}\colon V \to V^*$ such that $\dot{u} \in L^2(0,T;H)$ is not sufficient. Then, we search for a solution within the Sobolev-Bochner space $W^{2;2,2,2}(0,T;V,V,V^*)$. In this case, initial conditions with data $g \in V$ and $h \in H$ are required also for $\mathcal{F} \in L^2(0,T;V^*)$.

4.3. Operator DAEs. As ODEs lead to DAEs when adding an algebraic constraint, we can obtain abstract DAEs in a similar way. In the abstract setting, the role of the algebraic constraint may itself be a differential equation but without time derivatives. Alternatively, abstract DAEs may be characterized by the fact that a semi-discretization in space leads to a DAE. Note that the dimension of the resulting DAE depends on the level of discretization and may be very large.

In Section 6 we will consider systems of semi-explicit structure which generalize the DAE (2.2) to the abstract setting. For this, consider a system of the form

$$\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } V^*, \tag{4.4a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } Q^* \tag{4.4b}$$

which should hold a.e. in $[0,T]$ with initial condition

$$u(0) = g \in H. \tag{4.4c}$$

Here, V and Q denote reflexive and separable Banach spaces. For the right-hand sides we assume $\mathcal{F} \in L^q(0,T;V^*)$ and $\mathcal{G} \in W^{1;p}(0,T;Q^*)$. Because of the constraint (4.4b), which does not involve any time derivative, this system generalizes the notion of a semi-explicit DAE. Note that the parts of the solution $u(t)$ and $\lambda(t)$ are still part of an infinite-dimensional Banach space instead of being a vector in $\mathbb{R}^n$. As already mentioned, a spatial discretization of this system would lead to a semi-explicit DAE in the usual sense. This motivates to call system (4.4) an *abstract DAE* or, as we often refer to, an *operator DAE*.

The operators $\mathcal{K}$ and $\mathcal{B}$ should be Nemytskii mappings of the form

$$\mathcal{K}\colon L^p(0,T;V) \to L^q(0,T;V^*), \qquad \mathcal{B}\colon L^p(0,T;V) \to L^p(0,T;Q^*).$$

As for abstract ODEs above, we have to discuss the meaning of the initial condition if we assume a solution to satisfy $u \in W^{1;p,q}(0,T;V,V^*)$ with conjugate exponents p and q, i.e., $p \geq 2$ and $1 = 1/p + 1/q$. As mentioned in Section 4.2.1, the initial condition (4.4c) is meaningful if we assume an underlying Gelfand triple V, H, V^*. From the theory of DAEs it is known that initial conditions have to satisfy a consistency condition. Also in the abstract setting one directly obtains by equation (4.4b) that

$$\mathcal{B}u(0) = \mathcal{G}(0).$$

Note that $\mathcal{G}(0) \in Q^*$ is well-defined, since $\mathcal{G} \in W^{1;p}(0,T;Q^*) \hookrightarrow C([0,T];Q^*)$ by Lemma 3.45. We emphasize that the operator $\mathcal{B}$ on the left-hand side is not applicable for $u(0) \in H$. However, $\mathcal{B}u$ may be evaluated at $t = 0$. The equality shows that the initial data g cannot be chosen arbitrarily as for (abstract) ODEs. An exact characterization of admissible initial data is given in Section 6, using the therein performed regularization. Consistent initial conditions for operator DAEs are also discussed in [**EM13**].

REMARK 4.10. If the solution is sufficiently smooth, then it also has to satisfy the hidden constraint $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{B}u|_{t=0} = \dot{\mathcal{G}}(0)$.

As discussed already for abstract ODEs, the operator formulation corresponds to weak solutions of PDEs in time and space with time derivatives understood in a generalized sense. This remains valid for the here considered constrained PDEs. Hence, all operator equations of this subsection should be understood pointwise in L^1 as before.

For the convergence of the Lagrange multipliers in Part D of this thesis, we will rely on a weaker notion of solutions. For this, we note that λ is the (generalized) derivative of its primitive $\tilde{\lambda}$. As in [**EM13**] we then ask the pair $(u, \tilde{\lambda})$ to solve instead of (4.4a) the equation

$$\int_0^T -\big\langle u(t), v\big\rangle_{V^*,V}\dot{\phi}(t) + \big\langle \mathcal{K}u(t), v\big\rangle_{V^*,V}\phi(t) - \big\langle \mathcal{B}^*\tilde{\lambda}(t), v\big\rangle_{V^*,V}\dot{\phi}(t)\ \mathrm{d}t = \int_0^T \big\langle \mathcal{F}(t), v\big\rangle_{V^*,V}\phi(t)\ \mathrm{d}t$$

for all $v \in V$ and $\phi \in C_0^\infty(0,T)$. In this case, we say that $(u, \tilde{\lambda})$ solves system (4.4) in the *weak distributional sense*. Note that this formulation does not require $\dot{u} \in L^q(0,T;V^*)$ anymore.

In order to analyse applications of elastodynamics, we also consider second-order operator DAEs of semi-explicit structure. The here assumed structure generalizes the DAE of the form (2.5) to the infinite-dimensional case. Thus, we consider operator DAEs of the form

$$\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } V^*, \tag{4.5a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } Q^* \tag{4.5b}$$

which should hold for a.e. $t \in [0,T]$. Furthermore, we assume given initial conditions of the form $u(0) = g \in V$ and $\dot{u}(0) = h \in H$. For this particular problem class, we search for solutions in $W^{2;2,2,2}(0,T;V,V,V^*)$. Thus, the embeddings of Section 3.3.2 ensure that the initial conditions are meaningful, see also the discussion for operator ODEs above.

As for the equations of first order, there exist consistency conditions for the initial data g and h because of the constraint (4.5b). Obviously, g has to satisfy $\mathcal{B}g = \mathcal{G}(0)$. This means that only a part of g can be chosen arbitrarily. A detailed characterization of the admissible initial data is subject of Section 7. With more regular initial data of the form $h \in V$, we additionally obtain the constraint $\mathcal{B}h = \dot{\mathcal{G}}(0)$.

5. Discretization Schemes

For the simulation of time-dependent PDEs or its equivalent formulation as operator ODEs or DAEs, we need discretizations in time and space. Because of the special role of time we do not consider space-time schemes like the space-time finite element method [**HH90**]. Instead, we consider the approach of discretizing in space and time separately. This leads to the two possibilities of discretizing first in space (method of lines) or first in time (Rothe method).

For the spatial discretization we consider the finite element method for which we introduce common ansatz spaces. Special emphasis is placed on the discretization and stability of saddle point problems, so-called *mixed methods*. Such systems are of importance for the consideration of operator DAEs where the constraints are enforced by the Lagrangian method. For simplicity, we restrict the subsection on finite elements to two-dimensional domains $\Omega \subset \mathbb{R}^2$. Note, however, that most of the presented schemes and results are also valid for three space dimensions, if we consider triangulations out of tetrahedra.

For the time integration we restrict ourselves to the implicit Euler method. Thus, for second-order systems we consider the scheme which results from the Euler discretization to the corresponding first-order system. We shortly discuss the convergence properties of these methods when applied to DAEs. For this, we assume a semi-explicit structure of the system which is given for all the applications within this thesis.

Up to now, we have worked with Banach spaces V and H. In order to distinguish the infinite dimensional spaces from their approximation spaces, we use curly letters such as $\mathcal{V}$, $\mathcal{H}$, or $\mathcal{Q}$ for the general Banach spaces and V_h, H_h, or Q_h for its finite-dimensional counterpart.

5.1. Spatial Discretization. For the discretization in the space variable, we consider finite elements. For this, an infinite-dimensional ansatz or search space $\mathcal{V}$ is approximated by a finite-dimensional space V_h. We distinguish *conforming* ($V_h \subset \mathcal{V}$) and *nonconforming* methods ($V_h \not\subset \mathcal{V}$). Both approaches are of interest and will be considered in this section but also in the applications discussed in Section 6.3.

In general, the finite element spaces and resulting ansatz functions are based on a triangulation $\mathcal{T}$ of the domain Ω. We always assume that the given domain $\Omega \subset \mathbb{R}^2$ is a polygonal Lipschitz domain such that there exists a triangulation with $\bigcup_{T\in\mathcal{T}} T = \bar{\Omega}$. In the three-dimensional case we would accordingly assume a polyhedral domain. Furthermore, we only consider *regular triangulations* in the sense of Ciarlet [**Cia78**], i.e., we exclude hanging nodes. To avoid degenerate meshes, we also assume the mesh to be *shape regular* which means that the proportion of the diameter of each triangle and the radius of its interior sphere is bounded [**Bra07**, Ch. II.5].

In the sequel, we use the following notation: $\mathcal{N}$ denotes the set of nodes (or vertices) of the triangulation $\mathcal{T}$ and $\mathcal{E}$ the set of edges. The next subsection introduces common finite element spaces which will be used within this thesis. This includes the piecewise linear hat-functions as well as edge-bubble functions. After this, we shortly summarize how these spaces are used to approximate solutions of PDEs and analyse the stability of such methods for saddle point problems.

Further results on the history and convergence results of finite element methods can be found in the monographs [**Bra07, BS08**].

5.1.1. *Finite Element Spaces.* All considered finite element schemes are based on piecewise polynomials. Here, piecewise should be understood with respect to the given triangulation $\mathcal{T}$, i.e., the approximation space V_h consists of functions whose restriction to a single triangle is a polynomial. The space of piecewise polynomials of degree k is denoted

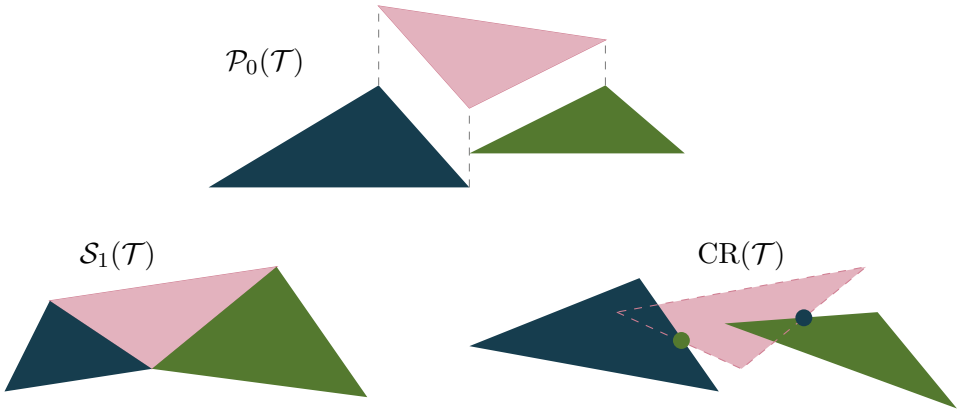


FIGURE 5.1. Illustration of the finite element spaces $\mathcal{P}_0(\mathcal{T})$ (top), $\mathcal{S}_1(\mathcal{T})$ (left), and $\mathrm{CR}(\mathcal{T})$ (right) in two space dimensions.

by $\mathcal{P}_k(\mathcal{T})$. If the functions are in addition globally continuous, we write

$$\mathcal{S}_k(\mathcal{T}) := \mathcal{P}_k(\mathcal{T}) \cap C(\bar{\Omega}) \subset H^1(\Omega).$$

In order to approximate the space $H_0^1(\Omega)$ we introduce the discrete space with zero boundary conditions, namely $\mathcal{S}_{k,0}(\mathcal{T})$.

In the linear case $k = 1$, the canonical basis functions are given by the usual nodal basis functions, also called *hat-functions*. An illustration of the space $\mathcal{S}_1(\mathcal{T})$ is given in Figure 5.1. These functions are node-oriented and have the value one at a certain node and vanish at any other node [**Bra07**, Ch. II]. We denote these functions by φ_i where the index i corresponds to a node of the triangulation. The dimension of $\mathcal{S}_1(\mathcal{T})$ equals the number of nodes of $\mathcal{T}$, whereas the dimension of $\mathcal{S}_{1,0}(\mathcal{T})$ equals the number of interior nodes.

REMARK 5.1. Nodal basis functions can also be defined on a tetrahedron [**Fla00**, Ch. 4.5]. Thus, the space $\mathcal{S}_1(\mathcal{T})$ can be defined in the same manner also for three-dimensional problems.

Another important function class are the so-called *edge-bubble functions* [**Ver96**, Ch. 1]. For an edge $E \in \mathcal{E}$ the edge-bubble function ψ_E is defined as the (scaled) product of the two nodal basis functions associated to the endpoints of E. These functions are quite popular because of their local support which equals the two triangles which share the edge E. Furthermore, their use has proven to be beneficial for stabilization purposes, cf. Section 5.1.3. Note that the space of edge-bubble functions $\mathcal{B}_2(\mathcal{T})$ is a subset of $\mathcal{S}_2(\mathcal{T})$.

A variant of the typical edge-bubble function was introduced by Bernardi and Raugel [**BR85**], see also [**GR86**, Ch. II]. For an interior edge $E \in \mathcal{E}_{\mathrm{int}}$, the corresponding basis function is given by

$$\Upsilon_E := \varphi_1 \varphi_2 \nu_E \in \mathbb{R}^2. \tag{5.1}$$

Therein, φ_1 and φ_2 denote the two hat-functions corresponding to the vertices of the edge E and ν_E equals the outer normal vector on E. For an illustration of the basis function we refer to Figure 5.2. These functions are used to construct a stable discretization scheme for saddle point problems, see Section 5.1.3 below.

REMARK 5.2. In three space dimensions, the outer normal vector is not well-defined along edges. Thus, we have to consider bubble functions w.r.t. faces instead. They are

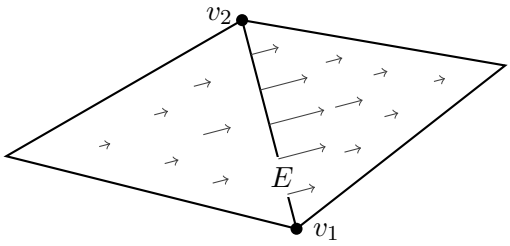


FIGURE 5.2. Illustration of the vector-valued basis function $\Upsilon_E = \varphi_1\varphi_2\nu_E$.

defined as the product of the three corresponding nodal basis functions and the outer normal vector of the face [**BR85**].

Piecewise polynomials can be used in a discontinuous manner as well. Such a non-conforming approximation space was introduced by Crouzeix and Raviart [**CR73**] and is defined by

$$\mathrm{CR}(\mathcal{T}) := \mathcal{P}_1(\mathcal{T}) \cap C(\{\mathrm{mid}(E) \mid E \in \mathcal{E}\}) \not\subset H^1(\Omega).$$

Thus, the space contains all piecewise linear functions which are continuous only at the midpoints of edges, cf. Figure 5.1. The degrees of freedom are hence the values taken in the midpoints of edges in $\mathcal{E}$. Because of this, we call the space $\mathrm{CR}(\mathcal{T})$ edge-oriented. Possible Dirichlet boundary conditions are introduced by setting the values at the midpoints of boundary edges to zero. This space is then denoted by $\mathrm{CR}_0(\mathcal{T})$. The Crouzeix-Raviart basis function corresponding to an edge E is denoted by ϕ_E. It takes the value one in the midpoint of E and vanishes in any other midpoint.

REMARK 5.3. On tetrahedra, the analogon is defined by piecewise affine functions which have a patch condition along the faces, cf. [**CR73**].

5.1.2. *Finite Element Discretization.* The finite-dimensional spaces introduced in the previous subsection can be used to achieve approximations of solutions of PDEs. We start with an elliptic problem in weak form, i.e., we look for $u \in \mathcal{V}$ such that for all test functions $v \in \mathcal{V}$ it holds that

$$a(u, v) = \langle f, v\rangle. \tag{5.2}$$

Therein, a denotes a bounded and coercive bilinear form and f a linear functional in $\mathcal{V}^*$. Solving for example the Laplace problem with homogeneous Dirichlet boundary conditions in a domain Ω, $-\Delta u = f$, we would set $\mathcal{V} = H^1_0(\Omega)$ and $a(u, v) := (\nabla u, \nabla v)_{L^2(\Omega)}$.

Let V_h denote a finite element space, e.g. from Section 5.1.1, which approximates the space $\mathcal{V}$. The finite element approximation is then obtained by replacing (5.2) by the finite-dimensional problem: find $u_h \in V_h$ such that for all $v_h \in V_h$ it holds that

$$a(u_h, v_h) = \langle f, v_h\rangle. \tag{5.3}$$

An equivalent formulation is given in terms of the coefficients of u_h w.r.t. a basis of V_h. Let $x = [x_i]$ denote the coefficient vector of u_h, i.e., for a given basis $\{\varphi_i\}_{i=1,\dots,n}$ of V_h we have $u_h = \sum_{i=1}^n x_i\varphi_i$. Then, (5.3) is equivalent to the linear system

$$Kx = b$$

with the *stiffness matrix* $[K_{ij}] := [a(\varphi_i, \varphi_j)]$ and the right-hand side given by $[b_i] := [\langle f, \varphi_i\rangle]$.

Working with operators instead of bilinear forms, i.e., assuming an equation of the form $\mathcal{K}u = \mathcal{F}$ in $\mathcal{V}^*$ in place of (5.2), we can rewrite this equivalently in the form: find

$u \in \mathcal{V}$ such that for all test functions $v \in \mathcal{V}$ it holds that $\langle \mathcal{K}u, v\rangle = \langle \mathcal{F}, v\rangle$. Then, the stiffness matrix reads accordingly $[K_{ij}] := [\langle \mathcal{K}\varphi_i, \varphi_j\rangle]$. In the case of a nonlinear operator $\mathcal{K}$, the stiffness matrix has to be replaced by a nonlinear function $K\colon \mathbb{R}^n \to \mathbb{R}^n$. The k-th component of this function is then given by

$$K(q)_k := \big\langle \mathcal{K}\big(\textstyle\sum_{i=1}^n q_i\varphi_i\big), \varphi_k\big\rangle.$$

In the case of a nonconforming discretization scheme, one has to assure that the involved operators are defined for the basis functions φ_i. Note that this is not automatically given since $V_h \not\subset \mathcal{V}$. For this, it may be necessary to generalize the application of the operator to a piecewise (w.r.t. the triangulation $\mathcal{T}$) application of it. This is normally sufficient since we work with piecewise polynomials and thus, piecewise smooth functions.

With regard to the operator DAEs analyzed in Part B, we also consider problems with a saddle point structure: find functions $u \in \mathcal{V}$ and $p \in \mathcal{Q}$ such that for all test functions $v \in \mathcal{V}$ and $q \in \mathcal{Q}$ it holds that

$$a(u,v) \;+\; b(v,p) = \langle f, v\rangle, \tag{5.4a}$$

$$b(u,q) \qquad\quad = \langle g, q\rangle. \tag{5.4b}$$

An example is given by the Stokes equation, describing the steady state motion of an incompressible viscous fluid,

$$-\Delta u + \nabla p = f, \qquad \nabla \cdot u = 0.$$

Therein, u denotes the velocity of the fluid and the scalar variable p indicates the pressure. In this case, the weak formulation works with the bilinear forms $a(u,v) := (\nabla u, \nabla v)_{L^2(\Omega)}$ and $b(v,p) := -(\nabla \cdot v, p)_{L^2(\Omega)}$. The corresponding Sobolev spaces are $\mathcal{V} = [H_0^1(\Omega)]^d$ and $\mathcal{Q} = L^2(\Omega)/\mathbb{R}$.

The discretization of (5.4) by finite elements then needs two discrete spaces V_h and Q_h. These methods are often called *mixed methods*, see [**BS08**, Ch. 12] for an introduction. The resulting linear system has the form

$$\begin{bmatrix} K & B^T \\ B & \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Therein, x and y denote the coefficient vectors of the finite element approximations u_h and p_h w.r.t a given basis $\{\varphi_i\}_{i=1,\dots,n}$ of V_h and $\{\psi_i\}_{i=1,\dots,m}$ of Q_h, respectively. The matrix B corresponds to the bilinear form b, i.e., $[B_{ij}] = [b(\varphi_j, \psi_i)]$. Besides the coercivity of the bilinear form a in the nullspace of B, the well-posedness of the discrete problem requires a stability condition of b of the form

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_{\mathcal{V}} \|q_h\|_{\mathcal{Q}}} = \beta_{\text{disc}} > 0. \tag{5.5}$$

Thereby, the bound β_{disc} should be independent of discretization parameters like the mesh size. This crucial property is called the *discrete inf-sup condition*, also called the *Ladyzhenskaya-Babuška-Brezzi* condition. It is especially related to the stability of the pressure variable p_h [**BF91**, Ch. VI.3]. Nowadays, this condition provides the right mathematical tool to analyse and prevent instabilities in the simulation of saddle point problems [**Bra07**, Chap. III.7]. Within this thesis, we are interested in two particular cases for the bilinear form b which appear in the applications of elastodynamics and fluid dynamics. The detailed analysis of these two cases is subject of the following subsection.

5.1.3. *Stability for Saddle Point Problems.* The first part of this subsection is devoted to the case where the bilinear form b corresponds to the trace operator of Definition 3.14. This particular case is of interest in the field of elastodynamics when there are constraints along a boundary part $\Gamma \subset \partial\Omega$, cf. Section 7.1.2 below. For this, consider the Sobolev spaces

$$\mathcal{V} := [H^1(\Omega)]^r, \qquad \mathcal{Q}^* := [H^{1/2}(\Gamma)]^r$$

and the bilinear form $b\colon \mathcal{V} \times \mathcal{Q} \to \mathbb{R}$, which is densely defined by

$$b(v, q) := \int_\Gamma v \cdot q \,\mathrm{d}x. \tag{5.6}$$

Note that the integral of $v \in \mathcal{V}$ over the boundary involves the trace operator γ of Section 3.1.4. The parameter r typically equals 1 (e.g. for the wave equation) or the dimension of the domain d (e.g. for applications in elastodynamics).

We give a particular example of a stable discretization. For the ansatz space V_h we choose the piecewise linear hat-functions in r dimensions. In order to ensure stability, this space has to be enriched by a certain number of edge-bubble functions. Let $\mathcal{B}_\Gamma(\mathcal{T})$ denote the subspace of $\mathcal{B}_2(\mathcal{T})$ which contains all edge-bubble functions corresponding to edges along Γ. Then, we set

$$V_h := [\mathcal{S}_{1,0}(\mathcal{T})]^r \oplus [\mathcal{B}_\Gamma(\mathcal{T})]^r.$$

On the other hand, the space $\mathcal{Q}$ is approximated by piecewise constant functions along Γ, i.e.,

$$Q_h := [\mathcal{P}_0(\mathcal{T})|_\Gamma]^r.$$

Note that functions in Q_h are discontinuous but satisfy $Q_h \subset \mathcal{Q}$.

Lemma 5.4 (Discrete inf-sup condition). *With the finite-dimensional spaces V_h and Q_h, the bilinear form b from* (5.6) *satisfies a* discrete inf-sup condition*, i.e., there exists a positive constant β_{disc}, independent of the mesh size of $\mathcal{T}$, which satisfies* (5.5).

Proof. This stability result is a special case of [**Lip04**, Th. 2.3.7] and is true for $d = 2$ as well as $d = 3$. □

Remark 5.5. The work of [**Lip04**] contains the more general stability result for $Q_h = [\mathcal{S}_k(\mathcal{T})|_\Gamma]^r$. In this case, V_h has to contain all products of edge-bubble functions with piecewise polynomials of degree k.

Remark 5.6 (Time-dependent Dirichlet boundary). For $d = 2$ the proposed discretization scheme V_h, Q_h remains stable if the boundary part Γ changes (smoothly) with time. The needed adjustments and assumptions can be found in [**Alt14**]. A possible application are flexible multibody systems where the boundary conditions are used to model the coupling. A movement of two connected domains then leads to time-dependent Dirichlet boundaries.

Remark 5.7 (Mortar elements). An alternative scheme can be adapted from the numerical analysis of contact problems, the so-called *mortar methods* [**BMP93, Woh99, BBBCM00**]. These elements were originally introduced to enforce weak continuity conditions in domain decomposition methods. The discrete ansatz space for the Lagrange multipliers are piecewise polynomials which are globally continuous on the boundary segment. However, the polynomial degree differs, depending on whether the edge is at the boundary of the segment or not.

The second example appears in the simulation of fluids dynamics, as already mentioned in the previous subsection. For the Stokes or Navier-Stokes equations, which we consider in Section 6.3, the bilinear form b corresponds to the divergence operator. With the spaces

$$\mathcal{V} := [H_0^1(\Omega)]^2, \qquad \mathcal{Q}^* := L^2(\Omega)/\mathbb{R},$$

for the two-dimensional case, we define the bilinear form $b\colon \mathcal{V} \times \mathcal{Q} \to \mathbb{R}$ by

$$b(v, q) := \int_\Omega \nabla \cdot v \; q \, \mathrm{d}x.$$

Because of the wide range of applications, several stable schemes are known for this bilinear form, see [**GR86**, Ch. II] or [**GS00**, Ch. 3]. A negative example in the sense of stability is given by the scheme $V_h = [\mathcal{S}_{1,0}(\mathcal{T})]^2$ with $Q_h = \mathcal{P}_0(\mathcal{T})/\mathbb{R}$, see [**BF91**, Ex. VI.3.1].

In [**BR85**] the ansatz space $[\mathcal{S}_{1,0}(\mathcal{T})]^2$ has been enriched by the variant of edge-bubble functions introduced in (5.1). Thus, the fluxes through interior edges yield additional degrees of freedom and serve as stabilization. This leads to the mixed scheme with spaces

$$V_h = [\mathcal{S}_{1,0}(\mathcal{T})]^2 \oplus \{\Upsilon_E \mid E \in \mathcal{E}_{\mathrm{int}}\}, \qquad Q_h = \mathcal{P}_0(\mathcal{T})/\mathbb{R}. \tag{5.7}$$

The proof of the corresponding inf-sup condition can be found in [**BR85**].

A stable scheme of lowest order is given by the nonconforming Crouzeix-Raviart element combined with a piecewise constant pressure approximation [**CR73, BM11**],

$$V_h = \big[\,\mathrm{CR}_0(\mathcal{T})\big]^2, \qquad Q_h = \mathcal{P}_0(\mathcal{T})/\mathbb{R}. \tag{5.8}$$

Thus, the piecewise linear ansatz is unstable in the conforming case and stable using discontinuities. It turns out that it is sufficient to consider the nonconforming ansatz in a single component. Thus, also the mixture of continuous and discontinuous velocity components, namely $V_h = \mathcal{S}_{1,0}(\mathcal{T}) \times \mathrm{CR}_0(\mathcal{T})$, leads to a stable discretization scheme [**KS95**].

Finally, we mention the popular schemes of Taylor-Hood type [**TH73**]. Therein, the velocities are approximated by polynomials of one degree higher than the pressure. The Taylor-Hood element of lowest order is given by

$$V_h = \big[\mathcal{S}_{2,0}(\mathcal{T})\big]^2, \qquad Q_h = \mathcal{S}_1(\mathcal{T})/\mathbb{R}.$$

One reason of the popularity is the continuity of the pressure ansatz which yields a more natural model. On the other hand, this scheme using at least second order polynomials is quite expensive, especially if one considers three-dimensional simulations of fluid flows.

5.2. Time Integration. This subsection is devoted to the temporal discretization methods used within this thesis. Using ODE methods for the discretization of DAEs calls for special care. Because of the numerical instabilities mentioned in Section 2, even simple linear DAE systems may not be integrated accurately by ODE methods [**LP86**]. However, index-1 systems can be solved without great difficulties whereas codes with automatic step size adaptation often fail for high-index DAEs [**GP84, KM06**]. This is certainly the best argument for index reduction techniques, cf. Section 2.3.

Examples from applications often provide a special structure which is advantageous for the numerical integration. In particular, we consider here semi-explicit systems as equations (2.2) and (2.3). For the temporal discretization, we simply consider the implicit Euler scheme but other schemes could be employed as well [**KM06**, Ch. 5].

5.2.1. *Implicit Euler Scheme.* Consider a semi-explicit DAE of first order

$$\dot{u} = f(u,p), \quad 0 = g(u). \tag{5.9}$$

We assume this system to be of index at most 2, cf. Lemma 2.2. The discretization by the *backward* or *implicit Euler scheme* with step size τ then reads

$$u_n - u_{n-1} = \tau f(u_n, p_n), \quad 0 = g(u_n).$$

It is shown in [**LP86**] that this scheme converges for systems which come e.g. from applications such as fluid dynamics. However, the analysis assumes a certain accuracy in solving the resulting (nonlinear) systems in every time step. This is necessary because of the high sensitivity of DAEs w.r.t. perturbations. More precisely, it is assumed that the differential equation in (5.9) is solved up to order $O(\tau)$ and the algebraic equation even up to order $O(\tau^2)$. A more detailed analysis, which does not rely on these assumptions, is given in [**Arn98b**, Ch. 2]. Therein, the different behavior of the differential variable u and algebraic variable p is stressed which shows that the index of a DAE provides no sharp estimates of the error of the single variables. For systems in Hessenberg form such as (5.9), the differential variable u is more robust to perturbations than its algebraic counterpart p.

In summary, numerical ODE methods may be applied to semi-explicit DAEs (5.9) of index up to 2 if the resulting algebraic systems are solved accurately enough. This counts for errors of iterative solvers as well as errors due to Newton iterations. Results on the implicit Euler scheme applied to operator DAEs of semi-explicit structure are subject of Section 10.

Also schemes of higher order such as BDF or Runge-Kutta schemes can be applied to DAEs [**Arn93, Ost93**]. However, these schemes are not used within this thesis.

5.2.2. *Schemes for Second-order Systems.* In this subsection we consider second-order DAEs of the form

$$\ddot{u} = f(u, \dot{u}, \lambda), \quad 0 = g(u). \tag{5.10}$$

The applications in view of second-order systems are mainly problems from elastodynamics or multibody dynamics. For such systems the Newmark scheme [**New59, GC01**] as well as further developments like the HHT [**HHT77**] or the generalized-α methods [**CH93, AB07**] are widely used. The two latter methods include numerical dissipation, i.e., an artificial loss of energy, in order to damp the high-frequency modes of the system, which often results from a finite element discretization [**YPR98**]. This is advantageous if one is interested only in the low-frequency response of a system. The Newmark scheme is of second order for ODEs but does not converge for the index-3 DAEs (5.10). More precisely, it is the Lagrange multiplier λ which does not converge. For systems from structural dynamics, the deformation u is not affected and still converges although one has to pay attention to the badly conditioned iteration matrix.

The Newmark scheme may also be applied to the index-1 DAEs which result from an index reduction. In this case, we regain the convergence of the method. Although Newmark-type schemes are popular, we do not consider them here since they are not suitable for the convergence of operator equations [**EŠT13**].

For the analysis of the Rothe method applied to second-order operator DAEs in Section 11, we consider the scheme which corresponds to the implicit Euler method applied to the equivalent first-order system. We only consider equidistant time steps with step size τ. Let u^n denote the approximation of u at time $t_n = n\tau$. For the temporal discretization

we then replace the derivatives $\dot{u}$ and $\ddot{u}$ at time t_n by

$$\dot{u} \to \frac{u^n - u^{n-1}}{\tau} =: Du^n, \qquad \ddot{u} \to \frac{u^n - 2u^{n-1} + u^{n-2}}{\tau^2} =: D^2u^n.$$

The convergence of this scheme for index-3 DAEs arising from multibody dynamics is discussed in [**LP86**]. We emphasize that the analysis used therein assumes the constraint to be solved with high accuracy, namely up to the order of $O(\tau^3)$.

Note that we obtain the Störmer-Verlet scheme if u is replaced by u^{n-1} instead of u^n. Also this scheme is used regularly in the analysis of second-order operator equations but is not considered within this thesis.

5.3. Discretization of Time-dependent PDEs. In Section 4 we have seen how to model time-dependent PDEs with the help of operator ODEs or, in the constrained case, operator DAEs. The subject of this section is the discretization of such systems in order to obtain appropriate simulations and approximate solutions. For this, we combine the techniques of Sections 5.1 and 5.2.

Since dynamical systems represent evolution problems, the time variable plays a special role which should result in a special treatment [**Emm04**, Ch. 6]. This also explains the consideration of abstract functions in the modeling part. One considers two main principles [**Rou05**, Ch. 8]:

(1) One possibility is to maintain the time as a continuous variable and use a discretization of Galerkin type (e.g. finite elements of Section 5.1). This ansatz is called the *method of lines* [**Hol07**, Ch. 3.4].
(2) The second ansatz is to discretize in time which leads to a stationary PDE in every time step. For this, a time discretization scheme such as in Section 5.2 is formally applied to the abstract differential equation. This method is often referred to as the *Rothe method* or the *reverse method of lines* [**Rot30**].

We discuss these two approaches in the following two subsections. Because of the mentioned significance of the time variable, we do not discuss methods such as space-time finite elements. An introduction of this approach by means of second-order systems can be found in [**HH90**], see also the more recent paper on parabolic problems [**NS11**].

5.3.1. *Method of Lines.* Applying the finite element method (or any other discretization scheme) to system (4.2), we obtain an ODE where t is the only remaining variable. Accordingly, the finite element method applied to system (4.4) leads to a DAE. Note that the size of the resulting systems may be arbitrary large and depends on the discretization level and the corresponding mesh size. However, using suitable basis functions with local support, we obtain sparse matrices in the ODE or DAE.

This ansatz can also be characterized in the following way: Approximate the solution $u \in L^2(0,T;\mathcal{V})$ by a function $u_h \in L^2(0,T;V_h)$ of the form

$$u_h(x,t) = \sum_{i=1}^{n} q_i(t)\varphi_i(x).$$

Therein, $\{\varphi_i\}_{i=1,\dots,n}$ denotes a basis of the finite-dimensional space V_h and $q_i \in L^2(0,T)$. A justification of this ansatz is subject of the following lemma which shows that functions of this form are dense in $L^2(0,T;\mathcal{V})$ if V_h is an appropriate approximation space of $\mathcal{V}$. By this we mean that the union of V_h over all mesh sizes is dense in $\mathcal{V}$, i.e.,

$$\overline{\bigcup_h V_h}^{\mathcal{V}} = \mathcal{V}. \tag{5.11}$$

For the following result, we restrict ourselves to the conforming case, i.e., $V_h \subset \mathcal{V}$. Nevertheless, analogous approximation properties also hold for nonconforming finite element schemes [**BS08**, Ch. 10.3].

LEMMA 5.8. *Consider a real Banach space $\mathcal{V}$ and a sequence of finite-dimensional approximation spaces V_h which satisfy property* (5.11). *Then, the union $\bigcup_h L^2(0,T;V_h)$ is dense in $L^2(0,T;\mathcal{V})$.*

PROOF. Consider $\varepsilon > 0$, an arbitrary function $u \in L^2(0,T;\mathcal{V})$, and an equidistant partition $0 = t_0 < t_1 < \cdots < t_n = T$ with step size τ. We define the piecewise constant function

$$u_\tau(t) := \frac{1}{\tau} \int_{t_{j-1}}^{t_j} u(t)\ \mathrm{d}t \in \mathcal{V} \quad \text{for } t \in]t_{j-1}, t_j].$$

Following the proof of Theorem 4.2.5 in [**Emm01**], we yield the convergence $u_\tau \to u$ in $L^2(0,T;\mathcal{V})$ as $\tau \to 0$. Hence, there exists a certain step size τ_ε with a corresponding partition of $[0,T]$ such that

$$\|u - u_{\tau_\varepsilon}\|_{L^2(0,T;\mathcal{V})} < \varepsilon/2.$$

With $u^j_{\tau_\varepsilon} := u_{\tau_\varepsilon}|_{]t_{j-1},t_j]} \in \mathcal{V}$ and the characteristic function $\chi^j := \chi_{]t_{j-1},t_j]}$, we can write u_{τ_ε} in the form

$$u_{\tau_\varepsilon}(t,x) = \sum_{j=1}^{n_\varepsilon} \chi^j(t) u^j_{\tau_\varepsilon}(x).$$

Note that the partition of $[0,T]$ and the step size τ_ε is fixed such that we have a finite number of subintervals. Since V_h is assumed to be an approximation space of $\mathcal{V}$ in the sense of (5.11), a sufficient fine triangulation yields for all $j = 1, \ldots, n_\varepsilon$ approximations $u^j_{h,\tau_\varepsilon} \in V_h$ which satisfy the bound

$$\big\|u^j_{\tau_\varepsilon} - u^j_{h,\tau_\varepsilon}\big\|_{\mathcal{V}} < \frac{\varepsilon}{2} T^{-1/2}.$$

Combining all the above estimates, we note that $u_h(t,x) := \sum_{j=1}^{n_\varepsilon} \chi^j(t) u^j_{h,\tau_\varepsilon}(x) \in L^2(0,T;V_h)$ satisfies

$$\begin{aligned}
\|u - u_h\|_{L^2(0,T;\mathcal{V})} &\le \|u - u_{\tau_\varepsilon}\|_{L^2(0,T;\mathcal{V})} + \|u_{\tau_\varepsilon} - u_h\|_{L^2(0,T;\mathcal{V})} \\
&< \frac{\varepsilon}{2} + \Big(\sum_{j=1}^{n_\varepsilon} \int_{t_{j-1}}^{t_j} \|u^j_{\tau_\varepsilon} - u^j_{h,\tau_\varepsilon}\|^2_{\mathcal{V}}\ \mathrm{d}t \Big)^{1/2} \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \Big(\sum_{j=1}^{n_\varepsilon} \int_{t_{j-1}}^{t_j} \frac{1}{T}\ \mathrm{d}t \Big)^{1/2} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \qquad \square
\end{aligned}$$

Lemma 5.8 shows that the ansatz of the method of lines is reasonable. In Part C of this thesis, we will analyse the index of the semi-discrete DAEs resulting from the method of lines applied to constrained operator equations of first and second order.

5.3.2. *Rothe Method.* Within the Rothe method, which goes back to Rothe [**Rot30**], one discretizes in time first. Thus, time integration schemes are formally applied to the abstract differential equation which then leads to a (stationary) PDE which has to be solved in each time step. In general, this method is favored by the finite element community since adaptive finite elements are easily applicable. The method of Rothe may also be used to prove the existence of solutions of operator ODEs, see e.g. [**Rou05**, Th. 8.9]. In some cases, this may require higher order schemes in time, cf. [**Emm01**].

We have seen in Section 4 that the right-hand sides of operator equations are not always assumed to be continuous. Thus, function evaluations of the right-hand side as used for the discretization of ODEs are not well-defined. In this case, $\mathcal{F}(t_n)$ may be replaced by an integral mean over one time step or any other local regularization such as the *Clément quasi-interpolant* [**Clé75**]. In this introductory subsection, we summarize convergence results for two kinds of discretizations of the right-hand side which are used within this thesis.

Consider a Bochner integrable function $\mathcal{F} \in L^2(0,T;\mathcal{V})$ with a real Banach space $\mathcal{V}$ and an equidistant partition $0 = t_0 < t_1 < \cdots < t_n = T$ of $[0,T]$. As in Lemma 5.8 we compute the Bochner integrals over one time step τ,

$$\mathcal{F}^j := \frac{1}{\tau} \int_{t_{j-1}}^{t_j} \mathcal{F}(t) \, \mathrm{d}t \in \mathcal{V}.$$

Recall that this is well-defined for $\mathcal{F} \in L^2(0,T;\mathcal{V})$. Therewith, we define the piecewise constant function $\mathcal{F}_\tau \colon [0,T] \to \mathcal{V}$ by

$$\mathcal{F}_\tau(t) := \mathcal{F}^j \quad \text{for } t \in]t_{j-1}, t_j] \tag{5.12}$$

and a continuous extension in $t = 0$. An easy calculation shows that $\mathcal{F}_\tau \in L^2(0,T;\mathcal{V})$,

$$\|\mathcal{F}_\tau\|^2_{L^2(0,T;\mathcal{V})} = \tau \sum_{j=1}^{n} \|\mathcal{F}^j\|^2_{\mathcal{V}} \leq \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \|\mathcal{F}(t)\|^2_{\mathcal{V}} \, \mathrm{d}t = \|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V})}. \tag{5.13}$$

One important property of $\mathcal{F}_\tau$, which we need within the convergence analysis of Part D, is the strong convergence to $\mathcal{F}$ as $\tau \to 0$.

Lemma 5.9 (Limit of $\mathcal{F}_\tau$). *Consider $\mathcal{F} \in L^2(0,T;\mathcal{V})$ with its approximation $\mathcal{F}_\tau$ as defined in* (5.12) *and let $A \colon \mathcal{V} \to \mathcal{W}$ denote a linear and bounded operator between the real Banach spaces $\mathcal{V}$ and $\mathcal{W}$. Then, $A\mathcal{F}_\tau \to A\mathcal{F}$ in $L^2(0,T;\mathcal{W})$ as $\tau \to 0$.*

Proof. The proof of the strong convergence $\mathcal{F}_\tau \to \mathcal{F}$ in $L^2(0,T;\mathcal{V})$ as $\tau \to 0$ is given in [**Tem77**, Ch. III, Lem. 4.9]. This then implies for $\tau \to 0$,

$$\|A\mathcal{F}_\tau - A\mathcal{F}\|^2_{L^2(0,T;\mathcal{W})} = \int_0^T \|A\big(\mathcal{F}_\tau(t) - \mathcal{F}(t)\big)\|^2_{\mathcal{W}} \, \mathrm{d}t \leq C_A^2 \|\mathcal{F}_\tau - \mathcal{F}\|^2_{L^2(0,T;\mathcal{V})} \to 0. \quad \square$$

For continuous functions $\mathcal{G} \in C([0,T];\mathcal{Q})$ function evaluations are well-defined. In this case, we may define

$$\mathcal{G}_\tau(t) := \mathcal{G}(t_j) \quad \text{for } t \in]t_{j-1}, t_j]. \tag{5.14}$$

Again we consider a continuous extension at $t = 0$. This then leads to the following convergence result.

Lemma 5.10 (Limit of $\mathcal{G}_\tau$). *Consider $\mathcal{G} \in C([0,T];\mathcal{Q})$ with its approximation $\mathcal{G}_\tau$ as defined in* (5.14) *and let $A \colon \mathcal{Q} \to \mathcal{R}$ denote a linear and bounded operator between the real Banach spaces $\mathcal{Q}$ and $\mathcal{R}$. Then, $A\mathcal{G}_\tau \to A\mathcal{G}$ in $L^2(0,T;\mathcal{R})$ as $\tau \to 0$.*

Proof. To prove the strong convergence $\mathcal{G}_\tau \to \mathcal{G}$ in $L^2(0,T;\mathcal{Q})$, we estimate

$$\|\mathcal{G}_\tau - \mathcal{G}\|^2_{L^2(0,T;\mathcal{Q})} = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \|\mathcal{G}_\tau(t) - \mathcal{G}(t)\|^2_{\mathcal{Q}} \, \mathrm{d}t \leq \sum_{j=1}^{n} \tau \max_{t \in [t_{j-1}, t_j]} \|\mathcal{G}(t_j) - \mathcal{G}(t)\|^2_{\mathcal{Q}}.$$

For an arbitrary $\varepsilon > 0$ the (uniform) continuity of $\mathcal{G}$ then implies that for a sufficiently small step size τ, it holds that

$$\|\mathcal{G}_\tau - \mathcal{G}\|^2_{L^2(0,T;\mathcal{Q})} \leq \sum_{j=1}^{n} \tau\varepsilon^2 = T\varepsilon^2.$$

The application of a linear and bounded operator does not influence the convergence. □

In Parts B and D we often deal with Bochner integrable functions of the form $\mathcal{G} \in W^{1,p}(0,T;\mathcal{Q})$. In this case, we propose to discretize $\mathcal{G}$ by means of function evaluations as in (5.14) and $\dot{\mathcal{G}}$ by the integral means as in (5.12). This approach has the nice property that the discrete derivative (in terms of the implicit Euler scheme) of $\mathcal{G}^j$ equals the approximation of the derivative, i.e.,

$$D\mathcal{G}^j := \frac{\mathcal{G}^j - \mathcal{G}^{j-1}}{\tau} = \dot{\mathcal{G}}^j.$$

The application of a nonlinear function to a convergent sequence requires a special treatment. In view of Section 10.5 below, we consider one particular result for trace functions.

LEMMA 5.11. *Consider a Lipschitz domain $\Omega \subset \mathbb{R}^d$ and a strongly converging sequence $\mathcal{G}_\tau$ with $\mathcal{G}_\tau \to \mathcal{G}$ in $L^2(0,T;H^{1/2}(\partial\Omega))$. If $f\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous, then also $f(\mathcal{G}_\tau) \to f(\mathcal{G})$ in $L^2(0,T;H^{1/2}(\partial\Omega))$.*

PROOF. For the proof we consider the norm of $H^{1/2}(\partial\Omega)$ given in [**Wlo87**, Ch. 4.2], i.e.,

$$\|u\|^2_{1/2} := \|u\|^2_{L^2(\partial\Omega)} + I_{1/2}(u) := \|u\|^2_{L^2(\partial\Omega)} + \int_{\partial\Omega}\int_{\partial\Omega} \frac{|u(x)-u(y)|}{|x-y|^d}\,\mathrm{d}x\,\mathrm{d}y.$$

The Lipschitz continuity of f with constant C_{Lip} yields

$$\|f\mathcal{G}_\tau - f\mathcal{G}\|_{L^2(\partial\Omega)} \leq C_{\mathrm{Lip}}\|\mathcal{G}_\tau - \mathcal{G}\|_{L^2(\partial\Omega)}.$$

Thus, it remains to show that $\int_0^T I_{1/2}(f\mathcal{G}_\tau - f\mathcal{G})\,\mathrm{d}t$ tends to zero as well when $\tau \to 0$. For this, consider an arbitrary $\varepsilon > 0$. With $\delta > 0$, which we fix below, we split the double integral into

$$\begin{aligned} I_{1/2}(f\mathcal{G}_\tau - f\mathcal{G}) &= I_{<\delta} + I_{\geq\delta} \\ &:= \int_{\partial\Omega}\int_{\substack{x\in\partial\Omega,\\ |x-y|<\delta}} \frac{|f\mathcal{G}_\tau(x) - f\mathcal{G}(x) - f\mathcal{G}_\tau(y) - f\mathcal{G}(y)|}{|x-y|^d}\,\mathrm{d}x\,\mathrm{d}y \\ &\quad + \int_{\partial\Omega}\int_{\substack{x\in\partial\Omega,\\ |x-y|\geq\delta}} \frac{|f\mathcal{G}_\tau(x) - f\mathcal{G}(x) - f\mathcal{G}_\tau(y) - f\mathcal{G}(y)|}{|x-y|^d}\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

For the first integral we use the triangle inequality in the numerator and the Lipschitz continuity of f. This yields

$$I_{<\delta} \leq C_{\mathrm{Lip}} \int_{\partial\Omega}\int_{\substack{x\in\partial\Omega,\\ |x-y|<\delta}} \frac{|\mathcal{G}_\tau(x) - \mathcal{G}_\tau(y)|}{|x-y|^d} + \frac{|\mathcal{G}(x) - \mathcal{G}(y)|}{|x-y|^d}\,\mathrm{d}x\,\mathrm{d}y$$

Since $\mathcal{G}_\tau$, $\mathcal{G} \in H^{1/2}(\partial\Omega)$, both fractions are integrable over $\partial\Omega \times \partial\Omega$ by the definition of the norm. Furthermore, the integration domain tends to zero as $\delta \to 0$. Thus, by

[**Bog07**, Th. 2.5.7], there exists a δ such that $\int_0^T I_{<\delta}\,\mathrm{d}t \leq \varepsilon/2$. For the second integral, the Lipschitz continuity and the lower bound of $|x-y|$ leads to the estimate

$$\begin{aligned} I_{\geq\delta} &\leq \frac{C_{\mathrm{Lip}}}{\delta^d} \int_{\partial\Omega} \int_{\substack{x\in\partial\Omega,\\ |x-y|\geq\delta}} |\mathcal{G}_\tau(x) - \mathcal{G}(x)| + |\mathcal{G}_\tau(y) - \mathcal{G}(y)|\,\mathrm{d}x\,\mathrm{d}y \\ &\leq 2|\partial\Omega| \frac{C_{\mathrm{Lip}}}{\delta^d} \int_{\partial\Omega} |\mathcal{G}_\tau(x) - \mathcal{G}(x)|\,\mathrm{d}x. \end{aligned}$$

Therein, $|\partial\Omega|$ denotes the $(d-1)$-dimensional measure of the boundary. Finally, the Cauchy-Schwarz inequality implies

$$I_{\geq\delta} \leq C(\delta)\|\mathcal{G}_\tau - \mathcal{G}\|_{L^2(\partial\Omega)}$$

with a constant $C(\delta)$ which depends on δ and the boundary. The strong convergence of $\mathcal{G}_\tau$ then ensures $\int_0^T I_{\geq\delta}\,\mathrm{d}t \leq \varepsilon/2$ for a sufficiently small step size τ. □

Part

B

Regularization of Operator DAEs

Within this part, a general framework for the regularization of constrained time-dependent PDEs is presented. We consider systems which can be written as semi-explicit and semi-linear operator DAEs of first or second order. Semi-explicit means that the equation which constrains the dynamics is explicitly given whereas semi-linear means that the time-derivative of the solution appears only linearly. The assumed structure is general enough to include several major applications from the fields of fluid dynamics as well as elastodynamics. In particular, we discuss the (linearized) Navier-Stokes equations, the two-phase Stefan problem, and flexible multibody systems. More generally, we consider operator DAEs which appear by the incorporation of a constraint via a Lagrange multiplier.

The main idea is to translate regularization techniques from the theory of DAEs to the infinite dimensional case. More precisely, we introduce a regularization process which is influenced by the index reduction technique of minimal extension. Here we use the fact that the system is assumed to be of semi-explicit structure. Note that this adoption is enabled by the formulation of the constrained PDE as operator DAE. By this we maintain the typical DAE structure although the system is stated in a Banach space formulation and we can translate the known results to the abstract framework.

We will call the presented procedure a regularization or index reduction on operator level. The reason for that, and thus, the justification of the reformulation, is discussed in detail in the two subsequent parts of this thesis. In Part C we analyse the beneficial effect on the spatial-discretized system whereas Part D investigates the impact on the convergence and stability of the temporal discretization.

This part is partitioned into two sections dealing with first- and second-order operator DAEs. In Section 6, we consider semi-explicit operator DAEs of first order as they appear within the Lagrangian method. Thus, we consider systems where the constraint is enforced

by the insertion of a Lagrange multiplier which leads to an operator DAE of saddle point structure. We gather assumptions on the constraint - of linear and nonlinear type - which allow a regularization in the sense described above. The section ends with a discussion of particular examples which fit in the given framework. This includes the Navier-Stokes equations, an example from PDE constrained optimization, and the regularized Stefan problem.

Second-order equations coming from the field of elastodynamics are subject of Section 7. Therein, we focus on a particular case with the constraint given by Dirichlet boundary conditions. For this example, existence and uniqueness results as well as the continuous dependency of the solution on the data are analysed for a particular problem including a nonlinear damping term.

6. Regularization of First-order Operator DAEs

This section is devoted to the regularization of first-order operator DAEs of semi-explicit structure. More precisely, we consider an operator ODE as introduced in Section 4.2,

$$\dot{u}(t) + \mathcal{K}(t)u(t) = \mathcal{F}(t),$$

stated in the dual space of a real, separable, and reflexive Banach space $\mathcal{V}$ which is constrained by an equation

$$\mathcal{B}(t)u(t) = \mathcal{G}(t).$$

Therein, the constraint is formulated in the dual space of a second real, separable, and reflexive Banach space $\mathcal{Q}$. We further introduce the separable Hilbert space $\mathcal{H}$ which is assumed to form a Gelfand triple $\mathcal{V} \hookrightarrow \mathcal{H} \cong \mathcal{H}^* \hookrightarrow \mathcal{V}^*$, see Section 3.3.1. The space $\mathcal{H}$ is of special importance for the incorporation of the initial condition. The application of the Lagrangian method [**Ste08**, Ch. 4.1.2] then leads to an operator DAE as in Section 4.3. The overall problem has the form:

Find abstract functions $u\colon [0,T] \to \mathcal{V}$ and $\lambda\colon [0,T] \to \mathcal{Q}$ such that

$$\dot{u}(t) + \mathcal{K}u(t) + \Big(\frac{\partial \mathcal{B}}{\partial u}(u)\Big)^* \lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{6.1a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^* \tag{6.1b}$$

holds for a.e. $t \in [0,T]$ with initial condition

$$u(0) = g \in \mathcal{H}. \tag{6.1c}$$

Note that the constraint in enforced by the Lagrange multiplier λ which is applied to the dual operator of the Fréchet derivative of $\mathcal{B}$, i.e.,

$$\Big\langle \Big(\frac{\partial \mathcal{B}}{\partial u}(u)\Big)^* \lambda(t), v \Big\rangle_{\mathcal{V}^*,\mathcal{V}} := \Big\langle \lambda(t), \frac{\partial \mathcal{B}}{\partial u}(u)v \Big\rangle_{\mathcal{Q},\mathcal{Q}^*}.$$

Recall that this system forms an operator DAE since equation (6.1b) generalizes an algebraic constraint in the DAE framework. Obviously, the system is semi-linear and semi-explicit in terms of the given constraint.

As discussed in Section 4, the choice of suitable ansatz spaces for a solution (u, λ) is crucial. The right Sobolev-Bochner spaces with the embeddings given in Section 3.3 then also provide an adequate meaning of the initial condition (6.1c). As in the finite-dimensional case, the initial condition has to satisfy a consistency condition as we discuss below. For the right-hand sides we assume that $\mathcal{F} \in L^q(0,T;\mathcal{V}^*)$ and $\mathcal{G} \in W^{1;p}(0,T;\mathcal{Q}^*)$. Note that due to the semi-explicit structure of system (6.1), we do not need generalized derivatives of $\mathcal{F}$. This corresponds to the finite-dimensional case. The discussion on

Nemytskii maps in Section 4.1 leads to the assumption that the operator $\mathcal{K}$ extends to the map $\mathcal{K}\colon L^p(0,T;\mathcal{V}) \to L^q(0,T;\mathcal{V}^*)$ with $1 < q \leq p < \infty$. Note that for the regularization we do not require the boundedness of the operator. Furthermore, we assume $p \geq 2$ which is typical for applications.

The proper search space for the solution u turns out to be $W^{1;p,q}(0,T;\mathcal{V},\mathcal{V}^*)$. However, we need to guarantee the well-posedness of the initial condition (6.1c).

REMARK 6.1. In the case $q \geq p'$ with the conjugate exponent $p' = 1-1/p$, Lemma 3.46 implies the regularity $u \in C([0,T],\mathcal{H})$. Thus, the initial condition (6.1c) is well-posed. However, since the analysis below is also valid for smaller exponents, we refrain from a restriction of the exponent.

For the regularization of operator DAEs we consider two cases. In Section 6.1 we analyse the case of linear constraints, i.e., the operator $\mathcal{B}$ is linear in u. However, we allow the constraint operator to be time-dependent. Nonlinear operators $\mathcal{B}$ are then subject of Section 6.2. Therein, we restrict ourselves to the time-independent case.

The results of this section are published within Sections 3 and 5 of [**AH14**].

6.1. Linear Constraints. Consider the operator DAE (6.1) with a linear constraint operator $\mathcal{B}(t)\colon \mathcal{V} \to \mathcal{Q}^*$. The linearity then implies that the Fréchet derivative of $\mathcal{B}$ is equal to itself. Thus, the operator DAE is simply given by

$$\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*(t)\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{6.2a}$$

$$\mathcal{B}(t)u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^* \tag{6.2b}$$

with a (consistent) initial condition as is (6.1c).

6.1.1. *Assumptions on $\mathcal{B}$.* The regularization of the operator DAE (6.2) requires several properties of the constraint operator $\mathcal{B}$. We summarize these requirements in the following assumption.

ASSUMPTION 6.2 (Properties of $\mathcal{B}$ [**AH14**]). *The constraint operator $\mathcal{B}(t)\colon \mathcal{V} \to \mathcal{Q}^*$ satisfies the following conditions:*

(a) $\mathcal{B}(t)$ is linear and uniformly bounded; $\mathcal{B}(\cdot)v$ is measurable for all $v \in \mathcal{V}$,

(b) $\mathcal{V}_\mathcal{B} := \ker \mathcal{B}(t)$ is independent of the time t,

(c) there exists a uniformly bounded right inverse of $\mathcal{B}(t)$, i.e., there exists a uniformly bounded operator $\mathcal{E}(t)\colon \mathcal{Q}^* \to \mathcal{V}$ such that for all $q \in \mathcal{Q}^*$ and a.e. $t \in [0,T]$ it holds that

$$\mathcal{B}(t)\mathcal{E}(t)q = q,$$

(d) the range of the right inverse $\mathcal{V}^c := \operatorname{range} \mathcal{E}(t)$ is independent of time t, and

(e) the time derivatives $\dot{\mathcal{B}}(t)\colon \mathcal{V} \to \mathcal{Q}^*$ and $\dot{\mathcal{E}}(t)\colon \mathcal{Q}^* \to \mathcal{V}$ are uniformly bounded.

REMARK 6.3. In the time-invariant case $\mathcal{B}(t) \equiv \mathcal{B}$, Assumption 6.2 reduces to the points (a) and (c). The continuity constants of $\mathcal{B}$ and $\mathcal{E}$ are denoted by $C_\mathcal{B}$ and $C_\mathcal{E}$, respectively.

REMARK 6.4. In the finite-dimensional case, it would be sufficient to assume that the range of the right-inverse has a constant rank, i.e., the spaces may change with time. Accordingly, one may think of a time-dependent Riesz basis of $\mathcal{V}$ which then defines time-dependent subspaces $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^c$. However, this is excluded here, since it requires an extension of the theory of Sobolev-Bochner spaces.

LEMMA 6.5 (Induced operators). *Let $\mathcal{B}(t)\colon \mathcal{V} \to \mathcal{Q}^*$ satisfy Assumption 6.2 with the pointwise right inverse $\mathcal{E}(t)$. Then, these operators induce the Nemytskii mappings*

$$\mathcal{B}\colon L^p(0,T;\mathcal{V}) \to L^p(0,T;\mathcal{Q}^*) \quad \text{and} \quad \mathcal{E}\colon L^p(0,T;\mathcal{Q}^*) \to L^p(0,T;\mathcal{V}^c),$$

where $\mathcal{E}$ is the right inverse of $\mathcal{B}$.

PROOF. The fact that $\mathcal{B}$ maps into the space $L^p(0,T;\mathcal{Q}^*)$ can be proved similar to the result in Corollary 4.4 and follows by the uniform boundedness of $\mathcal{B}(t)$. For this, consider an arbitrary $v \in L^p(0,T;\mathcal{V})$ and let p' denote the conjugate exponent of p. Then, every $q \in L^{p'}(0,T;\mathcal{Q})$ satisfies due to Hölders inequality, cf. see Section 3.2,

$$\begin{aligned}\big\langle \mathcal{B}v, q\big\rangle &= \int_0^T \big\langle \mathcal{B}(t)v(t), q(t)\big\rangle_{\mathcal{Q}^*,\mathcal{Q}} \,\mathrm{d}t \\ &\leq C_{\mathcal{B}} \int_0^T \|v(t)\|_{\mathcal{V}}\|q(t)\|_{\mathcal{Q}} \,\mathrm{d}t \leq C_{\mathcal{B}} \;\|v\|_{L^p(0,T;\mathcal{V})}\|q\|_{L^{p'}(0,T;\mathcal{Q})}.\end{aligned}$$

Thus, $\mathcal{B}v$ is bounded for all $q \in L^{p'}(0,T;\mathcal{Q})$ which implies that $\mathcal{B}v$ is itself an element of the dual space, namely $L^p(0,T;\mathcal{Q}^*)$. Since also $\mathcal{E}(t)$ is assumed to be uniformly bounded, we similarly obtain that the operator $\mathcal{E}$ maps functions from $L^p(0,T;\mathcal{Q}^*)$ to $L^p(0,T;\mathcal{V}^c)$.

Next, we show that $\mathcal{B}$ is surjective. For this, consider an arbitrarily element $q \in L^p(0,T;\mathcal{Q}^*)$, i.e.,

$$\int_0^T \|q(t)\|_{\mathcal{Q}^*}^p \,\mathrm{d}t < \infty.$$

In particular, $q(t) \in \mathcal{Q}^*$ for a.e. $t \in [0,T]$ which implies that $\mathcal{E}(t)$ is applicable and

$$v(t) := \mathcal{E}(t)q(t) \in \mathcal{V}^c.$$

Because of Assumption 6.2, $v(t)$ satisfies $\mathcal{B}(t)v(t) = q(t)$ and it only remains to prove that $v \in L^p(0,T;\mathcal{V})$. This then implies $\mathcal{B}v = q$ and thus, q is in the range of $\mathcal{B}$. Because of the uniform boundedness of $\mathcal{E}(t)$, we have the estimate

$$\int_0^T \|v(t)\|_{\mathcal{V}}^p \,\mathrm{d}t = \int_0^T \|\mathcal{E}(t)q(t)\|_{\mathcal{V}}^p \,\mathrm{d}t \leq C_{\mathcal{E}}^p \int_0^T \|q(t)\|_{\mathcal{Q}^*}^p \,\mathrm{d}t < \infty. \qquad \square$$

Note that the choice of the right inverse $\mathcal{E}$ (and therefore also $\mathcal{V}^c$) in Assumption 6.2 is not unique. For a Hilbert space $\mathcal{V}$ the canonical choice for the complement space is certainly the orthogonal complement of the kernel $\mathcal{V}_{\mathcal{B}}$. We proceed with a collection of properties of the right-inverse.

LEMMA 6.6 (Properties of $\mathcal{E}$ [**AH14**]). *Let $\mathcal{B}(t)$ satisfy Assumption 6.2. Then, the right inverse $\mathcal{E}(t)\colon \mathcal{Q}^* \to \mathcal{V}^c \subset \mathcal{V}$ is linear and one-to-one. Furthermore, $\mathcal{V}^c = \operatorname{range}\mathcal{E}(t)$ is a subspace of $\mathcal{V}$ and the operator $\mathcal{E}(t)\mathcal{B}(t)\colon \mathcal{V} \to \mathcal{V}$, restricted to $\mathcal{V}^c$, equals the identity.*

PROOF. The linearity of $\mathcal{E}(t)$ follows from the linearity of the operator $\mathcal{B}(t)$ [**RR04**, Ch. 8.1.2]. For the one-to-one relation, consider q_1, $q_2 \in \mathcal{Q}^*$ with $\mathcal{E}(t)q_1 = \mathcal{E}(t)q_2$. Then, the application of $\mathcal{B}(t)$ yields $q_1 = \mathcal{B}(t)\mathcal{E}(t)q_1 = \mathcal{B}(t)\mathcal{E}(t)q_2 = q_2$.

The linearity of $\mathcal{E}(t)$ implies that $\mathcal{V}^c$ is a subspace of $\mathcal{V}$. Finally, for $v \in \mathcal{V}^c$ and fixed $t \in [0,T]$ there exists an element $q \in \mathcal{Q}^*$ with $\mathcal{E}(t)q = v$. Then, Assumption 6.2 implies

$$v = \mathcal{E}(t)q = \mathcal{E}(t)\big(\mathcal{B}(t)\mathcal{E}(t)q\big) = \mathcal{E}(t)\mathcal{B}(t)v. \qquad \square$$

In particular, Lemma 6.6 implies that $\mathcal{E}(t)\mathcal{B}(t)\colon \mathcal{V} \to \mathcal{V}$ is a projection onto $\mathcal{V}^c$ for a.e. $t \in [0,T]$. As a result, the operator $\mathcal{B}(t)$ defines an isomorphism as operator $\mathcal{B}(t)\colon \mathcal{V}^c \to \mathcal{Q}^*$. Equivalently, by [**Bra07**, Lem. III.4.2], the dual operator

$$\mathcal{B}^*(t)\colon \mathcal{Q} \to \big\{ f \in \mathcal{V}^* \mid \langle f, v_0\rangle = 0 \text{ for all } v_0 \in \mathcal{V}_{\mathcal{B}} \big\} \subseteq \mathcal{V}^*$$

defines an isomorphism. A third equivalence is given by the well-known stability concept of the *inf-sup condition*, i.e., there exists a positive constant $\beta > 0$ such that

$$\inf_{q\in\mathcal{Q}} \sup_{v\in\mathcal{V}} \frac{\langle \mathcal{B}(t)v, q\rangle}{\|v\|_{\mathcal{V}}\|q\|_{\mathcal{Q}}} \geq \beta > 0.$$

Since $\mathcal{B}(t)$ and $\mathcal{E}(t)$ are assumed to be uniformly bounded, also the inf-sup constant β is independent of time. Note that the existence of a uniform inf-sup constant implies the existence of a right inverse $\mathcal{E}(t)$ but not point (d) of Assumption 6.2, i.e., the range of $\mathcal{E}(t)$ is not necessarily time-independent.

Another implication of Assumption 6.2 is the resulting decomposition of $L^p(0,T;\mathcal{V})$. This decomposition is necessary for the splitting of the variable u into an 'differential' and 'algebraic' part within the index reduction procedure of Section 6.1.2.

LEMMA 6.7 (Decomposition of $L^p(0,T;\mathcal{V})$ [**AH14**]). *Consider the subspaces $\mathcal{V}_{\mathcal{B}}$ and $\mathcal{V}^c$ of $\mathcal{V}$ from Assumption 6.2. Then, we have the decomposition*

$$L^p(0,T;\mathcal{V}) = L^p(0,T;\mathcal{V}_{\mathcal{B}}) \oplus L^p(0,T;\mathcal{V}^c).$$

PROOF. For given $v \in L^p(0,T;\mathcal{V})$, we define $r := \mathcal{B}v \in L^p(0,T;\mathcal{Q}^*)$, cf. Lemma 6.5. Then, a decomposition of $v \in L^p(0,T;\mathcal{V})$ is given by

$$v = v_0 + v^c := \big(v - \mathcal{E}r\big) + \mathcal{E}r. \tag{6.3}$$

Obviously, $v^c = \mathcal{E}r \in L^p(0,T;\mathcal{V}^c)$ and $v_0 \in L^p(0,T;\mathcal{V}_{\mathcal{B}})$ follows from Assumption 6.2 by $\mathcal{B}v_0 = \mathcal{B}v - \mathcal{B}\mathcal{E}\mathcal{B}v = 0$. We show that the decomposition in (6.3) is unique. For this, consider v_0, $w_0 \in L^p(0,T;\mathcal{V}_{\mathcal{B}})$ and v^c, $w^c \in L^p(0,T;\mathcal{V}^c)$ with $v = v_0 + v^c = w_0 + w^c$. The application of $\mathcal{B}$ yields $\mathcal{B}v^c = \mathcal{B}w^c$. Furthermore, there exist r_v, $r_w \in L^p(0,T;\mathcal{Q}^*)$ such that $v^c = \mathcal{E}r_v$ and $w^c = \mathcal{E}r_w$. By Assumption 6.2 we then obtain

$$r_v - r_w = \mathcal{B}\mathcal{E}r_v - \mathcal{B}\mathcal{E}r_w = \mathcal{B}v^c - \mathcal{B}w^c = 0.$$

Thus, it holds that $v^c = \mathcal{E}r_v = \mathcal{E}r_w = w^c$ and finally also $v_0 = w_0$. □

6.1.2. *Regularization.* As mentioned in the introduction of this section, the aim is a regularization of the operator DAE (6.2). The justification of this reformulation and the verification that this procedure is in fact an index reduction will be given in Section 8.2. Therein, the semi-discretization is performed which leads to a DAE of lower index.

The presented regularization is an adaptation of the index reduction technique of *minimal extension* [**KM04**], see also Section 2.3.2. In the sequel, we assume that the operator $\mathcal{B}$ satisfies Assumption 6.2.

As explained in Section 2, it is well-known from DAE theory that the existence of derivatives of the right-hand side are necessary for the solvability of a DAE. Since the operator DAE (6.2) results under semi-discretization in a DAE of index 2 (see Section 8.2), it can be expected that derivatives of the right-hand side are essential. Because of the semi-explicit structure of the operator DAE, only the derivative of $\mathcal{G}$ is necessary and we assume $\mathcal{G} \in W^{1;p}(0,T;\mathcal{Q}^*)$.

Consider the time derivative of the constraint (6.2b), which is formally given by

$$\mathcal{B}\dot{u} + \dot{\mathcal{B}}u = \dot{\mathcal{G}}.$$

In terms of the theory of DAEs, this is the hidden constraint of the operator DAE (6.2). At this point we assume sufficient regularity of the solution, namely $\dot{u} \in L^p(0,T;\mathcal{V})$. We will discover that this condition may be relaxed, since for the differential part of u it is sufficient to have a derivative in the dual space $\mathcal{V}^*$.

In order to split the variable u, we apply the decomposition of $L^p(0,T;\mathcal{V})$ into $L^p(0,T;\mathcal{V}_\mathcal{B})$ and $L^p(0,T;\mathcal{V}^c)$ from Lemma 6.7. Thus, we define $u_1 \in L^p(0,T;\mathcal{V}_\mathcal{B})$ and $u_2 \in L^p(0,T;\mathcal{V}^c)$ uniquely such that $u = u_1 + u_2$. Assuming that also u_1 and u_2 are differentiable in time, we can reduce the constraint (6.2b) and its derivative to

$$\mathcal{B}u_2 = \mathcal{G} \qquad \text{and} \qquad \mathcal{B}\dot{u}_2 + \dot{\mathcal{B}}u_2 = \dot{\mathcal{G}}.$$

Note that we make use of Lemma 3.47 at this point which implies that $\dot{u}_2 \in L^p(0,T;\mathcal{V}^c)$ and thus, the decomposition of $\dot{u}$ is given by $\dot{u} = \dot{u}_1 + \dot{u}_2$. The reduction of the constraint then causes that the operator $\mathcal{B}$ is only applied to the derivative of u_2. The assumed regularity of $\mathcal{G}$ implies with Assumption 6.2, Lemma 6.6, and equation (6.2b) that $u_2 \in W^{1;p}(0,T;\mathcal{V}^c)$. For the derivative of u_1 it is sufficient to have $\dot{u}_1 \in L^q(0,T;\mathcal{V}^*)$.

Following the procedure of minimal extension, we add an additional variable $v_2 := \dot{u}_2 \in L^p(0,T;\mathcal{V}^c)$. Since this variable is normally not of interest, it is often called a *dummy variable*. Note that the invertibility of $\mathcal{B}(t)$ on $\mathcal{V}^c$ from Lemma 6.6 corresponds to the full rank property of the Jacobian in the finite-dimensional case. This explains the choice of the new variable. Finally, we obtain again a balanced number of variables and equations.

Replacing all appearances of $\dot{u}_2$ by v_2, we note that in the resulting operator DAE u_2 is not differentiated anymore. Thus, the initial condition only concerns u_1. This corresponds to the fact that the initial condition (6.1c) has to satisfy a consistency condition, see Section 4.3. The initial condition for u_1 can be chosen arbitrarily in the closure of $\mathcal{V}_\mathcal{B}$ in $\mathcal{H}$, see the discussion in Remark 6.9. The initial data for u_2 is fixed by the right-hand side $\mathcal{G}$.

With these preparations, we are able to formulate the regularized operator DAE corresponding to (6.2). For this we neglect to write the time-dependence of $\mathcal{B}$. Given right-hand sides $\mathcal{F} \in L^q(0,T;\mathcal{V}^*)$, $\mathcal{G} \in W^{1;p}(0,T;\mathcal{Q}^*)$ and initial data $g \in \mathcal{H}$, find functions $u_1 \in W^{1;p,q}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}^*)$, $u_2, v_2 \in L^p(0,T;\mathcal{V}^c)$, and $\lambda \in L^{p'}(0,T;\mathcal{Q})$ such that

$$\dot{u}_1(t) + v_2(t) + \mathcal{K}\big(u_1(t) + u_2(t)\big) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{6.4a}$$

$$\mathcal{B}u_2(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*, \tag{6.4b}$$

$$\mathcal{B}v_2(t) + \dot{\mathcal{B}}u_2(t) = \dot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^* \tag{6.4c}$$

holds for a.e. $t \in [0,T]$ with initial condition

$$u_1(0) = g - \mathcal{E}\mathcal{G}(0) \in \mathcal{H}. \tag{6.4d}$$

The well-posedness of the initial condition (6.4d) is again guaranteed by the embedding due to the underlying Gelfand triple and the fact that $\mathcal{G}(0) \in \mathcal{Q}^*$ is well-defined. The latter is true, since $W^{1;p}(0,T;\mathcal{Q}^*)$ is continuously embedded in $C([0,T],\mathcal{Q}^*)$, see Section 3.3.2. Note that this does not imply that there exists a solution of system (6.4) for every initial data $g \in \mathcal{H}$, cf. Remark 6.9 below. The following theorem shows that the original system (6.2) and the regularized system (6.4) are equivalent.

THEOREM 6.8 (Equivalence of the reformulation [**AH14**]). *Consider exponents $1 < q \le p < \infty$ and p' with $1/p' + 1/p = 1$. Assume that $\mathcal{F} \in L^q(0,T;\mathcal{V}^*)$, $\mathcal{G} \in W^{1;p}(0,T;\mathcal{Q}^*)$, and $g \in \mathcal{H}$ as well as the operator $\mathcal{B}$ satisfying Assumption 6.2. Then, the operator DAE (6.2) has a solution (u,λ) with $u \in W^{1;p,q}(0,T;\mathcal{V},\mathcal{V}^*)$, and $\lambda \in L^{p'}(0,T;\mathcal{Q})$ if and only if system (6.4) has a solution (u_1,u_2,v_2,λ) with $u_1 \in W^{1;p,q}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}^*)$, u_2, $v_2 \in L^p(0,T;\mathcal{V}^c)$, and $\lambda \in L^{p'}(0,T;\mathcal{Q})$. Furthermore, it holds that $u = u_1 + u_2$ and $\dot{u}_2 = v_2$.*

PROOF. Since $\mathcal{Q}$ is separable, the dual space of $L^{p'}(0,T;\mathcal{Q})$ can be identified with $L^p(0,T;\mathcal{Q}^*)$, cf. Proposition 3.40. Thus, $L^{p'}(0,T;\mathcal{Q})$ is a suitable space for the Lagrange multiplier λ. Let (u,λ) be a solution of (6.2). We define

$$u_1 := u - \mathcal{E}\mathcal{B}u \in L^p(0,T;\mathcal{V}_\mathcal{B}) \quad \text{and} \quad u_2 := \mathcal{E}\mathcal{B}u \in L^p(0,T;\mathcal{V}^{\mathrm{c}}).$$

With (6.2b), we obtain $u_2 = \mathcal{E}\mathcal{G}$ and thus, by the regularity of $\mathcal{G}$ and Assumption 6.2, $\dot{u}_2 \in L^p(0,T;\mathcal{V}^{\mathrm{c}})$. With $v_2 := \dot{u}_2$ the quadruple (u_1,u_2,v_2,λ) satisfies equations (6.4a-c). The initial condition (6.4d) is satisfied because of

$$u_1(0) = u(0) - u_2(0) = g - \mathcal{E}\mathcal{G}(0).$$

For the reverse direction consider a solution of (6.4), namely (u_1,u_2,v_2,λ). Then, $u := u_1 + u_2 \in L^p(0,T;\mathcal{V})$ and because of the regularity of $\mathcal{G}$, equation (6.4b), and the boundedness of $\mathcal{B}$ and $\dot{\mathcal{B}}$, it holds that $\dot{u} = \dot{u}_1 + \dot{u}_2 \in L^q(0,T;\mathcal{V}^*)$. We show that $\dot{u}_2 = v_2$. Equation (6.4c) and the time derivative of equation (6.4b) yield

$$\mathcal{B}v_2 + \dot{\mathcal{B}}u_2 = \dot{\mathcal{G}} = \frac{\mathrm{d}}{\mathrm{d}t}\big(\mathcal{B}u_2\big) = \mathcal{B}\dot{u}_2 + \dot{\mathcal{B}}u_2.$$

Note that $\dot{u}_2 \in L^p(0,T;\mathcal{V}^{\mathrm{c}})$, as shown in the first part of the proof. The invertibility of $\mathcal{B}$ on $\mathcal{V}^{\mathrm{c}}$ (see Lemma 6.6) then gives $\dot{u}_2 = v_2$. Thus, the pair (u,λ) satisfies equations (6.2a) and (6.2b). For the initial condition (6.1c), we obtain

$$u(0) = u_1(0) + u_2(0) = g - \mathcal{E}\mathcal{G}(0) + \mathcal{E}\mathcal{G}(0) = g. \qquad \square$$

REMARK 6.9 (Consistent initial data). One conclusion of Theorem 6.8 is that not every initial data $g \in \mathcal{H}$ may lead to a solution to (6.2). More precisely, it provides the necessary condition that g can be decomposed into $g = g_0 + \mathcal{E}\mathcal{G}(0)$ with $\mathcal{E}\mathcal{G}(0) \in \mathcal{V}^{\mathrm{c}}$ and g_0 is in the closure of $\mathcal{V}_\mathcal{B}$ in $\mathcal{H}$. Thus, we suggest to write the initial condition (6.4d) in the form

$$u_1(0) = g_0 \in \overline{\mathcal{V}_\mathcal{B}}^{\mathcal{H}}. \tag{6.5}$$

Note that this provides no further restriction and goes along with the theory for abstract ODEs in Section 4.2 since the operator equation (6.4) contains no constraint for u_1. To see this, we consider the Gelfand triple corresponding to the kernel of $\mathcal{B}$, namely

$$\mathcal{V}_\mathcal{B} \hookrightarrow \mathcal{H}_\mathcal{B} \cong \mathcal{H}_\mathcal{B}^* \hookrightarrow \mathcal{V}_\mathcal{B}^*.$$

Since a dense embedding is required, $\mathcal{H}_\mathcal{B}$ has to equal the closure of $\mathcal{V}_\mathcal{B}$ in $\mathcal{H}$, i.e., $\mathcal{H}_\mathcal{B} := \overline{\mathcal{V}_\mathcal{B}}^{\mathcal{H}}$. Thus, $u_1 \in L^p(0,T;\mathcal{V}_\mathcal{B})$ with $\dot{u}_1 \in L^q(0,T;\mathcal{V}^*) \hookrightarrow L^q(0,T;\mathcal{V}_\mathcal{B}^*)$ implies $u_1 \in C([0,T];\mathcal{H}_\mathcal{B})$ for $q \geq p/(p-1)$ such that the initial condition (6.5) is indeed well-posed.

EXAMPLE 6.10. If the operator $\mathcal{B}$ equals the divergence operator and $\mathcal{V} = [H_0^1(\Omega)]^d$, then $\mathcal{V}_\mathcal{B}$ denotes the space of divergence-free functions in $\mathcal{V}$. In this case, the closure of $\mathcal{V}_\mathcal{B}$ w.r.t. $\mathcal{H} = [L^2(\Omega)]^d$ is a proper subspace of $\mathcal{H}$, cf. [**Tem77**, Ch. 1, Thm. 1.4],

$$\overline{\mathcal{V}_\mathcal{B}}^{\mathcal{H}} = \big\{ v \in \mathcal{H} \mid \nabla \cdot v = 0,\ v \cdot \nu_{\partial\Omega} = 0 \big\} \neq \mathcal{H}.$$

Note that the closure is even a subspace of $H(\mathrm{div},\Omega) = \{v \in \mathcal{H} \mid \nabla \cdot v \in L^2(\Omega)\}$. Thus, the initial value g_0 cannot be chosen arbitrarily in $\mathcal{H}$.

EXAMPLE 6.11. If $\mathcal{B}$ equals the trace operator from Section 3.1.4, i.e., $\mathcal{B}\colon \mathcal{V} = H^1(\Omega) \to H^{1/2}(\Omega)$, then we have $\mathcal{V}_\mathcal{B} = H_0^1(\Omega)$. Since the closure of $H_0^1(\Omega)$ in $\mathcal{H} = L^2(\Omega)$ equals $\mathcal{H}$ itself, the initial data only has to satisfy $g_0 \in \mathcal{H}$.

REMARK 6.12. The presented regularization is not the only possibility to obtain an operator DAE of index-1 type (in the sense of Section 8 below). Similar results can be obtained by including the hidden constraint with an additional Lagrange multiplier. This approach, which is influenced by the GGL-formulation [**GGL85**], needs no splitting of the variable u and maintains the saddle point structure. However, one has to require higher regularity assumptions on the solution. A regularization based on the *Baumgarte stabilization* [**Bau72**] is not recommended, since it strongly depends on the included parameter. Thus, a bad choice may lead to arbitrary inaccurate approximations [**Ost91, ACPR95**].

6.1.3. *Influence of Perturbations.* As mentioned in the introduction, we do not define an index for operator DAEs. However, the index concept for PDAEs from [**LMT01**], see also [**LMT13**, Ch. 12], is defined for the operator DAEs (6.2) and (6.4) if formulated in a stronger setting. Within this classification, the regularized operator DAE (6.4) is of index 1 which is not the case of the original system (6.2). This already provides a characterization of the behavior of the solution w.r.t. perturbations of the right-hand sides. Nevertheless, we analyse the influence of perturbations in detail within this subsection for the case $p = q = 2$. For this, we compare the exact solution (u, λ) with the solution of the perturbed problem

$$\dot{\hat{u}} + \mathcal{K}\hat{u} + \mathcal{B}^*\hat{\lambda} = \mathcal{F} + \delta \quad \text{in } \mathcal{V}^*, \tag{6.6a}$$

$$\mathcal{B}\hat{u} = \mathcal{G} + \theta \quad \text{in } \mathcal{Q}^*. \tag{6.6b}$$

Here, $(\hat{u}, \hat{\lambda})$ denotes the solution if we include perturbations $\delta \in L^2(0,T;\mathcal{V}^*)$ and $\theta \in W^{1;2}(0,T;\mathcal{Q}^*)$. For the regularized equations, the perturbed problem has the form

$$\dot{\hat{u}}_1 + \hat{v}_2 + \mathcal{K}(\hat{u}_1 + \hat{u}_2) + \mathcal{B}^*\hat{\lambda} = \mathcal{F} + \delta \quad \text{in } \mathcal{V}^*, \tag{6.7a}$$

$$\mathcal{B}\hat{u}_2 = \mathcal{G} + \theta \quad \text{in } \mathcal{Q}^*, \tag{6.7b}$$

$$\mathcal{B}\hat{v}_2 + \dot{\mathcal{B}}\hat{u}_2 = \dot{\mathcal{G}} + \xi \quad \text{in } \mathcal{Q}^*. \tag{6.7c}$$

For this system, we consider perturbations of the form $\delta \in L^2(0,T;\mathcal{V}^*)$ and $\theta,\ \xi \in L^2(0,T;\mathcal{Q}^*)$ and the solution is denoted by $(\hat{u}_1, \hat{u}_2, \hat{v}_2, \hat{\lambda})$. The initial condition is given by $\hat{u}_1(0) = u_1(0) - e_{1,0}$, i.e., $e_{1,0}$ contains the initial error. Because of Theorem 6.8, it is sufficient to consider the regularized system (6.7). Within the perturbation analysis of system (6.6) a splitting of u as in the regularization process would be necessary. This then leads to the equation

$$\mathcal{B}\hat{v}_2 + \dot{\mathcal{B}}\hat{u}_2 = \dot{\mathcal{G}} + \dot{\theta}.$$

Thus, the corresponding stability result for the original operator DAE follows if we replace ξ by $\dot{\theta}$. Although the operator $\mathcal{K}$ was not of great importance for the regularization process, we restrict the following analysis to linear, symmetric, on $\mathcal{V}_\mathcal{B}$ coercive, and bounded operators, i.e, for $u \in \mathcal{V}_\mathcal{B}$ and $v,\ w \in \mathcal{V}$ we assume

$$k_1\|u\|_\mathcal{V}^2 \le \langle \mathcal{K}u, u\rangle \quad \text{and} \quad \langle \mathcal{K}v, w\rangle \le k_2\|v\|_\mathcal{V}\|w\|_\mathcal{V}.$$

To simplify notation, we introduce the norms $\|\cdot\| := \|\cdot\|_\mathcal{V}$ and $|\cdot| := \|\cdot\|_\mathcal{H}$. By C_{emb} we denote the continuity constant of the embedding $\mathcal{V} \hookrightarrow \mathcal{H}$. Furthermore, we introduce the errors

$$e_1 := \hat{u}_1 - u_1, \qquad e_2 := \hat{u}_2 - u_2, \qquad e_v := \hat{v}_2 - v_2, \qquad e_\lambda := \hat{\lambda} - \lambda.$$

The errors e_2 and e_v can be easily estimated. Because $e_2 \in \mathcal{V}^c$, the continuity of the right-inverse of $\mathcal{B}$ directly implies with equations (6.4b) and (6.7b) that

$$\|e_2\|_{L^2(0,T;\mathcal{V})} \le C_\mathcal{E}\|\theta\|_{L^2(0,T;\mathcal{Q}^*)}.$$

Equation (6.7c) implies

$$\|e_v\|_{L^2(0,T;\mathcal{V})} \leq C_{\mathcal{E}}\|\xi\|_{L^2(0,T;\mathcal{Q}^*)} + C_{\mathcal{E}}^2 C_{\dot{\mathcal{B}}}\|\theta\|_{L^2(0,T;\mathcal{Q}^*)}.$$

Therein, $C_{\mathcal{E}}$ and $C_{\dot{\mathcal{B}}}$ denote the continuity constants of $\mathcal{E}$ and $\dot{\mathcal{B}}$, respectively. With these two estimates, we can derive an estimate for the error in the dynamic part u_1. For this, we test the difference of equations (6.7a) and (6.4a) with $e_1 \in \mathcal{V}_{\mathcal{B}}$. Thus, the term with the Lagrange multiplier vanishes and we obtain

$$\langle \dot{e}_1, e_1\rangle + \langle e_v, e_1\rangle + \langle \mathcal{K}(e_1+e_2), e_1\rangle = \langle \delta, e_1\rangle.$$

With this and the properties of $\mathcal{K}$, we obtain the estimate

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|e_1|^2 + k_1\|e_1\|^2 &\leq \langle \delta, e_1\rangle - \langle e_v, e_1\rangle - \langle \mathcal{K}e_2, e_1\rangle \\ &\leq \|\delta\|_{\mathcal{V}^*}\|e_1\| + C_{\mathrm{emb}}^2\|e_v\|\|e_1\| + k_2\|e_2\|\|e_1\|.\end{aligned}$$

With Young's inequality [**Eva98**, App. B], we further obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}|e_1|^2 + k_1\|e_1\|^2 \leq \frac{3}{k_1}\Big[\|\delta\|_{\mathcal{V}^*}^2 + C_{\mathrm{emb}}^4\|e_v\|^2 + k_2^2\|e_2\|^2\Big].$$

Integration over the interval $[0,t]$ for $t \leq T$ leads to

$$|e_1(t)|^2 - |e_1(0)|^2 + k_1\|e_1\|_{L^2(0,t;\mathcal{V})}^2 \leq \frac{3}{k_1}\Big[\|\delta\|_{L^2(0,T;\mathcal{V}^*)}^2 + C_{\mathrm{emb}}^4\|e_v\|_{L^2(0,T;\mathcal{V})}^2 + k_2^2\|e_2\|_{L^2(0,T;\mathcal{V})}^2\Big].$$

Note that this estimate holds for all $t \in [0,T]$. Thus, maximizing over t and inserting the estimates of e_2 and e_v, we obtain the following result.

THEOREM 6.13. *Consider the perturbed problem* (6.7) *with a linear, symmetric, coercive, and bounded operator $\mathcal{K}$, $\mathcal{B}$ satisfying Assumption 6.2, and perturbations $\delta \in L^2(0,T;\mathcal{V}^*)$ and θ, $\xi \in L^2(0,T;\mathcal{Q}^*)$. With $e_1 := \hat{u}_1 - u_1$, where u_1 denotes the solution of the unperturbed problem, the solution $\hat{u}_1$ then satisfies the estimate*

$$\|e_1\|_{C([0,T];\mathcal{H})}^2 + k_1\|e_1\|_{L^2(0,T;\mathcal{V})}^2 \leq |e_{1,0}|^2 + c\Big[\|\delta\|_{L^2(0,T;\mathcal{V}^*)}^2 + \|\theta\|_{L^2(0,T;\mathcal{Q}^*)}^2 + \|\xi\|_{L^2(0,T;\mathcal{Q}^*)}^2\Big].$$

REMARK 6.14. Note that the errors in u_1, u_2, and v_2 behave as in the operator ODE case, since we have a linear dependence on the perturbations. However, this is only true for the regularized system (6.4). For the original formulation (6.2) we have to insert $\xi = \dot{\theta}$. Thus, the error also depends on the derivative of the perturbation θ which leads to possible instabilities known from high-index DAEs, cf. Section 2. If the perturbation is not smooth enough, one may even obtain useless results.

REMARK 6.15. In the finite-dimensional setting, we obtain a continuous dependence of the error e_1 on θ also for the index-2 case if the constraint is linear [**Arn98b**, Ch. 2, Th. 3]. This does not work in the analysis of the operator equation. Here, we can only get rid of the $\dot{\theta}$-term if we integrate by parts which generates derivatives of e_1. However, the assumed regularity is not sufficient to bound this term such that we only manage to obtain an estimate involving $\dot{\theta}$.

In this weak setting of the evolution equations, it is not possible to gain similar estimates for e_λ. Estimates of the error in the Lagrange multiplier are only possible if we consider the primitive of e_λ or assume more regular data such that $\delta \in L^2(0,T;\mathcal{H}^*)$ and $e_{1,0} \in \mathcal{V}$, cf. Section 10.

We summarize the regularization of first-order operator DAEs with a linear constraint operator $\mathcal{B}$ in the following table. As before, we consider the Gelfand triples $\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}^*$ and $\mathcal{V}_{\mathcal{B}} \hookrightarrow \mathcal{H}_{\mathcal{B}} \hookrightarrow \mathcal{V}_{\mathcal{B}}^*$ with $\mathcal{H}_{\mathcal{B}} = \overline{\mathcal{V}_{\mathcal{B}}}^{\mathcal{H}}$. Furthermore, we use the abbreviations $L^2(\mathcal{V}^*) :=$

$L^2(0,T;\mathcal{V}^*)$ and $L^2(\mathcal{Q}^*) := L^2(0,T;\mathcal{Q}^*)$ and write $a \lesssim b$ for the existence of a positive constant $c \in \mathbb{R}$ such that $a \le cb$.

	original formulation	regularized formulation
system of equations	operator DAE (6.2)	operator DAE (6.4)
solution spaces	$u \in W^{1;p,q}(0,T;\mathcal{V},\mathcal{V}^*)$, $\lambda \in L^{p'}(0,T;\mathcal{Q})$	$u_1 \in W^{1;p,q}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}^*)$, $u_2,\, v_2 \in L^p(0,T;\mathcal{V}^c)$, $\lambda \in L^{p'}(0,T;\mathcal{Q})$
initial condition and consistency	$u(0) = g \in \mathcal{H}$ $g = g_0 + \mathcal{B}^-\mathcal{G}(0),\ g_0 \in \mathcal{H}_\mathcal{B}$	$u_1(0) = g_0 \in \mathcal{H}_\mathcal{B}$
spatial discretization	leads to DAE of index 2, cf. Section 8	leads to DAE of index 1, cf. Section 8
perturbations	$\|e_1\|^2_{C([0,T];\mathcal{H})} + \|e_1\|^2_{L^2(0,T;\mathcal{V})}$ $\lesssim \|e_{1,0}\|^2 + \|\delta\|^2_{L^2(\mathcal{V}^*)}$ $+\|\theta\|^2_{L^2(\mathcal{Q}^*)} + \|\dot\theta\|^2_{L^2(\mathcal{Q}^*)}$	$\|e_1\|^2_{C([0,T];\mathcal{H})} + \|e_1\|^2_{L^2(0,T;\mathcal{V})}$ $\lesssim \|e_{1,0}\|^2 + \|\delta\|^2_{L^2(\mathcal{V}^*)}$ $+\|\theta\|^2_{L^2(\mathcal{Q}^*)} + \|\xi\|^2_{L^2(\mathcal{Q}^*)}$

6.2. Nonlinear Constraints. The regularization from the previous subsection is also applicable to nonlinear constraint operators $\mathcal{B}\colon \mathcal{V} \to \mathcal{Q}^*$. Clearly, the assumptions on the operator have to be adapted accordingly. In order to focus on the changes due to the nonlinearity, we restrict ourselves to the time-independent case.

We denote the Fréchet derivative of $\mathcal{B}$ at $v \in \mathcal{V}$ by

$$\mathcal{C}_v := \frac{\partial \mathcal{B}}{\partial u}(v)\colon \mathcal{V} \to \mathcal{Q}^*.$$

The operator DAE (6.1) then reads

$$\dot u(t) + \mathcal{K}u(t) + \mathcal{C}_u^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{6.8a}$$

$$\mathcal{B}u(t) \qquad\quad = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^* \tag{6.8b}$$

for a.e. $t \in [0,T]$ with the initial condition

$$u(0) = g \in \mathcal{H}. \tag{6.8c}$$

Throughout this section, we suppose that there exists a solution of the constrained system (6.8) which satisfies $u \in W^{1;p,q}(0,T;\mathcal{V},\mathcal{V}^*)$ and $\lambda \in L^{p'}(0,T;\mathcal{Q})$. This includes the existence of the Fréchet derivative $\mathcal{C}_u$ along the solution u. As in the linear case, the existence of a solution requires higher derivatives of the right-hand side $\mathcal{G}$.

6.2.1. *Assumptions on $\mathcal{B}$.* Similar to the linear case, we gather properties of the operator $\mathcal{B}$ which allow a reformulation of system (6.8). Once again the aim is to obtain an equivalent but regularized formulation of the given operator DAE in terms of the semi-discretized system. For this, we assume that the constraint manifold is smooth in a certain sense and can be characterized by an implicit function.

Assumption 6.16 (Properties of $\mathcal{B}$ [**AH14**]). *Consider a function $u \in L^p(0,T;\mathcal{V})$ that satisfies $\mathcal{B}u = \mathcal{G}$ in $\mathcal{Q}^*$ for a.e. $t \in [0,T]$. There exists a splitting of $\mathcal{V}$ into subspaces $\mathcal{V}_1$ and $\mathcal{V}_2$, i.e., $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$, and a neighborhood $\mathcal{U}(t) \subseteq \mathcal{V}$ around $u(t)$ such that*

(a) $u = u_1 + u_2$ with $u_1 \in L^p(0,T;\mathcal{V}_1)$, $u_2 \in L^p(0,T;\mathcal{V}_2)$,

(b) the Fréchet derivative $\frac{\partial \mathcal{B}}{\partial u}$ exists in $\mathcal{U}(t)$,

(c) $\mathcal{C}_{2,u} := \frac{\partial \mathcal{B}}{\partial u_2}(u)\colon \mathcal{V}_2 \to \mathcal{Q}^*$ *is a homeomorphism, and*

(d) $\frac{\partial \mathcal{B}}{\partial u_2}(\cdot)$ *is continuous in* u.

REMARK 6.17. Later we use the fact that $\mathcal{C}_u v = \mathcal{C}_{2,u} v$ for $v \in \mathcal{V}_2$. This follows from the definition of the Fréchet derivative which implies that for all $w \in \mathcal{V}$ it holds that

$$\mathcal{B}(u+w) - \mathcal{B}(u) = \mathcal{C}_u w + o(\|w\|).$$

In particular, this equation is satisfied for all $v \in \mathcal{V}_2$ which then equals the definition of $\mathcal{C}_{2,u}$. The uniqueness of the Fréchet derivative finally yields the claim.

REMARK 6.18. The splitting of $\mathcal{V}$ into $\mathcal{V}_1$ and $\mathcal{V}_2$ only depends on the operator $\mathcal{B}$. Thus, the spaces $\mathcal{V}_1$ and $\mathcal{V}_2$ are independent of the right-hand side $\mathcal{G}$ and also independent of time. Note that this independence may restrict the length of the considered time interval, in order to guarantee point (c) of Assumption 6.16. For a longer time horizon or a time-dependent constraint operator $\mathcal{B}(t)$ an update of the decomposition may be necessary. In this case, $\mathcal{V}$ has to be split again and the ansatz space for the differential part u_1 changes.

Assumption 6.16 allows the application of the implicit function theorem for operators. Let $\mathcal{B}$ satisfy this assumption for a function $u \in L^p(0,T;\mathcal{V})$ which thus decomposes into $u = u_1 + u_2$ with $u_1 \in L^p(0,T;\mathcal{V}_1)$ and $u_2 \in L^p(0,T;\mathcal{V}_2)$. The implicit function theorem [**Ruž04**, Ch. 2.2] then implies the existence of a mapping $\eta(t)\colon \mathcal{V}_1 \to \mathcal{V}_2$, the so-called *implicit function*, and a neighborhood $\mathcal{U}_1 \subseteq \mathcal{V}_1$ around u_1 such that

$$\mathcal{B}(v_1 + \eta(v_1)) = \mathcal{G} \quad \text{in } \mathcal{Q}^*$$

for all $v_1 \in \mathcal{U}_1$. Thus, the constraint manifold given by (6.8b) can be locally described by the function η. Note that for the existence of the implicit function, point (b) of Assumption 6.16 can be weakened that only the Fréchet derivative with respect to u_2 exists. Nevertheless, the additional regularity ensures the Fréchet differentiability of η in a neighborhood of u_1 [**Ruž04**, Cor. 2.15]. This property is needed to ensure the differentiability of η in time as shown in the following lemma.

LEMMA 6.19 (Time derivative of η). *Consider* $\mathcal{B}$, $\mathcal{G}$, *and* u_1 *from Assumption 6.16 with* $\dot{u}_1 \in L^p(0,T;\mathcal{V}_1)$. *Furthermore, let* $\eta(t)\colon \mathcal{U}_1 \subseteq \mathcal{V}_1 \to \mathcal{V}_2$ *denote the implicit function defined above. Then, the function* $\eta(\cdot, u_1(\cdot))\colon [0,T] \to \mathcal{V}_2$ *is a.e. differentiable in time.*

PROOF. Note that η depends implicitly on time, since it depends on the right-hand side $\mathcal{G}$. We want to apply the implicit function theorem to the operator $\overline{\mathcal{B}}\colon \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{Q}^* \to \mathcal{Q}^*$, given by

$$\overline{\mathcal{B}}(u_1, u_2, \mathcal{G}) := \mathcal{B}(u_1 + u_2) - \mathcal{G}.$$

Obviously, $\overline{\mathcal{B}}(u_1, u_2, \mathcal{G}) = 0$ is equivalent to (u_1, u_2) being a solution of $\mathcal{B}(u_1 + u_2) = \mathcal{G}$. Because of Assumption 6.16, the implicit function theorem for operators [**Ruž04**, Ch. 2.2] also applies to the operator $\overline{\mathcal{B}}$. Thus, there exists a Fréchet differentiable mapping

$$\bar{\eta}\colon \mathcal{V}_1 \times \mathcal{Q}^* \to \mathcal{V}_2$$

which maps the pair $(u_1, \mathcal{G})$ to u_2 such that $\mathcal{B}(u_1, u_2) = \mathcal{G}$. In contrast to η, the dependence on $\mathcal{G}$ is explicitly included in the implicit function such that $\bar{\eta}$ is independent of t. By assumption, it holds that $\dot{u}_1 \in \mathcal{V}_1$ and $\dot{\mathcal{G}} \in \mathcal{Q}^*$ for a.e. $t \in [0,T]$. Thus, we may calculate

$$\frac{\mathrm{d}}{\mathrm{d}t}\eta(u_1) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{\eta}(u_1, \mathcal{G}) = \frac{\partial \bar{\eta}}{\partial u_1}\dot{u}_1 + \frac{\partial \bar{\eta}}{\partial \mathcal{G}}\dot{\mathcal{G}} \ \in \mathcal{V}_2. \qquad \square$$

Before we consider the regularized equations, we discuss the extensions of $\mathcal{B}$ and $\mathcal{C}_u$ to Bochner integrable functions. Since the Fréchet derivative $\mathcal{C}_u$ is linear, a sufficient condition to obtain a bounded operator of the form

$$\mathcal{C}_u \colon L^p(0,T;\mathcal{V}) \to L^p(0,T;\mathcal{Q}^*)$$

is the uniform boundedness of $\mathcal{C}_u$, cf. the proof of Lemma 6.5. Note that Assumption 6.16 only implies the pointwise continuity of the operator $\mathcal{C}_{2,u}$ and thus, with the help of Lemma 3.48, also for $\mathcal{C}_u$. A uniform continuity constant may be assumed in addition. For the nonlinear operator $\mathcal{B}$, Assumption 6.16 allows to perform the regularization process given in the next subsection. However, to make equation (6.8b) reasonable in $L^p(0,T;\mathcal{Q}^*)$, we need a Nemytskii map of the form

$$\mathcal{B} \colon L^p(0,T;\mathcal{V}) \to L^p(0,T;\mathcal{Q}^*).$$

A sufficient condition is given in Theorem 4.1 which is satisfied if, e.g., $\mathcal{B}$ is uniformly Lipschitz continuous.

6.2.2. *Regularization.* The regularization of the operator DAE (6.8) requires, as in the linear case of Section 6.1.2, the derivative of the constraint. With the Fréchet derivative of $\mathcal{B}$, namely $\mathcal{C}_u$, the derivative of (6.8b) has the form

$$\dot{\mathcal{G}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\big(\mathcal{B}u(t)\big) = \frac{\partial \mathcal{B}}{\partial u}(u)\dot{u}(t) = \mathcal{C}_u\dot{u}(t).$$

With the linearity of the Fréchet derivative [**Zei86**, Ch. 4.2] and the decomposition $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2$ from Assumption 6.16, we may also write, due to Remark 6.17,

$$\mathcal{C}_u\dot{u} = \mathcal{C}_u\dot{u}_1 + \mathcal{C}_{2,u}\dot{u}_2 = \dot{\mathcal{G}}(t).$$

Note that this calls for additional regularity of the form $\dot{u}_1$, $\dot{u}_2 \in L^p(0,T;\mathcal{V})$. Assuming this regularity, we actually obtain by Lemma 3.47 that $\dot{u}_2$ takes values in $\mathcal{V}_2$. Thus, we may define $v_2 := \dot{u}_2 \in L^p(0,T;\mathcal{V}_2)$.

With the additional (hidden) constraint and the new variable v_2, the extended operator DAE reads as follows: Find $u_1 \in W^{1;p}(0,T;\mathcal{V}_1)$, $u_2, v_2 \in L^p(0,T;\mathcal{V}_2)$, and $\lambda \in L^{p'}(0,T;\mathcal{Q})$ such that

$$\dot{u}_1(t) + v_2(t) + \mathcal{K}\big(u_1(t)+u_2(t)\big) + \mathcal{C}_u^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{6.9a}$$

$$\mathcal{B}\big(u_1(t)+u_2(t)\big) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*, \tag{6.9b}$$

$$\mathcal{C}_u\dot{u}_1(t) + \mathcal{C}_{2,u}v_2(t) = \dot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^* \tag{6.9c}$$

for a.e. $t \in [0,T]$ with the nonlinear initial condition

$$u_1(0) = g - \eta(u_1(0)) \in \mathcal{V}_1. \tag{6.9d}$$

Note that the assumed regularity $u_1 \in W^{1;p}(0,T;\mathcal{V}_1) \hookrightarrow C([0,T],\mathcal{V}_1)$ calls for initial data in $\mathcal{V}_1$ instead of $\mathcal{H}$ in (6.8c). In the example of Section 6.3.3 below, the term $\mathcal{C}_u\dot{u}_1(t)$ vanishes and it is sufficient to consider $u_1 \in W^{1;p,q}(0,T;\mathcal{V}_1,\mathcal{V}^*)$ and thus, $u_1(0) \in \mathcal{H}$.

Remark 6.20 (Regularity). If the operator $\mathcal{B}$ can be defined in a weaker sense, e.g., for functions in $\mathcal{H}$, then the regularity assumption on $\dot{u}_1$ may be weakened. For this case, equation (6.9c) has to be understood in a weaker topology. As illustrative example - although it is linear - consider the divergence operator $\operatorname{div}\colon [H_0^1(\Omega)]^d \to L_0^2(\Omega)$ which is also bounded as operator $\operatorname{div}\colon [L^2(\Omega)]^d \to H^{-1}(\Omega)$, cf. [**Hei14**, Sect. 3.2].

It remains to compare the extended system (6.9) with the original operator DAE (6.8). We show that the the two systems are basically equivalent. Nevertheless, one has to be aware of the additional smoothness assumptions on u_1.

THEOREM 6.21 (Equivalence of reformulation [**AH14**]). *Consider $\mathcal{F} \in L^q(0,T;\mathcal{V}^*)$, $\mathcal{G} \in W^{1;p}(0,T;\mathcal{Q}^*)$, $g \in \mathcal{V}$, and let $\mathcal{B}$ satisfy Assumption 6.16 for all $u \in L^p(0,T;\mathcal{V})$ that satisfy $\mathcal{B}u = \mathcal{G}$ in $\mathcal{Q}^*$. Then, there exists a solution $(u,\lambda) \in L^p(0,T;\mathcal{V}) \times L^{p'}(0,T;\mathcal{Q})$ of (6.8) with additional smoothness $\dot{u} \in L^p(0,T;\mathcal{V})$ if and only if (6.9) has a solution (u_1,u_2,v_2,λ) with $u_1 \in W^{1;p}(0,T;\mathcal{V}_1)$, $u_2, v_2 \in L^p(0,T;\mathcal{V}_2)$, and $\lambda \in L^{p'}(0,T;\mathcal{Q})$. Furthermore, we obtain the relations $u = u_1 + u_2$ and $\dot{u}_2 = v_2$.*

PROOF. Let (u,λ) be a solution of system (6.8) with initial condition $u(0) = g$. Assumption 6.16, Lemma 3.47, and the additional regularity of u allow for a decomposition $u = u_1 + u_2$ with $u_1 \in W^{1;p}(0,T;\mathcal{V}_1)$ and $u_2 \in W^{1;p}(0,T;\mathcal{V}_2)$. Then, the construction of the quadruple (u_1,u_2,v_2,λ) from this subsection with $v_2 := \dot{u}_2$ shows that it satisfies equations (6.9a)-(6.9c). Furthermore, we calculate $u_1(0)+\eta(u_1(0)) = u_1(0)+u_2(0) = u(0) = g$, which is the initial condition in (6.9d).

On the other hand, if (u_1,u_2,v_2,λ) is a solution of (6.9), we first define $u := u_1 + u_2 \in L^p(0,T;\mathcal{V})$. Because of (6.9b), Assumption 6.16 is satisfied for this function u. From the construction of u_1, u_2 in this subsection we see that $u = u_1+u_2$ is exactly the decomposition given by point (a) of Assumption 6.16. It remains to show that u_2 is time differentiable in the generalized sense and $v_2 = \dot{u}_2$. This then implies u_1, $u_2 \in W^{1;p}(0,T;\mathcal{V})$ and thus, the pair (u,λ) solves system (6.8).

By the implicit function theorem for operators [**Ruž04**, Ch. 2.2], we may locally write $u_2 = \eta(u_1)$. Since η is differentiable in time by Lemma 6.19, we obtain $\dot{u}_2 \in \mathcal{V}_2$ for a.e. $t \in [0,T]$. Equation (6.9c) and the time derivative of (6.9b) yield

$$\mathcal{C}_u\dot{u}_1 + \mathcal{C}_{2,u}\dot{u}_2 = \mathcal{C}_u\dot{u}_1 + \mathcal{C}_{2,u}v_2.$$

Since $\dot{u}_2$, $v_2 \in \mathcal{V}_2$, part (c) of Assumption 6.16 implies $\dot{u}_2 = v_2 \in L^p(0,T;\mathcal{V}_2)$. □

Theorem 6.21 provides a necessary condition on the initial data in (6.8c), cf. the end of Section 6.1.2. Note that the additional regularity $\dot{u} \in L^p(0,T;\mathcal{V})$ requires initial data $g \in \mathcal{V}$.

6.2.3. *Influence of Perturbations.* As for linear constraints in Section 6.1.3 we analyse the effect of perturbations in the right-hand sides on the solution behavior. We again restrict the analysis to linear, symmetric, coercive in $\mathcal{V}_1$, and bounded operators $\mathcal{K}$ and take Assumption 6.16 as given. Furthermore, we assume the inverse of $\mathcal{C}_{2,u}$ to be uniformly bounded.

Let (u_1,u_2,v_2,λ) denote the solution of the regularized system (6.9). As before, we consider perturbations $\delta \in L^2(0,T;\mathcal{V}^*)$ and θ, $\xi \in L^2(0,T;\mathcal{Q}^*)$ which yield a perturbed solution $(\hat{u}_1,\hat{u}_2,\hat{v}_2,\hat{\lambda})$. The errors

$$e_1 := \hat{u}_1 - u_1, \qquad e_2 := \hat{u}_2 - u_2, \qquad e_v := \hat{v}_2 - v_2, \qquad e_\lambda := \hat{\lambda} - \lambda$$

satisfy the system

$$\dot{e}_1 + e_v + \mathcal{K}\big(e_1+e_2\big) + \mathcal{C}_u^* e_\lambda = \delta \quad \text{in } \mathcal{V}^*, \tag{6.10a}$$

$$\mathcal{B}\big(\hat{u}_1+\hat{u}_2\big) - \mathcal{B}\big(u_1+u_2\big) = \theta \quad \text{in } \mathcal{Q}^*, \tag{6.10b}$$

$$\mathcal{C}_u\dot{e}_1 + \mathcal{C}_{2,u}e_v = \xi \quad \text{in } \mathcal{Q}^*. \tag{6.10c}$$

The initial error is denoted by $e_{1,0} := \hat{u}_1(0) - u_1(0)$. For the following computations we have to assume that $\hat{u}_1$ and $\hat{u}_2$ are 'close enough' to u_1 and u_2, respectively. By this we mean that there exist neighborhoods $\mathcal{U}_1 \subset \mathcal{V}_1$ and $\mathcal{U}_2 \subset \mathcal{V}_2$ with u_1, $\hat{u}_1 \in \mathcal{U}_1$ and u_2, $\hat{u}_2 \in \mathcal{U}_2$ such that the corresponding implicit functions $\eta\colon \mathcal{U}_1 \to \mathcal{U}_2$ and $\hat{\eta}\colon \mathcal{U}_1 \to \mathcal{U}_2$, cf.

[**Ruž04**, Ch. 2.2], satisfy

$$\mathcal{B}\big(u_1+\eta(u_1)\big)=\mathcal{B}\big(\hat{u}_1+\eta(\hat{u}_1)\big)=\mathcal{G},\quad \mathcal{B}\big(u_1+\hat{\eta}(u_1)\big)=\mathcal{B}\big(\hat{u}_1+\hat{\eta}(\hat{u}_1)\big)=\mathcal{G}+\theta. \tag{6.11}$$

Furthermore, we assume that the space $\mathcal{V}_1$ from Assumption 6.16 equals the kernel of $\mathcal{C}_u$, cf. Lemma 3.48. Exactly as in the linear case, we may test equation (6.10a) by e_1 which results in the estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}|e_1|^2+k_1\|e_1\|^2\leq\frac{3}{k_1}\Big(\|\delta\|_{\mathcal{V}^*}^2+C_{\mathrm{emb}}^4\|e_v\|^2+k_2^2\|e_2\|^2\Big). \tag{6.12}$$

Thus, it remains to find estimates of the errors e_2 and e_v. Since $\mathcal{C}_{2,u}$ is a homeomorphism and its inverse is assumed to be uniformly bounded, there exists a positive constant $c\in\mathbb{R}$ with

$$\|e_v\|\leq c\,\|\mathcal{C}_{2,u}e_v\|_{\mathcal{Q}^*}=c\,\|\mathcal{C}_u\dot{e}_1+\mathcal{C}_{2,u}e_v\|_{\mathcal{Q}^*}=c\,\|\xi\|_{\mathcal{Q}^*}.$$

By equation (6.11) and the definition of the Fréchet derivative, we obtain

$$\begin{aligned}\theta&=\mathcal{B}\big(\hat{u}_1+\hat{\eta}(\hat{u}_1)\big)-\mathcal{B}\big(u_1+\eta(u_1)\big)\\&=\mathcal{C}_u\big(\hat{u}_1-u_1+\hat{u}_2-u_2\big)+o(\|e\|)=\mathcal{C}_{2,u}\big(e_2\big)+o(\|e\|).\end{aligned}$$

Therein, we have used the abbreviation $e:=e_1+e_2$. Note that we benefit from the fact that $e_1\in\mathcal{V}_1=\ker\mathcal{C}_u$ and $e_2\in\mathcal{V}_2$. Furthermore, Remark 6.18 ensures that the spaces $\mathcal{V}_1$ and $\mathcal{V}_2$ are independent of the perturbation θ. Thus, we obtain up to a term of order $o(\|e\|)$ the estimate

$$\|e_2\|=\|\hat{u}_2-u_2\|=\|\hat{\eta}(\hat{u}_1)-\eta(u_1)\|\leq c\,\|\mathcal{C}_{2,u}\big(\hat{\eta}(\hat{u}_1)-\eta(u_1)\big)\|_{\mathcal{Q}^*}\approx\|\theta\|_{\mathcal{Q}^*}.$$

Integrating equation (6.12) over the interval $[0,t]$ for $t\leq T$ and using the gained estimates of this subsection, we obtain the following result.

THEOREM 6.22. *Consider the solution (u_1,u_2,v_2,λ) of the regularized system (6.9) with a linear, coercive, and bounded operator $\mathcal{K}$. Let $\mathcal{B}$ satisfy Assumption 6.16 and let $\delta\in L^2(0,T;\mathcal{V}^*)$ and θ, $\xi\in L^2(0,T;\mathcal{Q}^*)$ denote perturbations of the right-hand side. With $e_1:=\hat{u}_1-u_1$ and $e_2:=\hat{u}_2-u_2$, the solution of the perturbed problem $(\hat{u}_1,\hat{u}_2,\hat{v}_2,\hat{\lambda})$ satisfies, up to a term of order $o(\|e_1+e_2\|^2_{L^2(0,T;\mathcal{V})})$, the estimate*

$$\|e_1\|^2_{C([0,T];\mathcal{H})}+k_1\|e_1\|^2_{L^2(0,T;\mathcal{V})}\leq|e_{1,0}|^2+c\Big[\|\delta\|^2_{L^2(0,T;\mathcal{V}^*)}+\|\theta\|^2_{L^2(0,T;\mathcal{Q}^*)}+\|\xi\|^2_{L^2(0,T;\mathcal{Q}^*)}\Big].$$

REMARK 6.23. Results for the original system (6.8) are again obtained by setting $\xi=\dot{\theta}$. Thus, the error in u_1 is very sensitive to perturbations of the right-hand side $\mathcal{G}$.

6.3. Applications. In the previous subsections, we have provided a framework to regularize semi-explicit operator DAEs with linear as well as nonlinear constraints. We close the section on first-order problems with three applications which fit into the given framework and satisfy the postulated assumptions.

First, we consider the *Navier-Stokes equations* (or any linearized version) which contain a constraint on the divergence of the velocity. This gives an example for a linear constraint. As second example, we show that there are applications for which the divergence constraint is not homogeneous, i.e., we have a constraint of the form $\nabla\cdot u=\mathcal{G}\neq 0$. For this, we consider an *optimal control problem*, constrained by the Navier-Stokes equations, where the pressure appears in the cost functional. Finally, we consider the *Stefan problem* in a regularized version. This two phase flow example is constrained at the boundary in a nonlinear manner and provides an application of the framework of Section 6.2.

6.3.1. *Navier-Stokes Equations.* Consider a domain $\Omega \subset \mathbb{R}^d$ with sufficiently smooth boundary. The evolution of a velocity field $u(t)\colon \Omega \to \mathbb{R}^d$ and the pressure $p(t)\colon \Omega \to \mathbb{R}$ of an incompressible flow of a Newtonian fluid is described by the Navier-Stokes equations [**Tem77**]. Given initial data u_0, a volume force β, and a viscosity constant ν involving the Reynolds number, the equations of motion read

$$\begin{aligned} \dot{u} + (u \cdot \nabla)u - \nu\Delta u + \nabla p &= \beta && \text{in } \Omega \times [0,T], \\ \operatorname{div} u &= 0 && \text{in } \Omega \times [0,T], \\ u &= 0 && \text{on } \partial\Omega \times [0,T], \\ u(\cdot,0) &= u_0. \end{aligned}$$

For the weak formulation, the commonly used spaces are $\mathcal{V} := [H_0^1(\Omega)]^d$, $\mathcal{H} := [L^2(\Omega)]^d$, and $\mathcal{Q} := L^2(\Omega)/\mathbb{R}$. This already includes the homogeneous Dirichlet boundary conditions. Obviously, the dynamics are constrained by the incompressibility. Thus, the constraint operator has the form $\mathcal{B} = \operatorname{div}\colon \mathcal{V} \to \mathcal{Q}^*$ and the system can be written in the weak form as

$$\dot{u}(t) + \mathcal{K}u(t) - \mathcal{B}^* p(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{6.13a}$$

$$\mathcal{B}u(t) = 0 \quad \text{in } \mathcal{Q}^* \tag{6.13b}$$

with initial condition $u(0) = u_0 \in \mathcal{H}$. Therein, the operator $\mathcal{K}\colon \mathcal{V} \to \mathcal{V}^*$ is defined via

$$\langle \mathcal{K}u, v \rangle := \int_\Omega (u \cdot \nabla)u \cdot v \,\mathrm{d}x + \nu \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x.$$

Note that linearizations of the Navier-Stokes equations such as the *Stokes* or *Oseen* equations lead to the same structure. In the Stokes equations, $\mathcal{K}u$ corresponds to $-\nu\Delta u$ whereas the Oseen equations include the term $(u_\infty \cdot \nabla)u - \nu\Delta u$ with a given characteristic velocity u_∞. In this case, the unknown u describes the 'disturbance velocity'. Also the *Euler equations* may be modeled in the form of system (6.13).

In the nonlinear case and without any further assumptions, the operator $\mathcal{K}(t)$ only extends to $\mathcal{K}\colon L^2(0,T;\mathcal{V}) \to L^1(0,T;\mathcal{V}^*)$, cf. [**Tem77**, Lem. III.3.1]. This loss of regularity causes the main difficulties in the existence theory for the Navier-Stokes equations. For further remarks and results on the existence of a (unique) solution to the Navier-Stokes equations, we refer to [**Tem77**, Ch. III], [**Tar06**, Ch. 25], and [**HR90**]. Nevertheless, this has no influence on the regularization process presented in this section.

The operator DAE (6.13) has the structure of system (6.2) when the pressure p is interpreted as Lagrange multiplier to couple the incompressibility to the state equations. We show that Assumption 6.2 is satisfied. Clearly, $\mathcal{B}$ is linear and bounded. Furthermore, the space $\mathcal{V}$ allows the Helmholtz decomposition into divergence-free functions $\mathcal{V}_\mathcal{B} = \{v \in \mathcal{V} \mid \operatorname{div} v = 0\}$ and its orthogonal complement $\mathcal{V}^{\mathrm{c}} := \mathcal{V}_\mathcal{B}^{\perp_\mathcal{V}}$. Then, the divergence operator restricted to $\mathcal{V}^{\mathrm{c}}$ yields an isomorphism such that there exists a continuous right inverse of $\mathcal{B}$ [**Tar06**, Lem. I.4.1].

With regard to Remark 6.9 we emphasize that the closure of $\mathcal{V}_\mathcal{B}$ in $\mathcal{H}$ is a proper subspace of $\mathcal{H}$. Consequently, the initial data of the dynamic part of u (the divergence-free part) cannot be chosen arbitrarily in $\mathcal{H}$, see also Example 6.10.

The benefits of the possible regularization of (6.13) are presented in Section 8.4. Therein, we also discuss the implementation of the regularized equations in practical simulations. Furthermore, we analyse the influence of the regularization with regard to the stability of the Rothe method in Section 10. Note that the framework of the previous section also allows an inhomogeneity in the constraint (6.13b). Such an inhomogeneity

may appear within optimal control problems within the dual equations, as discussed in the following subsection.

6.3.2. *Optimal Control of Fluid Flows.* As a second example, we consider an optimal control problem where the minimization is constrained by the Navier-Stokes equations. For an introduction to PDE constrained optimization, we refer to [**Trö09**]. Minimization problems have a large variety of applications in the field of fluid dynamics such as finding optimal inputs in order to decrease the drag or increase the mixture of two fluids [**Hin00**]. Many numerical methods for the Navier-Stokes equations are tailored for the particular case of a vanishing divergence. Also in the analytical setting one often works in the space of divergence-free functions. The here presented regularization is not affected by an inhomogeneity in the constraint.

The following example presents an application where one is interested in solving a Navier-Stokes system with an inhomogeneous constraint on the divergence. Thus, we discuss an example of the form (6.2) with $\mathcal{G} \neq 0$. Since the control variable is traditionally denoted by u, we denote the velocity in this example by y. Then, the optimal control problem has the form: Find an input u which minimizes the cost functional

$$J(y, p, u) := \alpha \int_0^T \int_\Omega V(y, \nabla y, p) \, \mathrm{d}x \, \mathrm{d}t$$

subject to

$$\begin{aligned} \dot{y} - \nu \nabla y + (y \cdot \nabla) y + \nabla p &= \beta B u && \text{in } \Omega \times [0, T], \\ -\nabla \cdot y &= 0 && \text{in } \Omega \times [0, T] \end{aligned}$$

with homogeneous Dirichlet boundary condition $y = 0$ on $\partial\Omega$ and initial condition $y(0) = y_0$. The parameters α and β are assumed to be real and positive and B is a control extension operator. Following the general approach in [**Hin00**, App. A], we obtain the corresponding optimality system with dual variables z and q,

$$\dot{y} - \nu \nabla y + (y \cdot \nabla) y + \nabla p = \beta B u \qquad \text{in } \Omega \times [0, T], \tag{6.14a}$$

$$-\nabla \cdot y = 0 \qquad \text{in } \Omega \times [0, T], \tag{6.14b}$$

$$-\dot{z} - \nu \nabla z - (y \cdot \nabla) z + (\nabla y)^T z + \nabla q = -\alpha (V_1 - \nabla \cdot V_2) \qquad \text{in } \Omega \times [0, T], \tag{6.14c}$$

$$-\nabla \cdot z = -\alpha V_3 \qquad \text{in } \Omega \times [0, T]. \tag{6.14d}$$

Therein, V_i denotes the i-th partial derivative of $V(y, \nabla y, p)$ in the cost functional. In addition, we have the boundary conditions $y = z = 0$ on $\partial\Omega$ and the initial conditions $y(0) = y_0$ and $z(T) = 0$. Equations (6.14c) and (6.14d) then form a system of linear Navier-Stokes type for the dual variables z and q. However, the constraint is given by $-\nabla \cdot z = -\alpha V_3 \neq 0$, i.e., $\mathcal{G} \neq 0$ in the abstract setting, if the cost functional depends on the pressure.

6.3.3. *Regularized Stefan Problem.* We close this section with an example of an operator DAE with a nonlinear constraint which fits the framework of Section 6.2. For this, we consider the governing equations of a change of phase (e.g. water and ice), the so-called *Stefan problem* [**Fri68, And04**]. This problem includes a free boundary at the transition of the two phases. However, since we deal with the weak formulation of the problem, the explicit condition at the free boundary vanishes.

For the formulation as operator DAE, we use the enthalpy formulation as stated in [**DPVY13**]. In contrast to the temperature, the enthalpy jumps at the free boundary where the phase changes and contains all the information on the state of the material. More precisely, we consider the regularized version with an enthalpy-temperature function

β which satisfies the following conditions. First, we assume that $\beta\colon \mathbb{R} \to \mathbb{R}$ is strictly monotonically increasing and continuously differentiable with $\beta' \geq \varepsilon > 0$. Second, we state as in [**DPVY13**] that there exist constants c, $C > 0$ such that $\operatorname{sign}(s)\beta(s) \geq c|s| - C$. For the formulation as operator DAE we additionally assume that the derivative β' is Lipschitz continuous. Furthermore, we have to assume for the solution u of the regularized Stefan problem that $1/\beta'(\gamma u) \in \mathcal{Q}^*$ for a.e. $t \in [0,T]$. Recall that γ denotes the trace operator from Section 3.1.4. A sufficient condition would be the Lipschitz continuity of the inverse of β'.

The governing equations of the Stefan problem then have the form: Find the enthalpy $u\colon [0,T] \to \mathcal{V}$ which satisfies the system

$$\dot{u} - \nabla \cdot \big(\nabla \beta(u)\big) = f \qquad \text{in } \Omega \times [0,T], \tag{6.15a}$$

$$\beta(u) = g \qquad \text{on } \partial\Omega \times [0,T] \tag{6.15b}$$

with initial condition

$$u(0) = u_0. \tag{6.15c}$$

For the operator formulation, we pass to the weak formulation in which we search for $u \in W^{1;2,2}(0,T;\mathcal{V},\mathcal{V}^*)$. Furthermore, we regard the nonlinear boundary condition (6.15b) as constraint on u, which we enforce weakly by the Lagrangian method.

Remark 6.24. The one-to-one property of β implies the existence of an inverse such that equation (6.15b) can be written in the form $u = \beta^{-1}(g)$. In this case, the boundary constraint is linear. Nevertheless, assuming that β^{-1} is not given in explicit form, we would rely on some kind of Newton method for the boundary values. In the finite-dimensional case, i.e., in the case of DAEs, it is well-known that small perturbations in the constraints may lead to crucial instabilities, see the discussion in [**Arn98b**, Ch. 2.1]. Because of this, convergence proofs normally assume that the error in the constraint (and thus, the Newton error) is maximal of size $O(\tau^2)$ where τ equals the time step size. Because of this, we prefer to handle the nonlinear boundary condition (6.15b) with the help of a Lagrange multiplier.

For the formulation of system (6.15) as operator DAE, we define the spaces

$$\mathcal{V} := H^1(\Omega), \qquad \mathcal{V}_{\mathcal{B}} := H_0^1(\Omega), \qquad \mathcal{V}^{\mathrm{c}} := [H_0^1(\Omega)]^{\perp_{\mathcal{V}}}, \qquad \mathcal{Q}^* := H^{1/2}(\partial\Omega).$$

The constraint operator $\mathcal{B}\colon \mathcal{V} \to \mathcal{Q}^*$ has the form $\mathcal{B}u := \beta(\gamma u)$, i.e., for $q \in L^2(\partial\Omega)$ we have

$$\big\langle \mathcal{B}u, q \big\rangle_{\mathcal{Q}^*,\mathcal{Q}} = \int_{\partial\Omega} \beta(u) q \,\mathrm{d}x.$$

Its Fréchet derivative at some $\bar{u} \in \mathcal{V}$ is given by the linear map

$$\mathcal{C}_{\bar{u}} := \frac{\partial \mathcal{B}}{\partial u}(\bar{u})\colon \mathcal{V} \to \mathcal{Q}^*, \qquad v \mapsto \beta'(\gamma\bar{u}) \cdot \gamma v.$$

Note that this is well-defined, since β' is assumed to be Lipschitz continuous and thus, $\beta'(\gamma\bar{u}) \in \mathcal{Q}^*$. With the operator $\mathcal{B}$, the constraint (6.15b) has the weak form $\mathcal{B}u = \mathcal{G}$ in $\mathcal{Q}^*$ with right-hand side $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$ densely defined by

$$\big\langle \mathcal{G}(t), q \big\rangle_{\mathcal{Q}^*,\mathcal{Q}} := \int_{\partial\Omega} g(t) q \,\mathrm{d}x$$

for all $q \in L^2(\partial\Omega)$ and a.e. $t \in [0,T]$. Note that this requires a certain regularity of the given right-hand side g. This notion allows to formulate system (6.15) as operator DAE,

$$\dot{u} + \mathcal{K}u + \mathcal{C}_u^*\lambda = \mathcal{F} \quad \text{in } \mathcal{V}^*, \tag{6.16a}$$

$$\mathcal{B}u = \mathcal{G} \quad \text{in } \mathcal{Q}^* \tag{6.16b}$$

with initial condition as in (6.15c).

REMARK 6.25. For sufficiently regular data and a consistent initial value, system (6.15) has a unique weak solution $u \in W^{1;2,2}(0,T;\mathcal{V},\mathcal{V}_\mathcal{B}^*)$, see [**DPVY13**]. The properties of the constraint operator $\mathcal{B}$ and its Fréchet derivative then imply the unique solvability of the corresponding operator DAE (6.16) for $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$. The solution satisfies $u \in W^{1;2,2}(0,T;\mathcal{V},\mathcal{V}^*)$ and $\lambda \in L^2(0,T;\mathcal{Q})$.

We claim that system (6.16) fits the framework of Section 6.2 for nonlinear constraints. For this, we show that Assumption 6.16 is satisfied. The splitting of $\mathcal{V}$ is given by the trace-free functions $\mathcal{V}_1 := \mathcal{V}_\mathcal{B}$ and its orthogonal complement $\mathcal{V}_2 := \mathcal{V}^{\mathrm{c}}$. Furthermore, the Fréchet derivative $\mathcal{C}_u$ exists for all $u \in \mathcal{V}$ and is also continuous in u, since the trace operator is continuous and β is continuously differentiable. It remains to check whether the Fréchet derivative

$$\mathcal{C}_{2,\bar{u}} := \frac{\partial \mathcal{B}}{\partial u_2}(\bar{u})\colon \mathcal{V}^{\mathrm{c}} \to \mathcal{Q}^*, \qquad v \mapsto \beta'(\gamma\bar{u}) \cdot \gamma v,$$

gives a homeomorphism along the solution of the Stefan problem. For this, the key property is the fact that the trace operator is a homeomorphism as mapping from $\mathcal{V}^{\mathrm{c}}$ to $\mathcal{Q}^*$, see Lemma 7.1 below. We denote the continuity constant by C_{tr} and estimate

$$\|\mathcal{C}_{2,u}v\|_{\mathcal{Q}^*} \le \|\beta'(\gamma u)\|_{\mathcal{Q}^*} C_{\mathrm{tr}} \|v\|.$$

Note that $\|\beta'(\gamma u)\|_{\mathcal{Q}^*}$ is finite for $u \in \mathcal{V}$ because of the Lipschitz continuity of β'. Finally, the boundedness of the inverse operator follows from the inverse trace theorem [**Ste08**, Th. 2.22] and the assumption that $1/\beta'(\gamma u) \in \mathcal{Q}^*$. In summary, we have shown that the theory and regularization procedure of Section 6.2 is applicable to the regularized Stefan problem (6.15) if we formulate the boundary condition as nonlinear constraint.

7. Regularization of Second-order Operator DAEs

This section is devoted to the regularization of second-order operator DAEs. Again we consider dynamical systems of semi-explicit structure but here, the equations are of index-3 type. By this we mean that a suitable semi-discretization leads to DAEs of index 3 as shown in Section 9 below.

Nevertheless, we can apply similar techniques and follow the same ideas as in Section 6 with regard to the inclusion of dummy variables. Note that we do not consider the general case here. Instead we focus on a particular application from elastodynamics where the constraints are given on the boundary, i.e., the constraint operator equals the trace operator γ from Section 3.1.4. Such systems arise in the field of flexible multibody dynamics [**Sha97, GC01, Bau10, Sim13**], where deformable bodies are coupled as in the theory of multibody systems [**RS88**]. Common examples are given by the slider crank mechanism [**Sim96**] or the pantograph and catenary system [**AS00**].

Note that the considered approach also fits to a more general coupling of a flexible body with any other dynamical system as long as the coupling is modeled by Dirichlet boundary conditions. Thus, the presented framework may be used to model multi-physics applications, i.e., the coupling of systems including different kinds of physics such as chemical reactions, fluid flows, or electromagnetics. A simple model of a flexible body coupled with a mass-spring-damper system is given in [**Alt13b**].

The results of this section are published within Sections 2 and 4 of [**Alt13a**]. The therein used notation was adapted to the notation presented in Part A.

7.1. Equations of Motion in Elastodynamics. Before we formulate the equations of motion in the abstract setting as operator DAE, we review the governing equations for elastic media. Within this section, $\Omega \subset \mathbb{R}^d$ denotes a domain with Lipschitz boundary, cf. Section 3. Furthermore, $\Gamma_{\mathrm{D}} \subseteq \partial\Omega$ denotes the Dirichlet boundary and $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}$ the Neumann boundary. Note that we do not consider the pure Neumann problem, i.e., $\Gamma_{\mathrm{N}} = \partial\Omega$, since this would exclude the considered coupling throughout the boundary. In this case, the coupling takes place through the velocities which requires a different model.

7.1.1. *Principle of Virtual Work.* The equations of elastodynamics describe the evolution of a deformable body under the influence of applied forces based on Cauchy's theorem [**Cia88**, Ch. 2]. We consider the theory of linear elasticity for homogeneous and isotropic materials, i.e., we assume small deformations only. Note that the regularization performed in Section 7.2 can also be applied to the nonlinear case. However, the existence and uniqueness results of this section as well as the analysis performed in Section 11 is restricted to the linear case.

The deformation of the domain Ω is described by the time-dependent displacement field $u(t)\colon \Omega \to \mathbb{R}^d$ with $t \in [0,T]$. We define the linearized strain tensor $\varepsilon(u) \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ by

$$\varepsilon(u) := \frac{1}{2}\big[\nabla u + (\nabla u)^T\big].$$

Furthermore, Hooke's law [**Sad10**, Ch. 4.2] states that the stress tensor $\sigma(u) \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ depends linearly on the strain tensor and the material constants λ and μ, the so-called *Lamé parameters*,

$$\sigma(u) := \lambda \operatorname{trace} \varepsilon(u) I_d + 2\mu\varepsilon(u).$$

Therein, I_d denotes the $d \times d$ identity matrix. Let $\rho \in \mathbb{R}_{>0}$ denote the constant density of the material and n the outer normal vector along the boundary. Then, the corresponding

initial-boundary value problem in strong form with prescribed Dirichlet data u_D and applied forces β and τ reads

$$\rho\ddot{u} - \operatorname{div}(\sigma(u)) = \beta \qquad \text{in } \Omega, \tag{7.1a}$$

$$u = u_D \qquad \text{on } \Gamma_{\mathrm{D}} \subseteq \partial\Omega, \tag{7.1b}$$

$$\sigma(u) \cdot n = \tau \qquad \text{on } \Gamma_{\mathrm{N}} = \partial\Omega \setminus \Gamma_{\mathrm{D}}. \tag{7.1c}$$

with initial conditions

$$u(0) = g, \qquad \dot{u}(0) = h. \tag{7.1d}$$

For the formulation in operator form, we need the weak formulation which is obtained by the introduction of test functions and integration by parts. This then leads to the so-called *principle of virtual work* (in the reference configuration). With the notation of Section 3.1.3 for Sobolev spaces, we set

$$\mathcal{V} := \big[H^1(\Omega)\big]^d, \quad \mathcal{V}_{\mathcal{B}} := \big[H^1_{\Gamma_{\mathrm{D}}}(\Omega)\big]^d, \quad \mathcal{H} := \big[L^2(\Omega)\big]^d.$$

We denote the norms in $\mathcal{V}$ and $\mathcal{H}$ by $\|\cdot\| := \|\cdot\|_{\mathcal{V}}$ and $|\cdot| := \|\cdot\|_{\mathcal{H}}$. With the inner product for matrices, $A : B := \operatorname{trace}(AB^T) = \sum_{i,j} A_{ij}B_{ij}$, we define the symmetric bilinear form

$$a(u, v) := \int_\Omega \sigma(u) : \varepsilon(v) \,\mathrm{d}x.$$

This then defines a linear operator $\mathcal{K}\colon \mathcal{V} \to \mathcal{V}^*$, given by

$$\langle \mathcal{K}u, v\rangle_{\mathcal{V}^*,\mathcal{V}} := a(u, v). \tag{7.2}$$

By Korn's inequality [**BS08**, Ch. 11.2], a is a coercive and bounded bilinear form on $\mathcal{V}_{\mathcal{B}}$ if Γ_{D} has positive measure, i.e., if we are not in the pure Neumann case. Since $\varepsilon(\cdot)$ vanishes for constant functions, a cannot be coercive on the entire space $\mathcal{V}$. Obviously, the operator $\mathcal{K}$ inherits the properties of the bilinear form a.

The weak form (in terms of the space variable) of system (7.1) with homogeneous Dirichlet boundary conditions $u_D = 0$ has the form: Find $u(t) \in \mathcal{V}_{\mathcal{B}}$ such that for all $t \in [0, T]$ it holds that

$$(\rho\ddot{u}, v)_{L^2(\Omega)} + a(u, v) = (\beta, v)_{L^2(\Omega)} + (\tau, v)_{L^2(\Gamma_{\mathrm{N}})} \quad \text{for all } v \in \mathcal{V}_{\mathcal{B}}. \tag{7.3}$$

Note that the Dirichlet boundary condition on u is taken into account by the fact that we search $u(t)$ within the space $\mathcal{V}_{\mathcal{B}}$. Thus, we also assume the initial data to satisfy $g \in \mathcal{V}_{\mathcal{B}}$ and $h \in \mathcal{V}_{\mathcal{B}}$. The inclusion of inhomogeneous Dirichlet boundary conditions is part of Section 7.1.2.

Before discussing the inclusion of Dirichlet boundary conditions in detail, we shortly review the concept of *damping* or *dissipation*. In many applications one has to include dissipation such as friction to the mathematical model in order to obtain reasonable results. Often viscous damping [**Hug87**, Ch. 7.2] is considered which corresponds to a generalization of Hooke's law.

The popular generalization of the mass proportional and stiffness proportional damping is called *Rayleigh damping* [**CP03**, Ch. 12]. This concept combines frequency dependent and independent damping and is widespread in modeling internal structural damping. Let ζ_1, $\zeta_2 \geq 0$ be two real parameters. The first parameter ζ_1 regularizes the frequency dependent damping such that the stress tensor does not depend linearly on the strain tensor $\varepsilon(u)$ anymore. The linear proportionality of the damping and the response frequencies implies that the stiffness proportional damping acts stronger on the higher modes

of the structure. Unfortunately, this quite common approach has no physical justification [**Wil98**, Ch. 19]. The second parameter ζ_2 characterizes the frequency independent damping. In combination, the Rayleigh damping is given by the bilinear form

$$d(\dot{u}, v) := \zeta_1 a(\dot{u}, v) + \zeta_2 (\rho \dot{u}, v)_{L^2(\Omega)}.$$

Within this section we allow more general nonlinear damping terms. For this, consider a mapping $d\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ which is linear only in the second component. Based on this mapping, a nonlinear damping operator $\mathcal{D}\colon \mathcal{V} \to \mathcal{V}^*$ is defined by

$$\langle \mathcal{D}u, v \rangle_{\mathcal{V}^*,\mathcal{V}} := d(u, v). \tag{7.4}$$

Further assumptions will be given in Section 7.3 when the existence and uniqueness of solutions is analyzed.

7.1.2. *Dirichlet Boundary Conditions.* It remains to include the inhomogeneous Dirichlet boundary conditions which are prescribed on $\Gamma_{\mathrm{D}} \subseteq \partial\Omega$. Since we exclude the pure Neumann case, we assume that Γ_{D} has positive measure. Following the work of [**Sim00**], we incorporate the boundary conditions in a weak form, i.e., with the help of Lagrange multipliers. This then leads to a dynamic saddle point problem.

In text books on numerical analysis of PDEs, one often assumes homogeneous boundary data, since the inhomogeneities may be included in the right-hand side. However, this approach requires the construction of a function in $H^1(\Omega)$ with the given Dirichlet data. A second drawback arises if Γ_{D} is time-dependent, i.e., the position of the constraint changes with time. Then, also the ansatz space for the solution would be time-dependent.

According to Section 3.1.4 the Sobolev space $\mathcal{V}$ has a well-defined trace on Γ_{D}. We define the space $\mathcal{Q}$ by its dual space,

$$\mathcal{Q}^* := [H^{1/2}(\Gamma_{\mathrm{D}})]^d.$$

Recall that $\mathcal{Q}^*$ is a Hilbert space. Furthermore, the spaces $\mathcal{Q}^*$, $[L^2(\Gamma_{\mathrm{D}})]^d$, $\mathcal{Q}$ form a Gelfand triple such that the dual pairing $\langle \cdot, \cdot \rangle_{\mathcal{Q},\mathcal{Q}^*}$ is densely defined for $q \in [L^2(\Gamma_{\mathrm{D}})]^d$ and $\vartheta \in \mathcal{Q}^*$ by

$$\langle q, \vartheta \rangle_{\mathcal{Q},\mathcal{Q}^*} := \int_{\Gamma_{\mathrm{D}}} q \cdot \vartheta \, \mathrm{d}x. \tag{7.5}$$

Based on this duality pairing, we introduce the bilinear form $b\colon \mathcal{V} \times \mathcal{Q} \to \mathbb{R}$ by

$$b(u, q) := \langle q, u \rangle_{\mathcal{Q},\mathcal{Q}^*}. \tag{7.6}$$

We emphasize that this definition involves the trace operator γ from Section 3.1.4. Then, the classical form of the boundary condition $u(\cdot, t) = u_D(\cdot, t)$ on Γ_{D} for all $t \in [0, T]$ turns into

$$b(u(t), q) = \big\langle q, u_D(t) \big\rangle_{\mathcal{Q},\mathcal{Q}^*},$$

for all $q \in \mathcal{Q}$ and a.e. $t \in [0, T]$. A subtle but important property of b is the inf-sup condition, which is discussed within Lemma 7.1 below. Since b involves the boundary constraint, its analysis is a main part of the existence theory of solutions [**Bra07**, Ch. 3]. The bilinear form b defines the operator $\mathcal{B}\colon \mathcal{V} \to \mathcal{Q}^*$, given by

$$\langle q, \mathcal{B}u \rangle_{\mathcal{Q},\mathcal{Q}^*} := b(u, q). \tag{7.7}$$

According to Section 3.1.1, the dual operator $\mathcal{B}^*\colon \mathcal{Q} \to \mathcal{V}^*$ satisfies

$$\langle \mathcal{B}^* q, u \rangle_{\mathcal{V}^*,\mathcal{V}} := \langle q, \mathcal{B}u \rangle_{\mathcal{Q},\mathcal{Q}^*} = b(u, q). \tag{7.8}$$

We summarize important properties of these two operators (or rather the bilinear form b) in the following lemma. Therein, we consider the orthogonal decomposition $\mathcal{V} = \mathcal{V}_{\mathcal{B}} \oplus \mathcal{V}^{\mathrm{c}}$,

i.e., $\mathcal{V}^{\mathrm{c}} = (\mathcal{V}_{\mathcal{B}})^{\perp_{\mathcal{V}}}$. However, as for the systems of first order in Section 6 the orthogonal complement could be replaced by any other complement of $\mathcal{V}_{\mathcal{B}}$. Furthermore, we introduce the *polar set* or *annihilator* of $\mathcal{V}_{\mathcal{B}}$, namely

$$\mathcal{V}_{\mathcal{B}}^{\mathrm{o}} := \left\{ f \in \mathcal{V}^* \mid \langle f, v \rangle = 0 \text{ for all } v \in \mathcal{V}_{\mathcal{B}} \right\} \subset \mathcal{V}^*. \tag{7.9}$$

This set should not be mixed up with the dual $\mathcal{V}_{\mathcal{B}}^*$ which contains all linear and bounded functionals defined on $\mathcal{V}_{\mathcal{B}}$ und thus satisfies $\mathcal{V}^* \subset \mathcal{V}_{\mathcal{B}}^*$.

LEMMA 7.1 (Properties of $\mathcal{B}$ and $\mathcal{B}^*$ [**Alt13a**]). *Let Γ_D have positive measure and consider the decomposition $\mathcal{V} = \mathcal{V}_{\mathcal{B}} \oplus \mathcal{V}^c$. Then, the following assertions hold,*

(a) $\mathcal{B}$ vanishes on $\mathcal{V}_{\mathcal{B}}$,

(b) $\mathcal{B}$ satisfies the inf-sup condition, i.e., there exists a constant $\beta > 0$ with

$$\inf_{q \in \mathcal{Q}} \sup_{v \in \mathcal{V}} \frac{\langle \mathcal{B}v, q \rangle}{\|v\| \ \|q\|_{\mathcal{Q}}} = \beta > 0,$$

(c) $\mathcal{B}$ restricted to $\mathcal{V}^c$ is an isomorphism,

(d) $\mathcal{B}^* \colon \mathcal{Q} \to \mathcal{V}_{\mathcal{B}}^{\mathrm{o}}$ defines an isomorphism,

(e) $\beta \|v\| \leq \|\mathcal{B}v\|_{\mathcal{Q}^*}$ for all $v \in \mathcal{V}^c$,

(f) $\frac{d}{dt}(\mathcal{B}u) = \mathcal{B}\dot{u}$ for all $u \in H^1(0,T;\mathcal{V})$.

PROOF. (a) As for first-order systems, $\mathcal{V}_{\mathcal{B}}$ is expected to be the kernel of $\mathcal{B}$. We show this be a density argument. Consider an arbitrary $v \in \mathcal{V}_{\mathcal{B}}$ and $q \in [L^2(\Gamma_{\mathrm{D}})]^d$. Since v vanishes on Γ_{D},

$$\langle q, \mathcal{B}v \rangle_{\mathcal{Q},\mathcal{Q}^*} = b(v,q) = \int_{\Gamma_{\mathrm{D}}} v \cdot q \, \mathrm{d}x = 0.$$

For the general case consider $q \in \mathcal{Q}$. Since $[L^2(\Gamma_{\mathrm{D}})]^d$ is densely embedded in $\mathcal{Q}$, there exists a sequence $\{q_n\} \subset [L^2(\Gamma_{\mathrm{D}})]^d$ with $q_n \to q$ in $\mathcal{Q}$ as $n \to \infty$. Thus,

$$\langle q, \mathcal{B}v \rangle_{\mathcal{Q},\mathcal{Q}^*} = \lim_{n\to\infty} \langle q_n, \mathcal{B}v \rangle_{\mathcal{Q},\mathcal{Q}^*} = \lim_{n\to\infty} 0 = 0.$$

(b) The proof of the inf-sup condition can be found in [**Ste08**, Lemma 4.7] and is a consequence of Theorem 3.15.
(c) The assertion follows with the help of [**Bra07**, Ch. III, Th. 3.6]. It requires the continuity of b, which follows by the trace theorem [**Ste08**, Th. 2.21], the inf-sup condition from (b), and a non-degeneration condition of the form: for every $v \in \mathcal{V}^{\mathrm{c}}$, $v \neq 0$ there exists an element $q \in \mathcal{Q}$ such that $b(v,q) \neq 0$. To show the latter, assume there exists a $v \in \mathcal{V}^{\mathrm{c}}$ such that $b(v,q)$ vanishes for all $q \in \mathcal{Q}$. Thus, v has trace zero on Γ_{D} and hence $v \in \mathcal{V}_{\mathcal{B}} \cap \mathcal{V}^{\mathrm{c}} = \{0\}$.
(d) The claim that $\mathcal{B}^* \colon \mathcal{Q}^* \to \mathcal{V}_{\mathcal{B}}^{\mathrm{o}}$ is an isomorphism is equivalent to part (c), cf. [**Bra07**, Ch. III, Lem. 4.2].
(e) Since $\mathcal{B}^*$ is an isomorphism, b also fulfills an inf-sup condition of the form

$$\inf_{v \in \mathcal{V}^{\mathrm{c}}} \sup_{q \in \mathcal{Q}} \frac{b(v,q)}{\|v\| \ \|q\|_{\mathcal{Q}}} = \beta > 0.$$

Thus, for all $v \in \mathcal{V}^{\mathrm{c}}$ the following chain of inequalities holds,

$$\beta \|v\| \leq \sup_{q \in \mathcal{Q}} \frac{b(v,q)}{\|q\|_{\mathcal{Q}}} \leq \sup_{q \in \mathcal{Q}} \frac{\|\mathcal{B}v\|_{\mathcal{Q}^*} \|q\|_{\mathcal{Q}}}{\|q\|_{\mathcal{Q}}} = \|\mathcal{B}v\|_{\mathcal{Q}^*}.$$

(f) For every $q \in \mathcal{Q}$, the functional $\mathcal{B}^* q$ is bounded and thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\big\langle q, \mathcal{B}u(t)\big\rangle_{\mathcal{Q},\mathcal{Q}^*} = \frac{\mathrm{d}}{\mathrm{d}t}\big\langle \mathcal{B}^* q, u(t)\big\rangle_{\mathcal{V}^*,\mathcal{V}} = \big\langle \mathcal{B}^* q, \dot{u}(t)\big\rangle_{\mathcal{V}^*,\mathcal{V}} = \big\langle q, \mathcal{B}\dot{u}(t)\big\rangle_{\mathcal{Q},\mathcal{Q}^*}. \qquad \square$$

7.1.3. *Formulation as Operator DAE.* With the preparations from the previous two subsections, we are able to state system (7.1) with damping and inhomogeneous boundary conditions in a weak form. First, we consider the weak formulation only in terms of the space variable. For this, we introduce a Lagrange multiplier λ which has to be an element of the space $\mathcal{Q}$. The problem then reads: Determine $u(t) \in \mathcal{V}$ and $\lambda(t) \in \mathcal{Q}$ such that for all $t \in [0,T]$ and all test functions $v \in \mathcal{V}$ and $q \in \mathcal{Q}$ it holds that

$$\begin{aligned} (\rho\ddot{u}, v)_{L^2(\Omega)} + d(\dot{u}, v) + a(u, v) + b(v, \lambda) &= (\beta, v)_{L^2(\Omega)} + (\tau, v)_{L^2(\Gamma_{\mathrm{N}})}, \\ b(u, q) &= \langle q, u_D\rangle_{\mathcal{Q},\mathcal{Q}^*}. \end{aligned}$$

Remark 7.2. In contrast to the homogeneous case, where we search for $u \in \mathcal{V}_{\mathcal{B}}$, the Lagrange multiplier method extends the search space to $\mathcal{V}$. In view of this, the Sobolev space $\mathcal{V}$ is also used as test space in the first equation.

In the following, we derive the corresponding formulation with operators including time derivatives which are viewed in the generalized sense of Section 3.1.2. In addition, we have to discuss the Sobolev-Bochner spaces in which we search for solutions (u, λ). For the right-hand sides, we introduce two linear functionals. With the Dirichlet data u_D we define $\mathcal{G} \in L^2(0,T;\mathcal{Q}^*)$ by

$$\big\langle \mathcal{G}(t), q\big\rangle := \big\langle q, u_D(t)\big\rangle_{\mathcal{Q},\mathcal{Q}^*}. \tag{7.10}$$

As discussed in Section 4.3, we will need $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$, i.e., sufficiently smooth boundary data u_D. The applied forces β and the possible Neumann data τ are combined within the functional $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, which is defined by

$$\big\langle \mathcal{F}(t), v\big\rangle := \big\langle \beta(t), v\big\rangle_{\mathcal{V}^*,\mathcal{V}} + \big(\tau(t), v\big)_{L^2(\Gamma_{\mathrm{N}})}.$$

Finally, we introduce the operator $\mathcal{M}\colon \mathcal{V}^* \to \mathcal{V}^*$ which includes the density ρ of the given material. Recall that due to the Gelfand triple $\mathcal{V}$, $\mathcal{H}$, $\mathcal{V}^*$ the embeddings from Section 3.3.1 imply for $u \in \mathcal{H}$ and $v \in \mathcal{V}$, that

$$\langle \mathcal{M}u, v\rangle_{\mathcal{V}^*,\mathcal{V}} := \langle \rho u, v\rangle_{\mathcal{V}^*,\mathcal{V}} = (\rho u, v)_{\mathcal{H}}. \tag{7.11}$$

Otherwise, the action of $\mathcal{M}u$ is defined as the continuous extension of this map. To ensure that the introduced operators are defined for the solution, we assume that the deformation variable satisfies $u \in H^1(0,T;\mathcal{V})$ with derivative $\ddot{u} \in L^2(0,T;\mathcal{V}^*)$, i.e., we search for $u \in W^{2;2,2,2}(0,T;\mathcal{V},\mathcal{V},\mathcal{V}^*)$. Note that $\dot{u} \in L^2(0,T;\mathcal{H})$ is not sufficient because of the damping term. As search space for the Lagrange multiplier we consider $\lambda \in L^2(0,T;\mathcal{Q})$. Thus, the dynamic saddle point problem of elastodynamics in operator form reads: Find u and λ such that

$$\mathcal{M}\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{7.12a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^* \tag{7.12b}$$

is satisfied for a.e. $t \in [0,T]$ with initial conditions

$$u(0) = g \in \mathcal{V}, \qquad \dot{u}(0) = h \in \mathcal{H}. \tag{7.12c}$$

Remark 7.3 (Initial conditions). The assumed regularity of the solution implies $u \in C([0,T];\mathcal{V})$ and $\dot{u} \in C([0,T];\mathcal{H})$, see Section 3.3.2. Thus, the initial conditions are well-posed for $g \in \mathcal{V}$ and $h \in \mathcal{H}$. We discuss the consistency of the initial data in Remark 7.4 below.

7.2. Extension and Regularization. With the operator DAE (7.12) at hand, this section is devoted to the regularization of this system which corresponds to an index reduction in finite dimensions. In Section 9 we will see that a finite element discretization of (7.12) leads to a DAE of index 3 whereas the reformulated system results in a DAE of index 1. Since the regularization follows the same ideas as for the first order systems in Section 6.1, we keep this part short. As before, the reformulation is motivated by the index reduction technique of minimal extension, see Section 2.3.2.

As a first step, we add the first two time derivatives of the constraint (7.12b). According to Lemma 7.1 (f), these hidden constraints are given by

$$\mathcal{B}\dot{u} = \dot{\mathcal{G}} \quad \text{and} \quad \mathcal{B}\ddot{u} = \ddot{\mathcal{G}}.$$

For this to make sense, we require $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$. A sufficient condition is that the boundary values satisfy $u_D \in H^2(0,T;\mathcal{Q}^*)$ or, equivalently, there exists a function $v \in H^2(0,T;\mathcal{V})$ with $u_D = \gamma v$. Thereby, γ denotes the trace operator from Section 3.1.4. The extended operator DAE reads

$$\mathcal{M}\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{7.13a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*, \tag{7.13b}$$

$$\mathcal{B}\dot{u}(t) = \dot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^*, \tag{7.13c}$$

$$\mathcal{B}\ddot{u}(t) = \ddot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^*. \tag{7.13d}$$

Obviously, this system is equivalent to system (7.12a)-(7.12b) under the regularity assumption that $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$ and $u \in H^2(0,T;\mathcal{V})$. We show that the latter requirement can be relaxed as in the first-order case. According to Lemma 7.1 we split the deformation u into

$$u = u_1 + u_2 \quad \text{with} \quad u_1(t) \in \mathcal{V}_\mathcal{B},\ u_2(t) \in \mathcal{V}^c.$$

This then implies $\mathcal{B}u = \mathcal{B}u_2$. Furthermore, Lemma 3.47 yields $\mathcal{B}\dot{u} = \mathcal{B}\dot{u}_2$ and $\mathcal{B}\ddot{u} = \mathcal{B}\ddot{u}_2$.

The second step is to introduce new variables. For this, we define the two dummy variables $v_2 := \dot{u}_2 \in L^2(0,T;\mathcal{V}^c)$ and $w_2 := \ddot{u}_2 \in L^2(0,T;\mathcal{V}^c)$ which yield the two constraints

$$\mathcal{B}v_2 = \dot{\mathcal{G}} \quad \text{and} \quad \mathcal{B}w_2 = \ddot{\mathcal{G}}.$$

Note that because of Lemma 7.1 (b) the three constraints already fix the variables u_2, v_2, and w_2 in terms of $\mathcal{G}$ and its derivatives. Thus, the equation $\dot{u}_2 = v_2$ is redundant and does not have to be added to the system. The same applies for w_2. Replacing all appearances of $\dot{u}_2$ and $\ddot{u}_2$, we obtain the regularized operator DAE:

Find $u_1 \in W^{2;2,2,2}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}_\mathcal{B},\mathcal{V}^*)$ as well as u_2, v_2, $w_2 \in L^2(0,T;\mathcal{V}^c)$ and the multiplier $\lambda \in L^2(0,T;\mathcal{Q})$ such that

$$\mathcal{M}(\ddot{u}_1 + w_2) + \mathcal{D}(\dot{u}_1 + v_2) + \mathcal{K}(u_1 + u_2) + \mathcal{B}^*\lambda = \mathcal{F} \quad \text{in } \mathcal{V}^*, \tag{7.14a}$$

$$\mathcal{B}u_2 = \mathcal{G} \quad \text{in } \mathcal{Q}^*, \tag{7.14b}$$

$$\mathcal{B}v_2 = \dot{\mathcal{G}} \quad \text{in } \mathcal{Q}^*, \tag{7.14c}$$

$$\mathcal{B}w_2 = \ddot{\mathcal{G}} \quad \text{in } \mathcal{Q}^*, \tag{7.14d}$$

is satisfied for a.e. $t \in [0,T]$ with with initial conditions

$$u_1(0) = g_0 \in \mathcal{V}_\mathcal{B}, \qquad \dot{u}_1(0) = h_0 \in \mathcal{H}. \tag{7.14e}$$

Remark 7.4 (Consistent initial conditions). Similar to Remark 6.9 concerning first-order systems, the reformulation of the operator DAE provides necessary conditions for the initial data. As discussed in Section 4.3, the initial data in (7.12c) has to satisfy $\mathcal{B}g = \mathcal{G}(0)$. Thus, we obtain the decomposition $g = g_0 + \mathcal{B}^-\mathcal{G}(0)$ with $g_0 \in \mathcal{V}_\mathcal{B}$ and

$\mathcal{B}^-$ denoting a right-inverse of $\mathcal{B}$. For h we have, similarly to the first-order case, the decomposition $h = h_0 + \mathcal{B}^-\dot{\mathcal{G}}(0)$ with $h_0 \in \mathcal{H}$, cf. Example 6.11. For more regular data $h \in \mathcal{V}$ we obtain the same decomposition with $h_0 \in \mathcal{V}_\mathcal{B}$.

REMARK 7.5. Since u_1 is by definition in the kernel of $\mathcal{B}$, we may add $\mathcal{B}u_1$ to equation (7.14b). Assuming sufficient regularity, we may also add the vanishing terms $\mathcal{B}\dot{u}_1$ and $\mathcal{B}\ddot{u}_1$ to equations (7.14c) and (7.14d), respectively. This will be used for the semi-discretization in Section 9 and is necessary because of the application of nonconforming finite elements.

7.3. Existence Results and Well-posedness. For the operator DAEs of first order in Section 6 we have stated results on the equivalence of the original and reformulated systems. Analogous results apply for the equations of this section. Since we deal here with a more specific situation where we focus on elastodynamics, we prefer to state an explicit existence result instead. Thus, we study a precise problem class with nonlinear damping term for which we prove the well-posedness. To show the existence and uniqueness of a solution of the regularized operator DAE (7.14), we start with the homogeneous case, i.e., with $\mathcal{G}(t) \equiv 0$. We then show the existence of a unique Lagrange multiplier and the remaining variables. Finally, we show that the solution depends continuously on the data, i.e., on the initial values, the data of the right-hand side and the nonlinearity of the damping term.

Further existence results for different assumptions were discussed in Section 4.2.2.

7.3.1. *Homogeneous Problem.* The starting point for the existence and uniqueness of solutions of the operator DAE (7.14) is the homogeneous problem with $u_D = 0$. Thus, we consider the following problem: Given initial values $g \in \mathcal{V}_\mathcal{B}$, $h \in \mathcal{H}$ and a right-hand side $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, determine a function $u \in H^1(0,T;\mathcal{V}_\mathcal{B})$ with $\ddot{u} \in L^2(0,T;\mathcal{V}^*)$ such that for a.e. $t \in [0,T]$ it holds that

$$\mathcal{M}\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}_\mathcal{B}^* \tag{7.15}$$

and u satisfies the initial conditions $u(0) = g_0$ and $\dot{u}(0) = h_0$. Recall that equation (7.15) stated in the weaker space $\mathcal{V}_\mathcal{B}^*$ means that it has to be tested only by functions in $\mathcal{V}_\mathcal{B}$ and not for all $v \in \mathcal{V}$.

As stated in Section 7.1.1 we deal with linear elastodynamics with the operator $\mathcal{K}$ defined in (7.2). Thus, we assume the existence of a positive constant k_2 such that for all u, $v \in \mathcal{V}$ it holds that

$$a(u,v) = \langle \mathcal{K}u, v\rangle_{\mathcal{V}^*,\mathcal{V}} \le k_2\|u\|\|v\|.$$

Furthermore, $\mathcal{K}$ is coercive with constant k_1 on $\mathcal{V}_\mathcal{B}$. The operator $\mathcal{M}$ involves the density of the material and is defined in (7.11). It remains to specify the assumptions on the damping operator $\mathcal{D}$ introduced in (7.4). In the following, we assume strong monotonicity and Lipschitz continuity, i.e., we assume that there exist constants d_1, $d_2 > 0$ and $d_0 \ge 0$ such that for all u, v, $w \in \mathcal{V}$ it holds that

$$d_1\|u-v\|^2 \le \big\langle (\mathcal{D} + d_0\,\mathrm{id})u - (\mathcal{D} + d_0\,\mathrm{id})v, u - v\big\rangle_{\mathcal{V}^*,\mathcal{V}} \tag{7.16a}$$

and

$$\|\mathcal{D}u - \mathcal{D}v\|_{\mathcal{V}^*} \le d_2\|u - v\|. \tag{7.16b}$$

Note that the identity map id is used in terms of the embedding $\mathcal{V} \hookrightarrow \mathcal{V}^*$ given by the Gelfand triple $\mathcal{V}$, $\mathcal{H}$, $\mathcal{V}^*$, i.e., $\langle \mathrm{id}\,u, v\rangle_{\mathcal{V}^*,\mathcal{V}} = (u,v)_\mathcal{H}$. The existence of a unique solution under the given assumptions is matter of the following theorem.

THEOREM 7.6 (Homogeneous problem [**Alt13a**]). *Consider initial data $g_0 \in \mathcal{V}_\mathcal{B}$, $h_0 \in \mathcal{H}$ and $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$. Then, there exists a unique solution $u \in H^1(0,T;\mathcal{V}_\mathcal{B})$ of equation (7.15) with initial conditions $u(0) = g_0$ and $\dot{u}(0) = h_0$. Furthermore, the second time derivative satisfies $\ddot{u} \in L^2(0,T;\mathcal{V}^*)$.*

PROOF. Since $\mathcal{M}$ is just a multiplication by a constant, this theorem is a special case of a theorem in [**GGZ74**, Ch. 7]. The proof makes use of Korn's inequality for the operator $\mathcal{K}$ as well as the given properties of $\mathcal{D}$ which imply that the operator $(\mathcal{D} + d_0\,\mathrm{id})$ is continuous, monotone, and coercive. The condition $h \in \mathcal{H}$ is justified by Remark 7.3. The regularity of $\ddot{u}$ follows from the assumed regularity of the right-hand side $\mathcal{F}$. □

REMARK 7.7. Theorem 7.6 can be extended to nonlinear elastodynamics. For details and assumptions on a possibly nonlinear elasticity operator $\mathcal{K}$ we refer to [**GGZ74**, Ch. 7]. Note that a nonlinear operator is necessary if we model large deformations for which the assumption that the stress depends linearly on the strain is not reasonable.

REMARK 7.8. The corresponding existence result for the damping-free case, i.e., $\mathcal{D} \equiv 0$, can be found in [**Zei90a**, Ch. 24].

7.3.2. *Existence of the Lagrange Multiplier.* The inclusion of arbitrary Dirichlet data does not include further difficulties. As usual, the general case can be reduced to the homogeneous case from the previous subsection.

THEOREM 7.9 (Non-homogeneous problem [**Alt13a**]). *Let the Dirichlet data on Γ_D be given by $u_D \in W^{2;2,2,2}(0,T;\mathcal{V},\mathcal{V},\mathcal{V}^*)$. Furthermore, assume that $g \in \mathcal{V}$ with $g = u_D(0)$ on Γ_D, $h \in \mathcal{H}$, and $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$. Then, there exists a unique solution $u \in H^1(0,T;\mathcal{V})$ of*

$$\mathcal{M}\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}_\mathcal{B}^* \tag{7.17}$$

with $u(t) = u_D(t)$ on Γ_D for a.e. $t \in [0,T]$ and initial conditions $u(0) = g$ and $\dot{u}(0) = h$. Furthermore, the second time derivative satisfies $\ddot{u} \in L^2(0,T,\mathcal{V}^*)$.

PROOF. Instead of finding u as stated in the theorem, we consider the equivalent problem: Find $w = u - u_D \in H^1(0,T;\mathcal{V}_\mathcal{B})$ such that

$$\mathcal{M}\ddot{w}(t) + \hat{\mathcal{D}}\dot{w}(t) + \mathcal{K}w(t) = \mathcal{F}(t) - \mathcal{M}\ddot{u}_D(t) - \mathcal{K}u_D(t) \quad \text{in } \mathcal{V}_\mathcal{B}^* \tag{7.18a}$$

with initial conditions

$$w(0) = g - u_D(0) \in \mathcal{V}_\mathcal{B}, \tag{7.18b}$$

$$\dot{w}(0) = h - \dot{u}_D(0) \in \mathcal{H}. \tag{7.18c}$$

Therein, $\hat{\mathcal{D}}$ denotes the operator defined by $\hat{\mathcal{D}}\dot{w} := \mathcal{D}(\dot{w} + \dot{u}_D)$. It is easy to see that $\hat{\mathcal{D}}$ is Lipschitz continuous and strongly monotone with the same constants as $\mathcal{D}$. Thus, we apply Theorem 7.6 to equation (7.18) which states the existence of a unique solution w and hence the unique solvability of the original problem (7.17). Note that the special choice of the initial data in (7.18b)-(7.18c) implies that u satisfies the posed initial conditions. In addition, we obtain $\ddot{w} \in L^2(0,T;\mathcal{V}^*)$ and thus, the claimed regularity for u and its derivatives. □

In view of of the formulation as operator DAE, we follow [**Sim00**] and show that for a solution $u \in L^2(0,T;\mathcal{V})$ of the non-homogeneous operator ODE (7.17) there exists a unique Lagrange multiplier $\lambda \in L^2(0,T;\mathcal{Q})$.

THEOREM 7.10 (Existence of the Lagrange multiplier [**Alt13a**]). *Let g, h, $\mathcal{F}$, and u_D be as in Theorem 7.9 and $\mathcal{G}$ as defined in* (7.10). *Furthermore, let $u \in H^1(0,T;\mathcal{V})$ denote the unique solution from Theorem 7.9. Then, there exists a unique Lagrange multiplier $\lambda \in L^2(0,T;\mathcal{Q})$ such that (u,λ) is a solution of system* (7.12).

PROOF. Note that u fulfills the desired Dirichlet boundary condition $u = u_D$ along Γ_{D} and thus (7.12b). Since $\mathcal{B}^*\lambda$ vanishes for functions in $\mathcal{V}_{\mathcal{B}}$, the given solution u of (7.17) satisfies

$$\mathcal{M}\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}_{\mathcal{B}}^*.$$

In order to guarantee equation (7.12a) in $\mathcal{V}^*$, λ has to satisfy

$$\mathcal{B}^*\lambda(t) = \mathcal{F}(t) - \mathcal{M}\ddot{u}(t) - \mathcal{D}\dot{u}(t) - \mathcal{K}u(t) \quad \text{in } (\mathcal{V}^{\mathrm{c}})^*.$$

The functional on the right-hand side vanishes for all test functions in $\mathcal{V}_{\mathcal{B}}$ and thus, is an element of $\mathcal{V}_{\mathcal{B}}^{\mathrm{o}}$. As a result, Lemma 7.1 (d) implies that this equation has a unique solution $\lambda \in L^2(0,T;\mathcal{Q})$. Thus, the pair (u,λ) solves system (7.12). □

7.3.3. *Well-posedness of the Saddle Point Problem.* At the end of this section, we state the main result, namely the well-posedness of the operator DAE (7.14).

THEOREM 7.11 (Well-posedness [**Alt13a**]). *Consider $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G}$ from* (7.10) *with Dirichlet data $u_D \in H^2(0,T;\mathcal{V})$, and initial data $g_0 \in \mathcal{V}_{\mathcal{B}}$, $h_0 \in \mathcal{H}$. Then, problem* (7.14) *is well-posed in the following sense. First, there exists a unique solution $(u_1,u_2,v_2,w_2,\lambda)$ with $u_1 \in W^{2;2,2,2}(0,T;\mathcal{V}_{\mathcal{B}},\mathcal{V}_{\mathcal{B}},\mathcal{V}^*)$, u_2, v_2, $w_2 \in L^2(0,T;\mathcal{V}^{\mathrm{c}})$, and $\lambda \in L^2(0,T;\mathcal{Q})$. Second, the map*

$$\big(g_0, h_0, \mathcal{D}(0), \mathcal{F}, \mathcal{G}\big) \mapsto \big(u_1, u_2, v_2, w_2, \ddot{u}_1 + \mathcal{D}\dot{u}_1 + \mathcal{B}^*\lambda\big)$$

is a linear and continuous map of the form

$$\begin{aligned}\mathcal{V}_{\mathcal{B}} \times \mathcal{H} \times \mathcal{V}^* \times L^2(0,T;\mathcal{V}^*) \times H^2(0,T;\mathcal{Q}^*) \to \\ C([0,T],\mathcal{V}) \cap C^1([0,T],\mathcal{H}) \times L^2(0,T;\mathcal{V}^{\mathrm{c}})^3 \times L^2(0,T;\mathcal{V}^*).\end{aligned}$$

PROOF. Note that $u_D \in H^2(0,T;\mathcal{V})$ and the trace theorem implies $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$,

$$\langle q, \ddot{\mathcal{G}}(t)\rangle_{\mathcal{Q},\mathcal{Q}^*} := \langle q, \ddot{u}_D(t)\rangle_{\mathcal{Q},\mathcal{Q}^*}, \quad \|\ddot{\mathcal{G}}\|_{L^2(0,T;\mathcal{Q}^*)} \le C_{\mathrm{tr}}\|\ddot{u}_D\|_{L^2(0,T;\mathcal{V})}.$$

Uniqueness: Assume that $(u_1,u_2,v_2,w_2,\lambda)$ and $(U_1,U_2,V_2,W_2,\Lambda)$ are two solutions of problem (7.14). Equation (7.14b) provides $\mathcal{B}(u_2 - U_2) = 0$ in $\mathcal{Q}^*$. Using the isomorphism from Lemma 7.1 b), we obtain $u_2 = U_2$. With the same arguments, we achieve $v_2 = V_2$ and $w_2 = W_2$. With the differences $e := u_1 - U_1$ and $\mu := \lambda - \Lambda$, equation (7.14a) reads

$$\mathcal{M}\ddot{e} + \hat{\mathcal{D}}\dot{e} + \mathcal{K}e + \mathcal{B}^*\mu = 0 \text{ in } \mathcal{V}^*$$

with the operator $\hat{\mathcal{D}}(\dot{e}) := \mathcal{D}(\dot{e} + \dot{U}_1 + v_2) - \mathcal{D}(\dot{U}_1 + v_2)$ and initial conditions $e(0) = 0$ and $\dot{e}(0) = 0$. Obviously, $\hat{\mathcal{D}}$ is Lipschitz continuous and strongly monotone with the same constants as $\mathcal{D}$ such that Theorem 7.6 is applicable. Thus, testing only with functions in $\mathcal{V}_{\mathcal{B}}$, we obtain by Theorem 7.6 that $e = 0$. Since $\hat{\mathcal{D}}(0) = 0$, it remains the equation $b(\mu, v) = 0$ for all $v \in \mathcal{V}^{\mathrm{c}}$ which implies $\mu = 0$.

Existence: Let $\mathcal{P}$ be the projection onto $\mathcal{V}_{\mathcal{B}}$, i.e., $\mathcal{P}\colon \mathcal{V} \to \mathcal{V}_{\mathcal{B}}$. Further, let (u,λ) denote the solution from Theorem 7.10 with initial data $u(0) = g_0 + \mathcal{B}^-\mathcal{G}(0) = g_0 + (\mathrm{id} - \mathcal{P})u_D(0)$ and $\dot{u}(0) = h_0 + \mathcal{B}^-\dot{\mathcal{G}}(0)$. This then implies that $u_1 := \mathcal{P}u$ satisfies $u_1 \in \mathcal{C}([0,T],\mathcal{V}_{\mathcal{B}})$ with $\dot{u}_1 \in W^{1;2,2}(0,T;\mathcal{V}_{\mathcal{B}},\mathcal{V}^*)$. With the help of Lemma 7.1, we define

$$u_2 := (\mathrm{id} - \mathcal{P})u = \mathcal{B}^-\mathcal{G}, \quad v_2 := \dot{u}_2 = \mathcal{B}^-\dot{\mathcal{G}}, \quad w_2 := \ddot{u}_2 = \mathcal{B}^-\ddot{\mathcal{G}}.$$

The regularity of $\mathcal{G}$, namely $\mathcal{G}$, $\dot{\mathcal{G}}$, $\ddot{\mathcal{G}} \in L^2(0,T;\mathcal{Q}^*)$, implies that u_2, v_2, $w_2 \in L^2(0,T;\mathcal{V}^{\mathrm{c}})$. Obviously, the tuple $(u_1, u_2, v_2, w_2, \lambda)$ satisfies equations (7.14a)-(7.14d). The initial values satisfy

$$u_1(0) = \mathcal{P}u(0) = \mathcal{P}g_0 = g_0, \quad \dot{u}_1(0) = \dot{u}(0) - \dot{u}_2(0) = h_0 + \mathcal{B}^-\dot{\mathcal{G}}(0) - v_2(0) = h_0.$$

Continuous dependence on data: Recall that we use the abbreviations $\|\cdot\| = \|\cdot\|_{\mathcal{V}}$ and $|\cdot| = \|\cdot\|_{\mathcal{H}}$. Lemma 7.1 (e) implies the estimate

$$\|u_2(t)\| \le \frac{1}{\beta}\|\mathcal{B}u_2(t)\|_{\mathcal{Q}^*} = \frac{1}{\beta}\|\mathcal{G}(t)\|_{\mathcal{Q}^*}.$$

Similar estimates for v_2 and w_2 result in

$$\|u_2\|^2_{L^2(0,T;\mathcal{V})} + \|v_2\|^2_{L^2(0,T;\mathcal{V})} + \|w_2\|^2_{L^2(0,T;\mathcal{V})} \le \frac{1}{\beta^2}\|\mathcal{G}\|^2_{H^2(0,T;\mathcal{Q}^*)}. \tag{7.19}$$

For an estimate of u_1, we test equation (7.14a) with $\dot{u}_1(t) \in \mathcal{V}_{\mathcal{B}}$. We omit the explicit time-dependence and obtain

$$\begin{aligned}
&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big[\rho|\dot{u}_1|^2 + a(u_1,u_1)\Big] + \big\langle \mathcal{D}(\dot{u}_1 + v_2) - \mathcal{D}v_2, \dot{u}_1\big\rangle \\
&\qquad = \langle \mathcal{F}, \dot{u}_1\rangle - (\rho w_2, \dot{u}_1)_{\mathcal{H}} - \big\langle \mathcal{D}v_2, \dot{u}_1\big\rangle - a(u_2, \dot{u}_1) \\
&\qquad = \langle \mathcal{F}, \dot{u}_1\rangle - (\rho w_2, \dot{u}_1)_{\mathcal{H}} - \big\langle \mathcal{D}v_2 - \mathcal{D}(0), \dot{u}_1\big\rangle - \big\langle \mathcal{D}(0), \dot{u}_1\big\rangle - a(u_2, \dot{u}_1).
\end{aligned}$$

Recall that d_2 and k_2 denote the continuity constants of $\mathcal{D}$ and $\mathcal{K}$, respectively, and d_0, d_1 the monotonicity constants of $\mathcal{D}$. With $\eta(t) := \rho|\dot{u}_1(t)|^2 + a(u_1(t), u_1(t))$, the strong monotonicity of $\mathcal{D}$, and the Cauchy-Schwarz inequality we obtain the estimate

$$\begin{aligned}
&\dot{\eta} + 2d_1\|\dot{u}_1\|^2 - 2d_0|\dot{u}_1|^2 \\
&\qquad \le \rho\frac{\mathrm{d}}{\mathrm{d}t}|\dot{u}_1|^2 + \frac{\mathrm{d}}{\mathrm{d}t}a(u_1,u_1) + 2\big\langle \mathcal{D}(\dot{u}_1 + v_2) - \mathcal{D}v_2, \dot{u}_1\big\rangle \\
&\qquad \le 2\|\mathcal{F}\|_{\mathcal{V}^*}\|\dot{u}_1\| + 2\rho|w_2||\dot{u}_1| + 2d_2\|v_2\|\|\dot{u}_1\| + 2\|\mathcal{D}(0)\|_{\mathcal{V}^*}\|\dot{u}_1\| + 2k_2\|u_2\|\|\dot{u}_1\|.
\end{aligned}$$

By Young's inequality $2ab \le a^2/c + cb^2$ [**Eva98**, App. B] with appropriate choices of the constant $c > 0$, we obtain

$$\begin{aligned}
\dot{\eta} + 2d_1\|\dot{u}_1\|^2 &\le \frac{2}{d_1}\|\mathcal{F}\|^2_{\mathcal{V}^*} + \frac{d_1}{2}\|\dot{u}_1\|^2 + \rho|w_2|^2 + (\rho + 2d_0)|\dot{u}_1|^2 + \frac{2d_2^2}{d_1}\|v_2\|^2 \\
&\qquad + \frac{d_1}{2}\|\dot{u}_1\|^2 + \frac{2}{d_1}\|\mathcal{D}(0)\|^2_{\mathcal{V}^*} + \frac{d_1}{2}\|\dot{u}_1\|^2 + \frac{2k_2^2}{d_1}\|u_2\|^2 + \frac{d_1}{2}\|\dot{u}_1\|^2 \\
&\le \frac{\rho + 2d_0}{\rho}\eta + \frac{2}{d_1}\|\mathcal{F}\|^2_{\mathcal{V}^*} + \rho|w_2|^2 + \frac{2d_2^2}{d_1}\|v_2\|^2 + \frac{2}{d_1}\|\mathcal{D}(0)\|^2_{\mathcal{V}^*} \\
&\qquad + \frac{2k_2^2}{d_1}\|u_2\|^2 + 2d_1\|\dot{u}_1\|^2.
\end{aligned}$$

Thus, there exists a generic constant c, such that

$$\dot{\eta}(t) \le (1 + 2d_0/\rho)\eta(t) + c\xi(t)$$

with

$$\xi(t) = \|\mathcal{F}(t)\|^2_{\mathcal{V}^*} + \|w_2(t)\|^2 + \|v_2(t)\|^2 + \|\mathcal{D}(0)\|^2_{\mathcal{V}^*} + \|u_2(t)\|^2.$$

Thus, by the absolute continuity of η and Gronwall's lemma [**Eva98**, App. B] we obtain that η is bounded by

$$\eta(t) \le (1 + 2d_0/\rho)e^t\Big(\eta(0) + c\int_0^t \xi(s)\ ds\Big).$$

The initial value of η is given by the initial values in (7.14e),

$$\eta(0) = \rho|h_0|^2 + a(g_0, g_0) \leq \rho|h_0|^2 + k_2\|g_0\|^2.$$

The integral term can be bounded with the help of (7.19). Therewith, the existence of a positive constant c follows such that for all $t \in [0,T]$ it holds that

$$\eta(t) \leq c\,\left[\|g_0\|^2 + |h_0|^2 + \|\mathcal{D}(0)\|^2_{\mathcal{V}^*} + \|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}^*)} + \|\mathcal{G}\|^2_{H^2(0,T;\mathcal{Q}^*)}\right].$$

Since the right-hand side is independent of t, we can maximize over t and obtain bounds for u_1 and $\dot{u}_1$ in $C([0,T],\mathcal{V})$ and $C([0,T],\mathcal{H})$, respectively,

$$\begin{aligned}\|u_1\|_{C([0,T],\mathcal{V})} &+ \|\dot{u}_1\|_{C([0,T],\mathcal{H})} \\ &\leq c\,\left[\|g_0\| + |h_0|_{\mathcal{H}} + \|\mathcal{D}(0)\|_{\mathcal{V}^*} + \|\mathcal{F}\|_{L^2(0,T;\mathcal{V}^*)} + \|\mathcal{G}\|_{H^2(0,T;\mathcal{Q}^*)}\right].\end{aligned}$$

It remains to bound $\rho\ddot{u}_1 + \mathcal{D}\dot{u}_1 + \mathcal{B}^*\lambda$ in $L^2(0,T;\mathcal{V}^*)$. By the definition of the $\mathcal{V}^*$-norm and equation (7.14a), we achieve

$$\begin{aligned}\|\rho\ddot{u}_1(t) &+ \mathcal{D}\dot{u}_1(t) + \mathcal{B}^*\lambda(t)\|_{\mathcal{V}^*} \\ &= \sup_{v\in\mathcal{V}} \frac{\langle \mathcal{M}\ddot{u}_1(t), v\rangle + \langle \mathcal{D}\dot{u}_1(t), v\rangle + \langle \mathcal{B}^*\lambda(t), v\rangle}{\|v\|} \\ &= \sup_{v\in\mathcal{V}} \frac{\langle \mathcal{F}, v\rangle - \langle \mathcal{M}w_2, v\rangle - \langle \mathcal{K}(u_1+u_2), v\rangle + \langle \mathcal{D}\dot{u}_1 - \mathcal{D}(\dot{u}_1+v_2), v\rangle}{\|v\|} \\ &\leq \|\mathcal{F}(t)\|_{\mathcal{V}^*} + \rho|w_2(t)| + k_2\|u_1(t)+u_2(t)\| + d_2\|v_2(t)\|.\end{aligned}$$

Thus, by integration over the interval $[0,T]$, Young's inequality, and the estimates for u_1, u_2, v_2, and w_2 from above, we obtain a positive constant c with

$$\begin{aligned}\|\ddot{u}_1(t) &+ \mathcal{D}\dot{u}_1(t) + \mathcal{B}^*\lambda(t)\|_{L^2(0,T;\mathcal{V}^*)} \\ &\leq c\,\left[\|g\| + |h| + \|\mathcal{D}(0)\|_{\mathcal{V}^*} + \|\mathcal{F}\|_{L^2(0,T;\mathcal{V}^*)} + \|\mathcal{G}\|_{H^2(0,T;\mathcal{Q}^*)}\right].\end{aligned}$$

□

7.4. Influence of Perturbations. As for the operator DAEs of first order, we discuss the influence of small perturbations in the right-hand side on the solution behavior. We only consider the regularized system (7.14). The results for the original formulation (7.12) then follow as in Section 6.1.3 and include derivatives of the perturbations. Let $(\hat{u}_1, \hat{u}_2, \hat{v}_2, \hat{w}_2, \hat{\lambda})$ denote the solution of the perturbed problem

$$\mathcal{M}(\ddot{\hat{u}}_1 + \hat{w}_2) + \mathcal{D}(\dot{\hat{u}}_1 + \hat{v}_2) + \mathcal{K}(\hat{u}_1 + \hat{u}_2) + \mathcal{B}^*\hat{\lambda} = \mathcal{F} + \delta \quad \text{in } \mathcal{V}^*, \tag{7.20a}$$

$$\mathcal{B}\hat{u}_2 = \mathcal{G} + \theta \quad \text{in } \mathcal{Q}^*, \tag{7.20b}$$

$$\mathcal{B}\hat{v}_2 = \dot{\mathcal{G}} + \xi \quad \text{in } \mathcal{Q}^*, \tag{7.20c}$$

$$\mathcal{B}\hat{w}_2 = \ddot{\mathcal{G}} + \vartheta \quad \text{in } \mathcal{Q}^*. \tag{7.20d}$$

Therein, the perturbations satisfy $\delta \in L^2(0,T;\mathcal{V}^*)$ and $\theta,\ \xi,\ \vartheta \in L^2(0,T;\mathcal{Q}^*)$. Furthermore, we have perturbed initial data of the form $\hat{u}_1(0) = u_1(0) + e_{1,0}$ and $\dot{\hat{u}}_1(0) = \dot{u}_1(0) + \dot{e}_{1,0}$.

THEOREM 7.12. *Consider the perturbed problem* (7.20) *with operators $\mathcal{M}$, $\mathcal{D}$, $\mathcal{K}$, and $\mathcal{B}$ as described in the previous subsections. Then, the solution $(\hat{u}_1, \hat{u}_2, \hat{v}_2, \hat{w}_2, \hat{\lambda})$ satisfies with $e_1 := \hat{u}_1 - u_1$ the stability estimate*

$$\begin{aligned}\|\dot{e}_1\|^2_{C([0,T];\mathcal{H})} + \|e_1\|^2_{C([0,T];\mathcal{V})} \leq c\,\Big[&|\dot{e}_{1,0}|^2 + \|e_{1,0}\|^2 + \|\delta\|^2_{L^2(0,T;\mathcal{V}^*)} \\ &+ \|\theta\|^2_{L^2(0,T;\mathcal{Q}^*)} + \|\xi\|^2_{L^2(0,T;\mathcal{Q}^*)} + \|\vartheta\|^2_{L^2(0,T;\mathcal{Q}^*)}\Big].\end{aligned}$$

Furthermore, the $L^2(0,T;\mathcal{V})$ errors of $\hat{u}_2 - u_2$, $\hat{v}_2 - v_2$, and $\hat{w}_2 - w_2$ are bounded by the $L^2(0,T;\mathcal{Q}^*)$ norm of θ, ξ, and ϑ, respectively.

PROOF. We consider the difference of systems (7.20) and (7.14) which yields an operator DAE for the errors. The bounds for $e_2 := \hat{u}_2 - u_2$, $e_v := \hat{v}_2 - v_2$, and $e_w := \hat{w}_2 - w_2$ are obvious because of Lemma 7.1 (c). We proceed as in the proof of Theorem 7.11, i.e., we test the first equation of the system by $\dot{e}_1$ and use the Gronwall lemma. Testing the difference of equation (7.20a) and (7.14a) by $\dot{e}_1$, we obtain

$$\rho\frac{\mathrm{d}}{\mathrm{d}t}|\dot{e}_1|^2 + \frac{\mathrm{d}}{\mathrm{d}t}|\mathcal{K}^{1/2}e_1|^2 + 2\big\langle \mathcal{D}(\dot{\hat{u}}_1 + \hat{v}_2) - \mathcal{D}(\dot{u}_1 + v_2), \dot{e}_1\big\rangle = 2\big\langle \delta - \rho e_w - \mathcal{K}e_2, \dot{e}_1\big\rangle. \tag{7.21}$$

Considering the damping term, by the properties given in (7.16) we get

$$\begin{aligned}
&2\big\langle \mathcal{D}(\dot{\hat{u}}_1+\hat{v}_2) - \mathcal{D}(\dot{u}_1 + v_2), \dot{e}_1\big\rangle \\
&\qquad = 2\big\langle \mathcal{D}(\dot{\hat{u}}_1 + \hat{v}_2) - \mathcal{D}(\dot{u}_1 + v_2), \dot{e}_1 + e_v\big\rangle - 2\big\langle \mathcal{D}(\dot{\hat{u}}_1 + \hat{v}_2) - \mathcal{D}(\dot{u}_1 + v_2), e_v\big\rangle \\
&\qquad \geq 2d_1\|\dot{e}_1 + e_v\|^2 - 2d_0|\dot{e}_1 + e_v|^2 - 2d_2\|e_v\|\|\dot{e}_1 + e_v\| \\
&\qquad \geq 2d_1\|\dot{e}_1\|^2 - 2d_0|\dot{e}_1 + e_v|^2 - 2d_2\|e_v\|\|\dot{e}_1\| - 2d_2\|e_v\|^2.
\end{aligned}$$

Note that we have used the orthogonality of $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^\mathrm{c}$ in the last step which implies $\|\dot{e}_1 + e_v\|^2 = \|\dot{e}_1\|^2 + \|e_v\|^2$. For the right-hand side of equation (7.21) we estimate by the Cauchy-Schwarz inequality,

$$2\big\langle \delta - \rho e_w - \mathcal{K}e_2, \dot{e}_1\big\rangle \leq 2\|\delta\|_{\mathcal{V}^*}\|\dot{e}_1\| + \rho C_{\mathrm{emb}}^2\|e_w\|\|\dot{e}_1\| + 2k_2\|e_2\|\|\dot{e}_1\|.$$

Therewith, equation (7.21) turns into

$$\begin{aligned}
\rho\frac{\mathrm{d}}{\mathrm{d}t}|\dot{e}_1|^2 + \frac{\mathrm{d}}{\mathrm{d}t}|\mathcal{K}^{1/2}e_1|^2 + 2d_1\|\dot{e}_1\|^2 \leq{}& 2d_0|\dot{e}_1 + e_v|^2 + 2d_2\|e_v\|\|\dot{e}_1\| + 2d_2\|e_v\|^2 \\
&+ 2\|\delta\|_{\mathcal{V}^*}\|\dot{e}_1\| + \rho C_{\mathrm{emb}}^2\|e_w\|\|\dot{e}_1\| + 2k_2\|e_2\|\|\dot{e}_1\|.
\end{aligned}$$

An application of Young's inequality then yields with a generic constant c,

$$\begin{aligned}
\rho\frac{\mathrm{d}}{\mathrm{d}t}|\dot{e}_1|^2 + \frac{\mathrm{d}}{\mathrm{d}t}|\mathcal{K}^{1/2}e_1|^2 &\leq 4d_0|\dot{e}_1|^2 + c\left[\|\delta\|_{\mathcal{V}^*} + \|e_2\|^2 + \|e_v\|^2 + \|e_w\|^2\right] \\
&\leq 4d_0|\dot{e}_1|^2 + c\left[\|\delta\|_{\mathcal{V}^*}^2 + \|\theta\|_{\mathcal{Q}^*}^2 + \|\xi\|_{\mathcal{Q}^*}^2 + \|\vartheta\|_{\mathcal{Q}^*}^2\right].
\end{aligned}$$

The Gronwall lemma [**Eva98**, App. B] and maximizing over $t \in [0,T]$ then gives the claim, since $k_1\|e_1\|^2 \leq |\mathcal{K}^{1/2}e_1|^2$. □

The corresponding result for the original formulation (7.12) shows that the error in u_1 depends on δ and θ but also on the derivatives $\dot{\theta}$ and $\ddot{\theta}$. Thus, the operator DAE in its original formulation is much more sensitive with regard to perturbations than the regularized system. This result corresponds to the findings of Section 9 where we show that a spatial discretization of the regularized system leads to a DAE of lower index. Note, however, that in the finite-dimensional case with corresponding Hessenberg structure the error of the differential variable can be bounded in terms of the initial error, δ, θ, and $\dot{\theta}$, i.e., without the second derivative of θ, see [**Arn98b**, Ch. 2.3].

As in the section on first-order systems, we give a short overview of the properties of the regularized formulation in form of a table. Recall that we consider the Gelfand triple $\mathcal{V} \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{V}^*$ with $\mathcal{V}$ denoting the Sobolev space $H^1(\Omega)$ in d components. Since the kernel of the constraint operator $\mathcal{B}$ equals the space $H^1_{\Gamma_D}(\Omega)$, which is dense in $L^2(\Omega)$, we obtain $\overline{\mathcal{V}_\mathcal{B}}^{\mathcal{H}} = \mathcal{H}$. In the following, we use the abbreviations $L^2(\mathcal{V}^*) := L^2(0,T;\mathcal{V}^*)$ and $L^2(\mathcal{Q}^*) := L^2(0,T;\mathcal{Q}^*)$.

	original formulation	regularized formulation
system of equations	operator DAE (7.12)	operator DAE (7.14)
solution spaces	$u \in W^{2;2,2,2}(0,T;\mathcal{V},\mathcal{V},\mathcal{V}^*)$, $\lambda \in L^2(0,T;\mathcal{Q})$	$u_1 \in W^{2;2,2,2}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}_\mathcal{B},\mathcal{V}^*)$, $u_2,\, v_2,\, w_2 \in L^2(0,T;\mathcal{V}^c)$, $\lambda \in L^2(0,T;\mathcal{Q})$
initial conditions and consistency	$u(0) = g \in \mathcal{V}$, $\dot{u}(0) = h \in \mathcal{H}$ $g = g_0 + \mathcal{B}^- \mathcal{G}(0),\ g_0 \in \mathcal{V}_\mathcal{B}$	$u_1(0) = g_0 \in \mathcal{V}_\mathcal{B}$, $\dot{u}_1(0) = h_0 \in \mathcal{H}$
spatial discretization	leads to DAE of index 3, cf. Section 9	leads to DAE of index 1, cf. Section 9
perturbations	$\|\dot{e}_1\|^2_{C([0,T];\mathcal{H})} + \|e_1\|^2_{C([0,T];\mathcal{V})}$ $\lesssim \|\dot{e}_{1,0}\|^2 + \|e_{1,0}\|^2$ $+\|\delta\|^2_{L^2(\mathcal{V}^*)} + \|\theta\|^2_{L^2(\mathcal{Q}^*)}$ $+\|\dot{\theta}\|^2_{L^2(\mathcal{Q}^*)} + \|\ddot{\theta}\|^2_{L^2(\mathcal{Q}^*)}$	$\|\dot{e}_1\|^2_{C([0,T];\mathcal{H})} + \|e_1\|^2_{C([0,T];\mathcal{V})}$ $\lesssim \|\dot{e}_{1,0}\|^2 + \|e_{1,0}\|^2$ $+\|\delta\|^2_{L^2(\mathcal{V}^*)} + \|\theta\|^2_{L^2(\mathcal{Q}^*)}$ $+\|\xi\|^2_{L^2(\mathcal{Q}^*)} + \|\vartheta\|^2_{L^2(\mathcal{Q}^*)}$

7.5. Applications in Flexible Multibody Dynamics. In contrast to the first-order case in Section 6, we have considered in this section the reformulation of operator DAEs of second order for a specific case. More precisely, we have regarded the dynamics of an elastic body which is constrained at the boundary which is modeled in terms of Dirichlet boundary conditions. As mentioned before, the regularization also applies for different choices of the damping and stiffness operators. Even the trace operator could be replaced by some other constraint operator, cf. Section 6.

In this last subsection of Part B, we discuss further applications and extensions where the here presented methods apply directly. For this, we remove the restriction of a single body and consider *flexible multibody systems* [**GC01, Bau10**].

Multibody systems in which we allow the single parts to be deformable are called flexible multibody systems. These systems can be formulated as abstract differential equations and are currently very popular because of the wide range of applications. These models are necessary, since the traditional design of machines with a maximization of stiffness avoids elastic vibrations but leads to a drastic increase of mass and thus, of costs. Modern mechanisms need lightweight designs and thus, include bodies where the deformation cannot be neglected anymore. As a result, accurate and meaningful simulations need to include the vibrations and thus, model the systems in the form of flexible multibody systems [**SHD11**].

The presented model of a deformable body can be extended to flexible multibody systems if the coupling can be expressed in terms of Dirichlet boundary conditions. In this case, the coupled system can be written in the same structure as a single body constrained by Dirichlet boundary conditions [**Alt13b**]. Thus, the presented regularization technique from Section 7.2 also applies for flexible multibody systems of this form.

Example 7.13 (Slider crank mechanism). One popular benchmark example for multibody systems is the *slider crank mechanism* [**JPD93, Sim96**]. The simplest model consists of two rods connected by a revolute joint (also called pin joint). As usual, the discrete model equations yield a DAE of index 3. A possible extension of this model adds flexibility to the system. For this, we replace one rigid rod by a flexible body, cf. [**Sim06**] or [**Sim13**, Ch. 8.2]. An illustration is given in Figure 7.1, in which the rod on the right (red) is modeled as a deformable body. Details on the modeling part, following the *floating*

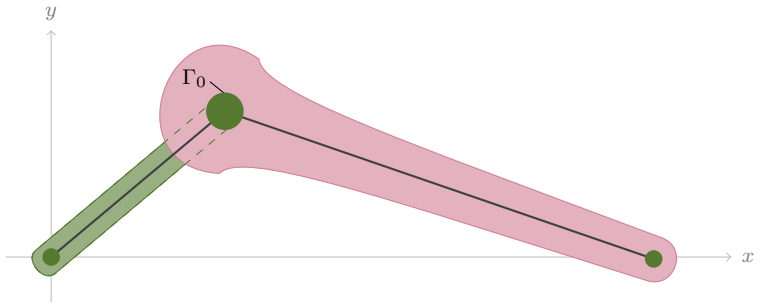


FIGURE 7.1. Illustration of the slider crank mechanism from Example 7.13 with a rigid rod (left) and a flexible rod (right).

reference approach, are given in [**Sim06**]. This then leads to a PDE with a constraint at the joint of the two rods, i.e., an operator DAE. The constraint may either be modeled by an inequality or by a nonlinear equation which - in the strong form - reads

$$\big(x + u(x,t)\big)^T \, \big(x + u(x,t)\big) = r^2. \tag{7.22}$$

Therein, the function u denotes the deformation along the boundary Γ_0 (where it is connected to the rigid rod, see Figure 7.1) and r equals the constant radius of the pin. Thus, the constraint (7.22) guarantees that the boundary Γ_0 of the flexible rod stays in shape of a circle of radius r. Assuming small deformations, we may neglect the quadratic term. Since $x^T x = r^2$ on the boundary, the linearization of equation (7.22) is given by $u(x,t)^T x = 0$. In the weak form, this constraint is given by

$$\big\langle \mathcal{B}u, q \big\rangle := \int_{\Gamma_0} q\big(u^T x\big) \, \mathrm{d}x = 0$$

for all test functions $q \in \mathcal{Q}$. For the formulation of the problem as operator DAE, it remains to find appropriate function spaces $\mathcal{V}$, $\mathcal{V}_\mathcal{B}$, $\mathcal{H}$, and $\mathcal{Q}$. For this, we set

$$\mathcal{V} := \big[H^1(\Omega)\big]^2, \quad \mathcal{V}_\mathcal{B} := \big[H^1_{\Gamma_0}(\Omega)\big]^2, \quad \mathcal{H} := \big[L^2(\Omega)\big]^2, \quad \mathcal{Q}^* := H^{1/2}(\Gamma_0).$$

Note that space $\mathcal{Q}$ has only one component whereas the space $\mathcal{V}$ is defined in two space dimensions.

Part

C

The Method of Lines

This part is devoted to justify the presented regularization of operator DAEs in Part B. We analyse the resulting benefits in the spatial discretized system which is in fact a DAE. Loosely speeking, we show that the performed reformulation is an index reduction in the abstract setting. Recall that we have not defined any index concept for operator equations such that one has to be cautious with the terminology. However, applying the finite element method within the method of lines as described in Section 5.3, we obtain a DAE for which the (differentiation) index is well-defined. We show that the DAEs resulting from the semi-discretization of the regularized operator DAEs turn out to be of index 1, whereas the semi-discretization of the original equations are of index 2 or 3, respectively. Thus, the reformulation on operator level positively effects the structure of the semi-discretized system.

As we have called for a splitting of the ansatz space in the regularization process, we also need a splitting of the finite-dimensional approximation space. This will mostly result in nonconforming discretization schemes, since we need to approximate the (orthogonal) complement of the kernel of the constraint operator.

Similiar as before, this part is divided into sections on first- and second-order systems. In Section 8, we discuss the regularized first-order operator DAEs from Section 6. Again we analyse the cases of linear and nonlinear constraints separately and discuss needed assumptions on the discretization. Afterwards, we focus on the particular example of flow equations, which includes the Navier-Stokes equations. For this, we provide specific algorithms which provide the needed splitting of the finite element space. As a numerical example we consider the Navier-Stokes equations within a cylinder wake. The performed regularization allows for reliable simulation results also for relatively large errors within the iterative solver routine solving the resulting algebraic systems.

The justification of the regularization for second-order operator DAEs is subject of Section 9. Recall that we consider here the case of elastodynamics with linear constraints along the boundary. For such systems, spatial discretizations normally lead to DAEs of index 3 which have the same structure as in multibody dynamics. On the other hand, the regularized operator DAE yields a DAE of index 1. Finally, we comment on the commutativity of semi-discretization and index reduction for suitable discretization schemes.

8. The Method of Lines for First-order Systems

In this section we determine the index of the semi-discretized systems corresponding to the operator DAE (6.1) and its regularized versions (6.4) and (6.9). This then shows one of the achieved benefits resulting from the regularization process. For the example of flow equations, as introduced in Section 6.3.1, we provide more details for the application of the proposed method in practice. In particular, we discuss how to find a regular block of the constraint matrix which corresponds to a splitting of the finite element space. For this, we consider two specific finite element schemes used in computational fluid dynamics.

The results of this section are published within Section 4 of [**AH14**] as well as in [**AH13**].

8.1. Preliminaries and Notation. In preparation for the index determination of the semi-discretized systems which occur by the method of lines, we recall the operator equations of the previous part. In Section 6 we have considered the original system (6.1) which (in the linear case) has the form

$$\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*(t)\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{8.1a}$$

$$\mathcal{B}(t)u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*. \tag{8.1b}$$

Recall that in the case of nonlinear constraints we have to replace $\mathcal{B}^*$ in equation (8.1a) by the dual operator of its Fréchet derivative along u, namely $\mathcal{C}_u$. For linear constraints, the proposed regularization results in system (6.4) which has the form

$$\dot{u}_1(t) + v_2(t) + \mathcal{K}\big(u_1(t) + u_2(t)\big) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{8.2a}$$

$$\mathcal{B}u_2(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*, \tag{8.2b}$$

$$\mathcal{B}v_2(t) + \dot{\mathcal{B}}u_2(t) = \dot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^*. \tag{8.2c}$$

In the nonlinear case, we have obtained the reformulated operator DAE (6.9), i.e.,

$$\dot{u}_1(t) + v_2(t) + \mathcal{K}\big(u_1(t) + u_2(t)\big) + \mathcal{C}_u^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{8.3a}$$

$$\mathcal{B}\big(u_1(t) + u_2(t)\big) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*, \tag{8.3b}$$

$$\mathcal{C}_u\dot{u}_1(t) + \mathcal{C}_{2,u}v_2(t) = \dot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^*. \tag{8.3c}$$

For the discretization in space, we consider finite element discretizations as described in Section 5.1. For the finite element spaces we use the following notation. The finite-dimensional spaces V_{1h}, V_{2h}, and Q_h approximate the Sobolev spaces $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{Q}$, respectively. In the linear case, we use the notion $\mathcal{V}_1 = \mathcal{V}_\mathcal{B}$ and $\mathcal{V}_2 = \mathcal{V}^{\mathrm{c}}$. We emphasize that we do not assume the finite element spaces to be subspaces of the corresponding continuous space, i.e., we allow nonconforming discretization schemes. Furthermore, we set $V_h = V_{1h} \oplus V_{2h}$ as approximation space of $\mathcal{V}$. The dimensions of the finite element spaces are given by

$$\dim V_h = n, \qquad \dim V_{1h} = n - m, \qquad \dim V_{2h} = \dim Q_h = m.$$

Given appropriate basis functions, we represent the discrete approximations of u_1, u_2, v_2, and λ by the corresponding coefficient vectors $q_1 \in \mathbb{R}^{n-m}$, q_2, $p_2 \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^m$, respectively. Furthermore, we denote by $q \in \mathbb{R}^n$ the vector $q = [q_1^T, q_2^T]^T$.

Accordingly, the basis functions allow to obtain the finite-dimensional representation of the operators $\mathcal{K}$ and $\mathcal{B}$ as well as the right-hand sides $\mathcal{F}$ and $\mathcal{G}$, see Section 5.1. We denote the finite-dimensional representations by $K\colon \mathbb{R}^n \to \mathbb{R}^n$, $B\colon \mathbb{R}^n \to \mathbb{R}^m$, $f\colon [0,T] \to \mathbb{R}^n$, and $g\colon [0,T] \to \mathbb{R}^m$, respectively. The mass matrix, which corresponds to the identity operator, is denoted by $M \in \mathbb{R}^{n,n}$.

REMARK 8.1. The definition of the discrete operators requires that the continuous operators can be applied to the basis functions of the discrete spaces. Since we allow nonconforming discretizations, this is not automatically satisfied and has to be assumed. Usually, this is no restriction for the standard finite element spaces, since the basis functions are piecewise smooth (w.r.t. the triangulation $\mathcal{T}$) such that a piecewise application of the operators is possible.

A reasonable assumption on the resulting mass matrix $M \in \mathbb{R}^{n,n}$ is the positive definiteness. In the linear constraint case, where B is a (time-dependent) $m \times n$ matrix, we assume that it is continuously differentiable which corresponds to Assumption 6.2. As discussed in Section 5.1.3, the given saddle point structure requires stability conditions. Thus, we assume B to satisfy an inf-sup condition of the form: There exists a constant $\beta_{\text{disc}} > 0$, independent of h and time t, such that

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{\langle \mathcal{B} v_h, q_h \rangle}{\|v_h\|_{\mathcal{V}} \|q_h\|_{\mathcal{Q}}} = \beta_{\text{disc}} > 0.$$

Note that for the results of the following sections it is sufficient to assume a full rank property of B. However, we strongly encourage to use discretization schemes satisfying the stronger stability condition to ensure stable approximations of the Lagrange multiplier λ w.r.t. the discretization parameter h, see also the discussion in [**Bra07**, Chap. III.7].

We emphasize that there are two possible representations of B which we do not distinguish in the sequel. First, the matrix representation used above,

$$B\colon \mathbb{R}^n \to \mathbb{R}^m, \qquad \big\langle B(q), e_j \big\rangle := \big\langle \mathcal{B}(\textstyle\sum_{i=1}^n q_i \varphi_i), \psi_j \big\rangle.$$

Therein, e_j denotes the j-th cananical basis vector in $\mathbb{R}^n$. Second, B can be the discrete operator acting on the discrete functions instead of the coefficients,

$$B\colon V_h \to Q_h^*, \qquad \big\langle B v_h, q_h \big\rangle := \big\langle \mathcal{B} v_h, q_h \big\rangle. \tag{8.4}$$

In the case of nonlinear constraints, we state the above assumptions including the discrete inf-sup condition on the Jacobian of the constraint operator.

8.2. Linear Constraints. In the case of a linear constraint operator $\mathcal{B}(t)$, the semi-discretization of system (8.1) by finite elements as described above leads to the DAE

$$M\dot{q} + K(q) + B^T(t)\mu = f, \tag{8.5a}$$

$$B(t)q \qquad\qquad = g. \tag{8.5b}$$

Therein, $q = [q_i] \in \mathbb{R}^n$ denotes the coefficient vector corresponding to the finite element discretization of u and $\mu = [\mu_i] \in \mathbb{R}^m$ corresponds to the approximation of λ. Furthermore, we obtain an initial condition for $q(0) \in \mathbb{R}^n$. We show that this DAE has index 2 with the help of Lemma 2.2. For this, it is sufficient that the matrix $BM^{-1}B^T$ is invertible, which directly follows from the properties of M and B.

It remains to determine the index of the DAE which results form a spatial discretization of the regularized operator DAE (8.2). Here we consider two cases. First, the case of conforming discretizations for which we have $V_{1h} \subset \mathcal{V}_1 = \mathcal{V}_\mathcal{B}$, $V_{2h} \subset \mathcal{V}_2 = \mathcal{V}^c$, and $Q_h \subset \mathcal{Q}$. Second, the more general case where we do not state these assumptions, namely the nonconforming ansatz.

REMARK 8.2. Conforming finite element schemes are naturally more convenient for the analysis, since properties of the continuous spaces are transferred to the discrete setting. However, the necessary splitting of $\mathcal{V}$ into $\mathcal{V}_1$ and $\mathcal{V}_2$ demands for approximation spaces of rather complicated subspaces. Thus, the used discretization schemes are in general of nonconforming type. Note that $V_h = V_{1h} \oplus V_{2h}$ still may satisfy $V_h \subset \mathcal{V}$. For an example we refer to Section 8.4 below where the splitting is performed for flow equations.

8.2.1. *Conforming Discretization.* The assumption $V_{1h} \subset \mathcal{V}_1$ implies that the basis functions of V_{1h} vanish under the action of $\mathcal{B}$. Thus, the matrix $B(t)$ has the block structure $B(t) = [0 \;\; B_2(t)]$. Because of the full rank property of B and the dimensions of V_{2h} and Q_h, the matrix $B_2(t)$ is square and invertible. The semi-discrete analogue to the operator DAE (8.2) then reads

$$M \begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} + K(q_1, q_2) + \begin{bmatrix} 0 \\ B_2^T(t) \end{bmatrix} \mu = f, \tag{8.6a}$$

$$B_2(t) q_2 = g, \tag{8.6b}$$

$$B_2(t) p_2 = \dot{g} - \dot{B}_2(t) q_2. \tag{8.6c}$$

We postpone the proof that this system is of index 1 to the more general case in Lemma 8.3 below.

As a result, we have seen in Theorem 6.8 that the operator DAEs (8.1) and (8.2) are equivalent but result in DAEs of different index. The semi-discretization of the regularized operator equation leads to a DAE of index 1 rather than index 2 as for the original formulation. This fact explains why we call the procedure from Part B an index reduction on operator level. The resulting benefits of the regularized operator DAE (8.2) are strongly related to the general advantages of a low-index formulation. Recall that the index quantifies the level of difficulty of the numerical simulation due to instabilities resulting from hidden constraints and needed differentiation steps.

8.2.2. *Nonconforming Discretization.* As mentioned before, $V_{1h} \subset \mathcal{V}_1$ may not be a reasonable assumption for the discretization scheme. In this case, we loose the property $\ker B \not\subset \ker \mathcal{B}$, cf. [**GR86**, Ch. 3]. Clearly, this also affects the structure of the matrix B such that we do not obtain a zero-block as in the conforming case. Instead, we assume w.l.o.g. that the matrix B has the block structure

$$B(t) = \begin{bmatrix} B_1(t) & B_2(t) \end{bmatrix}$$

with a non-singular block $B_2(t)$. Note that this is no restriction, since the full rank property of B implies the existence of a regular block. Thus, the discrete spaces V_{1h} and V_{2h} can be chosen such that the matrix B has the given block structure. As a result, we obtain the same structural benefits as before in the conforming case. We consider the DAE resulting from a nonconforming discretization of system (8.2). For this, we add the vanishing terms $\mathcal{B}u_1$ and $\mathcal{B}\dot{u}_1$ to the constraints again, cf. Remark 7.5. The resulting

semi-discrete system has the form

$$M \begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} + K(q_1, q_2) \; + \; B^T(t)\mu = f, \tag{8.7a}$$

$$B_2(t)q_2 = g - B_1(t)q_1, \tag{8.7b}$$

$$B_2(t)p_2 = \dot{g} - B_1(t)\dot{q}_1 - \dot{B}_1(t)q_1 - \dot{B}_2(t)q_2. \tag{8.7c}$$

We show that this DAE (and thus, also system (8.6)) is of index 1.

LEMMA 8.3 (Index-1 DAE [**AH14**]). *For a positive definite mass matrix M and a continuously differentiable constraint matrix B with a non-singular block B_2, the DAEs* (8.6) *and* (8.7) *are of index 1.*

PROOF. Similar to the proof of [**KM06**, Th. 6.12], we show that (8.7) is of index 1. The property then follows for system (8.6) as well because it is a special case.

Since the matrix $B_2(t)$ is of full rank, equations (8.7b) and (8.7c) yield direct expressions of q_2 and p_2 in terms of q_1 and $\dot{q}_1$. Furthermore, a multiplication of (8.7a) from the left by BM^{-1} provides a formula for μ in terms of q_1. Here we use the assumptions on M and B which imply that the matrix $BM^{-1}B^T$ is invertible. Finally, inserting all these expressions into equation (8.7a), we obtain an ODE in q_1. Thus, we can solve system (8.7) without any further differentiation steps. □

REMARK 8.4 (Commutativity). The presented regularization process of Section 6 followed by a finite element discretization is equivalent to the traditional approach of first discretizing and then performing the index reduction. Thus, an application of minimal extension to the DAE (8.5) leads to the index-1 DAE (8.7), assuming that corresponding discretization schemes are used. For this, assume that the DAE (8.5) results from the discretization scheme $V_h = V_{1h} \oplus V_{2h}$, Q_h. The assumption on the structure of the matrix B, namely the invertibility of the B_2 block, then implies that the splitting $q = [q_1^T,\ q_2^T]^T$ according to $V_h = V_{1h} \oplus V_{2h}$ satisfies the conditions from Section 2.3.2. Thus, minimal extension with the dummy variable $p_2 := q_2$ yields the index-1 DAE (8.7).

8.3. Nonlinear Constraints. As mentioned before, the assumed properties of B now pass to the Jacobian of the constraint. For this, we prefer to work with the discrete operator B in the form of (8.4). Similar to the assumptions on the nonlinear constraint operator in Assumption 6.16 for the regularization of the operator DAE, we formulate sufficient assumptions on B which ensure that the semi-discretized system is of index 1. Also here the assumptions are strongly connected to the implicit function theorem.

ASSUMPTION 8.5 (Properties of B [**AH14**]). *Consider $u_h \in V_h$ such that $\langle Bu_h, \psi_j\rangle = \langle \mathcal{G}, \psi_j\rangle$ for $j = 1, \ldots, m$. We assume that*

(a) there exist subspaces V_{1h} and V_{2h} with $V_h = V_{1h} \oplus V_{2h}$, $u_h = u_{h,1} + u_{h,2}$,

(b) B is continuously differentiable in a neighborhood of u_h,

(c) the matrix corresponding to $\partial B/\partial u_{h,2}(u_h)$ is invertible.

REMARK 8.6. In view of Assumption 8.5 with the splitting $V_h = V_{1h} \oplus V_{2h}$, we may assume an appropriate ordering of the basis $\{\varphi_i\}_{i=1,\ldots,n}$ of V_h which implies a decomposition of the coefficient vector q. More precisely, the coefficient vector $q \in \mathbb{R}^n$ decomposes into $q = [q_1^T, q_2^T]^T$ with $u_{h,1} = \sum_{i=1}^{n-m} q_{1i}\varphi_i \in V_{1h}$ and $u_{h,2} = \sum_{i=1}^{m} q_{2i}\varphi_{n-m+i} \in V_{2h}$.

To simplify notation, we introduce C as the Jacobian of B, i.e., $C := \partial B/\partial u_h(u_h)$. Note that the discretization of the Fréchet derivative $\partial\mathcal{B}/\partial u(\cdot)\colon \mathcal{V} \to (\mathcal{V} \to \mathcal{Q}^*)$ equals the

Jacobian of the discrete operator B. The spatial discretization of system (6.8) then leads to the system

$$M\dot{q} + K(q) + C^T\mu = f, \tag{8.8a}$$

$$B(q) = g. \tag{8.8b}$$

Following Lemma 2.2, we note that the invertibility of $CM^{-1}C^T$, which follows from the full rank property of the Jacobian, implies that the DAE (8.8) is of index 2. The same discretization scheme with the splitting $V_h = V_{1h} \oplus V_{2h}$ from Assumption 8.5 and basis functions as described in Remark 8.6 is applied to system (8.3) and yields

$$M \begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} + K(q_1, q_2) + C^T\mu = f, \tag{8.9a}$$

$$B(q_1, q_2) = g, \tag{8.9b}$$

$$C \begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} = \dot{g}. \tag{8.9c}$$

It remains to show that this DAE has index 1. This then justifies the regularization procedure of Part B also for the case with nonlinear constraints.

LEMMA 8.7 (Index-1 DAE [**AH14**]). *Let M be positive definite and the nonlinear function B satisfy Assumption 8.5 along u_h which corresponds to the solution q of* (8.9). *Then, the DAE* (8.9) *is of index 1.*

PROOF. Assumption 8.5 implies that the Jacobian C has the block structure

$$C = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$$

with an invertible matrix C_2. Because of the implicit function theorem, locally we obtain from (8.9b-c) expressions for q_2 and p_2 in terms of q_1 and $\dot{q}_1$. Furthermore, a multiplication of CM^{-1} from the left to (8.9a) yields an equation for μ in terms of q_1 and the right-hand side $\dot{g}$. Here we use the full rank property of C which implies that $CM^{-1}C^T$ is non-singular. Finally, we have algebraic equations for q_2, p_2, and μ and an ODE for q_1 without the application of any differentiation steps. □

8.4. Application to Flow Equations. The dynamics of incompressible flows are characterized by the Navier-Stokes equations or a corresponding linearized version such as the Stokes or Oseen equations, cf. Section 6.3.1. In any case, a spatial discretization by finite elements as described in the first part of this section leads to a DAE of the form

$$M\dot{q} + K(q) + B^T p = f, \qquad Bq = 0. \tag{8.10}$$

Due to the large interest in industrial applications, there exist several approaches in the field of computational fluid dynamics which can be roughly grouped in

- penalty methods [**HV95, She95**],
- pressure correction or projection methods [**GS00**], and
- methods using divergence-free discretization schemes.

A summary of those methods can also be found in [**Wei96**]. All these methods pay special attention to the treatment of the pressure variable. In fact, all these approaches try to avoid the instabilities coming form the index-2 structure of the given problem. Recall that flow equations have a saddle point structure and the pressure takes the role of the Lagrange multiplier, cf. Section 5.1.3. The first two methods decouple the velocity and pressure variable. Although this decoupling seems to be computationally beneficial because of the

splitting into smaller subsystems, the computation of the pressure from the velocity is ill-conditioned, since it involves a multiplication by the discrete divergence operator. This includes a factor h^{-1} where h denotes the spatial mesh parameter. Particularly, this ansatz is unfeasible if the coupling conditions include the pressure variable as, e.g., in levitated droplet problems in which one considers the effect of the surface tension on a fluid interface [**BKZ92, EGR10**].

The idea of *penalty methods* is to replace the discrete divergence constraint by $Bq = -\lambda^{-1}p$ with a penalty parameter $\lambda \gg 1$. Note that this penalization reduces the index of the DAE (8.10) to one and yields an ODE for the velocity, namely $M\dot{q} + K(q) - \lambda B^T Bq = f$. To gain an approximation of the pressure, one can use the so-called *pressure Poisson equation* including the calculated velocity [**SH90**]. The main disadvantages of this approach are the degenerating condition number of the resulting algebraic system and the difficulties for small velocities [**HV95**]. Furthermore, the method requires a suitable value for the (heuristic) penalty parameter λ [**She95**].

Projection methods are based on a guess for the pressure which is used to compute an approximate velocity update $\tilde{q}$ by equation (8.10). This step needs to be performed in every time step. Then, one splits $\tilde{q}$ in a discrete divergence-free component and a complement which results in an incomplete decoupling [**Wei97**]. Furthermore, this ansatz calls for boundary conditions for the pressure which are unphysical [**GS00**].

From a theoretical point of view, a complete decoupling by the use of divergence-free elements is optimal, since the DAE (8.10) automatically turns to an ODE. Hence, the discretization scheme would work on the subspace of divergence-free functions and yield an approximation which satisfies the algebraic constraint a priori. However, these methods are only rarely used because of the high computational costs. In order to avoid expensive computations, one may also use quasi divergence-free elements, see e.g. [**MS06**]. Unfortunately, this ansatz only reduces the size of the algebraic system but does not change the high-index structure of the DAE. Furthermore, divergence-free methods are restricted to constraints with a homogeneous right-hand side.

Within this thesis, we propose to combine the index-1 formulation of the flow equations with a decomposition of the finite element space used for the velocity approximation. This method does not depend on any heuristic parameter and requires no unfeasible boundary conditions for the pressure, since the equations are not decoupled. Thus, the pressure variable p remains part of the system. Furthermore, this approach is consistent with the infinite-dimensional setting in the sense that it has a valid representation on operator level [**AH13**]. Hence, there is no restriction on the size of the time steps.

Recall that the index reduction procedure (in the finite- as well as in the infinite-dimensional setting) adds the so-called hidden constraint $B\dot{q} = 0$ to the system which reduces instabilities. This gains particularly robustness w.r.t. perturbations in the right-hand side as they may appear due to the inexact solution of the algebraic equations. In addition, we stress the fact that the method does not rely on the vanishing divergence and allows for constraints of the form $Bq = g \neq 0$, see Section 6.3.2 for an example.

Throughout this section, we assume the computational domain $\Omega \subset \mathbb{R}^2$ to be connected with Lipschitz boundary. As in Section 5.1, we concentrate on the two-dimensional case but comment on the extension to three space dimensions. Let $\mathcal{T}$ denote a regular triangulation and $\mathcal{E}$ the set of edges, including the interior edges $\mathcal{E}_{\text{int}}$. A number of stable discretization schemes in the sense of a discrete inf-sup condition were already discussed in Section 5.1.3. Here, we focus on two examples for which we illustrate how to find the required decomposition of V_h which satisfies the properties stated in Section 8.2.

The general advance is to take a stable discretization scheme V_h, Q_h and construct the subspaces V_{1h} and V_{2h} of V_h. In other words, the task is to find a regular block in the matrix B which equals the discrete divergence operator. Note that the direct approach using a QR decomposition is not applicable for large systems, since it does not benefit from the sparsity of B. Instead of such an algebraic approach, we aim to find suitable subspaces based on the geometry given by the triangulation $\mathcal{T}$. Since the divergence constraint is not local (as e.g. a constraint on the boundary) the subspace V_{2h} cannot be read off directly. Since we exclude divergence-free elements in this discussion, the decomposition of V_h directly leads to discretizations of nonconforming nature.

As discussed in Section 8.2, we desire a splitting of V_h such that the resulting constraint matrix B has the block structure $B = [B_1 \;\; B_2]$. Therein, the square block B_2 corresponds to the subspace V_{2h} and is required to be non-singular. The aim of the following subsections is to find such a decomposition for specific discretization schemes used in computational fluid dynamics. This then guarantees the applicability of Lemma 8.3 and thus, the index-1 property of the resulting DAE. As shown in [**AH13**] this allows to apply half-explicit discretization schemes, i.e., to discretize the differential part of the system with an explicit scheme.

8.4.1. *Decomposition for Crouzeix-Raviart Elements.* The mixed scheme of Crouzeix and Raviart introduced in (5.8) of Section 5.1.3 is the most popular scheme of nonconforming type. This low-order scheme turns out to provide an efficient tool in computational fluid dynamics [**BM11**]. Recall that the ansatz functions of $\mathrm{CR}_0(\mathcal{T})$ are edge-oriented, i.e., the degrees of freedom are located on the edges rather than to the nodes of the triangulation. The basis functions in two dimensions are of the form $[\phi_E,\ 0]^T$ and $[0,\ \phi_E]^T$ where ϕ_E denotes a Crouzeix-Raviart basis function. Since we consider the space $\mathcal{P}_0(\mathcal{T})/\mathbb{R}$ for the pressure approximation, we choose one triangle, namely $T_0 \in \mathcal{T}$, where the pressure is fixed. This compensates the fact that the governing equations only determine the pressure up to a constant.

We define a mapping $\iota : \mathcal{T} \setminus \{T_0\} \to \mathcal{E}_{\mathrm{int}}$ which will provide a suitable way to find proper basis functions which then span the space V_{2h}. More precisely, the basis functions of interest are the functions corresponding to the edges in the range of ι. The definition of ι is part of the following algorithm, see also the illustration shown in Figure 8.1.

Algorithm 8.8 (Mapping ι [**AH13**]). Step 1: *Choose any $T \in \mathcal{T} \setminus \{T_0\}$ which shares an edge with T_0 and denote this edge by $E := T_0 \cap T \in \mathcal{E}_{int}$. Then, define $\iota(T) := E$ and $\mathcal{T}_R := \mathcal{T} \setminus \{T_0, T\}$. If $\mathcal{T}_R = \emptyset$, then stop.*
Step 2: *If T from the previous step has an edge-neighbour in $\mathcal{T}_R$, then continue with Step 2a. Otherwise, go to Step 2b.*
Step 2a: *Select such a neighbouring triangle $S \in \mathcal{T}_R$ and set $E := T \cap S \in \mathcal{E}_{int}$. Furthermore, set $\iota(S) := E$ and $\mathcal{T}_R := \mathcal{T}_R \setminus \{S\}$. If $\mathcal{T}_R = \emptyset$, then stop. Otherwise, return to Step 2 with $T := S$.*
Step 2b: *Reset $T \in \mathcal{T} \setminus \mathcal{T}_R$ such that there exists an edge-neighbour in $\mathcal{T}_R$ and return to Step 2.*

Remark 8.9. We show that Algorithm 8.8 terminates in a finite number of steps. Step 2a always reduces the (finite) set of triangles $\mathcal{T}_R$ by one and Step 2b is realizable, since $\mathcal{T}_R \neq \emptyset$ and the domain Ω is assumed to be connected with Lipschitz boundary.

Algorithm 8.8 provides besides the mapping ι also an order of the triangles in $\mathcal{T}$. For this, we number consecutively the triangles by their first appearance in the algorithm and obtain $\{T_j\}_{j=1,\dots,|\mathcal{T}|-1}$.

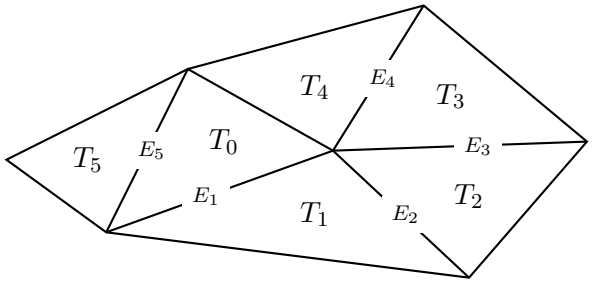


FIGURE 8.1. Illustration of Algorithm 8.8, $\iota(T_i) = E_i$ for $i = 1, \dots, 5$. Step 2b of the algorithm is applied once to reset $T := T_0$.

Consider an edge $E \in \operatorname{range}(\iota) \subset \mathcal{E}_{\text{int}}$ with corresponding Crouzeix-Raviart basis function ϕ_E. The construction given in Algorithm 8.8 then implies that $T = \iota^{-1}(E) \subset \operatorname{supp} \phi_E$ and thus, $\phi_E|_T$ is not constant. As a result, $\nabla \phi_E|_T \neq 0$ which means that either $\partial_x \phi_E|_T$ or $\partial_y \phi_E|_T$ does not vanish. In other words, the divergence of the ansatz function $[\phi_E,\ 0]^T$ or $[0,\ \phi_E]^T$ is constant but nonzero. Let Φ_E denote one of these two functions with $\operatorname{div}(\Phi_E|_T) \neq 0$. Repeating this procedure for all edges in $\operatorname{range}(\iota)$, we gain the ansatz space

$$V_{2h} := \operatorname{span}\{\Phi_E \mid E \in \operatorname{range}(\iota)\}. \tag{8.11}$$

The span of the remaining basis functions of V_h defines the subspace V_{1h} such that we have found a decomposition of the discrete space V_h. It remains to show that this decomposition satisfies the requested properties.

LEMMA 8.10 (Decomposition for Crouzeix-Raviart [**AH13**]). *The discretization scheme V_h, Q_h from* (5.8) *with the decomposition $V_h = V_{1h} \oplus V_{2h}$ defined in* (8.11) *yields the required block structure of B, i.e., $B = [B_1\ B_2]$ with a non-singular matrix B_2.*

PROOF. The matrix $B_2 \in \mathbb{R}^{m,m}$ corresponds to the discrete space V_{2h} and is defined by

$$B_{2,ij} = \int_\Omega \chi_i \operatorname{div} \Phi_j \, \mathrm{d}x = \int_{T_i} \operatorname{div} \Phi_j \, \mathrm{d}x.$$

Therein, $\{\Phi_j\}_{j=1,\dots,m}$ denote the basis functions of V_{2h} and $\{\chi_i\}_{i=1,\dots,m}$ the basis functions of Q_h, i.e., $\chi_i = 1$ on the triangle T_i and zero elsewhere. Since $\operatorname{div} \Phi_i \neq 0$ on T_i by construction, the diagonal entries of B_2 are nonzero. Furthermore, every column can only have two entries because of the support of edge-bubble functions. By the construction of Algorithm 8.8, the second entry can only be above the diagonal and thus, B_2 is upper triangular and non-singular. □

Lemma 8.10 shows that the regularization performed in Section 6.1 together with the splitting of V_h given by Algorithm 8.8 yields a stable numerical scheme. The proposed splitting is used in the numerical example in Section 8.4.4 below.

REMARK 8.11 (Condition number). The condition number of the matrix B_2 obtained by Algorithm 8.8 and Lemma 8.10 scales as h^{-1} where h denotes the mesh size. For a uniform mesh of the unit square where Algorithm 8.8 runs without reset, i.e., without

entering step 2b, the matrix B_2 has the structure

$$B_2 = \begin{bmatrix} h & h & & \\ & \ddots & \ddots & \\ & & \ddots & h \\ & & & h \end{bmatrix}, \qquad B_2 B_2^T = h \begin{bmatrix} h & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & h \end{bmatrix}.$$

The eigenvalues of $h^{-1} B_2 B_2^T$ are given by

$$\lambda_j = h + 2\cos\big(j\pi h^2/2\big), \quad j = 1, \dots, n = 2h^{-2} - 1.$$

Hence, a rough estimate of the condition number yields

$$\operatorname{cond} B_2 = \frac{\lambda_{\max}}{\lambda_{\min}} \approx \frac{h+2}{h} \approx \frac{2}{h}.$$

Note, however, that a degeneration of the mesh may lead to large deviations.

REMARK 8.12 (Outflow boundary conditions). For flow problems that have an outflow with *homogeneous Neumann* or *do nothing* boundary conditions, the pressure must not be fixed on T_0. In this case, Algorithm 8.8 defines V_{2h} if one starts with a T_0 that shares an edge E_0 with the outflow boundary. Then, the inclusion of χ_0 leads to a block $B_2 \in \mathbb{R}^{m,m-1}$ that is in *Hessenberg form* with the last column missing and with nonzero entries on the subdiagonal. Adding the basis funciton Φ_{E_0} to V_{2h}, for which $\operatorname{div}(\Phi_{E_0}|_{T_0}) \neq 0$, we add a column that is zero except from the first row's entry and that makes B_2 square and invertible.

REMARK 8.13 (Extension to three space dimensions). As discussed in Section 5.1, the finite element spaces V_h and Q_h of this subsection have a straightforward analogue in three space dimensions [**CR73**]. Also Algorithm 8.8 can easily be adapted by the use of tetrahedra and faces in place of triangles and edges. Hence, the given results also apply to three-dimensional simulations.

8.4.2. *Decomposition for Bernardi-Raugel Elements.* As second example we consider a mixed scheme of conforming type with a continuous approximation of the velocity. The discrete spaces V_h, Q_h of Bernardi-Raugel were introduced in (5.7) of Section 5.1.1. Recall that the space V_h is composed by hat-functions and vector-valued edge-bubble functions of the form $\Upsilon_E := \varphi_1\varphi_2\nu_E$. For the decomposition of V_h we propose to span V_{2h} by a number of edge-bubble functions. For this, we make again use of Algorithm 8.8 and the resulting map $\iota : \mathcal{T} \setminus \{T_0\} \to \mathcal{E}_{\text{int}}$. We define the subspace

$$V_{2h} = \operatorname{span}\{\Upsilon_E \mid E \in \operatorname{range}(\iota)\}. \tag{8.12}$$

The complement V_{1h} is defined as the span of the remaining basis functions of V_h. We show that this decomposition fulfills the desired properties.

LEMMA 8.14 (Decomposition for Bernardi-Raugel [**AH13**]). *The discretization scheme V_h, Q_h from* (5.7) *with the given decomposition $V_h = V_{1h} \oplus V_{2h}$ defined in* (8.12) *yields the required block structure of B, i.e., $B = [B_1\ B_2]$ with a non-singular matrix B_2.*

PROOF. Note that the structure of B_2 is as in Lemma 8.10. Thus, it remains to show that the integral of $\operatorname{div} \Upsilon_E$ for $\Upsilon_E \in V_{2h}$ does not vanish. By definition of Υ_E, it holds that

$$\operatorname{div} \Upsilon_E = \nabla(\varphi_1\varphi_2) \cdot \nu_E = \varphi_1 \nabla\varphi_2 \cdot \nu_E + \varphi_2 \nabla\varphi_1 \cdot \nu_E.$$

Hence, for a triangle T with edge E,

$$\begin{aligned}\int_T \operatorname{div} \Upsilon_E \,\mathrm{d}x &= \nabla\varphi_2 \cdot \nu_E \int_T \varphi_1 \,\mathrm{d}x + \nabla\varphi_1 \cdot \nu_E \int_T \varphi_2 \,\mathrm{d}x \\ &= \frac{|T|}{3}\big(\nabla\varphi_2 + \nabla\varphi_1\big) \cdot \nu_E.\end{aligned}$$

Let $[x_i, y_i]^T$, $i = 1, 2$, denote the coordinates of the nodes corresponding to φ_1 and φ_2, respectively. W.l.o.g, we assume that the third node is located in $[0, 0]^T$. Then, the outer normal vector for E is, up to a constant, given by $\nu_E = [y_1 - y_2,\ x_2 - x_1]^T$. The hat-functions are defined by

$$\varphi_1(x, y) = \frac{1}{d}\big(y_2 x - x_2 y\big), \qquad \varphi_2(x, y) = \frac{1}{d}\big(-y_1 x + x_1 y\big)$$

with $d = x_1 y_2 - x_2 y_1 \neq 0$, since the triangle is part of a regular triangulation. Thus, we obtain

$$\big(\nabla\varphi_2 + \nabla\varphi_1\big) \cdot \nu_E = -\frac{1}{d}\big((x_1 - x_2)^2 + (y_1 - y_2)^2\big) \neq 0$$

and therefore the claim $\int_T \operatorname{div} \Upsilon_E \,\mathrm{d}x \neq 0$. □

Remark 8.15 (Extension to three space dimensions). In Section 5.1 we have discussed the extension of the Bernardi-Raugel elements to the three-dimensional case [**BR85**]. As in Lemma 8.14, one can show that the integral of the divergence of the basis functions does not vanish on certain tetrahedra. The full-rank property of B_2 then follows as in the two-dimensional case.

8.4.3. *Further Elements.* In Section 5.1.3 we also mentioned the Taylor-Hood approach with continuous pressure approximation and higher order velocity fields. At this point we briefly review the decomposition for this scheme and refer to the details given in Section 3.5 of [**AH13**].

The decomposition is based on the idea of macro elements, i.e., the triangulation is grouped into sub-triangulations. Each macro element contains exactly one interior node with all its adjacent triangles. Then, a special algorithm defines a mapping $\jmath\colon \mathcal{N} \setminus \{v_0\} \to \mathcal{E}_{\text{int}}$, where v_0 denotes the node where the pressure is fixed, similar to the triangle T_0 in the previous sections. Edge-bubble functions corresponding to the range of $\jmath$ are chosen to span the subspace V_{2h}. The particular choice depends on the angles between the underlying edge and the axis of the coordinate system.

The proposed methods based on a splitting of the discrete velocity space V_h also work with triangulations $\mathcal{T}$ containing quadrilaterals. For this, consider a partition of $\bar{\Omega}$ into convex quadrilaterals. The here presented schemes of Taylor-Hood type and Bernardi-Raugel have a direct analogon for such cases, see [**GR86**, Ch. II.3]. The analogue of the discontinuous approach of Crouzeix-Raviart was introduced by Rannacher and Turek [**RT92**] and is given by

$$V_h = [\tilde{\mathcal{Q}}_{1,0}(\mathcal{T})]^2, \qquad Q_h = \mathcal{P}_0(\mathcal{T})/\mathbb{R}.$$

Therein, $\tilde{\mathcal{Q}}_{1,0}$ denotes the nonconforming space of piecewise polynomials of partial degree 1 which are globally continuous. This space has one degree of freedom per interior edge but is, in contrast to the Crouzeix-Raviart element, not piecewise affine. Nevertheless, the decomposition of V_h works exactly as in the triangular case with the help of Algorithm 8.8. Note that this nonconforming scheme was found superior over comparable conforming elements in terms of stability, accuracy, and efficiency [**Tur99**, Ch. 3.1.1]. The main reason for this is the given robustness of the discrete inf-sup constant against mesh deformations.

For more complicated schemes or discretizations of higher order the search of a good subspace V_{2h} may be very complex. In this case, one may favor to find a splitting of V_h in an algebraic manner e.g. by methods discussed in [**GOS$^+$10**].

8.4.4. *Numerical Example.* We consider the Navier-Stokes equations for the simulation of a cylinder wake as illustrated in Figure 8.2. We refer to Section 6.3.1 for the system equations. As boundary conditions we set *no-slip* conditions at the walls, a parabola as the inflow profile at the left boundary, and *do-nothing* conditions at the outflow at the right. We consider the flow at Reynolds number $Re = 60$, calculated with the cylinder diameter and the peak inflow velocity. We consider the time evolution of the flow in the time interval $[0, 0.2]$, starting with the steady-state Stokes solution.

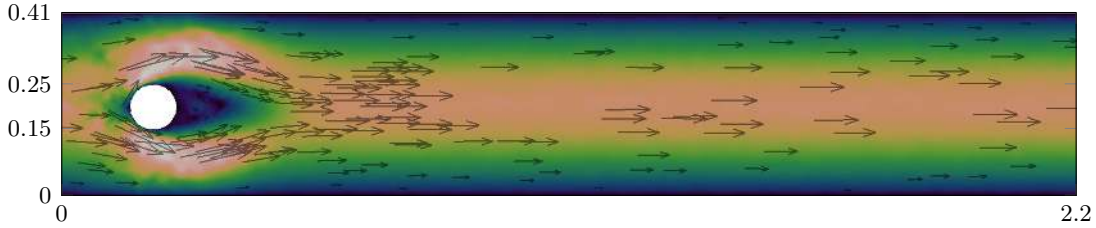


FIGURE 8.2. Illustration of the cylinder wake with Reynolds number $Re = 60$ at time $t = 0.2$, started at the steady-state Stokes solution.

For the spatial discretization, we use Crouzeix-Raviart elements from Section 5.1.1 on a nonuniform mesh with about 15000 velocity nodes and 5000 pressure nodes. We employ Algorithm 8.8 with the modification proposed in Remark 8.12 to compute the splitting $V_h = V_{1h} \oplus V_{2h}$ that we need for the index-1 formulation.

To account for the included stiffness in the system equations, we consider an *implicit-explicit* Euler scheme for the discretization which treats the linear diffusion implicitly and the nonlinear convection explicitly. We compute the approximation error for various time step sizes and for various accuracy levels tol for the iterative solution of the resulting linear systems. Since there is no analytical solution, we take the result of solving the spatial discretized Navier-Stokes equations with the implicit trapezoidal rule with direct solves and with step size $\tau = 0.2 \cdot 2^{-11} \approx 10^{-4}$ as the reference solution.

The results of the numerical investigation are illustrated in Figure 8.3. They clearly show the improvements of the index-1 formulation (right) for the pressure approximation. However, since for the cylinder wake the velocity is not discretely divergence free, i.e. due to a non-vanishing right-hand side, the poor pressure approximation directly affects the velocity approximation. As predicted by the theoretical considerations in [**AH13**], in the index-2 formulation, a numerical error in the algebraic constraints leads to a linear growth in the pressure error with decreasing time step sizes τ. A smaller residual in the continuity equation only postpones this instability. In the index-1 formulation, this systematic instability is not observed and we obtain the expected linear convergence with respect to the time discretization for the velocity and the pressure approximation. A breakdown due to the algebraic error is only observed for a rough tolerance for the linear solver.

The code used for the numerical investigations is available from the github account [**Hei15**]. The finite element implementation uses *FEniCS, Version 1.3.0* [**LORW12**] and the linear systems are solved with *Krypy* [**Gau14**].

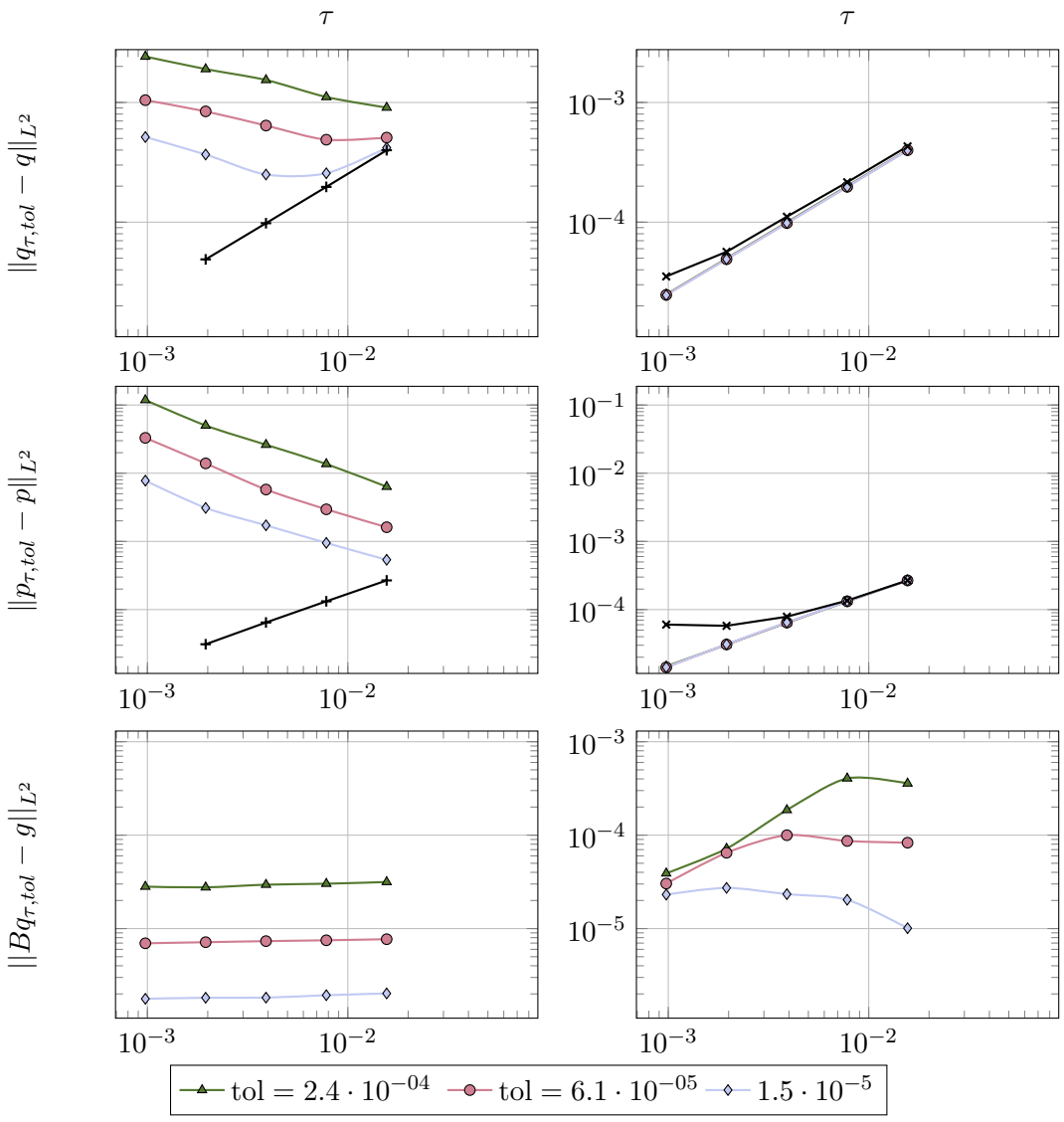


FIGURE 8.3. The evolution of the errors in the velocity (top) and the pressure (middle) or residuals of the constraint (bottom) of the index-2 (left) and index-1 (right) formulation for varying time discretization parameter τ and tol for the cylinder wake. The additional data points for the index-1 case are calculated for the much rougher tolerance tol $= 3.9 \cdot 10^{-3}$. The additional data points in the index-2 plots are the results for exact solutions of the algebraic equations.

9. The Method of Lines for Second-order Systems

Similar to the previous section, we show that the regularization presented in Section 7 can be interpreted as an index reduction on operator level. For this, we show that a spatial discretization of system (7.14) by finite elements leads to a DAE of index 1 rather than index 3 as for the original system (7.12). Furthermore, an appropriate choice of the finite element spaces results in the commutativity of semi-discretization and index reduction. Thus, the suggested reformulation of the operator DAE is eligible for adaptive simulations, since it allows to modify the triangulation as well as the discrete ansatz spaces. In contrast to the original formulation, changes of the spatial discretization scheme do not call for another index reduction step afterwards.

The results of this section are part of [**Alt13a**].

9.1. Recap and Notation. In this subsection we recall the operator equations which we want to discretize in space and recapitulate the most important properties of the applied finite element schemes. The dynamics of elastic media from Section 7.1.3 lead to the operator DAE (7.12) which has the form

$$\mathcal{M}\ddot{u}(t) + \mathcal{D}\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{9.1a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*. \tag{9.1b}$$

The regularization procedure of Section 7.2 then results in the extended system (7.14). As mentioned in Remark 7.5, we do not change the system if we include the vanishing terms $\mathcal{B}u_1$ and its derivatives. Since we allow nonconforming discretization schemes for which the discretization of u_1 may not vanish under the action of $\mathcal{B}$, we add these terms here. Hence, assuming sufficient regularity of u_1, we consider the operator DAE

$$\mathcal{M}(\ddot{u}_1 + w_2) + \mathcal{D}(\dot{u}_1 + v_2) + \mathcal{K}(u_1 + u_2) + \mathcal{B}^*\lambda = \mathcal{F} \quad \text{in } \mathcal{V}^*, \tag{9.2a}$$

$$\mathcal{B}(u_1 + u_2) = \mathcal{G} \quad \text{in } \mathcal{Q}^*, \tag{9.2b}$$

$$\mathcal{B}(\dot{u}_1 + v_2) = \dot{\mathcal{G}} \quad \text{in } \mathcal{Q}^*, \tag{9.2c}$$

$$\mathcal{B}(\ddot{u}_1 + w_2) = \ddot{\mathcal{G}} \quad \text{in } \mathcal{Q}^*. \tag{9.2d}$$

As before, the finite element method based on a triangulation $\mathcal{T}$ is used for the spatial discretization, see Section 5.1. This then leads to finite-dimensional approximation spaces of $\mathcal{V}$, its subspaces $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^\mathrm{c}$, as well as of $\mathcal{Q}$. As in the previous section, theses spaces are denoted by $V_h = V_{1h} \oplus V_{2h}$ and Q_h. Thereby, V_{1h} denotes the approximation space of $\mathcal{V}_1 := \mathcal{V}_\mathcal{B}$ and V_{2h} of $\mathcal{V}_2 := \mathcal{V}^\mathrm{c}$. Recall that we do not assume the finite-dimensional spaces to be subspaces of its continuous analogue. The dimensions are given by

$$\dim V_h = n, \qquad \dim V_{1h} = n - m, \qquad \dim V_{2h} = \dim Q_h = m.$$

In the sequel, we always assume that the discretization scheme satisfies a discrete inf-sup condition, see Section 5.1.3. An example of a stable scheme is given in Lemma 5.4.

The spatial discretization turns the operator DAEs (9.1) and (9.2) into semi-explicit (nonlinear) DAEs in terms of the coefficient vectors w.r.t. a given basis. Let $M \in \mathbb{R}^{n,n}$ be the resulting mass matrix, $D\colon \mathbb{R}^n \to \mathbb{R}^n$ the discrete damping function, $K \in \mathbb{R}^{n,n}$ the stiffness matrix, and $B \in \mathbb{R}^{m,n}$ the constraint matrix, cf. Section 5.1.2. Since the density ρ is assumed to be positive and the discretization scheme is stable, M is positive definite and B is of full rank. The discretized right-hand sides are denoted by f, g, $\dot{g}$, and $\ddot{g}$.

9.2. Determination of the Index. First, we analyse the index of the DAE coming from the spatial discretization of system (9.1). Let $q = [q_i] \in \mathbb{R}^n$ denote the coefficient vector of the finite element approximation of u and $\mu = [\mu_i] \in \mathbb{R}^m$ the corresponding vector for λ. Then, the semi-discrete problem can be written in the form

$$M\ddot{q}(t) + D(\dot{q}(t)) + Kq(t) + B^T\mu(t) = f(t), \tag{9.3a}$$

$$Bq(t) = g(t). \tag{9.3b}$$

In addition, the initial conditions for u and $\dot{u}$ provide initial conditions for q and $\dot{q}$. Regardless of the damping and stiffness terms, this DAE is of index 3. Note that the DAE has the typical structure of a constrained multibody system because of the positive definiteness of M and the full rank property of B, cf. Lemma 2.3.

The index-3 property can also be seen by a double differentiation of the algebraic constraint, namely $B\ddot{q} = \ddot{g}$. Replacing the algebraic constraint by this derivative, we can write the DAE in the form

$$\begin{bmatrix} M & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} \ddot{q} \\ \mu \end{bmatrix} = \begin{bmatrix} f - D(\dot{q}) - Kq \\ \ddot{g} \end{bmatrix}.$$

The properties of M and B then imply that the matrix on the left-hand side is invertible. Thus, the system decouples in an algebraic equation for μ and an ODE in q. One additional differentiation then provides an ODE for q and μ. Thus, the DAE is of index 3 according to Section 2.

Second, we analyse the DAE resulting from a spatial discretization of system (9.2). At this point we need the finite-dimensional approximations V_{1h} and V_{2h} of the spaces $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^\mathrm{c}$, respectively.

Example 9.1. Consider the stable scheme from Section 5.1.3, i.e., V_h contains the hat functions and edge-bubble functions at the boundary and Q_h is given by piecewise constant functions along the boundary,

$$V_h = [\mathcal{S}_{1,0}(\mathcal{T})]^2 \oplus [\mathcal{B}_\Gamma(\mathcal{T})]^2, \qquad Q_h = [\mathcal{P}_0(\mathcal{T})|_\Gamma]^2.$$

Then, one possible splitting of V_h is given by $V_{1h} := [\mathcal{S}_{1,0}(\mathcal{T})]^2$ and $V_{2h} := [\mathcal{B}_\Gamma(\mathcal{T})]^2$. Since the space $\mathcal{B}_\Gamma(\mathcal{T})$ contains one edge-bubble function per boundary edge, we have $\dim V_{2h} = \dim Q_h$. We emphasize that neither V_{1h} is a subspace of $\mathcal{V}_\mathcal{B}$ nor V_{2h} is a subspace of $\mathcal{V}^\mathrm{c}$. Thus, we obtain a nonconforming finite element scheme although $V_h \subset \mathcal{V}$.

As already mentioned in the previous section, the discretizations of the subspaces $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^\mathrm{c}$ are often of nonconforming type. This may seem contradictory at first but provides numerical benefits as we reduce the index of the DAE and thus, avoid singularities. Furthermore, the splitting appears naturally in the sense that $u_h \in V_h$ is equivalent to the existence of $u_{1,h} \in V_{1h}$ and $u_{2,h} \in V_{2h}$ with $u_h = u_{1,h} + u_{2,h}$.

In Section 9.3 we show that this splitting (of the deformation variable) corresponds to the needed splitting in the minimal extension procedure. Although we do not assume that $V_{2h} \subset \mathcal{V}^\mathrm{c}$, we still need that V_{2h} approximates $\mathcal{V}^\mathrm{c}$ sufficiently well. To be precise, we assume that the matrix B has the corresponding block structure $B = [B_1\ B_2]$ with an invertible matrix $B_2 \in \mathbb{R}^{m,m}$. This property is crucial to guarantee that the semi-discretization leads to a DAE of index 1.

Given appropriate basis functions, we represent the discrete approximations of u_1, u_2, v_2, w_2, and λ by the corresponding coefficient vectors $q_1 \in \mathbb{R}^{n-m}$, q_2, p_2, $r_2 \in \mathbb{R}^m$, and $\mu \in \mathbb{R}^m$, respectively. Then, the discrete variational formulation is equivalent to the DAE

$$M \begin{bmatrix} \ddot{q}_1 \\ r_2 \end{bmatrix} + D\left(\begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} \right) + K \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} + B^T \mu = f, \tag{9.4a}$$

$$B \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = g, \tag{9.4b}$$

$$B \begin{bmatrix} \dot{q}_1 \\ p_2 \end{bmatrix} = \dot{g}, \tag{9.4c}$$

$$B \begin{bmatrix} \ddot{q}_1 \\ r_2 \end{bmatrix} = \ddot{g}. \tag{9.4d}$$

We postpone the proof that the DAE (9.4) has index 1 to the following subsection. Therein, we show that this DAE equals the system one obtains by the application of minimal extension to the DAE (9.3) which is known to be of index 1. This then justifies the regularization presented in Section 7 as well as calling this procedure an index reduction on operator level.

9.3. Commutativity. We apply the index reduction method of minimal extension from Section 2.3.2 to the DAE (9.3). The aim of this subsection is to show that this leads to the DAE (9.4) which we have obtained from the spatial discretization of the regularized operator DAE. This then shows that regularization (respectively index reduction) and spatial discretization commute if we use assortative finite elements schemes.

Following the procedure of Section 2.3.2, we need to find a transformation of variables in order to find a regular block of the constraint matrix B. This choice is not unique but with regard to the previous subsection there is a canonical selection. If we assume that the basis functions of V_h are sorted such that the first $n-m$ functions form a basis of V_{1h} and the last m a basis of V_{2h}, then the last m columns of B are linearly independent.

Thus, the partition of variables is simply given by $q^T = [q_1^T \; q_2^T]$ with $q_1 \in \mathbb{R}^{n-m}$ and $q_2 \in \mathbb{R}^m$. Next, we add the two hidden constraints corresponding to equation (9.3b), i.e.,

$$B\dot{q} = \dot{g}, \qquad B\ddot{q} = \ddot{g}$$

and introduce the dummy variables $p_2 := \dot{q}_2$ and $r_2 := \ddot{q}_2$. Replacing all appearances of $\dot{q}_2$ and $\ddot{q}_2$, we finally arrive at system (9.4). Thus, the DAE is of index 1, cf. [**KM06**, Th. 6.12]. Clearly, the index reduction of the index-3 DAE and the semi-discretization of the regularized operator DAE only coincide if the underlying bases are equal. This then shows that the order of semi-discretization and index reduction are permutable as shown in the commutative diagram in Figure 9.1.

The commutativity of semi-discretization and regularization provides benefits for the adaptive simulation in the field of elastodynamics or any other physical problem for which the regularization of Part B is applicable. In a conventional simulation with the method of lines one would first discretize the operator DAE (9.1) in space which leads to the index-3 DAE (9.3). In order to obtain reasonable results, an index reduction (e.g. by minimal extension) would be advisable before starting the time integration. Then, every time some error estimator calls on refining the triangulation for the spatial discretization, the index reduction has to be repeated. In contrast, using the regularized operator equation (9.2), we obtain an index-1 DAE for any (suitable) spatial discretization. Thus, the change of the triangulation during the simulation does not call for additional regularization steps. Clearly, these lines also apply for the systems of first order analysed in Section 8.

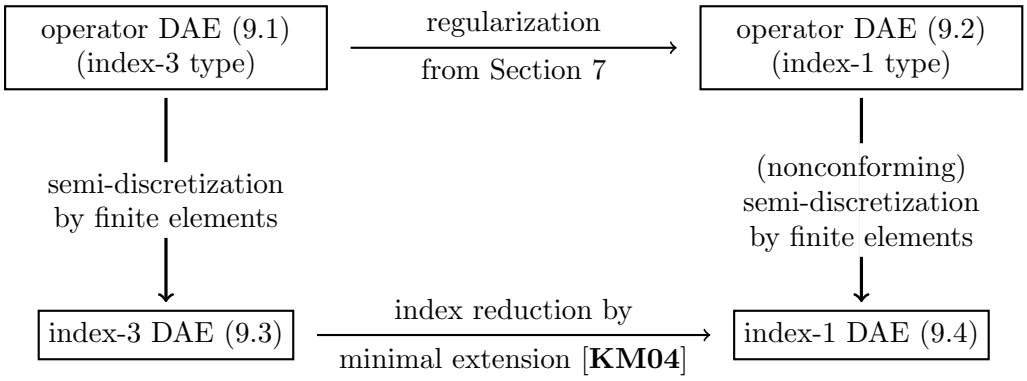


FIGURE 9.1. Commutative diagram showing the reversibility of semi-discretization and index reduction.

The gained potential for adaptivity is of special interest for the simulation of multi-physics systems arising from modern automatic modelling tools. Therein, one module of the big network could be an elasticity model as discussed in this section.

Part

D

The Rothe Method

This part deals with the application of the Rothe method to the regularized operator DAEs from Part B, i.e., we discretize the operator equations in time. We restrict ourselves to first-order time integration schemes. Note that high-order schemes may not pay off when the spatial error dominates. In particular, this is the case in the dynamics of elastic media [**LS09**].

We emphasize that the method of lines and the Rothe method are equivalent for linear problems when corresponding discretizations are used. Nevertheless, the inversion of the order of time and space discretization simplifies the insertion of adaptive strategies in space [**SB98, CDD$^+$14**]. Discretizing in time first, we obtain in each time step a stationary PDE which allows to use adaptive procedures. In particular, the underlying triangulation may be changed easily from time step to time step. Clearly, this includes new challenges such as mesh interpolations which is not topic of this thesis.

The application of the Rothe method for abstract ODEs is discussed is several papers. In this thesis, we mostly rely on the works [**Emm01, EM13**] for first-order systems as well as [**ET10a, Rou05**] for the applications of second order. The papers [**LO93, LO95**] discuss the application of Runge-Kutta methods to parabolic equations. However, these works mainly work in the framework of semigroups and not in the here presented setting with Gelfand triples. This then leads - due to the stronger regularity assumptions - to stronger results than presented here.

In this part, we mainly focus on the convergence of time discretization schemes applied to the regularized operator DAEs of Part B. For this, the general strategy is to construct bounded sequences which approximate the solution of the operator DAE and then use compactness results in the underlying Bochner spaces. Recall the two results from Section 3.1.6 for a reflexive and separable Banach space $\mathcal{V}$:

(a) If (u_n) is a bounded sequence in $L^p(0,T;\mathcal{V})$ and $1 < p < \infty$, then there exists an element $u \in L^p(0,T;\mathcal{V})$ and a subsequence which satisfies $u_{n'} \rightharpoonup u$ in $L^p(0,T;\mathcal{V})$.

(b) If (u_n) is a bounded sequence in $L^\infty(0,T;\mathcal{V})$, then there exists an element $u \in L^\infty(0,T;\mathcal{V})$ and a subsequence which satisfies $u_{n'} \overset{*}{\rightharpoonup} u$ in $L^\infty(0,T;\mathcal{V})$.

The second step of the Rothe method, the spatial discrization, is then included as a perturbation of the right-hand side. For this, we analyse the influence of perturbations for first- as well as for second-order systems.

We emphasize the fact that even the convergence of the Euler scheme without any spatial error is of practical importance. This then corresponds to the limit case as the mesh parameter h tends to zero. Results on the convergence then show the consistency of the discretization scheme to the infinite-dimensional setting and develop a better understanding of the scheme.

We maintain the organization of the previous two parts and start with the analysis of first-order systems before we consider the case of elastodynamics. In Section 10 we apply the Euler discretization in time whose convergence we then prove. For this, we concentrate on the linear case and show which techniques from the analysis of operator ODEs can be preserved. Here we use the regularization of Part B and the splitting of the variable u. Because of the semi-explicit structure of the equations, this splitting then leads to the expected results for u. However, the analysis will also show qualitative differences in the variable u and the Lagrange multiplier λ. Finally, we consider the example of a two-phase flow from Section 6.3.3 which includes a nonlinear constraint operator.

The section on second-order operator DAEs is again specialized to applications in elastodynamics. Thus, Section 11 is based on the (regularized) equations from Section 7. We then analyse the convergence of the time integration scheme from Section 5.2.2.

10. Convergence for First-order Systems

This section is devoted to the convergence analysis of the implicit Euler method applied to the first-order operator DAEs discussed in Section 6. The temporal discretization then leads to a stationary PDE which has to be solved in every time step. In order to obtain convergence results for the Rothe method, we then include spatial errors as perturbations to the system.

In Sections 10.1-10.4 we focus on the purely linear case, i.e., with a linear constraint operator $\mathcal{B}$ as well as a linear operator $\mathcal{K}$. For this case, we show the convergence of the Euler scheme and comment on the influence of perturbations in the right-hand sides. This then shows the advantage of the regularization performed in Part B. Finally, we comment in Section 10.5 on the nonlinear case by means of the two-phase flow example.

10.1. Setting. Before we apply the temporal discretization to the operator DAE, we recall the system equations and summarize the assumptions on the operators used in this section. We restrict the analysis to the linear case with $p = q = 2$, i.e., we consider linear operators $\mathcal{K}$ and $\mathcal{B}$. Furthermore, we assume these operators to be independent of time. The time-dependence could be included if the assumptions below hold uniformly in time. We consider these restrictions in order to focus on the differential-algebraic structure within the convergence proof. The inclusion of nonlinear operators is then subject of Section 10.5 as well as Section 11 for second-order systems.

The original linear operator DAE (6.1) from Section 6 has the form

$$\dot{u}(t) + \mathcal{K}u(t) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{10.1a}$$

$$\mathcal{B}u(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^* \tag{10.1b}$$

with a consistent initial condition $u(0) = g \in \mathcal{H}$ and an underlying Gelfand triple $\mathcal{V}$, $\mathcal{H}$, $\mathcal{V}^*$. As in Part B, $\mathcal{V}_\mathcal{B}$ denotes the kernel of the constraint operator $\mathcal{B}$ and $\mathcal{V}^c$ a complement in $\mathcal{V}$ on which $\mathcal{B}$ is invertible.

If the right-hand side of the constraint vanishes, i.e., $\mathcal{G} = 0$, then the operator DAE (10.1) reduces to an operator ODE on the kernel of the constraint operator $\mathcal{B}$, i.e., on the space $\mathcal{V}_\mathcal{B}$. A typical example are the (Navier)-Stokes equations as already discussed in Section 6.3. In this case, standard methods for the convergence of time discretization schemes of operator ODEs can be applied as e.g. in [**Emm01**].

The regularized formulation presented in Section 6 allows to perform the convergence analysis similarly also in the non-homogeneous case. Recall that the regularized system (6.4) is given by

$$\dot{u}_1(t) + v_2(t) + \mathcal{K}\big(u_1(t) + u_2(t)\big) + \mathcal{B}^*\lambda(t) = \mathcal{F}(t) \quad \text{in } \mathcal{V}^*, \tag{10.2a}$$

$$\mathcal{B}u_2(t) = \mathcal{G}(t) \quad \text{in } \mathcal{Q}^*, \tag{10.2b}$$

$$\mathcal{B}v_2(t) = \dot{\mathcal{G}}(t) \quad \text{in } \mathcal{Q}^* \tag{10.2c}$$

with initial condition $u_1(0) = g_0 \in \overline{\mathcal{V}_\mathcal{B}}^{\mathcal{H}}$. The desired solution of system (10.2) should satisfy $u_1 \in W^{1;2,2}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}^*)$, $u_2, v_2 \in L^2(0,T;\mathcal{V}^c)$, and $\lambda \in L^2(0,T;\mathcal{Q})$.

We retain the notion from Section 6 and use the following abbreviations for the inner product in $\mathcal{H}$ and the norms in $\mathcal{H}$ and $\mathcal{V}$, namely

$$(u,v) := (u,v)_\mathcal{H}, \qquad |u| := \|u\|_\mathcal{H}, \qquad \|u\| := \|u\|_\mathcal{V}.$$

The constant of the continuous embedding $\mathcal{V} \hookrightarrow \mathcal{H}$, which is implied by the Gelfand structure, is given by C_{emb}, i.e, $|\cdot| \le C_{\text{emb}}\|\cdot\|$. The setting with the Gelfand triple implies for the kernel $\mathcal{V}_\mathcal{B}$ that

$$\mathcal{V}_\mathcal{B} \subset \mathcal{V} \subset \mathcal{H} = \mathcal{H}^* \subset \mathcal{V}^* \subset \mathcal{V}_\mathcal{B}^* := (\mathcal{V}_\mathcal{B})^*.$$

Recall that the polar set $\mathcal{V}_\mathcal{B}^\circ \subset \mathcal{V}^*$, see its definition in (7.9), should be distinguished from the dual space $\mathcal{V}_\mathcal{B}^*$.

For the right-hand sides we assume $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$ and $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$ as discussed in Section 6. As mentioned before, we consider the case with a linear and symmetric operator $\mathcal{K}\colon \mathcal{V} \to \mathcal{V}^*$ which is assumed to be positive on $\mathcal{V}_\mathcal{B}$ and continuous, i.e., there exist positive constants k_1, $k_2 \in \mathbb{R}$ such that for all $u \in \mathcal{V}_\mathcal{B}$ and v, $w \in \mathcal{V}$ it holds that

$$k_1\|u\|^2 \le \langle \mathcal{K}u, u\rangle, \qquad \langle \mathcal{K}v, w\rangle \le k_2\|v\|\|w\|.$$

Remark 10.1. The given assumptions on the operator $\mathcal{K}$ already imply that $\mathcal{V}_\mathcal{B}$ is a Hilbert space. In many applications, $(\cdot,\cdot) + \langle\mathcal{K}\cdot,\cdot\rangle$ defines an inner product in $\mathcal{V}$. In this case, we may assume that the splitting $\mathcal{V} = \mathcal{V}_\mathcal{B} \oplus \mathcal{V}^c$ is orthogonal w.r.t. this inner product. This then allows to prove stronger convergence results for more regular data, cf. Theorem 10.10.

The operator $\mathcal{B}$ is assumed to be independent of time and to satisfy Assumption 6.2, i.e., $\mathcal{B}$ is linear, continuous, and there exists a continuous right-inverse $\mathcal{B}^-\colon \mathcal{Q}^* \to \mathcal{V}^c$ with continuity constant $C_{\mathcal{B}^-} := \|\mathcal{B}^-\|$. Recall that this also implies that $\mathcal{B}$ satisfies an inf-sup condition with a constant $\beta_{\text{inf}} > 0$ according to [**Bra07**, Lem. III.4.2].

In the following subsection, we apply the Euler scheme to system (10.2). The analysis is then separately performed for the variable in the kernel of $\mathcal{B}$ (namely u_1) and the remaining variables u_2, v_2, and λ. Note that the latter variables correspond to the algebraic variables in the finite-dimensional DAE case.

10.2. Temporal Discretization. We denote the discrete approximation of u_1, u_2, v_2, and λ at time $t_j = j\tau$ by u_1^j, u_2^j, v_2^j, and λ^j, respectively. Hence, we consider the partition $0 = t_0 < t_1 < \cdots < t_n = T$ of the interval $[0,T]$ with equidistant time step size τ. For the application of the implicit Euler scheme to system (10.2) we have to replace the temporal derivative $\dot{u}_1$ by the *discrete derivative* $Du_1^j := (u_1^j - u_1^{j-1})/\tau$. This then leads to the semi-discrete equations which have to be solved for all time steps, i.e., for $j = 1, \dots, n$.

We consider first the discretization of the constraints (10.2b) and (10.2c). Therein, we have to find $u_2^j \in \mathcal{V}^{\mathrm{c}}$ and $v_2^j \in \mathcal{V}^{\mathrm{c}}$ such that for all test functions $q \in \mathcal{Q}$ it holds that

$$\langle \mathcal{B}u_2^j, q\rangle = \langle \mathcal{G}^j, q\rangle, \qquad \langle \mathcal{B}v_2^j, q\rangle = \langle \dot{\mathcal{G}}^j, q\rangle. \tag{10.3}$$

Since there exists a right-inverse of the operator $\mathcal{B}$ on $\mathcal{V}^{\mathrm{c}}$, we may also write $u_2^j = \mathcal{B}^- \mathcal{G}^j$ and $v_2^j = \mathcal{B}^- \dot{\mathcal{G}}^j$. Recall that $\dot{\mathcal{G}}^j$ cannot be the function evaluation of $\dot{\mathcal{G}}$ at time t_j, since this is not well-defined. Instead, we use integral means over a time interval as introduced in Section 5.3.2.

Second, we consider the discretization of equation (10.2a). If this equation is only tested with functions in the kernel $\mathcal{V}_{\mathcal{B}}$, then we obtain the following problem: Given $u_1^{j-1} \in \mathcal{H}$, search for $u_1^j \in \mathcal{V}_{\mathcal{B}}$ such that for all $v \in \mathcal{V}_{\mathcal{B}}$ it holds that

$$(Du_1^j, v) + \langle \mathcal{K}u_1^j, v\rangle = \langle \mathcal{F}^j, v\rangle - (\mathcal{B}^- \dot{\mathcal{G}}^j, v) - \langle \mathcal{K}\mathcal{B}^- \mathcal{G}^j, v\rangle. \tag{10.4}$$

Note that $\mathcal{F}$ is not continuous such that $\mathcal{F}^j$ again cannot equal a function evaluation at time t_j. The piecewise constant approximations of $\mathcal{F}$, $\mathcal{G}$, and $\dot{\mathcal{G}}$ are denoted by $\mathcal{F}_\tau$, $\mathcal{G}_\tau$, and $\dot{\mathcal{G}}_\tau$, respectively. The precise definition is given in (5.12). We emphasize once more that $\dot{\mathcal{G}}_\tau$ does not denote the derivative of $\mathcal{G}_\tau$. In the sequel we assume these approximations to satisfy

$$\mathcal{F}_\tau \to \mathcal{F} \text{ in } L^2(0,T;\mathcal{V}^*), \qquad \mathcal{G}_\tau \to \mathcal{G},\ \dot{\mathcal{G}}_\tau \to \dot{\mathcal{G}} \text{ in } L^2(0,T;\mathcal{Q}^*).$$

Furthermore, we assume $\mathcal{F}_\tau$, $\mathcal{G}_\tau$, and $\dot{\mathcal{G}}_\tau$ to be continuous in $t = 0$. In order to obtain equation (10.4), we have applied the explicit formulae for u_2^j and v_2^j given by (10.3). Finally, the equation for the discrete Lagrange multiplier is given by the discretization of (10.2a) with test functions in the complement space $\mathcal{V}^{\mathrm{c}}$. The task is then to find $\lambda^j \in \mathcal{Q}$ such that for all $v \in \mathcal{V}^{\mathrm{c}}$ we have

$$\langle \mathcal{B}^* \lambda^j, v\rangle = \langle \mathcal{F}^j, v\rangle - (\mathcal{B}^- \dot{\mathcal{G}}^j, v) - \langle \mathcal{K}\mathcal{B}^- \mathcal{G}^j, v\rangle - (Du_1^j, v) - \langle \mathcal{K}u_1^j, v\rangle. \tag{10.5}$$

Before we use the discrete approximations to construct global approximations in $L^2(0,T;\mathcal{V})$ and $L^2(0,T;\mathcal{Q})$, we need to discuss the solvability of the equations (10.4) and (10.5). Afterwards, we have to find a priori bounds of the approximations in order to extract converging subsequences. The final task is then to show that the resulting limits are (in some sense) solutions of the operator DAE (10.2).

10.2.1. *Existence of Solutions.* We have already discussed the solvability of the equations in (10.3) due to the existence of a right-inverse of the operator $\mathcal{B}$. Next, we comment on the solvability of (10.4) to ensure the existence of the sequence u_1^j. With the bilinear form $c\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, given by

$$c(u,v) := \frac{1}{\tau}(u,v) + \langle \mathcal{K}u, v\rangle,$$

and the functional $F \in \mathcal{V}^*$,

$$\langle F, v\rangle := \langle \mathcal{F}^j, v\rangle - (\mathcal{B}^- \dot{\mathcal{G}}^j, v) - \langle \mathcal{K}\mathcal{B}^- \mathcal{G}^j, v\rangle + \frac{1}{\tau}(u_1^{j-1}, v),$$

equation (10.4) can be written as $c(u_1^j, v) = F(v)$ for all $v \in \mathcal{V}_\mathcal{B}$. The unique solvability of (10.4) then follows by the Lax-Milgram lemma [**Eva98**, Sect. 6.2.1]. The needed properties of the bilinear form c such as the coercivity on $\mathcal{V}_\mathcal{B}$ follow directly from the assumptions on $\mathcal{K}$.

It remains to show that equation (10.5) obtains a unique solution λ^j. Obviously, the right-hand side defines a functional in $\mathcal{V}^*$. Equation (10.4) even implies that the right-hand side vanishes for all functions in $\mathcal{V}_\mathcal{B}$. Thus, the functional is an element of the polar set $\mathcal{V}_\mathcal{B}^\circ$ on which the operator $\mathcal{B}^*$ is invertible [**Bra07**, Ch. III, Lem. 4.2].

10.2.2. *A Priori Estimates.* This subsection provides stability or a priori bounds of the discrete approximations defined above. Since the equation for u_1^j comes essentially from an operator ODE, the given proofs follow the lines of the stability results in [**Emm01**, Ch. 4], see also [**Tem77**, Ch. III.4]. Amongst others, we will make use of the equality

$$2(Du^j, u^j) = D|u^j|^2 + \tau|Du^j|^2. \tag{10.6}$$

This identity of the discrete derivative follows by a simple calculation and can also be found in [**Emm01**, Lem. 3.2.2].

Recall that u_1^0 is given and represents the approximation of u_1 at time $t = 0$, i.e., the initial data g_0. However, we do not assume that u_1^0 and g_0 coincide at this point. In order to obtain a prioi estimates of the approximation of the differential variable, namely u_1^j, we test equation (10.4) by $u_1^j \in \mathcal{V}_\mathcal{B}$. Since the Lagrange multiplier is not present in this equation, we can maintain most of the techniques used for operator ODEs. This then leads to the following result.

LEMMA 10.2 (Stability I). *Assume $\mathcal{F} \in L^2(0,T;\mathcal{V}_\mathcal{B}^*)$ and $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$. Then, the approximations $u_1^j \in \mathcal{V}_\mathcal{B}$ given by the Euler scheme* (10.4) *with $u_1^0 \in \mathcal{H}$ satisfy for all $1 \le k \le n$ the estimate*

$$|u_1^k|^2 + \tau^2 \sum_{j=1}^k |Du_1^j|^2 + \tau k_1 \sum_{j=1}^k \|u_1^j\|^2 \le M^2 \tag{10.7}$$

with constant $M^2 := |u_1^0|^2 + 3\big(\|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}_\mathcal{B}^*)} + C^2_{\mathcal{B}^-}(C^4_{emb} + k_2^2)\|\mathcal{G}\|^2_{H^1(0,T;\mathcal{Q}^*)}\big)/k_1$.

PROOF. Using as test function $v = u_1^j \in \mathcal{V}_\mathcal{B}$, $j \ge 1$, in the Euler scheme (10.4), we obtain

$$(Du_1^j, u_1^j) + \langle \mathcal{K}u_1^j, u_1^j\rangle = \langle \mathcal{F}^j, u_1^j\rangle - (\mathcal{B}^-\dot{\mathcal{G}}^j, u_1^j) - \langle \mathcal{K}\mathcal{B}^-\mathcal{G}^j, u_1^j\rangle. \tag{10.8}$$

Summation over $j = 1, \dots, k$, together with property (10.6), the Cauchy-Schwarz inequality, and the continuous embedding $\mathcal{V} \hookrightarrow \mathcal{H}$ yield

$$\begin{aligned}
|u_1^k|^2 - |u_1^0|^2 + \tau^2 \sum_{j=1}^k |Du_1^j|^2 + 2\tau k_1 \sum_{j=1}^k \|u_1^j\|^2 & \\
&\overset{(10.6)}{\le} 2\tau \sum_{j=1}^k (Du_1^j, u_1^j) + 2\tau \sum_{j=1}^k \langle \mathcal{K}u_1^j, u_1^j\rangle \\
&\overset{(10.8)}{\le} 2\tau \sum_{j=1}^k \Big(\|\mathcal{F}^j\|_{\mathcal{V}_\mathcal{B}^*} + C_{\mathrm{emb}}|\mathcal{B}^-\dot{\mathcal{G}}^j| + k_2\|\mathcal{B}^-\mathcal{G}^j\|\Big)\|u_1^j\|.
\end{aligned}$$

By Young's inequality, the last line is bounded from above by

$$\frac{3\tau}{k_1}\sum_{j=1}^{k}\Big(\|\mathcal{F}^j\|^2_{\mathcal{V}^*_{\mathcal{B}}}+C^2_{\mathrm{emb}}|\mathcal{B}^-\dot{\mathcal{G}}^j|^2+k_2^2\|\mathcal{B}^-\mathcal{G}^j\|^2\Big)+\tau k_1\sum_{j=1}^{k}\|u_1^j\|^2.$$

Thus, by the boundedness of the right-inverse of $\mathcal{B}$, we obtain

$$\begin{aligned}|u_1^k|^2-|u_1^0|^2+\tau^2\sum_{j=1}^{k}|Du_1^j|^2+\tau k_1\sum_{j=1}^{k}\|u_1^j\|^2&\\ \le\frac{3\tau}{k_1}\sum_{j=1}^{k}\Big(\|\mathcal{F}^j\|^2_{\mathcal{V}^*_{\mathcal{B}}}+C^2_{\mathcal{B}^-}C^4_{\mathrm{emb}}\|\dot{\mathcal{G}}^j\|^2_{\mathcal{Q}^*}+C^2_{\mathcal{B}^-}k_2^2\|\mathcal{G}^j\|^2_{\mathcal{Q}^*}\Big).&\end{aligned}$$

Finally, property (5.13) for the right-hand sides implies that

$$\begin{aligned}|u_1^k|^2-|u_1^0|^2+\tau^2\sum_{j=1}^{k}|Du_1^j|^2+\tau k_1\sum_{j=1}^{k}\|u_1^j\|^2&\\ \le\frac{3}{k_1}\|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}^*_{\mathcal{B}})}+\frac{3}{k_1}C^2_{\mathcal{B}^-}C^4_{\mathrm{emb}}\|\dot{\mathcal{G}}\|^2_{L^2(0,T;\mathcal{Q}^*)}+\frac{3}{k_1}C^2_{\mathcal{B}^-}k_2^2\|\mathcal{G}\|^2_{L^2(0,T;\mathcal{Q}^*)}.&\end{aligned}$$ □

For the discrete derivative of u_1^j, namely Du_1^j, we obtain the following bound.

Lemma 10.3 (Stability II). *Consider the same assumptions as in Lemma 10.2. Then, there exists a positive constant c such that the approximations u_1^j given by* (10.4) *satisfy*

$$\tau\sum_{j=1}^{n}\|Du_1^j\|^2_{\mathcal{V}^*_{\mathcal{B}}}\le cM^2. \tag{10.9}$$

Proof. From equation (10.4) we obtain for $j\ge 1$,

$$\begin{aligned}\|Du_1^j\|_{\mathcal{V}^*_{\mathcal{B}}}&:=\sup_{v\in\mathcal{V}_{\mathcal{B}},\ \|v\|=1}\big|\langle\mathcal{F}^j,v\rangle-(\mathcal{B}^-\dot{\mathcal{G}}^j,v)-\langle\mathcal{K}\mathcal{B}^-\mathcal{G}^j,v\rangle-\langle\mathcal{K}u_1^j,v\rangle\big|\\ &\le\|\mathcal{F}^j\|_{\mathcal{V}^*_{\mathcal{B}}}+C_{\mathcal{B}^-}C^2_{\mathrm{emb}}\|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}+k_2C_{\mathcal{B}^-}\|\mathcal{G}^j\|_{\mathcal{Q}^*}+k_2\|u_1^j\|.\end{aligned}$$

Thus, with Young's inequality and Lemma 10.2, the summation over j yields

$$\begin{aligned}\tau\sum_{j=1}^{n}\|Du_1^j\|^2_{\mathcal{V}^*_{\mathcal{B}}}&\le 4\|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}^*_{\mathcal{B}})}+4C^2_{\mathcal{B}^-}(C^4_{\mathrm{emb}}+k_2^2)\|\mathcal{G}\|^2_{H^1(0,T;\mathcal{Q}^*)}+4\tau k_2^2\sum_{j=1}^{n}\|u_1^j\|^2\\ &\overset{(10.7)}{\le}\frac{4k_1}{3}M^2+\frac{4k_2^2}{k_1}M^2.\end{aligned}$$ □

Remark 10.4. Assume that $(\cdot,\cdot)+\langle\mathcal{K}\cdot,\cdot\rangle$ defines an inner product in $\mathcal{V}$ with which the decomposition $\mathcal{V}=\mathcal{V}_{\mathcal{B}}\oplus\mathcal{V}^{\mathrm{c}}$ is orthogonal. If we assume more regularity of the given data in the form of $\mathcal{F}\in L^2(0,T;\mathcal{H}^*)$ and $u_1^0\in\mathcal{V}_{\mathcal{B}}$, then we even obtain the estimate

$$\tau\sum_{j=1}^{n}|Du_1^j|^2\le M^2_{\mathrm{reg}}\big(\|u_1^0\|,\|\mathcal{F}\|_{L^2(0,T;\mathcal{H}^*)}\big).$$

This estimate can be obtained by testing equation (10.4) by $Du_1^j\in\mathcal{V}_{\mathcal{B}}$. Note that this equals the result in Lemma 10.3 but in a stronger norm. We will see in the sequel that this difference is crucial in view of the Lagrange multiplier.

10.3. Global Approximations and Convergence. The stability estimates of the previous subsection are the basis for the proof of convergence of the Euler scheme. With this, we can prove that the global approximations of u_1, u_2, v_2, and λ, which we define within this subsection, are uniformly bounded.

10.3.1. *Definition of $U_{1,\tau}$, $U_{2,\tau}$, and $V_{2,\tau}$.* With the discrete approximations given by the Euler scheme (10.3)-(10.5) we define piecewise constant and piecewise linear functions on the interval $[0,T]$. Given u_1^j, $j=1,\dots,n$, we define $U_{1,\tau}$, $\hat{U}_{1,\tau}\colon [0,T] \to \mathcal{V}_\mathcal{B}$ by

$$U_{1,\tau}(t) := \begin{cases} u_1^0 & \text{if } t = 0, \\ u_1^j & \text{if } t \in]t_{j-1}, t_j] \end{cases}, \quad \hat{U}_{1,\tau}(t) := \begin{cases} u_1^0 & \text{if } t = 0, \\ u_1^j + (t-t_j)Du_1^j & \text{if } t \in]t_{j-1}, t_j] \end{cases}. \tag{10.10}$$

One aim of this section is to show that the sequences $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ converge to the solution of the operator equation (10.2) as $\tau \to 0$. For this, we have to show the uniform boundedness of the sequences in order to obtain a converging subsequence, see also [**Emm01**, Ch. 4].

LEMMA 10.5 (Boundedness of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$). *Assume $\mathcal{F} \in L^2(0,T;\mathcal{V}_\mathcal{B}^*)$, $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$, and $u_1^0 \in \mathcal{V}_\mathcal{B}$. Then, the sequences $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ are uniformly bounded in $L^\infty(0,T;\mathcal{H})$ and $L^2(0,T;\mathcal{V}_\mathcal{B})$. Furthermore, the sequence of derivatives $\dot{\hat{U}}_{1,\tau}$ is bounded in $L^2(0,T;\mathcal{V}_\mathcal{B}^*)$.*

PROOF. The boundedness in $L^\infty(0,T;\mathcal{H})$ and $L^2(0,T;\mathcal{V}_\mathcal{B})$ follows directly from (10.7) together with $u_1^0 \in \mathcal{V}_\mathcal{B}$. The details can be found in [**Emm01**, Lem. 4.2.1]. The boundedness of $\dot{\hat{U}}_{1,\tau}$ follows from the second stability estimate (10.9), namely,

$$\big\|\dot{\hat{U}}_{1,\tau}\big\|^2_{L^2(0,T;\mathcal{V}_\mathcal{B}^*)} = \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \|Du_1^j\|^2_{\mathcal{V}_\mathcal{B}^*} \,\mathrm{d}t = \tau \sum_{j=1}^n \|Du_1^j\|^2_{\mathcal{V}_\mathcal{B}^*} \le cM^2. \qquad \square$$

With the shown boundedness of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ in Lemma 10.5, by Theorem 3.31 we obtain that there exist weakly convergent subsequences in $L^2(0,T;\mathcal{V}_\mathcal{B})$. Furthermore, the estimate (10.7) implies that

$$\big\|U_{1,\tau} - \hat{U}_{1,\tau}\big\|^2_{L^2(0,T;\mathcal{H})} \le \tau \sum_{j=1}^n \big|u_1^j - u_1^{j-1}\big|^2 \le \tau M^2 \to 0.$$

Thus, the limits coincide in $L^2(0,T;\mathcal{H})$ and the continuous embedding $\mathcal{V} \hookrightarrow \mathcal{H}$ implies that the same is true for the limit in $L^2(0,T;\mathcal{V}_\mathcal{B})$. We denote the joined limit of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ by U_1, i.e., $U_{1,\tau}$, $\hat{U}_{1,\tau} \rightharpoonup U_1$ in $L^2(0,T;\mathcal{V}_\mathcal{B})$. Note that the result is true for the entire sequence, since the considered linear operator DAE has a unique solution.

Next, we consider the approximations of u_2 and v_2. For this, similar as before, we define the piecewise constant functions

$$U_{2,\tau}(t) := u_2^j \ \text{ if } t \in]t_{j-1}, t_j], \qquad V_{2,\tau}(t) := v_2^j \ \text{ if } t \in]t_{j-1}, t_j] \tag{10.11}$$

with a continuous extension in $t=0$. Note that this definition implies that $U_{2,\tau} = \mathcal{B}^-\mathcal{G}_\tau$ and $V_{2,\tau} = \mathcal{B}^-\dot{\mathcal{G}}_\tau$. Thus, with the help of Lemma 5.9, we obtain that

$$U_{2,\tau} \to U_2 := \mathcal{B}^-\mathcal{G}, \qquad V_{2,\tau} \to V_2 := \mathcal{B}^-\dot{\mathcal{G}} \quad \text{ in } L^2(0,T;\mathcal{V}^\mathrm{c}).$$

The embedding $H^1(0,T;\mathcal{Q}^*) \hookrightarrow C([0,T];\mathcal{Q}^*)$ implies additionally that U_2 satisfies the consistency condition $U_2(0) = \mathcal{B}^-\mathcal{G}(0)$. In the case of sufficiently regular data, for which $\dot{\mathcal{G}}(0)$ is well-defined in $\mathcal{Q}^*$, we also obtain the consistency condition corresponding to the hidden constraint, i.e., $V_2(0) = \mathcal{B}^-\dot{\mathcal{G}}(0)$.

10.3.2. *Definition of* Λ_τ. As global approximation of the Lagrange multiplier we define $\Lambda_\tau\colon]0,T] \to \mathcal{Q}$ by

$$\Lambda_\tau(t) := \lambda^j \quad \text{if } t \in]t_{j-1}, t_j]. \tag{10.12}$$

Therein, λ^j denote the discrete approximations of the Lagrange multiplier from equation (10.5). The desired boundedness of Λ_τ in $L^2(0,T;\mathcal{Q})$ requires a uniform upper bound of

$$\|\Lambda_\tau\|^2_{L^2(0,T;\mathcal{Q})} = \int_0^T \|\Lambda_\tau(t)\|^2_{\mathcal{Q}} \,\mathrm{d}t = \tau \sum_{j=1}^n \|\lambda^j\|^2_{\mathcal{Q}}.$$

With the inf-sup condition of the operator $\mathcal{B}$, by equation (10.5), we can estimate

$$\beta_{\inf}\|\lambda^j\|_{\mathcal{Q}} \le \sup_{v\in\mathcal{V}} \frac{\langle \mathcal{B}^*\lambda^j, v\rangle}{\|v\|_{\mathcal{V}}} \le \|\mathcal{F}^j\|_{\mathcal{V}^*} + C_{\mathrm{emb}}|\mathcal{B}^-\dot{\mathcal{G}}^j| + k_2\|\mathcal{B}^-\mathcal{G}^j\| + \|Du_1^j\|_{\mathcal{V}^*} + k_2\|u_1^j\|.$$

Thus, we can only show the boundedness of Λ_τ if we find an estimate of $\tau\sum_{j=1}^n \|Du_1^j\|^2_{\mathcal{V}^*}$. This, however, is problematic since the Lemmata 10.2 and 10.3 only provide estimates of $\tau^2\sum_{j=1}^n |Du_1^j|^2$ and $\tau\sum_{j=1}^n \|Du_1^j\|^2_{\mathcal{V}_\mathcal{B}^*}$. As a consequence, we are not be able to show the convergence of Λ_τ to the solution of the operator DAE (10.2).

Remark 10.6. The lack of convergence of the Lagrange multiplier is a regularity problem. In the finite-dimensional setting, this difficulty does not occur, since all norms are equivalent. In addition, the application of $\mathcal{B}$ requires a certain regularity whereas in the discrete setting the discrete analogon of $\mathcal{B}$ equals a matrix whose application is always possible. Note that the additional regularity assumptions in Remark 10.4 would suffice to prove the boundedness of the sequence of Lagrange multipliers and thus, the existence of a weak limit $\Lambda \in L^2(0,T;\mathcal{Q})$.

One possible way out, without assuming more regular right-hand sides, is to consider solutions of (10.2) in the weak distributional sense, see Section 4.3. We follow [**EM13**] and show the convergence (of the primitive) to the tuple $(u_1, u_2, v_2, \tilde{\lambda})$ which solves the operator DAE in a weaker sense. For this, we define the primitive $\tilde{\Lambda}_\tau \in AC([0,T];\mathcal{Q})$ by

$$\tilde{\Lambda}_\tau(t) := \int_0^t \Lambda_\tau(s)\,\mathrm{d}s. \tag{10.13}$$

The formulation of (10.5) with the primitive has the advantage that the problematic term including Du_1^j drops out. First, we consider an equivalent form of equation (10.5) in terms of $U_{1,\tau}$, $U_{2,\tau}$, $V_{2,\tau}$, and Λ_τ, namely

$$\langle \mathcal{B}^*\Lambda_\tau, v\rangle = \langle \mathcal{F}_\tau, v\rangle - (\mathcal{B}^-\dot{\mathcal{G}}_\tau, v) - \langle \mathcal{K}\mathcal{B}^-\mathcal{G}_\tau, v\rangle - (\dot{U}_{1,\tau}, v) - \langle \mathcal{K}U_{1,\tau}, v\rangle \tag{10.14}$$

for a.e. point in time. Integrating this equation over $[0,t]$, we obtain the equation for the primitive of the Lagrange multiplier,

$$\langle \mathcal{B}^*\tilde{\Lambda}_\tau, v\rangle = \langle \tilde{\mathcal{F}}_\tau, v\rangle - (\mathcal{B}^-\tilde{\dot{\mathcal{G}}}_\tau, v) - \langle \mathcal{K}\mathcal{B}^-\tilde{\mathcal{G}}_\tau, v\rangle - (\hat{U}_{1,\tau}, v) - \langle \mathcal{K}\tilde{U}_{1,\tau}, v\rangle + c(v). \tag{10.15}$$

Therein, $\tilde{\mathcal{F}}_\tau$, $\tilde{\mathcal{G}}_\tau$, $\tilde{\dot{\mathcal{G}}}_\tau$, and $\tilde{U}_{1,\tau}$ denote the primitives of $\mathcal{F}_\tau$, $\mathcal{G}_\tau$, $\dot{\mathcal{G}}_\tau$, and $U_{1,\tau}$, respectively. Furthermore, the term $c(v)$, which occurs because of the integration step, is constant in time and equals $c(v) = (u_1^0, v)$. In contrast to the Lagrange multiplier Λ_τ, we can prove an a priori bound of the primitive $\tilde{\Lambda}_\tau$ independent of the step size τ.

Lemma 10.7 (Boundedness of $\tilde{\Lambda}_\tau$). *Assume $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$, and $u_1^0 \in \mathcal{V}_\mathcal{B}$. Then, the sequence $\tilde{\Lambda}_\tau$ is bounded in $C([0,T];\mathcal{Q})$.*

PROOF. We make use of the inf-sup condition of the operator $\mathcal{B}$ and, by equation (10.15), we obtain the estimate

$$\begin{aligned}\beta_{\inf}\|\tilde{\Lambda}_\tau\|_{C([0,T];\mathcal{Q})} &= \beta_{\inf}\max_{t\in[0,T]}\|\tilde{\Lambda}_\tau(t)\|_{\mathcal{Q}}\\ &\le \max_{t\in[0,T]}\sup_{v\in\mathcal{V}}\frac{\langle\mathcal{B}^*\tilde{\Lambda}_\tau(t),v\rangle}{\|v\|}\\ &\overset{(10.15)}{\le} \max_{t\in[0,T]}\Big[\|\tilde{\mathcal{F}}_\tau(t)\|_{\mathcal{V}^*}+C_{\mathrm{emb}}|\mathcal{B}^-\dot{\tilde{\mathcal{G}}}_\tau(t)|+k_2\|\mathcal{B}^-\tilde{\mathcal{G}}_\tau(t)\|\\ &\qquad + C_{\mathrm{emb}}|\hat{U}_{1,\tau}(t)|+k_2\|\tilde{U}_{1,\tau}(t)\|+C_{\mathrm{emb}}|u_1^0|\Big].\end{aligned}$$

The properties of the Bochner integral from Section 3.2 and the Cauchy-Schwarz inequality yield, together with Lemma 10.2, the estimates

$$\begin{aligned}&\max_{t\in[0,T]}\|\tilde{\mathcal{F}}_\tau(t)\|_{\mathcal{V}^*}\le\int_0^T\|\mathcal{F}_\tau(t)\|_{\mathcal{V}^*}\,\mathrm{d}t\overset{(5.13)}{\le}T^{1/2}\|\mathcal{F}\|_{L^2(0,T;\mathcal{V}^*)},\\ &\max_{t\in[0,T]}|\mathcal{B}^-\dot{\tilde{\mathcal{G}}}_\tau(t)|\le C_{\mathrm{emb}}C_{\mathcal{B}^-}\max_{t\in[0,T]}\|\dot{\tilde{\mathcal{G}}}_\tau(t)\|_{\mathcal{Q}^*}\overset{(5.13)}{\le}C_{\mathrm{emb}}C_{\mathcal{B}^-}T^{1/2}\|\dot{\mathcal{G}}\|_{L^2(0,T;\mathcal{Q}^*)},\\ &\max_{t\in[0,T]}\|\mathcal{B}^-\tilde{\mathcal{G}}_\tau(t)\|_{\mathcal{V}}\le C_{\mathcal{B}^-}\int_0^T\|\mathcal{G}_\tau(t)\|_{\mathcal{Q}^*}\,\mathrm{d}t\overset{(5.13)}{\le}C_{\mathcal{B}^-}T^{1/2}\|\mathcal{G}\|_{L^2(0,T;\mathcal{Q}^*)},\\ &\max_{t\in[0,T]}|\hat{U}_{1,\tau}(t)|\le\max_j|u_1^j|\overset{(10.7)}{\le}M,\\ &\max_{t\in[0,T]}\|\tilde{U}_{1,\tau}(t)\|\le\int_0^T\|U_{1,\tau}(t)\|\,\mathrm{d}t\le\tau n^{1/2}\Big(\sum_{j=1}^n\|u_1^j\|^2\Big)^{1/2}\overset{(10.7)}{\le}T^{1/2}k_1^{-1/2}M.\end{aligned}$$

Thus, $\|\tilde{\Lambda}_\tau\|_{C([0,T];\mathcal{Q})}$ is uniformly bounded in terms of T, the initial data, and the right-hand sides. □

A direct consequence of Lemma 10.7 is the existence of a weak limit $\tilde{\Lambda}$ of a subsequence of $\tilde{\Lambda}_\tau$, i.e.,

$$\tilde{\Lambda}_\tau\rightharpoonup\tilde{\Lambda}\quad\text{in } L^p(0,T;\mathcal{Q})$$

for all $1<p<\infty$. In the following subsection we analyse in which sense the obtained limits U_1, U_2, V_2, and $\tilde{\Lambda}$ solve the operator DAE (10.2).

10.3.3. *Convergence Results.* In the subsections above we have only assumed u_1^0 to be bounded. In order to show that the obtained limits solve the operator DAE (10.2), we assume in the sequel that $u_1^0=g_0\in\mathcal{V}_{\mathcal{B}}$. Note that this assumption could be weakened to $u_1^0\to g_0$ in $\mathcal{V}_{\mathcal{B}}$ as $\tau\to0$. Since $\mathcal{G}_\tau\to\mathcal{G}$ and $\dot{\mathcal{G}}_\tau\to\dot{\mathcal{G}}$ in $L^2(0,T;\mathcal{Q}^*)$ as shown in Section 5.3.2, we know that the limits U_2 and V_2 solve equations (10.2b) and (10.2c). The following result is devoted to the behavior of the limit U_1.

THEOREM 10.8. *Assume $\mathcal{F}\in L^2(0,T;\mathcal{V}_{\mathcal{B}}^*)$, $\mathcal{G}\in H^1(0,T;\mathcal{Q}^*)$, and $u_1^0=g_0\in\mathcal{V}_{\mathcal{B}}$. Then, the weak limit $U_1\in L^2(0,T;\mathcal{V}_{\mathcal{B}})$ of the sequence $U_{1,\tau}$ solves equation* (10.2a) *in $\mathcal{V}_{\mathcal{B}}^*$, i.e., for test functions in $\mathcal{V}_{\mathcal{B}}$. Furthermore, U_1 has a generalized derivative which satisfies $\dot{U}_1\in L^2(0,T;\mathcal{V}_{\mathcal{B}}^*)$.*

PROOF. As in [**Emm01**, Ch. 4], we consider equation (10.4) in terms of the global approximations, for which we know the existence of (weak) limits. Thus, for test functions

$v \in \mathcal{V}_\mathcal{B}$, we consider the equation

$$\frac{\mathrm{d}}{\mathrm{d}t}(\hat{U}_{1,\tau}, v) + \langle \mathcal{K} U_{1,\tau}, v\rangle = \langle \mathcal{F}_\tau, v\rangle - (\mathcal{B}^- \dot{\mathcal{G}}_\tau, v) - \langle \mathcal{K}\mathcal{B}^- \mathcal{G}_\tau, v\rangle.$$

In order to show that U_1 solves the operator DAE, we advance to the limit $\tau \to 0$. For this, we consider the integral formulation with $\Phi \in C_0^\infty(0,T)$ and integrate by parts,

$$\begin{aligned}\int_0^T -\big(\hat{U}_{1,\tau}, v\big)\dot{\Phi}(t) + \big\langle \mathcal{K} U_{1,\tau}, v\big\rangle \Phi(t)\ \mathrm{d}t \\ = \int_0^T \big\langle \mathcal{F}_\tau, v\big\rangle \Phi(t) - \big(\mathcal{B}^- \dot{\mathcal{G}}_\tau, v\big)\Phi(t) - \big\langle \mathcal{K}\mathcal{B}^- \mathcal{G}_\tau, v\big\rangle \Phi(t)\ \mathrm{d}t.\end{aligned}$$

With the limit functions U_2 and V_2 from Section 10.3.1 and the convergence of $\mathcal{F}_\tau$ shown in Section 5.3.2, the right-hand side converges for $\tau \to 0$ to

$$\begin{aligned}\int_0^T \big\langle \mathcal{F}_\tau, v\big\rangle \Phi(t) - \big(\mathcal{B}^- \dot{\mathcal{G}}_\tau, v\big)\Phi(t) - \big\langle \mathcal{K}\mathcal{B}^- \mathcal{G}_\tau, v\big\rangle \Phi(t)\ \mathrm{d}t \\ \longrightarrow \int_0^T \big\langle \mathcal{F}, v\big\rangle \Phi(t) - \big(V_2, v\big)\Phi(t) - \big\langle \mathcal{K} U_2, v\big\rangle \Phi(t)\ \mathrm{d}t.\end{aligned}$$

Furthermore, the weak convergence of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ in $L^2(0,T;\mathcal{V}_\mathcal{B})$ is sufficient to obtain

$$\int_0^T -\big(\hat{U}_{1,\tau}, v\big)\dot{\Phi} + \big\langle \mathcal{K} U_{1,\tau}, v\big\rangle \Phi\ \mathrm{d}t \longrightarrow \int_0^T -\big(U_1, v\big)\dot{\Phi} + \big\langle \mathcal{K} U_1, v\big\rangle \Phi\ \mathrm{d}t.$$

As a result, the obtained limit $U_1 \in L^2(0,T;\mathcal{V}_\mathcal{B})$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}(U_1, v) + (U_2, v) + \langle \mathcal{K}(U_1 + U_2), v\rangle = \langle \mathcal{F}, v\rangle \tag{10.16}$$

for all $v \in \mathcal{V}_\mathcal{B}$. Next, we show that U_1 has a generalized derivative. From the definition of $\hat{U}_{1,\tau}$ in (10.10) we know that its time derivative equals Du_1^j for $t \in]t_{j-1}, t_j[$. Further, recall that $\frac{\mathrm{d}}{\mathrm{d}t}\hat{U}_{1,\tau}$ is bounded in $L^2(0,T;\mathcal{V}_\mathcal{B}^*)$ due to Lemma 10.5. Thus, there exists a subsequence which weakly converges to a limit $V_1 \in L^2(0,T;\mathcal{V}_\mathcal{B}^*)$. For every $\Phi \in C_0^\infty(0,T)$ and $v \in \mathcal{V}_\mathcal{B}$ this limit satisfies the equality

$$\begin{aligned}\int_0^T \big\langle U_1(t), v\big\rangle\ \dot{\Phi}(t)\ \mathrm{d}t &= \lim_{\tau\to 0} \int_0^T \big\langle \hat{U}_{1,\tau}(t), v\big\rangle\ \dot{\Phi}(t)\ \mathrm{d}t \\ &= \lim_{\tau\to 0} -\int_0^T \big\langle \dot{\hat{U}}_{1,\tau}(t), v\big\rangle\ \Phi(t)\ \mathrm{d}t = -\int_0^T \big\langle V_1(t), v\big\rangle\ \Phi(t)\ \mathrm{d}t.\end{aligned}$$

This shows that $\dot{U}_1 = V_1 \in L^2(0,T;\mathcal{V}_\mathcal{B}^*)$ in the generalized sense. As a result, U_1 solves equation (10.2a) if tested only with functions in $\mathcal{V}_\mathcal{B}$. Finally, we have to check whether U_1 satisfies the stated initial condition $U_1(0) = u_1^0 = g_0 \in \mathcal{V}_\mathcal{B}$. Since $U_1 \in W^{1;2,2}(0,T;\mathcal{V}_\mathcal{B},\mathcal{V}_\mathcal{B}^*)$ and $\hat{U}_{1,\tau} \rightharpoonup U_1$ as well as $\frac{\mathrm{d}}{\mathrm{d}t}\hat{U}_{1,\tau} \rightharpoonup \dot{U}_1 = V_1$, for $\Phi \in C^1([0,T])$ with $\Phi(T) = 0$ and arbitrary $v \in \mathcal{V}_\mathcal{B}$, we derive that

$$\begin{aligned}0 &= \lim_{\tau\to 0} \int_0^T \big\langle \dot{\hat{U}}_{1,\tau} - \dot{U}_1, v\big\rangle \Phi\ \mathrm{d}t \\ &= \lim_{\tau\to 0} -\int_0^T \big\langle \hat{U}_{1,\tau} - U_1, v\big\rangle \dot{\Phi}\ \mathrm{d}t - \big(\hat{U}_{1,\tau}(0) - U_1(0), v\big)\Phi(0) \\ &= -\big(g_0 - U_1(0), v\big)\Phi(0).\end{aligned}$$

Since $\mathcal{V}_\mathcal{B}$ is dense in $\mathcal{H}_\mathcal{B} := \overline{\mathcal{V}_\mathcal{B}}^{\mathcal{H}}$, this implies $U_1(0) = g_0$ in $\mathcal{H}_\mathcal{B}$. Finally, the injectivity of the embedding $\mathcal{V}_\mathcal{B} \hookrightarrow \mathcal{H}_\mathcal{B}$ yields $U_1(0) = g_0$ also in $\mathcal{V}_\mathcal{B}$. □

It remains to analyse in which sense $\tilde{\Lambda}$ from Section 10.3.2 solves the operator DAE. We show that this only holds in the weak distributional sense, see Section 4.3 for the used terminology.

THEOREM 10.9. *Assume $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$, and $u_1^0 = g_0 \in \mathcal{V}_\mathcal{B}$. Then, for any sequence of step sizes with $\tau \to 0$ the sequence $\tilde{\Lambda}_\tau$ converges weakly to $\tilde{\Lambda}$ in $L^2(0,T;\mathcal{Q})$ such that $(U_1, U_2, V_2, \tilde{\Lambda})$ solves system (10.2) in the weak distributional sense.*

PROOF. We have already seen that the boundedness of the sequence $\tilde{\Lambda}_\tau$ shown in Lemma 10.7 implies the existence of a weak limit $\tilde{\Lambda}$ in $L^p(0,T;\mathcal{Q})$. Thus, for all $\Phi \in C_0^\infty(0,T)$ and $v \in \mathcal{V}$, we obtain that

$$\int_0^T \langle \mathcal{B}^* \tilde{\Lambda}_\tau, v\rangle \dot{\Phi} \, \mathrm{d}t \;\to\; \int_0^T \langle \mathcal{B}^* \tilde{\Lambda}, v\rangle \dot{\Phi} \, \mathrm{d}t.$$

Considering equation (10.14) and test functions $\Phi \in C_0^\infty(0,T)$, we obtain by the integration by parts formula

$$-\int_0^T \langle \mathcal{B}^* \tilde{\Lambda}_\tau, v\rangle \dot{\Phi} \, \mathrm{d}t = \int_0^T \Big[\langle \mathcal{F}_\tau, v\rangle - \big(\mathcal{B}^- \dot{\mathcal{G}}_\tau, v\big) - \langle \mathcal{K}\mathcal{B}^- \mathcal{G}_\tau, v\rangle - \big\langle \mathcal{K} U_{1,\tau}, v\big\rangle\Big] \Phi + \big(\hat{U}_{1,\tau}, v\big) \dot{\Phi} \, \mathrm{d}t.$$

Since we already know that $U_2 = \mathcal{B}^- \mathcal{G}$ and $V_2 = \mathcal{B}^- \dot{\mathcal{G}}$, we may pass to the limit as in the proof of Theorem 10.8 and obtain the equation

(10.17)
$$\int_0^T -\big(U_1, v\big)\dot{\Phi} + \big(V_2, v\big)\Phi + \big\langle \mathcal{K}(U_1 + U_2), v\big\rangle \Phi - \big\langle \mathcal{B}^* \tilde{\Lambda}, v\big\rangle \dot{\Phi} \, \mathrm{d}t = \int_0^T \big\langle \mathcal{F}, v\big\rangle \Phi \, \mathrm{d}t.$$

From Theorem 10.8 we know that U_1 satisfies the initial condition such that $(U_1, U_2, V_2, \tilde{\Lambda})$ by (10.17) solves the operator DAE (10.2) in the weak distributional sense defined in Section 4.3. □

Summarizing the above, we have seen that the approximations of u_1, u_2, and v_2 given by the implicit Euler scheme converge weakly to the solution of the regularized operator DAE (10.2) whereas for the Lagrange multiplier we only obtain the convergence in a weaker sense, namely in the weak distributional sense. To obtain the convergence of the sequence Λ_τ one may either use a time discretization of higher order or assume a higher regularity of the given data. For completeness we state the following result with additional regularity assumptions as in Remark 10.4. Since the argumentation for the convergence is as before, merely with stronger norms, we leave out the proof.

THEOREM 10.10 (Convergence for additional regularity). *Let the decomposition $\mathcal{V} = \mathcal{V}_\mathcal{B} \oplus \mathcal{V}^c$ be orthogonal w.r.t. the inner product $(\cdot,\cdot) + \langle \mathcal{K}\cdot, \cdot\rangle$ and assume $\mathcal{F} \in L^2(0,T;\mathcal{H}^*)$, $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$, and $u_1^0 = g_0 \in \mathcal{V}_\mathcal{B}$. Then, the (weak) limits $U_1 \in L^2(0,T;\mathcal{V}_\mathcal{B})$, U_2, $V_2 \in L^2(0,T;\mathcal{V}^c)$, and $\Lambda \in L^2(0,T;\mathcal{Q})$ solve the regularized operator DAE (10.2). In addition, we have $\dot{U}_1 \in L^2(0,T;\mathcal{H})$.*

To obtain conclusions on the convergence of the Rothe method, we have to include spatial discretization errors as well. For this, these errors may be interpreted as perturbations of the right-hand sides.

10.4. Influence of Perturbations. In this subsection we consider the semi-discrete version of system (10.2) with additional perturbations in the right-hand sides. We show that these perturbations can then be interpreted as the errors coming from a spatial discretization. Note that we still assume the operators to satisfy the assumptions from Section 10.1. The differences of the exact and perturbed solution $(\hat{u}_1^j, \hat{u}_2^j, \hat{v}_2^j, \hat{\lambda}^j)$, namely,

$$e_1^j := \hat{u}_1^j - u_1^j \in \mathcal{V}_\mathcal{B}, \quad e_2^j := \hat{u}_2^j - u_2^j \in \mathcal{V}^\mathrm{c}, \quad e_v^j := \hat{v}_2^j - v_2^j \in \mathcal{V}^\mathrm{c}, \quad e_\lambda^j := \hat{\lambda}^j - \lambda^j \in \mathcal{Q}$$

satisfy the equations

$$De_1^j + e_v^j + \mathcal{K}\big(e_1^j + e_2^j\big) + \mathcal{B}^* e_\lambda^j = \delta^j \quad \text{in } \mathcal{V}^*, \tag{10.18a}$$

$$\mathcal{B}e_2^j = \theta^j \quad \text{in } \mathcal{Q}^*, \tag{10.18b}$$

$$\mathcal{B}e_v^j = \xi^j \quad \text{in } \mathcal{Q}^*. \tag{10.18c}$$

Therein, we assume perturbations $\delta^j \in \mathcal{H}^*$ and $\theta^j, \xi^j \in \mathcal{Q}^*$. We analyse the positive effects of the regularization from Part B in terms of perturbations. Furthermore, we assume the spaces $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^\mathrm{c}$ to be orthogonal w.r.t. the inner product defined by $(\cdot,\cdot) + \langle\mathcal{K}\cdot,\cdot\rangle$. This property is needed to obtain an estimate of the Lagrange multiplier in terms of the perturbations. For the remaining variables it is sufficient to assume $\delta^j \in \mathcal{V}^*$.

Remark 10.11. If we assume that the perturbations are of the same order of magnitude for each time step, i.e., $\delta^j \approx \delta \in \mathcal{H}^*$, $\theta^j \approx \theta \in \mathcal{Q}^*$, and $\xi^j \approx \xi \in \mathcal{Q}^*$ for all $j = 1, \dots, n$, then we may summarize, e.g.,

$$\tau \sum_{j=1}^{n} \|\delta^j\|_{\mathcal{V}^*} \approx \tau n \|\delta\|_{\mathcal{V}^*} = T\|\delta\|_{\mathcal{V}^*}.$$

Remark 10.12 (Index-2 formulation). If we consider the index-2 type formulation instead of the regularized operator DAE, then equation (10.18c) has to be replaced by $\mathcal{B}De_2^j = D\theta^j$. Thus, the perturbation ξ^j has to be replaced by the discrete derivative of θ^j which then leads to an additional $1/\tau$ term in the error estimates.

10.4.1. *Error Analysis.* By equations (10.18b) and (10.18c) we directly obtain the estimates

$$\|e_2^j\| \le C_{\mathcal{B}^-}\|\theta^j\|_{\mathcal{Q}^*}, \qquad \|e_v^j\| \le C_{\mathcal{B}^-}\|\xi^j\|_{\mathcal{Q}^*}. \tag{10.19}$$

With the same calculation as in Lemma 10.2, i.e., testing equation (10.18a) by e_1^j, we obtain for $k = 1, \dots, n$ the estimate

$$\begin{aligned} &|e_1^k|^2 + \tau^2 \sum_{j=1}^{k} |De_1^j|^2 + \tau k_1 \sum_{j=1}^{k} \|e_1^j\|^2 \\ &\qquad \le |e_1^0|^2 + \frac{3\tau}{k_1} \sum_{j=1}^{k} \Big(\|\delta^j\|_{\mathcal{V}^*}^2 + k_2^2 C_{\mathcal{B}^-}^2 \|\theta^j\|_{\mathcal{Q}^*}^2 + C_{\mathcal{B}^-}^2 C_{\mathrm{emb}}^4 \|\xi^j\|_{\mathcal{Q}^*}^2 \Big). \end{aligned} \tag{10.20}$$

Since we have assumed $\delta \in \mathcal{H}^*$, cf. Remark 10.4, we obtain an additional result if we test (10.18a) by De_1^j, namely

$$|De_1^j|^2 + \langle\mathcal{K}e_1^j, De_1^j\rangle = \langle\delta^j, De_1^j\rangle - (e_v^j, De_1^j) + (e_2^j, De_1^j).$$

Note that we have used here the assumed orthogonality of $\mathcal{V}_\mathcal{B}$ and $\mathcal{V}^\mathrm{c}$ w.r.t. $(\cdot,\cdot) + \langle\mathcal{K}\cdot,\cdot\rangle$. Using the equality $2\langle\mathcal{K}e_1^j, De_1^j\rangle = D\langle\mathcal{K}e_1^j, e_1^j\rangle + \tau\langle\mathcal{K}De_1^j, De_1^j\rangle$, cf. equation (10.6), and the

Cauchy-Schwarz inequality, we further obtain

$$2|De_1^j|^2 + D\langle \mathcal{K}e_1^j, e_1^j\rangle + \tau k_1 \|De_1^j\|^2 \le 2\|\delta^j\|_{\mathcal{H}^*}|De_1^j| + 2C_{\mathrm{emb}}\|e_v^j\||De_1^j| + 2C_{\mathrm{emb}}\|e_2^j\||De_1^j|.$$

Next, we apply Young's inequality on the right-hand side and get

$$|De_1^j|^2 + D\langle \mathcal{K}e_1^j, e_1^j\rangle + \tau k_1 \|De_1^j\|^2 \le 3\|\delta^j\|_{\mathcal{H}^*}^2 + 3C_{\mathrm{emb}}^2\|e_v^j\|^2 + 3C_{\mathrm{emb}}^2\|e_2^j\|^2.$$

A summation for $j = 1, \dots, k$ and a multiplication by τ finally leads to

$$\begin{aligned} &k_1\|e_1^k\|^2 + \tau\sum_{j=1}^k |De_1^j|^2 + \tau^2 k_1 \sum_{j=1}^k \|De_1^j\|^2 \\ &\qquad \le k_2\|e_1^0\|^2 + 3\tau\sum_{j=1}^k \Big(\|\delta^j\|_{\mathcal{H}^*}^2 + C_{\mathrm{emb}}^2\|e_2^j\|^2 + C_{\mathrm{emb}}^2\|e_v^j\|^2\Big). \end{aligned} \tag{10.21}$$

For an estimate of the Lagrange multiplier e_λ^j we use equation (10.18a) and the inf-sup property of $\mathcal{B}$ to obtain

$$\beta_{\mathrm{inf}}\|e_\lambda^j\|_{\mathcal{Q}} \le \sup_{v\in\mathcal{V}^{\mathrm{c}}} \frac{\langle \mathcal{B}^* e_\lambda^j, v\rangle}{\|v\|} \le \|\delta^j\|_{\mathcal{V}^*} + k_2\|e_1^j\| + k_2\|e_2^j\| + C_{\mathrm{emb}}^2\|e_v^j\| + C_{\mathrm{emb}}|De_1^j|.$$

Combining this estimate with the results in (10.19) and (10.21), we obtain with a generic constant, which we express with the relation symbol $\lesssim$, that

$$\begin{aligned} \tau\beta_{\mathrm{inf}}^2 \sum_{j=1}^k \|e_\lambda^j\|_{\mathcal{Q}}^2 &\overset{(10.19)}{\lesssim} \tau\sum_{j=1}^k \big(\|\delta^j\|_{\mathcal{V}^*}^2 + \|\theta^j\|_{\mathcal{Q}^*}^2 + \|\xi^j\|_{\mathcal{Q}^*}^2\big) + \tau\sum_{j=1}^k \big(\|e_1^j\|^2 + |De_1^j|^2\big) \\ &\overset{(10.21)}{\lesssim} \|e_1^0\|^2 + \tau\sum_{j=1}^k \big(\|\delta^j\|_{\mathcal{H}^*}^2 + \|\theta^j\|_{\mathcal{Q}^*}^2 + \|\xi^j\|_{\mathcal{Q}^*}^2\big). \end{aligned} \tag{10.22}$$

We summarize the results of this section in a theorem. For this, we define similarly as in Section 10.3 piecewise constant functions E_1, E_2, $E_v\colon [0,T] \to \mathcal{V}$ and $E_\lambda\colon [0,T] \to \mathcal{Q}$ by

$$E_1(t) = e_1^j, \qquad E_2(t) = e_2^j, \qquad E_v(t) = e_v^j, \qquad E_\lambda(t) = e_\lambda^j \tag{10.23}$$

for $t \in]t_{j-1}, t_j]$.

Theorem 10.13. *Consider system* (10.18) *where the operators $\mathcal{K}$ and $\mathcal{B}$ satisfy the assumptions stated in Section 10.1 and with perturbations $\delta^j \in \mathcal{H}^*$ and θ^j, $\xi^j \in \mathcal{Q}^*$ which are all of the same order of magnitude as in Remark 10.11. Furthermore, let $\mathcal{V}_{\mathcal{B}}$ and $\mathcal{V}^c$ be orthogonal w.r.t. the inner product $(\cdot,\cdot) + \langle \mathcal{K}\cdot,\cdot\rangle$ and let the initial error satisfy $e_1^0 \in \mathcal{V}_{\mathcal{B}}$. Then, E_1, E_2, E_v, and E_λ from* (10.23) *satisfy*

$$\begin{aligned} \|E_1\|_{L^2(0,T;\mathcal{V})} &\lesssim |e_1^0| + \sqrt{T}\big(\|\delta\|_{\mathcal{V}^*} + \|\theta\|_{\mathcal{Q}^*} + \|\xi\|_{\mathcal{Q}^*}\big), \\ \|E_1\|_{L^\infty(0,T;\mathcal{V})} &\lesssim \|e_1^0\| + \sqrt{T}\big(\|\delta\|_{\mathcal{H}^*} + \|\theta\|_{\mathcal{Q}^*} + \|\xi\|_{\mathcal{Q}^*}\big), \\ \|E_2\|_{L^2(0,T;\mathcal{V})} &\lesssim \sqrt{T}\|\theta\|_{Q^*}, \qquad \|E_v\|_{L^2(0,T;\mathcal{V})} \lesssim \sqrt{T}\|\xi\|_{Q^*}, \\ \beta_{inf}\,\|E_\lambda\|_{L^2(0,T;\mathcal{Q})} &\lesssim \|e_1^0\| + \sqrt{T}\big(\|\delta\|_{\mathcal{H}^*} + \|\theta\|_{\mathcal{Q}^*} + \|\xi\|_{\mathcal{Q}^*}\big). \end{aligned}$$

Proof. By (10.19) we directly obtain

$$\|E_2\|_{L^2(0,T;\mathcal{V})}^2 = \tau\sum_{j=1}^n \|e_2^j\|^2 \lesssim \tau\sum_{j=1}^n \|\theta^j\|_{\mathcal{Q}^*}^2 \approx T\|\theta\|_{\mathcal{Q}^*}^2.$$

The result for E_v follows accordingly. The two estimates for E_1 are implied by (10.20),

$$\|E_1\|^2_{L^2(0,T;\mathcal{V})} = \tau \sum_{j=1}^{n} \|e_1^j\|^2 \lesssim |e_1^0|^2 + \tau \sum_{j=1}^{n} \Big(\|\delta^j\|^2_{\mathcal{V}^*} + \|\theta^j\|^2_{\mathcal{Q}^*} + \|\xi^j\|^2_{\mathcal{Q}^*} \Big)$$

and (10.21),

$$\|E_1\|^2_{L^\infty(0,T;\mathcal{V})} = \max_k \|e_1^k\|^2 \lesssim \|e_1^0\|^2 + \tau \sum_{j=1}^{n} \Big(\|\delta^j\|^2_{\mathcal{H}^*} + \|\theta^j\|^2_{\mathcal{Q}^*} + \|\xi^j\|^2_{\mathcal{Q}^*} \Big).$$

Finally, by (10.22), we gain the result for the Lagrange multiplier in the same manner. □

10.4.2. *Spatial Discretization as Perturbation.* In Theorem 10.13 we have shown that the convergence of the semi-discrete solution is maintained when the perturbations tend to zero as $\tau \to 0$. In order to solve the semi-discrete equations (10.3)-(10.5), we need a spatial discretization which again produces numerical errors. In the sequel we show that this discretization error can be seen as perturbation of the semi-discrete system.

Remark 10.14. We emphasize that we have not proven any order of convergence for the Euler discretization. Thus, we cannot provide specific requirements on the needed accuracy of the spatial discretization. This, however, is essential for efficient computations to ensure that the discretization in space is neither too fine, i.e., too expensive, nor too coarse.

Let $(u_1^j, u_2^j, v_2^j, \lambda^j)$ denote the (exact in space) solution of the semi-discretized operator DAE as discussed in Section 10.2. Solving the PDEs by a *conform* finite element scheme, we obtain fully discrete approximations $(u_{1,h}^j, u_{2,h}^j, v_{2,h}^j, \lambda_h^j)$. Thereby, the index h denotes that we have applied a discretization scheme based on a triangulation with mesh parameter h. These approximations are given by the discrete variational problem

$$\begin{aligned} \big(Du_{1,h}^j, v_h\big) + \quad \big(v_{2,h}^j, v_h\big) \;+ \big\langle \mathcal{K}\big(u_{1,h}^j + u_{2,h}^j\big), v_h \big\rangle + \big\langle \mathcal{B}^* \lambda_h^j, v_h \big\rangle &= \big\langle \mathcal{F}^j, v_h \big\rangle, \\ \big\langle \mathcal{B} u_{2,h}^j, q_h \big\rangle &= \big\langle \mathcal{G}^j, q_h \big\rangle, \\ \big\langle \mathcal{B} v_{2,h}^j, q_h \big\rangle &= \big\langle \dot{\mathcal{G}}^j, q_h \big\rangle \end{aligned}$$

for all discrete test functions $v_h \in V_h \subset \mathcal{V}$ and $q_h \in Q_h \subset \mathcal{Q}$. For the error analysis of the spatial discretization error, one often looks at the residuals which are given as functionals of the form

$$\begin{aligned} \text{(10.24a)} \quad \big\langle \operatorname{Res}_1^j, v \big\rangle &:= \big\langle \mathcal{F}^j, v \big\rangle - \big(Du_{1,h}^j, v\big) - \big(v_{2,h}^j, v\big) - \big\langle \mathcal{K}\big(u_{1,h}^j + u_{2,h}^j\big), v \big\rangle - \big\langle \mathcal{B}^* \lambda_h^j, v \big\rangle \\ &= \big(De_1^j, v\big) + \big(e_v^j, v\big) + \big\langle \mathcal{K}\big(e_1^j + e_2^j\big), v \big\rangle + \big\langle \mathcal{B}^* e_\lambda^j, v \big\rangle, \\ \text{(10.24b)} \quad \big\langle \operatorname{Res}_2^j, q \big\rangle &:= \big\langle \mathcal{G}^j, q \big\rangle - \big\langle \mathcal{B} u_{2,h}^j, q \big\rangle = \big\langle \mathcal{B} e_2^j, q \big\rangle, \\ \text{(10.24c)} \quad \big\langle \operatorname{Res}_v^j, q \big\rangle &:= \big\langle \dot{\mathcal{G}}^j, q \big\rangle - \big\langle \mathcal{B} v_{2,h}^j, q \big\rangle = \big\langle \mathcal{B} e_v^j, q \big\rangle. \end{aligned}$$

Therein, we use the abbreviations $e_1^j := u_1^j - u_{1,h}^j$, $e_2^j := u_2^j - u_{2,h}^j$, $e_v^j := v_2^j - v_{2,h}^j$, and $e_\lambda^j := \lambda^j - \lambda_h^j$. Note that the residuals vanish on the discrete test spaces, which is also known as the *Galerkin orthogonality*.

Considering the definition of the residuals in (10.24), we note that they may be interpreted as the perturbations δ^j, θ^j and ξ^j from the beginning of Section 10.4. From this definition, we directly see that $\operatorname{Res}_1^j \in \mathcal{V}^*$ as well as Res_2^j, $\operatorname{Res}_v^j \in \mathcal{Q}^*$. Thus, we may apply the results from Section 10.4.1 for the errors e_1^j, e_2^j, and e_v^j. However, in order to obtain the corresponding results for the Lagrange multiplier from Theorem 10.13, we need

$\mathrm{Res}_1^j \in \mathcal{H}^*$. This assumption is certainly not given for all kinds of discretizations but may be reached with an appropriate discretization scheme of higher order.

REMARK 10.15. For spatial discretizations of nonconforming type, i.e., $V_h \not\subset \mathcal{V}$, we may have $e_1^j := u_1^j - u_{1,h}^j \notin \mathcal{V}_\mathcal{B}$. In this case, the achieved perturbation results from the previous subsection are not applicable and a different treatment is necessary.

10.5. Nonlinear Constraints. In this final section on first-order systems, we consider the operator DAEs from Section 6.2 with a nonlinear constraint operator $\mathcal{B}$. To show general results on the convergence of the Euler scheme for this case, one would need specific assumptions on the constraint operator such as weak-weak continuity or the strong convergence of $U_{1,\tau}$ in the energy norm. However, it is not clear whether these assumptions are reasonable or realistic in applications. In order to stay within a reasonable framework, we consider here the example of the regularized Stefan problem from Section 6.3.3.

The equations of motion are given in system (6.15). Written as operator DAE with the spaces

$$\mathcal{V} := H^1(\Omega), \quad \mathcal{H} := L^2(\Omega), \quad \mathcal{V}_1 := H_0^1(\Omega), \quad \mathcal{V}_2 := \big[H_0^1(\Omega)\big]^{\perp_\mathcal{V}}, \quad \mathcal{Q}^* := H^{1/2}(\partial\Omega),$$

we obtain system (6.16) which has the form

$$\dot{u} + \mathcal{K}u + \mathcal{C}_u^*\lambda = \mathcal{F} \quad \text{in } \mathcal{V}^*, \tag{10.25a}$$

$$\mathcal{B}u = \mathcal{G} \quad \text{in } \mathcal{Q}^*. \tag{10.25b}$$

According to Section 6.2 and equation (6.9), the regularized operator DAE has the form

$$\dot{u}_1 + v_2 + \mathcal{K}(u_1 + u_2) + \mathcal{C}_u^*\lambda = \mathcal{F} \quad \text{in } \mathcal{V}^*, \tag{10.26a}$$

$$\mathcal{B}u_2 = \mathcal{G} \quad \text{in } \mathcal{Q}^*, \tag{10.26b}$$

$$\mathcal{C}_{2,u}v_2 = \dot{\mathcal{G}} \quad \text{in } \mathcal{Q}^* \tag{10.26c}$$

with initial condition

$$u_1(0) = g_0 \in \mathcal{V}_1. \tag{10.26d}$$

The solution (u_1, u_2, v_2, λ) should satisfy $u_1 \in W^{1;2,2}(0,T;\mathcal{V}_1;\mathcal{V}^*)$, u_2, $v_2 \in L^2(0,T;\mathcal{V}_2)$, and $\lambda \in L^2(0,T;\mathcal{Q})$. Note that we have needed $u_1 \in H^1(0,T;\mathcal{V}_1)$ in the general case of Section 6.2. This is not necessary here, since the nonlinear constraint operator $\mathcal{B}$ vanishes on $\mathcal{V}_1$ such that $\dot{u}_1 \in L^2(0,T;\mathcal{V}^*)$ is sufficient. However, we stay with the assumption that the initial data g_0 is given in $\mathcal{V}_1$. The included operators $\mathcal{B}\colon \mathcal{V} \to \mathcal{Q}^*$ and $\mathcal{K}\colon \mathcal{V} \to \mathcal{V}^*$ are given by

$$\big\langle \mathcal{B}u, q\big\rangle_{\mathcal{Q}^*,\mathcal{Q}} = \int_{\partial\Omega} \beta(u) q \,\mathrm{d}x, \qquad \big\langle \mathcal{K}u, v\big\rangle_{\mathcal{V}^*,\mathcal{V}} = \int_\Omega \nabla\beta(u)\cdot\nabla v \,\mathrm{d}x. \tag{10.27}$$

The nonlinear enthalpy-temperature function $\beta\colon \mathbb{R} \to \mathbb{R}$ was assumed to be strictly monotonically increasing and continuously differentiable with $\beta' \geq \varepsilon > 0$. Furthermore, there exist positive constants c and C such that $\mathrm{sign}(s)\beta(s) \geq c|s| - C$. As a result, the inverse of β satisfies

$$\big|\beta^{-1}(s)\big| \leq c^{-1}|s| + c^{-1}C \tag{10.28}$$

and

$$\big|\beta^{-1}(x) - \beta^{-1}(y)\big| \leq \max_{\xi\in\mathbb{R}} \frac{1}{\beta'(\beta^{-1}(\xi))}|x-y| \leq \frac{1}{\varepsilon}|x-y|. \tag{10.29}$$

As in Section 6.3.3 we need some additional assumptions for the analysis of the Euler scheme. We assume that β' and β^{-1} are Lipschitz continuous and a bound of the form

$\|1/\beta'(\gamma u)\|_{\mathcal{Q}^*} \le C_{\beta\gamma}$ for a.e. $t \in [0,T]$ and u denoting the solution of the Stefan problem. These assumptions on the enthalpy-temperature function imply several properties of the operator $\mathcal{B}$ and its Fréchet derivative $\mathcal{C}_{\bar u}\colon \mathcal{V} \to \mathcal{Q}^*$,

$$\mathcal{C}_{\bar u} := \frac{\partial \mathcal{B}}{\partial u}(\bar u)\colon \mathcal{V} \to \mathcal{Q}^*, \qquad \mathcal{C}_{\bar u} v := \beta'(\gamma \bar u)\cdot \gamma v.$$

Recall that γ denotes the trace operator from Section 3.1.4. The restriction of $\mathcal{C}_{\bar u}$ to the subspace $\mathcal{V}_2$ is again denoted by $\mathcal{C}_{2,\bar u}$.

LEMMA 10.16. *Consider the operator $\mathcal{B}\colon \mathcal{V} \to \mathcal{Q}^*$, its Fréchet derivative $\mathcal{C}_{\bar u}\colon \mathcal{V} \to \mathcal{Q}^*$, as well as $\mathcal{K}\colon \mathcal{V} \to \mathcal{V}^*$ from* (10.27) *with the enthalpy-temperature function β. Furthermore, let $\mathcal{U}_M \subset \mathcal{V}$ denote the ball with functions satisfying $\|u\|_{L^\infty(\Omega)} \le M$ and M large enough such that the solution of the regularized Stefan problem* (10.25) *satisfies $u(t) \in \mathcal{U}_M$ for a.e. $t \in [0,T]$. Then,*

(a) the Fréchet derivative $\mathcal{C}_{\bar u}$ is continuous with constant $C_{tr}\|\beta'(\gamma\bar u)\|_{\mathcal{Q}^*}$,

(b) along the solution of the Stefan problem u, the operator $\mathcal{C}_{2,u}$ has a continuous inverse, i.e., there exists a constant $C_{\mathcal{C}_2 inv}$ such that $\|{\mathcal{C}_{2,u}}^{-1} q\| \le C_{\mathcal{C}_2 inv}\|q\|_{\mathcal{Q}^*}$,

(c) the operator $\mathcal{K}$ is monotone and positive on $\mathcal{V}_1$, i.e., we have $k_1\|u\|^2 \le \langle \mathcal{K}u, u\rangle$ for all $u \in \mathcal{V}_1$, and

(d) there exists a constant $k_2 > 0$ such that for all $u \in \mathcal{U}_M$ and v, $w \in \mathcal{V}$ it holds that

$$\int_\Omega \beta'(u)\nabla v \cdot \nabla w \, dx \le k_2 \|v\|\|w\|.$$

In particular, the operator $\mathcal{K}$ is continuous in $\mathcal{U}_M$ with constant k_2.

PROOF. (a) Since $\bar u \in \mathcal{V}$ implies $\beta'(\gamma\bar u) \in \mathcal{Q}^*$ by the assumed Lipschitz continuity of β', for $v \in \mathcal{V}$ we obtain

$$\|\mathcal{C}_{\bar u} v\|_{\mathcal{Q}^*} = \|\beta'(\gamma\bar u)\cdot\gamma v\|_{\mathcal{Q}^*} \le \|\beta'(\gamma\bar u)\|_{\mathcal{Q}^*}\|\gamma v\|_{\mathcal{Q}^*} \le \|\beta'(\gamma\bar u)\|_{\mathcal{Q}^*} C_{\mathrm{tr}}\|v\|.$$

(b) The inverse of $\mathcal{C}_{2,\bar u}$ is given by $q \mapsto \gamma^{-1}\big(\frac{1}{\beta'(\gamma\bar u)} q\big)$ where $\gamma^{-1}\colon \mathcal{Q}^* \to \mathcal{V}_2$ denotes the inverse trace operator, cf. Section 3.1.4. This operator is linear and from the continuity of the inverse trace operator, cf. Theorem 3.15, we obtain

$$\|{\mathcal{C}_{2,u}}^{-1} q\| \le C_{\mathrm{invTr}}\|1/\beta'(\gamma u)\|_{\mathcal{Q}^*}\|q\|_{\mathcal{Q}^*} \le C_{\mathrm{invTr}} C_{\beta\gamma}\|q\|_{\mathcal{Q}^*} =: C_{\mathcal{C}_2 \mathrm{inv}}\|q\|_{\mathcal{Q}^*}.$$

(c) The monotonicity of the operator $\mathcal{K}$ follows from the (strict) monotonicity of β. For this, we may define $w \in \mathcal{V}$ pointwise by u or v such that

$$\begin{aligned}\big\langle \mathcal{K}u - \mathcal{K}v, u - v\big\rangle &= \int_\Omega \beta'(u)\nabla u\cdot\nabla(u-v) - \beta'(v)\nabla v\cdot\nabla(u-v)\,\mathrm{d}x\\ &\ge \int_\Omega \beta'(w)\nabla(u-v)\cdot\nabla(u-v)\,\mathrm{d}x\\ &\ge \varepsilon\,\big|\nabla(u-v)\big|^2 \ge 0.\end{aligned}$$

Also the positivity on $\mathcal{V}_1$ follows from the strict monotonicity of β, namely

$$\big\langle \mathcal{K}u, u\big\rangle = \int_\Omega \nabla\beta(u)\cdot\nabla u\,\mathrm{d}x = \int_\Omega \beta'(u)\nabla u\cdot\nabla u\,\mathrm{d}x \ge \varepsilon|\nabla u|^2.$$

For $u \in \mathcal{V}_1$ the term $|\nabla u|$ is bounded (up to a constant) from below by $\|u\|$. Thus, there exists a positive constant k_1 such that the right-hand side is bounded by $k_1\|u\|^2$.
(d) By the definition of the set $\mathcal{U}_M$ we obtain

$$\int_\Omega \beta'(u)\nabla v\cdot\nabla w\,\mathrm{d}x \le \|\beta'(u)\|_{L^\infty(\Omega)}\|v\|\|w\|.$$

Thus, the claim follows with $k_2 := \beta'(M) < \infty$. □

10.5.1. *Temporal Discretization.* For the discretization we consider again the regularized version of the operator DAE (10.25). Thus, we apply the implicit Euler scheme to system (10.26), where we search for approximations u_1^j, u_2^j, v_2^j, and λ^j of u_1, u_2, v_2, and λ at time $t = t_j$, respectively. Throughout this section, we assume that the discrete approximation is close enough to the exact solution in the sense that $u_1^j + u_2^j \in \mathcal{U}_M$ with the set $\mathcal{U}_M$ introduced in Lemma 10.16. Furthermore, we have to assume that

$$\|1/\beta'(\gamma u_2^j)\|_{\mathcal{Q}^*} = \|1/\beta'(\beta^{-1}(\mathcal{G}^j))\|_{\mathcal{Q}^*} \tag{10.30}$$

is uniformly bounded, i.e., the property from Lemma 10.16 (b) also applies along the discrete solution $u_1^j + u_2^j$.

The semi-discrete system for one time step of the Stefan problem has the form

$$Du_1^j + \quad v_2^j \quad + \mathcal{K}(u_1^j + u_2^j) + \mathcal{C}_{u^j}^* \lambda^j = \mathcal{F}^j \quad \text{in } \mathcal{V}^*, \tag{10.31a}$$

$$\mathcal{B}u_2^j \qquad\qquad = \mathcal{G}^j \quad \text{in } \mathcal{Q}^*, \tag{10.31b}$$

$$\mathcal{C}_{2,u^j} v_2^j \qquad\qquad = \dot{\mathcal{G}}^j \quad \text{in } \mathcal{Q}^*. \tag{10.31c}$$

Note that we write $\mathcal{C}_{u^j}$ and $\mathcal{C}_{2,u^j}$ for the Fréchet derivative of $\mathcal{B}$ at $u_1^j + u_2^j$ for the purpose of notation. The unique solvability of system (10.31) and thus, the existence of a discrete approximation is shown in the following lemma.

Lemma 10.17. *With the assumptions introduced in this subsection, system* (10.31) *has a unique solution* $(u_1^j, u_2^j, v_2^j, \lambda^j)$ *for each time step* $j = 1, \dots, n$.

Proof. Equation (10.31b) is uniquely solvable, since β is injective and the trace operator is invertible as operator from $\mathcal{V}_2$ to $\mathcal{Q}^*$. The invertibility of $\mathcal{C}_{2,u^j}$ leads to a unique solution v_2^j. Next, consider equation (10.31a) restricted to test functions in $\mathcal{V}_1$. With the operator $\mathcal{A}\colon \mathcal{V}_1 \to \mathcal{V}_1^*$,

$$\langle \mathcal{A}u, v\rangle := \frac{1}{\tau}(u, v) + \langle \mathcal{K}(u + u_2^j), v\rangle$$

we may write this equation in the form $\mathcal{A}u_1^j = \hat{\mathcal{F}}^j := \mathcal{F}^j - v_2^j + u_1^{j-1}/\tau$. The continuity assumptions on $\mathcal{K}$ imply that $\mathcal{A}$ is hemicontinuous. Furthermore, $\mathcal{A}$ is strongly monotone because of Lemma 10.16 (c) and the fact that $|\nabla \cdot|$ is equivalent to $\|\cdot\|$ for functions in $\mathcal{V}_1$. This also implies the coercivity of the operator, cf. [**GGZ74**, Ch. III, Rem. 1.4]. Finally, the Browder-Minty theorem [**GGZ74**, Ch. III, Th. 2.1] yields the existence of a solution u_1^j. The uniqueness follows again from the strong monotonicity of $\mathcal{A}$. If we consider equation (10.31a) tested by functions in $\mathcal{V}_2$, we obtain a unique solution λ^j again by the invertibility of $\mathcal{C}_{2,u^j}$. □

To obtain an estimate of u_2^j we consider equation (10.31b). Let $|\partial\Omega|$ denote the $(d-1)$-dimensional measure of the boundary. Then, the Lipschitz continuity of β^{-1} and properties (10.28) and (10.29) yield the estimate

$$\|u_2^j\| \overset{(10.31b)}{\leq} C_{\mathrm{invTr}} \big\|\beta^{-1}\mathcal{G}^j\big\|_{\mathcal{Q}^*} \overset{(10.28),(10.29)}{\leq} c\,\big(\|\mathcal{G}^j\|_{\mathcal{Q}^*} + |\partial\Omega|^{1/2}\big).$$

Note that c denotes here a generic constant and that the proof of this estimate uses the definition of the $\mathcal{Q}^*$-norm as given in Lemma 5.11. By the second constraint (10.31c) and Lemma 10.16 (b) for the discrete solution, we calculate

$$\|v_2^j\| \leq \big\|{\mathcal{C}_{2,u^j}}^{-1}\dot{\mathcal{G}}^j\big\| \leq C_{\mathcal{C}_2\mathrm{inv}} \|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}.$$

Note that we have used here that the discrete solution is closed enough to the exact solution in the sense that (10.30) is uniformly bounded.

To obtain a stability estimate of u_1^j we proceed similarly as in the linear case in Section 10.2.2 and test equation (10.31a) by u_1^j. Note that in this example the operator $\mathcal{K}$ is nonlinear such that slight modifications are necessary as shown in the following lemma.

LEMMA 10.18. *Assume $\mathcal{F} \in L^2(0,T;\mathcal{V}_1^*)$, $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$, $u_1^0 \in \mathcal{H}$, and the operators $\mathcal{B}$ and $\mathcal{K}$ from* (10.27)*. Furthermore, let the discrete approximation satisfy $u_1^j + u_2^j \in \mathcal{U}_M$ for all $j = 1, \dots, n$ with the set $\mathcal{U}_M$ introduced in Lemma 10.16. Then, there exists a positive constant c such that for all $1 \le k \le n$ it holds that*

$$|u_1^k|^2 + \tau^2 \sum_{j=1}^{k} |Du_1^j|^2 + \tau k_1 \sum_{j=1}^{k} \|u_1^j\|^2 \le |u_1^0|^2 + c\Big[\|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}_1^*)} + \|\mathcal{G}\|^2_{H^1(0,T;\mathcal{Q}^*)} + T|\partial\Omega|\Big].$$

PROOF. Since we mainly follow the lines of the proof of Lemma 10.2, we only stress the differences compared to the linear case. Applying the test function $v = u_1^j \in \mathcal{V}_1$, $j \ge 1$ in equation (10.31a), we obtain

$$\big(Du_1^j, u_1^j\big) + \big\langle \mathcal{K}(u_1^j + u_2^j), u_1^j \big\rangle = \big\langle \mathcal{F}^j, u_1^j \big\rangle - \big(v_2^j, u_1^j\big).$$

Although the operator $\mathcal{K}$ is nonlinear, we may split

$$\big\langle \mathcal{K}(u_1^j + u_2^j), u_1^j \big\rangle = \int_\Omega \beta'(u_1^j + u_2^j) \nabla u_1^j \cdot \nabla u_1^j \,\mathrm{d}x + \int_\Omega \beta'(u_1^j + u_2^j) \nabla u_2^j \cdot \nabla u_1^j \,\mathrm{d}x.$$

As in Lemma 10.16 (c), a lower bound of the first term is given by $k_1 \|u_1^j\|^2$. For the second term we apply Lemma 10.16 (d) which yields an upper bound of the form $k_2 \|u_2^j\| \|u_1^j\|$. Thus, we obtain the overall estimate

$$D|u_1^j|^2 + \tau |Du_1^j|^2 + 2k_1 \|u_1^j\|^2 \le 2\|\mathcal{F}^j\|_{\mathcal{V}_1^*} \|u_1^j\| + 2C_{\mathrm{emb}}^2 \|v_2^j\| \|u_1^j\| + 2k_2 \|u_2^j\| \|u_1^j\|.$$

Up to the constants, this estimate is as in the linear case. Thus, we may again sum over $j = 1, \dots, k$ and use the Cauchy-Schwarz inequality which then finally leads to a constant $c > 0$ such that

$$|u_1^k|^2 + \tau^2 \sum_{j=1}^{k} |Du_1^j|^2 + \tau k_1 \sum_{j=1}^{k} \|u_1^j\|^2 \le |u_1^0|^2 + c\Big[\|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}_1^*)} + \|\mathcal{G}\|^2_{H^1(0,T;\mathcal{Q}^*)} + T|\partial\Omega|\Big].$$

Note that we have used here the estimates of u_2^j and v_2^j from the beginning of this subsection as well as property (5.13) for the right-hand sides. □

REMARK 10.19. Analog to Lemma 10.3 in the linear case, with the same assumptions as in Lemma 10.18, we obtain the boundedness of $\tau \sum_{j=1}^{n} \|Du_1^j\|^2_{\mathcal{V}_1^*}$.

10.5.2. *Convergence Results.* We define the global approximations $U_{1,\tau}, \hat{U}_{1,\tau} \colon [0,T] \to \mathcal{V}_1$ and $U_{2,\tau}, V_{2,\tau} \colon [0,T] \to \mathcal{V}_2$ as before in Section 10.3.1 and $\Lambda_\tau \colon [0,T] \to \mathcal{Q}$ as in Section 10.3.2. Furthermore, $\mathcal{F}_\tau \colon [0,T] \to \mathcal{V}^*$ and $\mathcal{G}_\tau, \dot{\mathcal{G}}_\tau \colon [0,T] \to \mathcal{Q}^*$ are defined as in Section 10.2, i.e., as piecewise constant approximations of $\mathcal{F}$, $\mathcal{G}$, and $\dot{\mathcal{G}}$, respectively.

Then, system (10.31) can then be expressed in the form

$$\dot{\hat{U}}_{1,\tau} + \quad V_{2,\tau} \quad + \mathcal{K}\big(U_{1,\tau} + U_{2,\tau}\big) + \mathcal{C}_{U_\tau}^* \Lambda_\tau = \mathcal{F}_\tau \quad \text{in } \mathcal{V}^*, \tag{10.32a}$$

$$\mathcal{B}U_{2,\tau} \qquad\qquad = \mathcal{G}_\tau \quad \text{in } \mathcal{Q}^*, \tag{10.32b}$$

$$\mathcal{C}_{2,U_\tau} V_{2,\tau} \qquad\qquad = \dot{\mathcal{G}}_\tau \quad \text{in } \mathcal{Q}^*. \tag{10.32c}$$

Again we use the short notation $\mathcal{C}_{U_\tau}$ and $\mathcal{C}_{2,U_\tau}$ for the Fréchet derivative of $\mathcal{B}$ in $U_{1,\tau}+U_{2,\tau}$. As a first result, we show that $U_{2,\tau}$ and $V_{2,\tau}$ converge to the solution of the constraint (10.26b) and its derivative, respectively.

THEOREM 10.20. *Assume $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$ and $U_{2,\tau}$, $V_{2,\tau}$ given by equations* (10.32b) *and* (10.32c)*, respectively. Then, $U_{2,\tau} \to U_2$ and $V_{2,\tau} \to V_2$ in $L^2(0,T;\mathcal{V}_2)$ where U_2 and V_2 solve the equations $\mathcal{B}U_2 = \mathcal{G}$ and $\mathcal{C}_{2,U_2}V_2 = \dot{\mathcal{G}}$ in $\mathcal{Q}^*$, i.e., equations* (10.26b) *and* (10.26c).

PROOF. By the definition of the operator $\mathcal{B}$ and the approximation of the right-hand side $\mathcal{G}_\tau$, we have

$$\mathcal{B}U_{2,\tau} = \beta\big(\gamma U_{2,\tau}\big) = \mathcal{G}_\tau \to \mathcal{G} \quad \text{in } L^2(0,T;\mathcal{Q}^*).$$

It follows from Lemma 5.11 that also $\beta^{-1}\mathcal{G}_\tau \to \beta^{-1}\mathcal{G}$ in $L^2(0,T;\mathcal{Q}^*)$. The linearity of the inverse trace operator then gives $U_{2,\tau} = \gamma^{-1}(\beta^{-1}\mathcal{G}_\tau) \to U_2 := \gamma^{-1}(\beta^{-1}\mathcal{G})$ in $L^2(0,T;\mathcal{V}_2)$. For the second claim we have by assumption

$$\mathcal{C}_{2,U_\tau}V_{2,\tau} = \beta'(\gamma U_{2,\tau})\cdot\gamma V_{2,\tau} = \dot{\mathcal{G}}_\tau, \qquad \mathcal{C}_{2,U_2}V_2 = \beta'(\gamma U_2)\cdot\gamma V_2 = \dot{\mathcal{G}}.$$

Using the continuity of the inverse of $\mathcal{C}_{2,U_2}$ from Lemma 10.16 (b), we obtain

$$\big\|V_{2,\tau} - V_2\big\|_{L^2(0,T;\mathcal{V})} \le C_{\mathcal{C}_2\mathrm{inv}}\big\|\mathcal{C}_{2,U_2}V_{2,\tau} - \mathcal{C}_{2,U_2}V_2\big\|_{L^2(0,T;\mathcal{Q}^*)}.$$

The triangle inequality then leads to

$$\begin{aligned}\big\|V_{2,\tau} - V_2\big\|_{L^2(0,T;\mathcal{V})} &\lesssim \big\|\mathcal{C}_{2,U_2}V_{2,\tau} - \mathcal{C}_{2,U_{2,\tau}}V_{2,\tau}\big\|_{L^2(0,T;\mathcal{Q}^*)} + \big\|\mathcal{C}_{2,U_{2,\tau}}V_{2,\tau} - \mathcal{C}_{2,U_2}V_2\big\|_{L^2(0,T;\mathcal{Q}^*)}\\ &= \big\|(\mathcal{C}_{2,U_2} - \mathcal{C}_{2,U_{2,\tau}})V_{2,\tau}\big\|_{L^2(0,T;\mathcal{Q}^*)} + \big\|\dot{\mathcal{G}}_\tau - \dot{\mathcal{G}}\big\|_{L^2(0,T;\mathcal{Q}^*)}.\end{aligned}$$

The second term tends to zero as $\tau \to 0$, since $\dot{\mathcal{G}}_\tau \to \dot{\mathcal{G}}$ in $L^2(0,T;\mathcal{Q}^*)$. For the first term we estimate the operator norm of $\mathcal{C}_{2,U_2} - \mathcal{C}_{2,U_{2,\tau}}$ by the continuity of the trace operator,

$$\sup_{v\in\mathcal{V}} \frac{1}{\|v\|}\big\|\mathcal{C}_{2,U_2}v - \mathcal{C}_{2,U_{2,\tau}}v\big\|_{\mathcal{Q}^*} \le C_{\mathrm{tr}}\|\beta'(\gamma U_2) - \beta'(\gamma U_{2,\tau})\|_{\mathcal{Q}^*}.$$

Thus, the strong convergence $U_{2,\tau} \to U_2$ and Lemma 5.11 together imply that $V_{2,\tau} \to V_2$ in $L^2(0,T;\mathcal{V}_2)$. □

In the remaining part of this subsection we analyse the limiting behavior of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$. As in the linear case, we show that there exists a common limit function $U_1 \in L^2(0,T;\mathcal{V})$. Because of the nonlinearity of the operator $\mathcal{K}$, the challenge is then to show that $\mathcal{K}(U_{1,\tau}+U_{2,\tau})$ converges to $\mathcal{K}(U_1+U_2)$. For this, we follow the procedure used in the proof of [**ET10b**, Th. 5.1].

By the stability estimate of Lemma 10.18 we obtain the boundedness of the sequences $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ in $L^\infty(0,T;\mathcal{H})$. The same lemma also implies the boundedness of $U_{1,\tau}$ in $L^2(0,T;\mathcal{V}_1)$. Note that $\hat{U}_{1,\tau}$ is only bounded in $L^2(0,T;\mathcal{V}_1)$ if we assume additionally that $u_1^0 \in \mathcal{V}_1$. The uniform boundedness and Theorem 3.31 then imply the existence of weakly converging subsequences. By the estimate

$$\big\|U_{1,\tau} - \hat{U}_{1,\tau}\big\|^2_{L^2(0,T;\mathcal{H})} \le \tau\sum_{j=1}^{n}\big|u_1^j - u_1^{j-1}\big|^2 \le \tau M^2 \to 0.$$

we conclude as in Section 10.3.1 that that limits of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ coincide. Therein, M denotes the upper bound from Lemma 10.18. Assuming $u_1^0 \in \mathcal{V}_1$, we have

$$U_{1,\tau},\ \hat{U}_{1,\tau} \rightharpoonup U_1 \quad \text{in } L^2(0,T,\mathcal{V}_1).$$

At the end of this section we will show that the limit U_1 solves the operator DAE (10.26). Since the solution is unique according to Remark 6.25, we do not need the restriction to subsequences. Also the derivatives $\frac{\mathrm{d}}{\mathrm{d}t}\hat{U}_{1,\tau}$ are uniformly bounded in $L^2(0,T;\mathcal{V}_1^*)$ by the stability estimate in Remark 10.19. Thus, there exists a weak limit which we denote by $V_1 \in L^2(0,T;\mathcal{V}_1^*)$. Exactly as in Theorem 10.8 one can show that V_1 equals the derivative of U_1 in the generalized sense, i.e., $\dot{U}_1 = V_1 \in L^2(0,T;\mathcal{V}_1^*)$.

In the following calculation we mainly need two equations. First, the semi-discrete system given by the Euler scheme, i.e., equation (10.32a), tested by functions in $\mathcal{V}_1$,

$$\dot{\hat{U}}_{1,\tau} + V_{2,\tau} + \mathcal{K}\big(U_{1,\tau} + U_{2,\tau}\big) = \mathcal{F}_\tau \quad \text{in } \mathcal{V}_1^*. \tag{10.33}$$

Second, we consider the limiting equation for $\tau \to 0$. Note that Lemma 10.16 (c) implies that $\mathcal{K}(U_{1,\tau} + U_{2,\tau})$ is bounded in $L^2(0,T;\mathcal{V}^*)$ such that there exists a weak limit a which satisfies (up to a subsequence) $\mathcal{K}(U_{1,\tau} + U_{2,\tau}) \rightharpoonup a$ in $L^2(0,T;\mathcal{V}^*)$. The resulting equation then reads

$$\dot{U}_1 + V_2 + a = \mathcal{F} \quad \text{in } \mathcal{V}_1^*. \tag{10.34}$$

Using equations (10.33) and (10.34) as well as the integration by parts formula, we obtain the following lemma.

LEMMA 10.21. *Consider the assumptions from Lemma 10.18 with $u_1^0 \in \mathcal{V}_1$ in addition. Then, it holds that $U_1(0) = u_1^0$ and $\liminf_{\tau\to 0}\langle \dot{\hat{U}}_{1,\tau}, U_{1,\tau}\rangle \geq \langle \dot{U}_1, U_1\rangle$.*

PROOF. We omit to give the proof here, since it can be found in the proof of [**ET10b**, Th. 5.1]. However, the details can also be found in the proof of Lemma 7.1 where a similar result for the second-order case is given. □

With this result and the convergence of $U_{2,\tau}$ and $V_{2,\tau}$ from Theorem 10.20, we are able to state the following convergence theorem.

THEOREM 10.22. *Assume $\mathcal{F} \in L^2(0,T;\mathcal{V}_1^*)$, $\mathcal{G} \in H^1(0,T;\mathcal{Q}^*)$, $u_1^0 = g_0 \in \mathcal{V}_1$, as well as $u_1^j + u_2^j \in \mathcal{U}_M$ with the set $\mathcal{U}_M$ introduced in Lemma 10.16. Then, the weak limit $a \in L^2(0,T;\mathcal{V}^*)$ equals $\mathcal{K}(U_1+U_2)$ and thus, U_1, U_2, and V_2 solve equation* (10.26a) *in $\mathcal{V}_1^*$, i.e., they satisfy $\dot{U}_1 + V_2 + \mathcal{K}(U_1+U_2) = \mathcal{F}$ for all test functions in $\mathcal{V}_1$ and $U_1(0) = g_0$.*

PROOF. We make use of the monotonicity of the operator $\mathcal{K}$, see Lemma 10.16 (c), which means that $\langle \mathcal{K}(U_{1,\tau}+U_{2,\tau}) - \mathcal{K}w, U_{1,\tau}+U_{2,\tau}-w\rangle \geq 0$ for any $w \in L^2(0,T;\mathcal{V})$. With this, by equation (10.33) tested by $U_{1,\tau}$, we obtain that

$$0 \geq \big\langle \dot{\hat{U}}_{1,\tau} + V_{2,\tau} - \mathcal{F}_\tau, U_{1,\tau}\big\rangle + \big\langle \mathcal{K}\big(U_{1,\tau}+U_{2,\tau}\big), w - U_{2,\tau}\big\rangle + \big\langle \mathcal{K}w, U_{1,\tau}+U_{2,\tau}-w\big\rangle.$$

The application of the limes inferior then yields the estimate

$$\begin{aligned} 0 \;&\geq\; \big\langle \dot{U}_1 + V_2 - \mathcal{F}, U_1\big\rangle + \big\langle a, w - U_2\big\rangle + \big\langle \mathcal{K}w, U_1+U_2-w\big\rangle \\ &\overset{(10.34)}{=} \big\langle a, w - U_1 - U_2\big\rangle + \big\langle \mathcal{K}w, U_1+U_2-w\big\rangle. \end{aligned}$$

Note that we have used the strong convergence of $U_{2,\tau}$, $V_{2,\tau}$, and $\mathcal{F}_\tau$ as well as Lemma 10.21. Following the so-called *Minty trick* [**RR04**, Lem. 10.47], we first set $w := U_1 + U_2 + sv$ with $s \in]0,1]$ and an arbitrary function $v \in L^2(0,T;\mathcal{V})$. This then gives the estimate

$$s\,\big\langle \mathcal{K}\big(U_1+U_2+sv\big), v\big\rangle \geq s\,\big\langle a, v\big\rangle.$$

Dividing by s and making the same ansatz for $w := U_1 + U_2 - sv$, we obtain

$$\big\langle \mathcal{K}\big(U_1+U_2+sv\big), v\big\rangle \geq \big\langle a, v\big\rangle, \qquad \big\langle \mathcal{K}\big(U_1+U_2-sv\big), v\big\rangle \leq \big\langle a, v\big\rangle.$$

In the limit $s \to 0$, we then conclude that $a = \mathcal{K}(U_1 + U_2)$, since v was arbitrary. Thus, the limiting equation (10.34) turns into

$$\dot{U}_1 + V_2 + \mathcal{K}(U_1 + U_2) = \mathcal{F} \quad \text{in } \mathcal{V}_1^*.$$

Finally, U_1 satisfies the initial condition due to $U_1(0) = u_1^0 = g_0$, see Lemma 10.21. □

REMARK 10.23 (Lagrange Multiplier). As in the linear case, we are not able to bound the approximation of the Lagrange multiplier independently of the step size. Due to the nonlinearity we cannot even prove the convergence in the weak distributional sense. The reason for this is the dependence of $\mathcal{C}_u^*$ on the solution of the operator DAE and hence, the dependence on time. Thus, one may only prove the convergence of $\int_0^T \mathcal{C}_U^* \Lambda \, \mathrm{d}t$.

10.5.3. *Influence of Perturbations.* Let $(\hat{u}_1^j, \hat{u}_2^j, \hat{v}_2^j, \hat{\lambda}^j)$ denote the solution of a perturbed problem with perturbations $\delta^j \in \mathcal{H}^*$ and θ^j, $\xi^j \in \mathcal{Q}^*$ of the right-hand sides. Then, the differences

$$e_1^j := \hat{u}_1^j - u_1^j, \qquad e_2^j := \hat{u}_2^j - u_2^j, \qquad e_v^j := \hat{v}_2^j - v_2^j, \qquad e_\lambda^j := \hat{\lambda}^j - \lambda^j$$

satisfy the system

$$De_1^j + \quad e_v^j \quad + \mathcal{K}(\hat{u}_1^j + \hat{u}_2^j) - \mathcal{K}(u_1^j + u_2^j) + \mathcal{C}_{u^j}^* e_\lambda^j = \delta^j \quad \text{in } \mathcal{V}^*, \tag{10.35a}$$

$$\mathcal{B}(\hat{u}_2^j) - \mathcal{B}(u_2^j) = \theta^j \quad \text{in } \mathcal{Q}^*, \tag{10.35b}$$

$$\mathcal{C}_{2,u^j} e_v^j = \xi^j \quad \text{in } \mathcal{Q}^*. \tag{10.35c}$$

The impact of these perturbations is analyzed in the following theorem. Therein, we use again the abbreviation $a \lesssim b$ for the existence of a positive constant $c \in \mathbb{R}$ such that $a \le cb$.

THEOREM 10.24. *Consider perturbations $\delta^j \in \mathcal{H}^*$ and θ^j, $\xi^j \in \mathcal{Q}^*$ which are of the same magnitude, i.e., $\delta^j \approx \delta$, $\theta^j \approx \theta$, and $\xi^j \approx \xi$ for all $j = 1, \dots, n$. Furthermore, assume $\tau \le 1/2$. Then, for all $1 \le k \le n$ it holds up to a term of order $o(\max_j \|e_2^j\|)$ that $\|e_2^k\| \lesssim \|\theta\|_{\mathcal{Q}^*}$, $\|e_v^k\| \lesssim \|\xi\|_{\mathcal{Q}^*}$, and*

$$|e_1^k|^2 + \tau^2 \sum_{j=1}^{k} |De_1^j|^2 \lesssim 4^T |e_1^0|^2 + T4^T \Big[\|\delta\|_{\mathcal{H}^*}^2 + \|\theta\|_{\mathcal{Q}^*}^2 + \|\xi\|_{\mathcal{Q}^*}^2 \Big].$$

PROOF. We start with the estimates of e_2^j and e_v^j. By Lemma 10.16 (b) and (10.35c) we directly obtain that $\|e_v^j\| \le C_{\mathcal{C}_2 \mathrm{inv}} \|\xi^j\|_{\mathcal{Q}^*}$. For an estimate of e_2^j we proceed as in Section 6.2.3, i.e., we use the definition of the Fréchet derivative. Therewith, we get

$$\|e_2^j\| \le C_{\mathcal{C}_2 \mathrm{inv}} \big\| \mathcal{C}_{2,u} e_2^j \big\|_{\mathcal{Q}^*} \approx C_{\mathcal{C}_2 \mathrm{inv}} \big\| \mathcal{B}(\hat{u}_2^j) - \mathcal{B}(u_2^j) \big\|_{\mathcal{Q}^*} = C_{\mathcal{C}_2 \mathrm{inv}} \|\theta^j\|_{\mathcal{Q}^*}$$

up to a term of order $o(\|e_2^j\|)$. Next, we test equation (10.35a) by e_1^j which leads to

$$2\big(De_1^j, e_1^j\big) + 2\big\langle \mathcal{K}(\hat{u}_1^j + \hat{u}_2^j) - \mathcal{K}(u_1^j + u_2^j), e_1^j \big\rangle = 2\big\langle \delta^j, e_1^j \big\rangle - 2\big(e_v^j, e_1^j\big). \tag{10.36}$$

By (10.6) we can write the first term as $2(De_1^j, e_1^j) = D|e_1^j|^2 + \tau |De_1^j|^2$. The second term is bounded from below by Lemma 10.16,

$$\begin{aligned} &\big\langle \mathcal{K}(\hat{u}_1^j + \hat{u}_2^j) - \mathcal{K}(u_1^j + u_2^j), e_1^j \big\rangle \\ &\qquad = \big\langle \mathcal{K}(\hat{u}_1^j + \hat{u}_2^j) - \mathcal{K}(u_1^j + u_2^j), e_1^j + e_2^j \big\rangle - \big\langle \mathcal{K}(\hat{u}_1^j + \hat{u}_2^j) - \mathcal{K}(u_1^j + u_2^j), e_2^j \big\rangle \\ &\qquad \ge \varepsilon \big| \nabla(e_1^j + e_2^j) \big|^2 - k_2 \big| \nabla(e_1^j + e_2^j) \big| \|e_2^j\|. \end{aligned}$$

Thus, by Young's inequality, we can bound this term from below by $-\frac{1}{4\varepsilon}k_2^2\|e_2^j\|^2$. In summary, equation (10.36), together with the Cauchy-Schwarz inequality, leads to the estimate

$$D|e_1^j|^2 + \tau|De_1^j|^2 \leq \frac{1}{2\varepsilon}k_2^2\|e_2^j\|^2 + 2\|\delta^j\|_{\mathcal{H}^*}|e_1^j| + 2C_{\mathrm{emb}}\|e_v^j\||e_1^j|.$$

Another application of Young's inequality and a multiplication by τ then yields the existence of a positive constant c such that

$$|e_1^j|^2 - |e_1^{j-1}|^2 + \tau^2|De_1^j|^2 \leq \tau c\Big(\|\delta^j\|_{\mathcal{H}^*}^2 + \|e_2^j\|^2 + \|e_v^j\|^2\Big) + \tau|e_1^j|^2. \tag{10.37}$$

With $a^j := (1-\tau)^j$, we estimate

$$\begin{aligned} a^j|e_1^j|^2 - a^{j-1}|e_1^{j-1}|^2 + \tau^2 a^{j-1}|De_1^j|^2 &= a^{j-1}\Big((1-\tau)|e_1^j|^2 - |e_1^{j-1}|^2 + \tau^2|De_1^j|^2\Big) \\ &\overset{(10.37)}{\leq} a^{j-1}\tau c\Big(\|\delta^j\|_{\mathcal{H}^*}^2 + \|e_2^j\|^2 + \|e_v^j\|^2\Big). \end{aligned}$$

Because of the assumption $0 < \tau < 1$, the coefficients satisfy $0 < a^j < 1$ as well as $a^j > a^k$ for $j < k$. Thus, the summation of the latter estimate yields

$$a^k|e_1^k|^2 - |e_1^0|^2 + \tau^2\sum_{j=1}^{k} a^{j-1}|De_1^j|^2 \leq \tau c\sum_{j=1}^{k}\Big(\|\delta^j\|_{\mathcal{H}^*}^2 + \|e_2^j\|^2 + \|e_v^j\|^2\Big).$$

Finally, a division by a^k and the assumptions on the perturbations yield (up to terms of higher order)

$$|e_1^k|^2 + \tau^2\sum_{j=1}^{k}|De_1^j|^2 \leq a^{-k}|e_1^0|^2 + a^{-k}cT\Big[\|\delta\|_{\mathcal{H}^*}^2 + \|\theta\|_{\mathcal{Q}^*}^2 + \|\xi\|_{\mathcal{Q}^*}^2\Big].$$

It remains to show that $a^{-k} \leq 4^T$. For this, note that $\tau \leq 1/2$ implies $n \geq 2T$ and that the monotonicity of the sequence $(1 + x/n)^n$ for $n > -x$ gives

$$a^k = (1-\tau)^k = (1 - T/n)^k \geq (1 - T/n)^n \geq (1 - T/2T)^{2T} = 4^{-T}. \qquad \square$$

Note that Theorem 10.24 does not include any statement about the influence of the perturbation on the Lagrange multiplier. As discussed in Remark 10.23, this is caused by the included nonlinearity. Furthermore, we have assumed $\delta^j \in \mathcal{H}^*$ in contrast to the linear case in which $\delta^j \in \mathcal{V}^*$ has been sufficient, cf. Section 10.4.1.

11. Convergence for Second-order Systems

In the simulation of flexible multibody systems the Rothe method has not established itself yet [**LS09**]. Instead, the method of lines is preferred which leads to very large DAEs. However, discretizing in time first allows adaptive procedures, especially in the space variable, since the underlying grid may be changed easily from time step to time step. The practical application of the Rothe method in view of flexible multibody dynamics is discussed in [**LS09**]. Therein, the same operator DAE as in Section 7 is used but without the damping term and only in its original form of index-3 type.

Within this section, we analyse the convergence of the Rothe method for the second-order operator DAEs introduced in Section 7. For the temporal discretization we restrict ourselves to the implicit method introduced in Section 5.2.2. Note that high-order schemes in time often do not pay off, since the spatial error dominates.

For the a priori estimates and the resulting convergence proofs, we apply the standard techniques as used in [**ET10a**] for abstract ODEs of second order. For this, we construct piecewise constant and linear (in time) approximations of the variables of interest. The a priori estimates then show the boundedness of the approximation independent of the step size such that a weakly convergent subsequence can be extracted.

11.1. Setting and Discretization. We retain the setting and notion from Section 7, i.e., we consider the Sobolev spaces

$$\mathcal{V} := \big[H^1(\Omega)\big]^d, \qquad \mathcal{V}_{\mathcal{B}} := \big[H^1_{\Gamma_{\mathrm{D}}}(\Omega)\big]^d, \qquad \mathcal{H} := \big[L^2(\Omega)\big]^d, \qquad \mathcal{Q}^* := [H^{1/2}(\Gamma_{\mathrm{D}})]^d$$

and use for the inner product in $\mathcal{H}$ and the norms in $\mathcal{H}$ and $\mathcal{V}$ the abbreviations

$$(u,v) := (u,v)_{\mathcal{H}}, \qquad |u| := \|u\|_{\mathcal{H}}, \qquad \|u\| := \|u\|_{\mathcal{V}}.$$

Throughout this section, we only consider equidistant time steps with step size τ. Furthermore, u^j denotes the approximation of u at time $t_j = j\tau$. For the discretization we use the scheme introduced in the end of Section 5.2.2, i.e., we replace $\dot{u}$ and $\ddot{u}$ by the discrete derivatives $\dot{u}(t_j) \approx Du^j$ and $\ddot{u}(t_j) \approx D^2u^j$. Recall that this scheme is based on the Euler scheme and is fully implicit. Applied to the regularized operator DAE (7.14), the first equation turns to

(11.1a)
$$\frac{\rho}{\tau^2}\big(u_1^j - 2u_1^{j-1} + u_1^{j-2}\big) + \rho w_2^j + \mathcal{D}\big(\frac{1}{\tau}u_1^j - \frac{1}{\tau}u_1^{j-1} + v_2^j\big) + \mathcal{K}\big(u_1^j + u_2^j\big) + \mathcal{B}^*\lambda^j = \mathcal{F}^j.$$

This equation has to be solved for $j = 2, \dots, n$ and is still stated in the dual space of $\mathcal{V}$ and thus, equals a PDE in the weak formulation. The three constraints (7.14b)-(7.14d) result in

(11.1b)
$$\mathcal{B}u_2^j = \mathcal{G}^j, \qquad \mathcal{B}v_2^j = \dot{\mathcal{G}}^j, \qquad \mathcal{B}w_2^j = \ddot{\mathcal{G}}^j \quad \text{in } \mathcal{Q}^*.$$

As discussed in Section 5.3.2, the definition of the right-hand sides $\mathcal{F}^j$, $\mathcal{G}^j$, $\dot{\mathcal{G}}^j$, and $\ddot{\mathcal{G}}^j$ has to be clarified, since we only assume $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$ and $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$. Furthermore, we define the piecewise constant approximations $\mathcal{F}_\tau$, $\mathcal{G}_\tau$, $\dot{\mathcal{G}}_\tau$, and $\ddot{\mathcal{G}}_\tau$ as in (5.12) for which we assume that $\mathcal{F}_\tau \to \mathcal{F}$ in $L^2(0,T;\mathcal{V}^*)$ as well as $\mathcal{G}_\tau \to \mathcal{G}$, $\dot{\mathcal{G}}_\tau \to \dot{\mathcal{G}}$ and $\ddot{\mathcal{G}}_\tau \to \ddot{\mathcal{G}}$ in $L^2(0,T;\mathcal{Q}^*)$.

Remark 11.1 (Special case $\mathcal{G} \equiv 0$). Consider the case where $\mathcal{G}$ vanishes on $[0,T]$ and thus, $u_2^j = v_2^j = w_2^j = 0$. Then, the problem reduces to an operator ODE and (11.1a) reads

$$\frac{\rho}{\tau^2}\big(u_1^j - 2u_1^{j-1} + u_1^{j-2}\big) + \mathcal{D}\big(\frac{1}{\tau}u_1^j - \frac{1}{\tau}u_1^{j-1}\big) + \mathcal{K}u_1^j = \mathcal{F}^j$$

with test functions in $\mathcal{V}_\mathcal{B}$.

In this section on second-order operator DAEs, we remain with the given framework on linear elasticity with nonlinear damping term as discussed in Section 7. Thus, we consider a linear and symmetric stiffness operator $\mathcal{K}\colon \mathcal{V} \to \mathcal{V}^*$. The operator is assumed to be positive on $\mathcal{V}_\mathcal{B}$ and bounded, i.e., there exist positive constants k_1 and k_2 such that for all $u \in \mathcal{V}_\mathcal{B}$ and v, $w \in \mathcal{V}$ it holds that

$$k_1\|u\|^2 \le \langle \mathcal{K}u, u\rangle_{\mathcal{V}^*,\mathcal{V}}, \qquad \langle \mathcal{K}v, w\rangle_{\mathcal{V}^*,\mathcal{V}} \le k_2\|v\|\|w\|. \tag{11.2}$$

Note that the symmetry of the operator implies that we may write $\langle \mathcal{K}u, u\rangle = |\mathcal{K}^{1/2}u|^2$. The nonlinear damping operator $\mathcal{D}\colon \mathcal{V} \to \mathcal{V}^*$ is assumed to be Lipschitz continuous and strongly monotone, i.e., there exist constants d_0, d_1, and d_2 such that for all u, $v \in \mathcal{V}$ it holds that

$$\|\mathcal{D}u - \mathcal{D}v\|_{\mathcal{V}*} \le d_2\|u - v\|, \quad d_1\|u - v\|^2 - d_0|u - v|^2 \le \langle \mathcal{D}u - \mathcal{D}v, u - v\rangle_{\mathcal{V}^*,\mathcal{V}}. \tag{11.3}$$

Furthermore, we may assume w.l.o.g. $\mathcal{D}(0) = 0$, see [**ET10a**, p. 181], and thus,

$$\|\mathcal{D}u\|_{\mathcal{V}^*} \le d_2\|u\|, \quad d_1\|u\|^2 - d_0|u|^2 \le \langle \mathcal{D}u, u\rangle_{\mathcal{V}^*,\mathcal{V}}.$$

Remark 11.2. Because of the continuous embedding $V \hookrightarrow H$, we have $|\cdot| \le C_{\mathrm{emb}}\|\cdot\|$. In the case $d_0C_{\mathrm{emb}}^2 < d_1$, we can write

$$\langle \mathcal{D}u, u\rangle_{\mathcal{V}^*,\mathcal{V}} \ge d_1\|u\|^2 - d_0|u|^2 \ge \big(d_1 - d_0C_{\mathrm{emb}}^2\big)\|u\|^2.$$

Thus, we may assume either $d_0 = 0$ or $d_0C_{\mathrm{emb}}^2 \ge d_1$.

Before we derive stability results for the discrete approximations, we have to discuss the solvability of the semi-discrete system (11.1).

Lemma 11.3. *With the assumptions introduced in this subsection, system* (11.1) *has a unique solution* $(u_1^j, u_2^j, v_2^j, w_2^j, \lambda^j)$ *for each time step* $j = 2, \dots, n$ *if the step size satisfies* $\tau < \rho/d_0$.

Proof. The invertibility of the trace operator for functions in $\mathcal{V}^{\mathrm{c}}$ implies that the equations in (11.1b) give unique approximations u_2^j, v_2^j, and w_2^j. Consider equation (11.1a) restricted to test functions in $\mathcal{V}_\mathcal{B}$. We define the operator $\mathcal{A}\colon \mathcal{V}_\mathcal{B} \to \mathcal{V}_\mathcal{B}^*$ and the functional $\hat{\mathcal{F}}^j \in \mathcal{V}^*$ by

$$\mathcal{A}u := \frac{\rho}{\tau^2}u + \mathcal{D}\Big(\frac{u - u_1^{j-1}}{\tau} + v_2^j\Big) + \mathcal{K}u, \quad \hat{\mathcal{F}}^j := \mathcal{F}^j + \frac{\rho}{\tau^2}\Big(2u_1^{j-1} - u_1^{j-2}\Big) - \rho w_2^j - \mathcal{K}u_2^j.$$

Then, equation (11.1a) can be written in the form $\mathcal{A}u_1^j = \hat{\mathcal{F}}^j$ in $\mathcal{V}_\mathcal{B}^*$. The operator $\mathcal{A}$ is continuous and for the monotonicity we obtain by (11.2) and (11.3),

$$\begin{aligned}\big\langle \mathcal{A}u - \mathcal{A}v, u - v\big\rangle &\ge \frac{\rho}{\tau^2}|u - v|^2 + \frac{d_1}{\tau}\|u - v\|^2 - \frac{d_0}{\tau}|u - v|^2 + k_1\|u - v\|^2 \\ &= \big(d_1/\tau + k_1\big)\|u - v\|^2 + \big(\rho/\tau^2 - d_0/\tau\big)|u - v|^2.\end{aligned}$$

This shows $\langle \mathcal{A}u - \mathcal{A}v, u - v\rangle \ge k_1\|u - v\|^2$ for $\tau < \rho/d_0$ and thus, the existence of a solution $u_1^j \in \mathcal{V}_\mathcal{B}$ due to the Browder-Minty theorem [**GGZ74**, Ch. III, Th. 2.1]. The strong monotonicity of $\mathcal{A}$ also implies the uniqueness of the solution. Finally, the unique solvability for λ^j follows from Lemma 7.1 (d). □

11.2. Stability and Convergence. As for first-order systems, we use the approximations from system (11.1) to define global approximations and show the uniform boundedness. In order to avoid too long terms for the discrete derivative, we use the abbreviation

$$v_1^j := Du_1^j = \frac{u_1^j - u_1^{j-1}}{\tau}.$$

Furthermore, we assume u_1^1 and v_1^1 to be the fixed initial data of the semi-discrete solution, i.e., approximations of the initial data $u_1(0) = g_0$ and $\dot{u}_1(0) = h_0$. Clearly, this also defines u_1^0 which - in the limit - coincides with u_1^1.

In the sequel we will take several times advantage of the equality

$$2(a-b)a = a^2 - b^2 + (a-b)^2. \tag{11.4}$$

11.2.1. *Stability Estimate.* As in Section 10, we need a stability estimate which is given in the following lemma. Note that this includes a step size restriction due to the nonlinear damping term.

LEMMA 11.4 (Stability). *Assume right-hand sides $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$ and initial approximations $u_1^1 \in \mathcal{V}_\mathcal{B}$, $v_1^1 \in \mathcal{H}$. Let the approximations u_1^j, u_2^j, v_2^j, and w_2^j be given by the semi-discrete system* (11.1) *and let the step size satisfy $\tau < \rho/8d_0$. Then, there exists a constant $c > 0$ such that for all $k \geq 2$ it holds that*

$$\rho\big|v_1^k\big|^2 + \rho\sum_{j=2}^{k}\big|v_1^j - v_1^{j-1}\big|^2 + \tau d_1\sum_{j=2}^{k}\big\|v_1^j\big\|^2 + k_1\big\|u_1^k\big\|^2 \leq c\ 2^{8d_0T/\rho}M^2 \tag{11.5}$$

with the constant

$$M^2 = |v_1^1|^2 + \|u_1^1\|^2 + \|\mathcal{F}\|^2_{L^2(0,T;\mathcal{V}^*)} + \|\mathcal{G}\|^2_{H^2(0,T;\mathcal{Q}^*)}.$$

PROOF. Equation (11.1b) directly leads to the estimates

$$\|u_2^j\| \leq C_{\mathcal{B}^-}\|\mathcal{G}^j\|_{\mathcal{Q}^*}, \quad \|v_2^j\| \leq C_{\mathcal{B}^-}\|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}, \quad \|w_2^j\| \leq C_{\mathcal{B}^-}\|\ddot{\mathcal{G}}^j\|_{\mathcal{Q}^*}. \tag{11.6}$$

The rest of the proof mainly follows the ideas of the proof of [**ET10a**, Th. 1] although a different time discretization scheme is used. We consider the case $d_0 > 0$. The proof with $d_0 = 0$ works in the same manner but with less difficulties. In the semi-discrete setting we test equation (11.1a) with the discrete derivative $v_1^j \in \mathcal{V}_\mathcal{B}$, $j \geq 2$. This leads to

$$\rho\big\langle Dv_1^j, v_1^j\big\rangle + \big\langle \mathcal{D}\big(v_1^j + v_2^j\big), v_1^j\big\rangle + \big\langle \mathcal{K}u_1^j, v_1^j\big\rangle = \big\langle \mathcal{F}^j, v_1^j\big\rangle - \rho\big\langle w_2^j, v_1^j\big\rangle - \big\langle \mathcal{K}u_2^j, v_1^j\big\rangle. \tag{11.7}$$

For the terms on the left-hand side, we estimate separately

$$\rho\big\langle Dv_1^j, v_1^j\big\rangle = \frac{\rho}{\tau}\big\langle v_1^j - v_1^{j-1}, v_1^j\big\rangle \overset{(11.4)}{=} \frac{\rho}{2\tau}\Big[\big|v_1^j\big|^2 - \big|v_1^{j-1}\big|^2 + \big|v_1^j - v_1^{j-1}\big|^2\Big],$$

for the damping term

$$\begin{aligned}\big\langle \mathcal{D}\big(v_1^j + v_2^j\big), v_1^j\big\rangle &= \big\langle \mathcal{D}\big(v_1^j + v_2^j\big) - \mathcal{D}v_2^j, v_1^j\big\rangle + \big\langle \mathcal{D}v_2^j, v_1^j\big\rangle \\ &\overset{(11.3)}{\geq} d_1\|v_1^j\|^2 - d_0|v_1^j|^2 - d_2\|v_1^j\|\|v_2^j\| \\ &\geq d_1\|v_1^j\|^2 - d_0|v_1^j|^2 - \frac{d_1}{6}\|v_1^j\|^2 - \frac{3d_2^2}{2d_1}\|v_2^j\|^2,\end{aligned}$$

and finally for the stiffness term

$$\big\langle \mathcal{K}u_1^j, v_1^j\big\rangle = \frac{1}{\tau}\big\langle \mathcal{K}u_1^j, u_1^j - u_1^{j-1}\big\rangle \overset{(11.4)}{\geq} \frac{1}{2\tau}\big|\mathcal{K}^{1/2}u_1^j\big|^2 - \frac{1}{2\tau}\big|\mathcal{K}^{1/2}u_1^{j-1}\big|^2.$$

For the right-hand side of (11.7) we obtain with the Cauchy-Schwarz inequality, followed by an application of Youngs inequality,

$$\begin{aligned}\big\langle \mathcal{F}^j, v_1^j\big\rangle - \rho\big\langle w_2^j, v_1^j\big\rangle - \big\langle \mathcal{K}u_2^j, v_1^j\big\rangle \\ &\le \|\mathcal{F}^j\|_{\mathcal{V}^*}\|v_1^j\| + \rho|w_2^j||v_1^j| + k_2\|u_2^j\|\|v_1^j\| \\ &\le \frac{3}{2d_1}\|\mathcal{F}^j\|_{\mathcal{V}^*}^2 + \frac{d_1}{6}\|v_1^j\|^2 + \frac{\rho^2}{4d_0}|w_2^j|^2 + d_0|v_1^j|^2 + \frac{3k_2^2}{2d_1}\|u_2^j\|^2 + \frac{d_1}{6}\|v_1^j\|^2.\end{aligned}$$

Summarizing, we multiply (11.7) by 2τ and use the above estimates to obtain

$$\begin{aligned}&\rho\Big[|v_1^j|^2 - |v_1^{j-1}|^2 + |v_1^j - v_1^{j-1}|^2\Big] + \tau d_1\|v_1^j\|^2 - 4\tau d_0|v_1^j|^2 + \big|\mathcal{K}^{1/2}u_1^j\big|^2 - \big|\mathcal{K}^{1/2}u_1^{j-1}\big|^2 \\ &\qquad \le \tau\left[\frac{3}{d_1}\|\mathcal{F}^j\|_*^2 + \frac{3k_2^2}{d_1}\|u_2^j\|^2 + \frac{3d_2^2}{d_1}\|v_2^j\|^2 + \frac{\rho^2}{2d_0}|w_2^j|^2\right].\end{aligned} \tag{11.8}$$

With the estimates of u_2^j, v_2^j, and w_2^j from equation (11.6) we can bound the right-hand side of the latter estimate by $c\tau(\|\mathcal{F}^j\|_{\mathcal{V}^*}^2 + \|\mathcal{G}^j\|_{\mathcal{Q}^*}^2 + \|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2 + \|\ddot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2)$. Therein, $c > 0$ denotes a generic constant which depends on $C_{\mathcal{B}^-}$, ρ, d_0, d_1, d_2, and k_2.

Before we sum over j and make benefit of several telescope sums, we have to deal with the term $4\tau d_0|v_1^j|^2$ on the left-hand side of (11.8). For this, we use arguments which are used to prove discrete versions of the Gronwall lemma [**Emm99**]. With $\kappa := 4d_0/\rho$ and $a^j := (1-\kappa\tau)^j$, we estimate

$$\begin{aligned}&\rho\Big[a^j|v_1^j|^2 - a^{j-1}|v_1^{j-1}|^2 + a^{j-1}|v_1^j - v_1^{j-1}|^2\Big] + \tau d_1 a^{j-1}\|v_1^j\|^2 + a^j\big|\mathcal{K}^{1/2}u_1^j\big|^2 - a^{j-1}\big|\mathcal{K}^{1/2}u_1^{j-1}\big|^2 \\ &\quad = a^{j-1}\Big[\rho(1-\kappa\tau)|v_1^j|^2 - \rho|v_1^{j-1}|^2 + \rho|v_1^j - v_1^{j-1}|^2 + \tau d_1\|v_1^j\|^2 \\ &\qquad\quad + (1-\kappa\tau)\big|\mathcal{K}^{1/2}u_1^j\big|^2 - \big|\mathcal{K}^{1/2}u_1^{j-1}\big|^2\Big] \\ &\quad \le a^{j-1}\Big[\rho|v_1^j|^2 - \rho|v_1^{j-1}|^2 + \rho|v_1^j - v_1^{j-1}|^2 + \tau d_1\|v_1^j\|^2 - 4\tau d_0|v_1^j|^2 \\ &\qquad\quad + \big|\mathcal{K}^{1/2}u_1^j\big|^2 - \big|\mathcal{K}^{1/2}u_1^{j-1}\big|^2\Big] \\ &\quad \overset{(11.8)}{\le} a^{j-1}\tau c\left(\|\mathcal{F}^j\|_{\mathcal{V}^*}^2 + \|\mathcal{G}^j\|_{\mathcal{Q}^*}^2 + \|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2 + \|\ddot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2\right).\end{aligned}$$

Note that we have used the fact that, due to the assumption on the step size τ, $0 < a^j < 1$ for all $j \ge 1$ and $\kappa \ge 0$. The summation of this estimate for $j = 2, \dots, k$ then yields

$$\begin{aligned}&\rho a^k\big|v_1^k\big|^2 + \rho\sum_{j=2}^k a^{j-1}\big|v_1^j - v_1^{j-1}\big|^2 + \tau d_1\sum_{j=2}^k a^{j-1}\|v_1^j\|^2 + a^k\big|\mathcal{K}^{1/2}u_1^k\big|^2 \\ &\qquad \le \rho a^1\big|v_1^1\big|^2 + a^1\big|\mathcal{K}^{1/2}u_1^1\big|^2 + \tau c\sum_{j=2}^k a^{j-1}\Big(\|\mathcal{F}^j\|_{\mathcal{V}^*}^2 + \|\mathcal{G}^j\|_{\mathcal{Q}^*}^2 + \|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2 + \|\ddot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2\Big).\end{aligned}$$

Finally, we divide by a^k and use the estimates $a^j > a^k$ for $j < k$ and $a^{-k} \leq 4^{\kappa T}$, cf. the proof of Theorem 10.24. This then leads to the final result

$$\rho\big|v_1^k\big|^2 + \rho\sum_{j=2}^{k}\big|v_1^j - v_1^{j-1}\big|^2 + \tau d_1\sum_{j=2}^{k}\big\|v_1^j\big\|^2 + k_1\big\|u_1^k\big\|^2$$
$$\leq 4^{\kappa T}\Big\{\rho\big|v_1^1\big|^2 + k_2\big\|u_1^1\big\|^2 + \tau c\sum_{j=2}^{k}\Big(\|\mathcal{F}^j\|_{\mathcal{V}^*}^2 + \|\mathcal{G}^j\|_{\mathcal{Q}^*}^2 + \|\dot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2 + \|\ddot{\mathcal{G}}^j\|_{\mathcal{Q}^*}^2\Big)\Big\}. \square$$

11.2.2. *Definition of Global Approximations.* In this section, we define the global approximations of u_1, u_2, v_2, and w_2. First, we define $U_{1,\tau}$, $\hat{U}_{1,\tau}\colon [0,T] \to \mathcal{V}_{\mathcal{B}}$ by

$$U_{1,\tau}(t) := u_1^j, \qquad \hat{U}_{1,\tau}(t) := u_1^j + (t - t_j)v_1^j$$

if $t \in]t_{j-1}, t_j]$ for $j \geq 2$ with $U_{1,\tau} \equiv \hat{U}_{1,\tau} \equiv u_1^1$ on $[0, t_1]$. By the stability estimate (11.5) of Lemma 11.4 we directly obtain the boundedness of $U_{1,\tau}$ and $\hat{U}_{1,\tau}$ in $L^\infty(0,T;\mathcal{V}_{\mathcal{B}})$ uniformly in τ. Thus, there exists an element $U_1 \in L^\infty(0,T;\mathcal{V}_{\mathcal{B}})$ with $U_{1,\tau}, \hat{U}_{1,\tau} \overset{*}{\rightharpoonup} U_1$ in $L^\infty(0,T;\mathcal{V}_{\mathcal{B}})$ as well as $U_{1,\tau}, \hat{U}_{1,\tau} \rightharpoonup U_1$ in $L^2(0,T;\mathcal{V}_{\mathcal{B}})$. Note that the limits of the two sequences coincide, again because of Lemma 11.4, since

$$\big\|\hat{U}_{1,\tau} - U_{1,\tau}\big\|_{L^2(0,T;\mathcal{H})}^2 = \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\big|(t - t_j)Du_1^j\big|^2\,\mathrm{d}t \leq \sum_{j=1}^{n}\tau^3\big|v_1^j\big|^2 \leq c\tau^2 M^2 \to 0.$$

In an analogous way, we define the piecewise constant functions $U_{2,\tau}, V_{2,\tau}, W_{2,\tau}\colon [0,T] \to \mathcal{V}^{\mathrm{c}}$. We set

$$U_{2,\tau}(t) := u_2^j, \qquad V_{2,\tau}(t) := v_2^j, \qquad W_{2,\tau}(t) := w_2^j$$

if $t \in]t_{j-1}, t_j[$ for $j \geq 1$ with a continuous extension in $t = 0$. By equation (11.1b) we have $\mathcal{B}U_{2,\tau} = \mathcal{G}_\tau$, $\mathcal{B}V_{2,\tau} = \dot{\mathcal{G}}_\tau$, and $\mathcal{B}W_{2,\tau} = \ddot{\mathcal{G}}_\tau$. Thus, Lemma 5.9 implies that

$$U_{2,\tau} \to U_2, \qquad V_{2,\tau} \to V_2, \qquad W_{2,\tau} \to W_2 \ \text{ in } L^2(0,T,\mathcal{V})$$

where U_2, V_2, and W_2 solve the equations $\mathcal{B}U_2 = \mathcal{G}$, $\mathcal{B}V_2 = \dot{\mathcal{G}}$, and $\mathcal{B}W_2 = \ddot{\mathcal{G}}$, respectively. This means nothing else than the (strong) convergence of $U_{2,\tau}$, $V_{2,\tau}$, and $W_{2,\tau}$ to the solutions of (7.14b)-(7.14d).

Finally, we define two approximations of the velocity in form of a piecewise constant and a piecewise linear approximation, namely

$$V_{1,\tau}(t) := v_1^j, \qquad \hat{V}_{1,\tau}(t) := v_1^j + (t - t_j)Dv_1^j$$

if $t \in]t_{j-1}, t_j]$ for $j \geq 2$ with $V_{1,\tau} \equiv \hat{V}_{1,\tau} \equiv v_1^1$ on $[0, t_1]$. An illustration can be seen in Figure 11.1. For the piecewise constant approximation, by Lemma 11.4, we obtain the

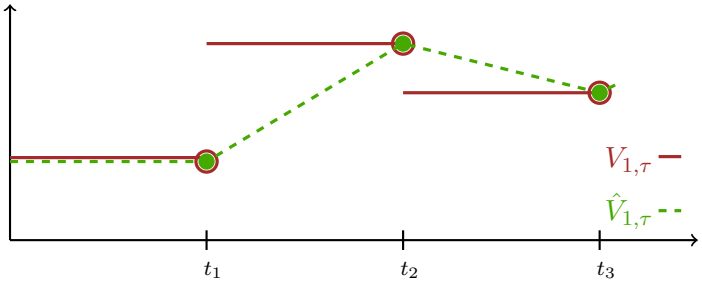


FIGURE 11.1. Illustration of the global approximations $V_{1,\tau}$ and $\hat{V}_{1,\tau}$ of $\dot{u}_1$.

estimate

$$\|V_{1,\tau}\|^2_{L^2(0,T;\mathcal{V})} = \int_0^T \|V_{1,\tau}(t)\|^2 \,\mathrm{d}t = \tau \sum_{j=1}^n \left\|v_1^j\right\|^2 \overset{(11.5)}{\leq} \tau \|v_1^1\|^2 + cM^2.$$

Thus, for $v_1^1 \in \mathcal{V}$ we have found a uniform bound which implies the existence of $V_1 \in L^2(0,T;\mathcal{V}_\mathcal{B})$ with $V_{1,\tau} \rightharpoonup V_1$ in $L^2(0,T;\mathcal{V}_\mathcal{B})$. In the same manner we obtain a bound of the piecewise linear approximation, since

$$\left\|\hat{V}_{1,\tau}\right\|^2_{L^2(0,T;\mathcal{V})} = \tau\|v_1^1\|^2 + \sum_{j=2}^n \int_{t_{j-1}}^{t^j} \left\|v_1^j + (t-t_j)Dv_1^j\right\|^2 \,\mathrm{d}t \leq 4\tau \sum_{j=1}^n \left\|v_1^j\right\|^2.$$

As before, we show that $V_{1,\tau}$ and $\hat{V}_{1,\tau}$ have the same limit V_1. For this, by Lemma 11.4 we calculate that

$$\left\|\hat{V}_{1,\tau} - V_{1,\tau}\right\|^2_{L^2(0,T;\mathcal{H})} = \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \left|\hat{V}_{1,\tau}(t) - V_{1,\tau}(t)\right|^2 \,\mathrm{d}t \leq \tau \sum_{j=2}^n \left|v_1^j - v_1^{j-1}\right|^2 \leq \tau cM^2 \to 0.$$

The agreement of the limits in $L^2(0,T;\mathcal{V})$ then follows from the assumed embedding $\mathcal{V} \hookrightarrow \mathcal{H}$ due to the Gelfand triple. In the following we show that the limit function V_1 equals the derivative of U_1 in the generalized sense. For this, we use the limits $\hat{U}_{1,\tau} \rightharpoonup U_1$ and $V_{1,\tau} \rightharpoonup V_1$ in $L^2(0,T;\mathcal{V}_\mathcal{B})$. Note, however, that $\frac{\mathrm{d}}{\mathrm{d}t}\hat{U}_{1,\tau} = V_{1,\tau}$ a.e. but not in the interval $[0,\tau]$. Applying the integration by parts formula with an arbitrary functional $f \in \mathcal{V}_\mathcal{B}^*$ and $\Phi \in C_0^\infty(0,T)$, we may write

$$\begin{aligned}\int_0^T \langle f, U_1\rangle \dot{\Phi} \,\mathrm{d}t &= \lim_{\tau\to 0} \int_0^T \langle f, \hat{U}_{1,\tau}\rangle \dot{\Phi} \,\mathrm{d}t = -\lim_{\tau\to 0} \int_0^T \langle f, \dot{\hat{U}}_{1,\tau}\rangle \Phi \,\mathrm{d}t \\ &= -\lim_{\tau\to 0} \int_0^T \langle f, V_{1,\tau}\rangle \Phi \,\mathrm{d}t - \int_0^\tau \langle f, v_1^1\rangle \Phi \,\mathrm{d}t = -\int_0^T \langle f, V_1\rangle \Phi \,\mathrm{d}t.\end{aligned}$$

Note that the integral over $[0,\tau]$ vanishes in the limit, since the integrand is bounded independently of the step size. As a result, the limit function U_1 has a generalized derivative and $\dot{U}_1 = V_1 \in L^2(0,T;\mathcal{V}_\mathcal{B})$.

Finally, we mention that also $\mathcal{D}(V_{1,\tau} + V_{2,\tau})$ gives a uniformly bounded sequence in $L^2(0,T;\mathcal{V}^*)$ due to the continuity of the damping operator $\mathcal{D}$. Thus, there exists a weak limit $a \in L^2(0,T;\mathcal{V}^*)$ with

$$\mathcal{D}(V_{1,\tau} + V_{2,\tau}) \rightharpoonup a \quad \text{in } L^2(0,T;\mathcal{V}^*).$$

One aim of the next subsection is to show that a equals $\mathcal{D}(V_1 + V_2)$. Before we pass to the limit and analyse the behavior of the (weak) limits, we summarize the convergence results of this subsection:

Assume right-hand sides $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$ and initial approximations u_1^1, $v_1^1 \in \mathcal{V}_\mathcal{B}$. Then, we have

$$U_{1,\tau},\ \hat{U}_{1,\tau} \rightharpoonup U_1, \quad V_{1,\tau},\ \hat{V}_{1,\tau} \rightharpoonup V_1 \ \text{ in } L^2(0,T;\mathcal{V}_\mathcal{B}), \tag{11.9a}$$

$$U_{2,\tau} \to U_2 = \mathcal{B}^-\mathcal{G}, \ \ V_{2,\tau} \to V_2 = \mathcal{B}^-\dot{\mathcal{G}}, \ \ W_{2,\tau} \to W_2 = \mathcal{B}^-\ddot{\mathcal{G}} \ \text{ in } L^2(0,T;\mathcal{V}^{\mathrm{c}}), \tag{11.9b}$$

$$\mathcal{D}(V_{1,\tau} + V_{2,\tau}) \rightharpoonup a \ \text{ in } L^2(0,T;\mathcal{V}^*). \tag{11.9c}$$

11.2.3. *Passing to the Limit.* In order to pass to the limit for $\tau \to 0$ it is beneficial to rewrite the semi-discretized equation (11.1a) in terms of the global approximations from Section 11.2.2. The space of test functions is still restricted to the space $\mathcal{V}_{\mathcal{B}}$ in order to remove the Lagrange multiplier from the system. The semi-discrete system has the form

$$\rho\big(\dot{\hat{V}}_{1,\tau} + W_{2,\tau}\big) + \mathcal{D}\big(V_{1,\tau} + V_{2,\tau}\big) + \mathcal{K}\big(U_{1,\tau} + U_{2,\tau}\big) = \mathcal{F}_\tau \tag{11.10}$$

for a.e. $t \in]\tau, T]$. Writing equation (11.10) in its actual meaning with test functions $v \in \mathcal{V}_{\mathcal{B}}$ and $\Phi \in C_0^\infty(0,T)$, cf. Section 4, and applying the integration by parts formula once, we get

$$\int_0^T -\big\langle \rho \hat{V}_{1,\tau}, v\big\rangle \dot{\Phi} + \big\langle \rho W_{2,\tau}, v\big\rangle \Phi + \big\langle \mathcal{D}\big(V_{1,\tau} + V_{2,\tau}\big), v\big\rangle \Phi + \big\langle \mathcal{K}\big(U_{1,\tau} + U_{2,\tau}\big), v\big\rangle \Phi \, \mathrm{d}t \\ = \int_0^T \big\langle \mathcal{F}_\tau, v\big\rangle \Phi \, \mathrm{d}t.$$

Passing to the limit for $\tau \to 0$, we then obtain by the achievements of the previous subsection, see the summary in (11.9), that

$$\int_0^T \big\langle \rho V_1, v\big\rangle \dot{\Phi} \, \mathrm{d}t = \int_0^T \big\langle \rho W_2 + a + \mathcal{K}\big(U_1 + U_2\big) - \mathcal{F}, v\big\rangle \Phi \, \mathrm{d}t.$$

Recall that a denotes the weak limit of $\mathcal{D}(V_{1,\tau} + V_{2,\tau})$ in $L^2(0,T;\mathcal{V}^*)$. This implies that V_1 has a generalized derivative $\dot{V}_1 \in L^2(0,T;\mathcal{V}_{\mathcal{B}}^*)$ which satisfies the equation

$$\rho \dot{V}_1 + \rho W_2 + a + \mathcal{K}(U_1 + U_2) = \mathcal{F} \quad \text{in } \mathcal{V}_{\mathcal{B}}^*. \tag{11.11}$$

The remaining part of this section is devoted to that proof that the weak limits U_1, U_2, V_2, and W_2 solve the operator DAE (7.14a) in $\mathcal{V}_{\mathcal{B}}^*$. With equation (11.11) at hand, it remains to show that a equals $\mathcal{D}(V_1 + V_2)$. In order to show this, we give two preparatory lemmata.

Lemma 11.5. *At the final point in time, the sequence $\hat{V}_{1,\tau}$ satisfies $\hat{V}_{1,\tau}(T) \rightharpoonup V_1(T)$ in $\mathcal{H}$. Furthermore, we obtain the estimate*

$$\liminf_{\tau \to 0} \big\langle \dot{\hat{V}}_{1,\tau}, V_{1,\tau}\big\rangle \geq \big\langle \dot{V}_1, V_1\big\rangle.$$

Proof. The proof follows the ideas of the proof of [**ET10b**, Th. 5.1] adapted to the given operator equation. First we show that $\hat{V}_{1,\tau}(T) \rightharpoonup V_1(T)$ in $\mathcal{H}$ as well as $\hat{V}_{1,\tau}(0) = V_1(0)$. Because of the stability estimate in Lemma 11.4, the final approximation $\hat{V}_{1,\tau}(T) = v_1^n$ is uniformly bounded in $\mathcal{H}$. Thus, there exists a weak limit $\xi \in \mathcal{H}$ which satisfies

$$v_1^n = \hat{V}_{1,\tau}(T) \rightharpoonup \xi \quad \text{in } \mathcal{H}.$$

Through the integration by parts formula and with $w \in \mathcal{V}_{\mathcal{B}}$ and $\Phi \in C^1([0,T];\mathbb{R})$, we obtain

$$\begin{aligned}
&\rho\big(V_1(T),w\big)\Phi(T) - \rho\big(V_1(0),w\big)\Phi(0)\\
&\qquad = \quad \big\langle \rho\dot{V}_1, w\Phi\big\rangle + \big\langle \rho V_1, w\dot{\Phi}\big\rangle\\
&\qquad \overset{(11.11)}{=} \big\langle \mathcal{F} - \rho W_2 - a - \mathcal{K}(U_1+U_2), w\Phi\big\rangle + \big\langle \rho V_1, w\dot{\Phi}\big\rangle\\
&\qquad \overset{(11.10)}{=} \big\langle \mathcal{F} - \mathcal{F}_\tau, w\Phi\big\rangle - \rho\big\langle W_2 - W_{2,\tau}, w\Phi\big\rangle - \big\langle a - \mathcal{D}(V_{1,\tau}+V_{2,\tau}), w\Phi\big\rangle\\
&\qquad\quad - \big\langle \mathcal{K}(U_1+U_2) - \mathcal{K}(U_{1,\tau}+U_{2,\tau}), w\Phi\big\rangle + \big\langle \rho V_1, w\dot{\Phi}\big\rangle + \big\langle \rho\dot{\hat{V}}_{1,\tau}, w\Phi\big\rangle\\
&\qquad = \quad \big\langle \mathcal{F} - \mathcal{F}_\tau, w\Phi\big\rangle - \rho\big\langle W_2 - W_{2,\tau}, w\Phi\big\rangle - \big\langle a - \mathcal{D}(V_{1,\tau}+V_{2,\tau}), w\Phi\big\rangle\\
&\qquad\quad - \big\langle \mathcal{K}(U_1+U_2) - \mathcal{K}(U_{1,\tau}+U_{2,\tau}), w\Phi\big\rangle + \rho\big\langle V_1 - \hat{V}_{1,\tau}, w\dot{\Phi}\big\rangle\\
&\qquad\quad + \rho\big(\hat{V}_{1,\tau}(T), w\big)\Phi(T) - \rho\big(\hat{V}_{1,\tau}(0), w\big)\Phi(0)\\
&\qquad \to \quad \rho\big(\xi, w\big)\Phi(T) - \rho\big(v_1^1, w\big)\Phi(0).
\end{aligned}$$

Thus, we have $v_1^n = \hat{V}_{1,\tau}(T) \rightharpoonup \xi = V_1(T)$ in $\mathcal{H}$ and $V(0) = v_1^1$. Note that at this point we need that the embedding $\mathcal{V}_{\mathcal{B}} \hookrightarrow \mathcal{H}$ is dense. A direct consequence of the weak convergence is that $|V_1(T)| \le \liminf_{\tau\to 0} |v_1^n|$. With the calculation

$$\big\langle \dot{\hat{V}}_{1,\tau}, V_{1,\tau}\big\rangle = \sum_{j=1}^n \big\langle v_1^j - v_1^{j-1}, v_1^j\big\rangle \ge -\frac{1}{2}\sum_{j=1}^n \Big(\big|v_1^{j-1}\big|^2 - \big|v_1^j\big|^2\Big) = \frac{1}{2}|v_1^n|^2 - \frac{1}{2}|v_1^1|^2$$

we finally conclude

$$\liminf_{\tau\to 0}\big\langle \dot{\hat{V}}_{1,\tau}, V_{1,\tau}\big\rangle \ge \frac{1}{2}\liminf_{\tau\to 0}\big(|v_1^n|^2 - |v_1^1|^2\big) \ge \frac{1}{2}\big|V_1(T)\big|^2 - \frac{1}{2}\big|V_1(0)\big|^2 = \big\langle \dot{V}_1, V_1\big\rangle. \quad \square$$

Remark 11.6. The fact that $\hat{V}_{1,\tau}(T) \rightharpoonup V_1(T)$ in $\mathcal{H}$ and $\hat{V}_{1,\tau}(0) = V_1(0)$, as shown in Lemma 11.5, implies with the integration by parts formula that for $w \in \mathcal{V}_{\mathcal{B}}$ and $\Phi \in C^2([0,T];\mathbb{R})$ it holds that

$$\begin{aligned}
\lim_{\tau\to 0}\big\langle \dot{\hat{V}}_{1,\tau}, w\dot{\Phi}\big\rangle &= \lim_{\tau\to 0} -\big\langle \hat{V}_{1,\tau}, w\ddot{\Phi}\big\rangle + \big(\hat{V}_{1,\tau}(T), w\big)\dot{\Phi}(T) - \big(\hat{V}_{1,\tau}(0), w\big)\dot{\Phi}(0)\\
&= -\big\langle V_1, w\ddot{\Phi}\big\rangle + \big(V_1(T), w\big)\dot{\Phi}(T) - \big(V_1(0), w\big)\dot{\Phi}(0) = \big\langle \dot{V}_1, w\dot{\Phi}\big\rangle.
\end{aligned}$$

The following lemma contains a similar result as in Lemma 11.5 for the stiffness operator $\mathcal{K}$.

Lemma 11.7. *The sequences $U_{1,\tau}$, $U_{2,\tau}$, and $V_{1,\tau}$ satisfy the estimate*

$$\liminf_{\tau\to 0}\big\langle \mathcal{K}(U_{1,\tau}+U_{2,\tau}), V_{1,\tau}\big\rangle \ge \big\langle \mathcal{K}(U_1+U_2), V_1\big\rangle.$$

Proof. Because of the linearity of $\mathcal{K}$ and the strong convergence of $U_{2,\tau}$ it is sufficient to analyse the limes inferior of $\langle \mathcal{K}U_{1,\tau}, V_{1,\tau}\rangle$ and show that $\liminf_{\tau\to 0}\langle \mathcal{K}U_{1,\tau}, V_{1,\tau}\rangle \ge \langle \mathcal{K}U_1, V_1\rangle$. For this, we proceed as in the proof of Lemma 11.5.

Lemma 11.4 implies the boundedness of $\hat{U}_{1,\tau}(T) = u_1^n$ in $\mathcal{V}$ such that there exists an element $\xi \in \mathcal{V}_{\mathcal{B}}$ with $\hat{U}_{1,\tau}(T) \rightharpoonup \xi$ in $\mathcal{V}_{\mathcal{B}}$. We show that $\mathcal{K}^{1/2}\xi = \mathcal{K}^{1/2}U_1(T)$ and $\mathcal{K}^{1/2}u_1^1 = \mathcal{K}^{1/2}U_1(0)$. Using the limit equation (11.11) and the semi-discrete equation

(11.10) with test functions $w \in \mathcal{V}_{\mathcal{B}}$ and $\Phi \in C^2([0,T];\mathbb{R})$, we obtain

$$\begin{aligned}
&\big\langle \mathcal{K}U_1(T),w\big\rangle\Phi(T) - \big\langle \mathcal{K}U_1(0),w\big\rangle\Phi(0)\\
&\qquad = \quad \big\langle \mathcal{K}\dot U_1, w\Phi\big\rangle + \big\langle \mathcal{K}U_1, w\dot\Phi\big\rangle\\
&\qquad \overset{(11.11)}{=} \big\langle \mathcal{K}\dot U_1, w\Phi\big\rangle + \big\langle \mathcal{F} - \rho W_2 - a - \mathcal{K}U_2 - \rho\dot V_1, w\dot\Phi\big\rangle\\
&\qquad \overset{(11.10)}{=} \big\langle \mathcal{F} - \mathcal{F}_\tau, w\dot\Phi\big\rangle - \rho\big\langle W_2 - W_{2,\tau}, w\dot\Phi\big\rangle - \big\langle a - \mathcal{D}(V_{1,\tau}+V_{2,\tau}), w\dot\Phi\big\rangle\\
&\qquad\qquad - \big\langle \mathcal{K}U_2 - \mathcal{K}U_{2,\tau}, w\dot\Phi\big\rangle - \rho\big\langle \dot V_1 - \dot{\hat V}_{1,\tau}, w\dot\Phi\big\rangle + \big\langle \mathcal{K}\dot U_1, w\Phi\big\rangle + \big\langle \mathcal{K}U_{1,\tau}, w\dot\Phi\big\rangle.
\end{aligned}$$

Passing to the limit with $\tau \to 0$, we make use of Remark 11.6 which implies that the term including $\dot V_1$ vanishes. In addition, we use the fact that, passing to the limit, we may replace $U_{1,\tau}$ by $\hat U_{1,\tau}$ since they have the same weak limit. Thus, another application of the integration by parts formula then leads to

$$\big\langle \mathcal{K}U_1(T), w\big\rangle\Phi(T) - \big\langle \mathcal{K}U_1(0), w\big\rangle\Phi(0) = \big\langle \mathcal{K}\xi, w\big\rangle\Phi(T) - \big\langle \mathcal{K}u_1^1, w\big\rangle\Phi(0).$$

Since $\langle \mathcal{K}\cdot,\cdot\rangle$ defines an inner product in $\mathcal{V}_{\mathcal{B}}$, we conclude that $U_1(T) = \xi$ and $U_1(0) = u_1^1$ in $\mathcal{V}_{\mathcal{B}}$. As a result, we obtain $\mathcal{K}^{1/2}u_1^n \rightharpoonup \mathcal{K}^{1/2}\xi = \mathcal{K}^{1/2}U_1(T)$ in $\mathcal{H}$ and $\mathcal{K}^{1/2}u_1^1 = \mathcal{K}^{1/2}U_1(0)$. Since $U_{1,\tau}$ and $V_{1,\tau}$ are both piecewise linear, as in the proof of Lemma 11.5, we may calculate that

$$\begin{aligned}
\big\langle \mathcal{K}U_{1,\tau}, V_{1,\tau}\big\rangle = \sum_{j=1}^{n} \big\langle \mathcal{K}u_1^j, u_1^j - u_1^{j-1}\big\rangle &\geq \frac12\big\langle \mathcal{K}u_1^n, u_1^n\big\rangle - \frac12\big\langle \mathcal{K}u_1^1, u_1^1\big\rangle + \tau\big\langle \mathcal{K}u_1^1, v_1^1\big\rangle\\
&= \frac12\big|\mathcal{K}^{1/2}u_1^n\big|^2 - \frac12\big|\mathcal{K}^{1/2}u_1^0\big|^2 + \tau\big\langle \mathcal{K}u_1^1, v_1^1\big\rangle.
\end{aligned}$$

Note that the term $\tau\langle \mathcal{K}u_1^1, v_1^1\rangle$ vanishes as $\tau \to 0$, since u_1^1 and v_1^1 are fixed. By the property $|\mathcal{K}^{1/2}U_1(T)| \leq \liminf_{\tau\to 0}|\mathcal{K}^{1/2}u_1^n|$ we finally summarize the partial results to

$$\begin{aligned}
\liminf_{\tau\to 0}\big\langle \mathcal{K}U_{1,\tau}, V_{1,\tau}\big\rangle &\geq \liminf_{\tau\to 0}\frac12\big|\mathcal{K}^{1/2}u_1^n\big|^2 - \frac12\big|\mathcal{K}^{1/2}u_1^1\big|^2\\
&\geq \frac12\big|\mathcal{K}^{1/2}U_1(T)\big|^2 - \frac12\big|\mathcal{K}^{1/2}U_1(0)\big|^2 = \big\langle \mathcal{K}U_1, \dot U_1\big\rangle = \big\langle \mathcal{K}U_1, V_1\big\rangle. \quad \square
\end{aligned}$$

With the previous two lemmata we are now able to prove that the limit of the damping term equals the damping operator applied to the limit functions.

THEOREM 11.8. *Assume right-hand sides $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$ and initial approximations $u_1^1 = g_0$, $v_1^1 = h_0 \in \mathcal{V}_{\mathcal{B}}$. Then, we have $a = \mathcal{D}(V_1 + V_2)$ and thus, the (weak) limits U_1, U_2, V_2, and W_2 solve the operator DAE* (7.14a) *for test functions $v \in \mathcal{V}_{\mathcal{B}}$.*

PROOF. Again we follow the ideas of [**ET10b**, Th. 5.1] where a first-order system is analyzed. We consider the semi-discrete equation (11.10) tested by $V_{1,\tau}$ and subtract the term $\langle \mathcal{D}(V_{1,\tau}+V_{2,\tau}) - \mathcal{D}w, V_{1,\tau}+V_{2,\tau}-w\rangle$ with $w \in L^2(0,T;\mathcal{V})$, which is non-negative because of the monotonicity of the damping operator. This then leads to

$$\begin{aligned}
0 \geq \big\langle \rho\dot{\hat V}_{1,\tau}, V_{1,\tau}\big\rangle + \big\langle \rho W_{2,\tau}, V_{1,\tau}\big\rangle + \big\langle \mathcal{K}\big(U_{1,\tau}+U_{2,\tau}\big), V_{1,\tau}\big\rangle - \big\langle \mathcal{F}_\tau, V_{1,\tau}\big\rangle\\
+ \big\langle \mathcal{D}\big(V_{1,\tau}+V_{2,\tau}\big), w - V_{2,\tau}\big\rangle + \big\langle \mathcal{D}w, V_{1,\tau}+V_{2,\tau}-w\big\rangle.
\end{aligned}$$

The application of the limes inferior on both sides in combination with Lemmata 11.5 and 11.7 then leads to

$$\begin{aligned} 0 \geq \big\langle \rho \dot{V}_1, V_1 \big\rangle + \big\langle \rho W_2, V_1 \big\rangle + \big\langle \mathcal{K}(U_1 + U_2), V_1 \big\rangle - \big\langle \mathcal{F}, V_1 \big\rangle \\ + \big\langle a, w - V_2 \big\rangle + \big\langle \mathcal{D} w, V_1 + V_2 - w \big\rangle. \end{aligned}$$

Note that we have used the fact that the sequences $V_{2,\tau}$ and $W_{2,\tau}$ converge strongly in $L^2(0,T;\mathcal{V})$ and that a equals the weak limit of $\mathcal{D}(V_{1,\tau} + V_{2,\tau})$. Rearranging the terms and applying the limit equation (11.11), we then obtain

$$\begin{aligned} \big\langle \mathcal{D} w, w - V_1 - V_2 \big\rangle &\geq \big\langle \rho \dot{V}_1 + \rho W_2 + \mathcal{K}(U_1 + U_2) - \mathcal{F}, V_1 \big\rangle + \big\langle a, w - V_2 \big\rangle \\ &\overset{(11.11)}{=} -\big\langle a, V_1 \big\rangle + \big\langle a, w - V_2 \big\rangle \\ &= \big\langle a, w - V_1 - V_2 \big\rangle. \end{aligned}$$

Following again the *Minty trick* as in the proof of Theorem 10.22, i.e., choosing $w := V_1 + V_2 + sv$ with an arbitrary function $v \in L^2(0,T;\mathcal{V})$ and different signs for s, we conclude that $a = \mathcal{D}(V_1 + V_2)$. Thus, with $V_1 = \dot{U}_1$ the limit equation (11.11) turns to

$$\rho \ddot{U}_1 + \rho W_2 + \mathcal{D}(\dot{U}_1 + V_2) + \mathcal{K}(U_1 + U_2) = \mathcal{F} \quad \text{in } \mathcal{V}_{\mathcal{B}}^*.$$

It remains to check whether U_1 satisfies the initial conditions. Note that $\dot{U}_1(0) = V_1(0) = v_1^1 = h_0$ was shown within the proof of Lemma 11.7, whereas $U_1(0) = u_1^1 = g_0$ was proved in Lemma 11.5. □

11.2.4. *Lagrange Multiplier.* In this subsection we analyse the limiting behavior of the Lagrange multiplier. Recall that the approximation of the Lagrange multiplier, namely λ^j, is given by equation (11.1a), i.e.,

$$\rho D v_1^j + \rho w_2^j + \mathcal{D}\big(v_1^j + v_2^j\big) + \mathcal{K}\big(u_1^j + u_2^j\big) + \mathcal{B}^* \lambda^j = \mathcal{F}^j \quad \text{in } \mathcal{V}^*.$$

In terms of the global approximations from Section 11.2.2 and with $\Lambda_\tau(t) := \lambda^j$ for $t \in]t_{j-1}, t_j]$, this equation can be written in the form

$$\rho\big(\dot{\hat{V}}_{1,\tau} + W_{2,\tau}\big) + \mathcal{D}\big(V_{1,\tau} + V_{2,\tau}\big) + \mathcal{K}\big(U_{1,\tau} + U_{2,\tau}\big) + \mathcal{B}^* \Lambda_\tau = \mathcal{F}_\tau \quad \text{in } \mathcal{V}^*. \tag{11.12}$$

As in Section 10 for first-order systems, we are not able to find a uniform bound of Λ_τ in $L^2(0,T;\mathcal{Q})$. This is caused by the missing upper bound of $\tau \sum_{j=1}^n \|D v_1^j\|_{\mathcal{V}^*}^2$. Hence, we show that the primitive of Λ_τ, namely $\tilde{\Lambda}_\tau$, converges to the solution of the considered operator DAE in a weaker sense.

In order to obtain an equation for $\tilde{\Lambda}_\tau$, we have to integrate equation (11.12) over the interval $[0,t]$. For an arbitrary test function $v \in \mathcal{V}$, this then leads to the equation

$$\big\langle \rho\big(\hat{V}_{1,\tau} + \tilde{W}_{2,\tau}\big), v \big\rangle + \big\langle \tilde{\mathcal{D}}, v \big\rangle + \big\langle \mathcal{K}\big(\tilde{U}_{1,\tau} + \tilde{U}_{2,\tau}\big), v \big\rangle + \big\langle \mathcal{B}^* \tilde{\Lambda}_\tau, v \big\rangle = \big\langle \tilde{\mathcal{F}}_\tau, v \big\rangle + \big\langle \rho v_1^1, v \big\rangle$$

with $\tilde{\mathcal{F}}_\tau$, $\tilde{U}_{1,\tau}$, $\tilde{U}_{2,\tau}$, and $\tilde{W}_{2,\tau}$ denoting the primitives of $\mathcal{F}_\tau$, $U_{1,\tau}$, $U_{2,\tau}$, and $W_{2,\tau}$, respectively, and

$$\big\langle \tilde{\mathcal{D}}(t), v \big\rangle := \int_0^t \big\langle \mathcal{D}\big(V_{1,\tau}(s) + V_{2,\tau}(s)\big), v \big\rangle \, \mathrm{d}s.$$

Note that the term $\rho v_1^1 = \rho \hat{V}_{1,\tau}(0)$ is independent of time and occurs due to the integration of $\dot{\hat{V}}_{1,\tau}$.

As for first-order systems, we are now able to bound $\tilde{\Lambda}_\tau$ in $C([0,T];\mathcal{Q})$. Because of (5.13), $\mathcal{F}_\tau$ is bounded in $L^2(0,T;\mathcal{V}^*)$ which implies that its primitive $\tilde{\mathcal{F}}_\tau$ is uniformly

bounded in $C([0,T];\mathcal{V}^*)$. Furthermore, we have shown in Section 11.2.2 the boundedness of $U_{1,\tau}$, $U_{2,\tau}$, and $W_{2,\tau}$ in $L^2(0,T;\mathcal{V})$. Thus, the primitives $\tilde{U}_{1,\tau}$, $\tilde{U}_{2,\tau}$, and $\tilde{W}_{2,\tau}$ are bounded in $C([0,T];\mathcal{V})$. With the Cauchy-Schwarz inequality, we calculate

$$\max_{t\in[0,T]} \left|\left\langle \tilde{\mathcal{D}}(t), v\right\rangle\right| \overset{(11.3)}{\leq} d_2 \int_0^T \|V_{1,\tau}(s)+V_{2,\tau}(s)\|\|v\| \,\mathrm{d}s \leq d_2 T^{1/2}\|V_{1,\tau}+V_{2,\tau}\|_{L^2(0,T;\mathcal{V})}\|v\|.$$

The boundedness of $V_{1,\tau}+V_{2,\tau}$ in $L^2(0,T;\mathcal{V})$ was already shown in Section 11.2.2. Finally, the estimate

$$\max_{t\in[0,T]} |\hat{V}_{1,\tau}(t)| \leq \max_j |v_1^j| \overset{(11.5)}{\leq} ce^{2d_0T/\rho}M$$

shows with the inf-sup constant β from Lemma 7.1 the boundedness of

$$\left\|\tilde{\Lambda}_\tau\right\|_{C([0,T];\mathcal{Q})} \leq \frac{1}{\beta} \max_{t\in[0,T]} \sup_{v\in\mathcal{V}} \frac{\left\langle \mathcal{B}^*\tilde{\Lambda}_\tau(t), v\right\rangle}{\|v\|}.$$

As a result, there exists a limit function $\tilde{\Lambda} \in L^p(0,T;\mathcal{Q})$ such that

$$\tilde{\Lambda}_\tau \rightharpoonup \tilde{\Lambda} \quad \text{in } L^p(0,T;\mathcal{Q})$$

for all $1 < p < \infty$. This then leads to the following convergence result.

THEOREM 11.9. *Assume right-hand sides $\mathcal{F} \in L^2(0,T;\mathcal{V}^*)$, $\mathcal{G} \in H^2(0,T;\mathcal{Q}^*)$ and initial data $u_1^1 = g_0$, $v_1^1 = h_0 \in \mathcal{V}_\mathcal{B}$. Then, the weak limit $\tilde{\Lambda}$ of the sequence $\tilde{\Lambda}_\tau$ in $L^2(0,T;\mathcal{Q})$ solves together with U_1, U_2, V_2, and W_2 system (7.14) in the weak distributional sense, meaning that for all $v \in \mathcal{V}$ and $\Phi \in C_0^\infty(0,T)$ it holds that*

$$\int_0^T -\rho\big\langle \dot{U}_1, v\big\rangle\dot{\Phi} + \big\langle \rho W_2 + \mathcal{D}\big(\dot{U}_1+V_2\big) + \mathcal{K}\big(U_1+U_2\big) - \mathcal{F}, v\big\rangle\Phi - \big\langle \mathcal{B}^*\tilde{\Lambda}, v\big\rangle\dot{\Phi}\ dt = 0$$

as well as $\mathcal{B}U_2 = \mathcal{G}$, $\mathcal{B}V_2 = \dot{\mathcal{G}}$, and $\mathcal{B}W_2 = \ddot{\mathcal{G}}$. Furthermore, U_1 satisfies the initial conditions $U_1(0) = g_0$ and $\dot{U}_1(0) = h_0$.

PROOF. Considering once more equation (11.12) and integrating by parts, for all $v \in \mathcal{V}$ and $\Phi \in C_0^\infty(0,T)$ we obtain

$$\int_0^T -\rho\big\langle \hat{V}_{1,\tau}, v\big\rangle\dot{\Phi} + \big\langle \rho W_{2,\tau} + \mathcal{D}\big(V_{1,\tau}+V_{2,\tau}\big) + \mathcal{K}\big(U_{1,\tau}+U_{2,\tau}\big) - \mathcal{F}_\tau, v\big\rangle\Phi - \big\langle \mathcal{B}^*\tilde{\Lambda}_\tau, v\big\rangle\dot{\Phi}\ \mathrm{d}t = 0$$

By the weak convergence of $\tilde{\Lambda}_\tau$, we conclude that

$$\int_0^T \left\langle \mathcal{B}^*\tilde{\Lambda}_\tau, v\right\rangle\dot{\Phi}\ \mathrm{d}t \ \rightarrow\ \int_0^T \left\langle \mathcal{B}^*\tilde{\Lambda}, v\right\rangle\dot{\Phi}\ \mathrm{d}t.$$

The convergence of all the remaining terms - also for test functions $v \in \mathcal{V}$ - as well as the satisfaction of the initial conditions was already shown in Theorem 11.8. □

In summary, we could prove the strong convergence of u_2, v_2, and w_2, the weak convergence of the differential variable u_1, and the convergence in the weak distributional sense of the Lagrange multiplier λ. This result emphasizes that the Lagrange multiplier behaves qualitatively different than the deformation variables.

11.3. Influence of Perturbations. As for first-order systems in Section 10.4, we analyse in this subsection the influence of perturbations in the right-hand sides. For this, we consider $\delta^j \in \mathcal{V}^*$ as well as θ^j, ξ^j, $\vartheta^j \in \mathcal{Q}^*$. As before, we indicate the solution of the perturbed problem by $\hat{\cdot}$. The differences of the exact and perturbed solution are denoted by

$$e_1^j := \hat{u}_1^j - u_1^j, \qquad e_2^j := \hat{u}_2^j - u_2^j, \qquad e_v^j := \hat{v}_2^j - v_2^j, \qquad e_w^j := \hat{w}_2^j - w_2^j.$$

The initial errors in u_1^1 and v_1^1 are denoted by e_1^1 and $\dot{e}_1^1$, respectively. Considering only test functions in $\mathcal{V}_\mathcal{B}$, these errors then satisfy the equation

$$\rho D D e_1^j + \rho e_w^j + \mathcal{D}\big(\hat{v}_1^j + \hat{v}_2^j\big) - \mathcal{D}\big(v_1^j + v_2^j\big) + \mathcal{K}\big(e_1^j + e_2^j\big) = \delta^j. \tag{11.13a}$$

Furthermore, e_2^j, e_v^j, and e_w^j satisfy in $\mathcal{Q}^*$ the equations

$$\mathcal{B} e_2^j = \theta^j, \qquad \mathcal{B} e_v^j = \xi^j, \qquad \mathcal{B} e_w^j = \vartheta^j. \tag{11.13b}$$

Equations (11.13b) directly lead to the estimates

$$\|e_2^j\| \le C_{\mathcal{B}^-} \|\theta^j\|_{\mathcal{Q}^*}, \quad \|e_v^j\| \le C_{\mathcal{B}^-} \|\xi^j\|_{\mathcal{Q}^*}, \quad \|e_w^j\| \le C_{\mathcal{B}^-} \|\vartheta^j\|_{\mathcal{Q}^*}.$$

From equation (11.13a) we obtain an estimate of the resulting error e_1^j. For this, we may follow again the lines of Lemma 11.4 and test the equation by $D e_1^j$. The only difference takes place is the estimate of the damping term for which we obtain here

$$\begin{aligned}
&\big\langle \mathcal{D}\big(\hat{v}_1^j + \hat{v}_2^j\big) - \mathcal{D}\big(v_1^j + v_2^j\big), D e_1^j \big\rangle \\
&\qquad = \big\langle \mathcal{D}\big(\hat{v}_1^j + \hat{v}_2^j\big) - \mathcal{D}\big(v_1^j + v_2^j\big), D e_1^j + e_v^j \big\rangle - \big\langle \mathcal{D}\big(\hat{v}_1^j + \hat{v}_2^j\big) - \mathcal{D}\big(v_1^j + v_2^j\big), e_v^j \big\rangle \\
&\qquad \overset{(11.3)}{\ge} d_1 \big\| D e_1^j + e_v^j \big\|^2 - d_0 \big| D e_1^j + e_v^j \big|^2 - d_2 \big\| D e_1^j + e_v^j \big\| \big\| e_v^j \big\|.
\end{aligned}$$

Following the remaining parts of the proof of Lemma 11.4, for $k \ge 2$ we then yield an estimate of the form

$$\rho \big| D e_1^k \big|^2 + \rho \sum_{j=2}^{k} \big| D e_1^j - D e_1^{j-1} \big|^2 + \tau d_1 \sum_{j=2}^{k} \big\| D e_1^j + e_v^j \big\|^2 + k_1 \big\| e_1^k \big\|^2 \le c e^{4 d_0 T/\rho} M_e^2.$$

Note that the calculation includes a restriction on the step size. The constant M_e then includes the initial errors as well as the perturbations. More precisely, assuming perturbations of comparable magnitude as in Remark 10.11, we have

$$M_e^2 = |\dot{e}_1^1|^2 + \|e_1^1\|^2 + T \Big[\|\delta\|_{\mathcal{V}_\mathcal{B}^*}^2 + \|\theta\|_{\mathcal{Q}^*}^2 + \|\xi\|_{\mathcal{Q}^*}^2 + \|\vartheta\|_{\mathcal{Q}^*}^2 \Big]. \tag{11.14}$$

Summarizing the estimates of this subsection, we obtain the following theorem.

THEOREM 11.10. *Consider system* (11.1) *with the operators $\mathcal{K}$, $\mathcal{D}$, and $\mathcal{B}$ from Section 11.1 and perturbations $\delta^j \in \mathcal{V}^*$ and θ^j, ξ^j, $\vartheta^j \in \mathcal{Q}^*$ which are all of the same order of magnitude. With the constant M_e from* (11.14) *and a sufficiently small step size τ the errors e_1^k, e_2^k, e_v^k, and e_w^k then satisfy*

$$\|e_1^k\|^2 + \|e_2^k\|^2 + \|e_v^k\|^2 + \|e_w^k\|^2 \le c e^{4 d_0 T/\rho} M_e^2.$$

As already seen in the previous sections, estimates of the Lagrange multiplier are more involved. Only in the linear case of first-order systems in Section 10.4 we were able to bound the error in the Lagrange multiplier in terms of the perturbations. For this, we have assumed $\delta \in \mathcal{H}^*$ and an orthogonality which ensures that $D e_1^j$ only appeares in the weaker norm of the space $\mathcal{H}$ instead of $\mathcal{V}$. Such an assumption seems unfeasible here because of the nonlinear damping operator. As a consequence, we are not able to provide comparable results for the present nonlinear case.

12. Summary and Outlook

Within this thesis we have introduced a regularization technique for semi-explicit operator DAEs as they appear in the dynamics of fluid flows or elastodynamics. We have shown that this reformulation does not change the solution set and that it can be seen as an index reduction (as known for DAEs) on operator level. Besides the well-posedness of the resulting operator DAE, the advantages of the regularization in terms of the numerical simulation have been displayed in detail.

Following the method of lines, i.e., discretizing in space first, we have obtained a DAE which is of lower index compared to the DAE we would get from the spatial discretization of the original equations. As known from the theory of DAEs, a lower index implies more robustness in terms of perturbations. The numerical example stresses the obtained robustness of the regularized system as we could gain a stable approximation of the pressure variable with relatively large errors within an iterative solver routine. Finally, we have observed that a semi-discretization of the regularized operator DAE leads to the same index-1 DAE as an application of minimal extension to the DAE resulting from the original operator DAE. Thus, the use of the regularization process facilitates the implementation of adaptive schemes as the index remains one, independent of the underlying finite element mesh. This means that an adaptation of the mesh does not call for another index reduction step.

Applying the method of Rothe, i.e., discretizing in time first, we have obtained a sequence of stationary PDEs which have to be solved in every time step. Due to the absence of the time-dependence, the underlying DAE structure looses its visibility. Nevertheless, the regularized operator equations are less sensible to perturbations in the right-hand sides. Note that this is of enormous practical importance, since spatial discretization errors may be interpreted as such perturbations. Furthermore, we have proved the convergence of the Euler scheme for first-order operator DAEs in the linear case. For a second-order system, as it appears in the dynamics of elastic media including a nonlinear damping term, we have proved the convergence of an analogous time integration scheme.

As the field of operator DAEs is wide and still not well-understood in several aspects, there remain many open problems. Linked to this thesis, the regularization procedure may be extended to further applications such as electromagnetics or generalized to a larger class of systems. In particular, this may contain coupled systems which maintain the semi-explicit structure if the coupling is realized with the help of the Lagrangian method. Since the system structure has been crucial for the regularization process, more general systems may call for different strategies.

Also in the analysis of temporal discretization schemes for operator DAEs there exists a great potential for improvements. One may apply other discretization schemes such as Runge-Kutta schemes in order to obtain the convergence of the Lagrange multiplier not only in the weak distributional sense. Furthermore, it would be preferable to detect the order of convergence of the discretization scheme in order to implement efficient simulation tools. For this, the accuracy of the spatial discretization has to be adjusted to the estimated error of the temporal discretization.

Bibliography

[AF03] R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*. Elsevier, Amsterdam, second edition, 2003.

[Alt92] H. W. Alt. *Lineare Funktionalanalysis*. Springer-Verlag, Heidelberg, second edition, 1992.

[Alt13a] R. Altmann. Index reduction for operator differential-algebraic equations in elastodynamics. *Z. Angew. Math. Mech. (ZAMM)*, 93(9):648–664, 2013.

[Alt13b] R. Altmann. Modeling flexible multibody systems by moving Dirichlet boundary conditions. In *Proceedings of Multibody Dynamics 2013 - ECCOMAS Thematic Conference (Zagreb, Croatia)*, 2013.

[Alt14] R. Altmann. Moving Dirichlet boundary conditions. *ESAIM Math. Model. Numer. Anal.*, 48:1859–1876, 11 2014.

[AH13] R. Altmann and J. Heiland. Finite element decomposition and minimal extension for flow equations. Preprint 2013–11, Technische Universität Berlin, Germany, 2013. accepted for publication in M2AN.

[AH14] R. Altmann and J. Heiland. Regularization of constrained PDEs of semi-explicit structure. Preprint 2014–05, Technische Universität Berlin, Germany, 2014.

[And04] D. Andreucci. Lecture notes on the Stefan problem. Lecture notes, Universitá da Roma La Sapienza, Italy, 2004.

[Arn93] M. Arnold. Stability of numerical methods for differential-algebraic equations of higher index. *Appl. Numer. Math.*, 13(1-3):5–14, 1993.

[Arn98a] M. Arnold. Half-explicit Runge-Kutta methods with explicit stages for differential-algebraic systems of index 2. *BIT*, 38(3):415–438, 1998.

[Arn98b] M. Arnold. *Zur Theorie und zur numerischen Lösung von Anfangswertproblemen für differentiell-algebraische Systeme von höherem Index*. VDI Verlag, Düsseldorf, 1998.

[AB07] M. Arnold and O. Brüls. Convergence of the generalized-α scheme for constrained mechanical systems. *Multibody Syst. Dyn.*, 18(2):185–202, 2007.

[AS00] M. Arnold and B. Simeon. Pantograph and catenary dynamics: A benchmark problem and its numerical solution. *Appl. Numer. Math.*, 34(4):345–362, 2000.

[ACPR95] U. M. Ascher, H. Chin, L. R. Petzold, and S. Reich. Stabilization of constrained mechanical systems with DAEs and invariant manifolds. *Mech. Structures Mach.*, 23(2):135–157, 1995.

[Bau10] O. A. Bauchau. *Flexible Multibody Dynamics*. Solid Mechanics and Its Applications. Springer-Verlag, 2010.

[Bau72] J. Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. *Comput. Methods Appl. Mech. Engrg.*, 1:1–16, 1972.

[BM11] R. Becker and S. Mao. Quasi-optimality of adaptive nonconforming finite element methods for the Stokes equations. *SIAM J. Numer. Anal.*, 49(3):970–991, 2011.

[BBBCM00] F. Ben Belgacem, C. Bernardi, N. Chorfi, and Y. Maday. Inf-sup conditions for the mortar spectral element discretization of the Stokes problem. *Numer. Math.*, 85(2):257–281, 2000.

[BMP93] C. Bernardi, Y. Maday, and A. T. Patera. Domain decomposition by the mortar element method. In *Asymptotic and numerical methods for partial differential equations with critical parameters (Beaune, 1992)*, pages 269–286. Kluwer Acad. Publ., Dordrecht, 1993.

[BR85] C. Bernardi and G. Raugel. Analysis of some finite elements for the Stokes problem. *Math. Comp.*, 44(169):71–79, 1985.

[BK04] W. Blajer and K. Kołodziejczyk. A geometric approach to solving problems of control constraints: theory and a DAE framework. *Multibody Syst. Dyn.*, 11(4):343–364, 2004.

[Bog07] V. I. Bogachev. *Measure Theory Vol. 1*. Springer-Verlag, Berlin, 2007.

[BKZ92] J. U. Brackbill, D. B. Kothe, and C. Zemach. A continuum method for modeling surface tension. *J. Comput. Phys.*, 100(2):335–354, 1992.

[Bra07] D. Braess. *Finite Elements - Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, New York, third edition, 2007.

[BCP96] K.E. Brenan, S.L. Campbell, and L. R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[BS08] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, New York, third edition, 2008.

[BF91] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, New York, 1991.

[CM99] S. L. Campbell and W. Marszalek. The index of an infinite-dimensional implicit system. *Math. Comput. Model. Dyn. Syst.*, 5(1):18–42, 1999.

[CH93] J. Chung and G. M. Hulbert. A time integration algorithm for structural dynamics with improved numerical dissipation: the generalized-α method. *Trans. ASME J. Appl. Mech.*, 60(2):371–375, 1993.

[Cia78] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978.

[Cia88] P. G. Ciarlet. *Mathematical Elasticity, Vol. 1*. North-Holland, Amsterdam, 1988.

[CDD+14] P. A. Cioica, S. Dahlke, N. Dhring, U. Friedrich, S. Kinzel, F. Lindner, T. Raasch, K. Ritter, and R. L Schilling. Convergence analysis of spatially adaptive Rothe methods. *Found. Comput. Math.*, 14(5):863–912, 2014.

[Clé75] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9(2):77–84, 1975.

[CP03] R. W. Clough and J. Penzien. *Dynamics of Structures*. McGraw-Hill, third edition, 2003.

[CR73] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Franc. Automat. Inform. Rech. Operat.*, 7(R-3):33–75, 1973.

[DPVY13] D. A. Di Pietro, M. Vohralík, and S. Yousef. Adaptive regularization, linearization, and discretization and a posteriori error control for the two-phase Stefan problem. *Math. Comp.*, 2013.

[ESF98] E. Eich-Soellner and C. Führer. *Numerical methods in multibody dynamics*. B. G. Teubner, Stuttgart, 1998.

[Emm99] E. Emmrich. Discrete versions of Gronwall's lemma and their application to the numerical analysis of parabolic problems. Preprint 637, Technische Universität Berlin, Germany, 1999.

[Emm01] E. Emmrich. *Analysis von Zeitdiskretisierungen des inkompressiblen Navier-Stokes-Problems*. Cuvillier, Göttingen, 2001.

[Emm04] E. Emmrich. *Gewöhnliche und Operator-Differentialgleichungen: Eine Integrierte Einführung in Randwertprobleme und Evolutionsgleichnugen für Studierende*. Vieweg, Wiesbaden, 2004.

[EM13] E. Emmrich and V. Mehrmann. Operator differential-algebraic equations arising in fluid dynamics. *Comput. Methods Appl. Math.*, 13(4):443–470, 2013.

[EŠT13] E. Emmrich, D. Šiška, and M. Thalhammer. On a full discretisation for nonlinear second-order evolution equations with monotone damping: construction, convergence, and error estimates. Technical report, University of Liverpool, 2013.

[ET10a] E. Emmrich and M. Thalhammer. Convergence of a time discretisation for doubly nonlinear evolution equations of second order. *Found. Comput. Math.*, 10(2):171–190, 2010.

[ET10b] E. Emmrich and M. Thalhammer. Stiffly accurate Runge-Kutta methods for nonlinear evolution problems governed by a monotone operator. *Math. Comp.*, 79(270):785–806, 2010.

[EGR10] P. Esser, J. Grande, and A. Reusken. An extended finite element method applied to levitated droplet problems. *Internat. J. Numer. Methods Engrg.*, 84(7):757–773, 2010.

[Eva98] L. C. Evans. *Partial Differential Equations*. American Mathematical Society (AMS), Providence, second edition, 1998.

[Fat85] H. O. Fattorini. *Second order linear differential equations in Banach spaces*. North-Holland, Amsterdam, 1985.

[Fla00] J. E. Flaherty. Finite Element Analysis. Lecture Notes, Math 6860, Rensselaer Polytechnic Institute, 2000.

[Fri68] A. Friedman. The Stefan problem in several space variables. *Trans. Amer. Math. Soc.*, 133:51–87, 1968.

[GGZ74] H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferential-Gleichungen*. Akademie-Verlag, 1974.

[Gau14] A. Gaul. Krypy. Iterative Solvers for Linear Systems. Public Git Repository, Commit: 110a1fb756fb, `https://github.com/andrenarchy/krypy`, 2014.

[GGL85] C. W. Gear, G. K. Gupta, and B. Leimkuhler. Automatic integration of Euler-Lagrange equations with constraints. *J. Comput. Appl. Math.*, 12-13:77–90, 1985.

[GP84] C. W. Gear and L. R. Petzold. ODE methods for the solution of differential/algebraic systems. *SIAM J. Numer. Anal.*, 21(4):716–728, 1984.

[GC01] M. Géradin and A. Cardona. *Flexible Multibody Dynamics: A Finite Element Approach*. John Wiley, Chichester, 2001.

[GR86] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations*. Springer-Verlag, Berlin, 1986.

[GOS+10] S. A. Goreinov, I. V. Oseledets, D. V. Savostyanov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. How to find a good submatrix. In *Matrix methods: theory, algorithms and applications*, pages 247–256. World Sci. Publ., Hackensack, 2010.

[GS00] P. M. Gresho and R. L. Sani. *Incompressible Flow and the Finite Element Method. Vol. 2: Isothermal Laminar Flow*. Wiley, Chichester, 2000.

[GM86] E. Griepentrog and R. März. *Differential-algebraic equations and their numerical treatment*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1986.

[GJH+13] S. Grundel, L. Jansen, N. Hornung, T. Clees, C. Tischendorf, and P. Benner. Model order reduction of differential algebraic equations arising from the simulation of gas transport networks. Preprint MPIMD/13-09, Max Planck Institute Magdeburg, Germany, 2013.

[Gün01] M. Günther. *Partielle differential-algebraische Systeme in der numerischen Zeitbereichsanalyse elektrischer Schaltungen*. VDI-Verlag, Düsseldorf, 2001.

[HLR89] E. Hairer, C. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge-Kutta methods*. Springer-Verlag, Berlin, 1989.

[HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, second edition, 1996.

[Hei14] J. Heiland. *Decoupling, Semi-discretization, and Optimal Control of Semi-linear Semi-explicit Index-2 Abstract Differential-Algebraic Equations and Application in Optimal Flow Control*. PhD thesis, Technische Universität Berlin, 2014.

[Hei15] J. Heiland. TayHoodMinExtForFlowEqns. Solution of Time-dependent 2D Nonviscous Flow with Nonconforming Minimal Extension. Public Git Repository, Commit: 8eb641f21d, `https://github.com/highlando/TayHoodMinExtForFlowEqns`, 2015.

[HV95] J. C. Heinrich and C. A. Vionnet. The penalty method for the Navier-Stokes equations. *Arch. Comput. Method E.*, 2:51–65, 1995.

[HR90] J. G. Heywood and R. Rannacher. Finite-element approximation of the nonstationary Navier-Stokes problem. IV: Error analysis for second-order time discretization. *SIAM J. Numer. Anal.*, 27(2):353–384, 1990.

[HHT77] H. M. Hilber, T. J. R. Hughes, and R. L. Taylor. Improved numerical dissipation for time integration algorithms in structural dynamics. *Earthquake Eng. Struc.*, 5(3):283–292, 1977.

[Hin00] M. Hinze. *Optimal and instantaneous control of the instationary Navier-Stokes equations*. Habilitationsschrift, Technische Universität Berlin, Institut für Mathematik, 2000.

[Hol07] M. H. Holmes. *Introduction to Numerical Methods in Differential Equations*. Springer-Verlag, New York, 2007.

[Hug87] T. J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Dover Publications, 1987.

[HH90] G. M. Hulbert and T. J. R. Hughes. Space-time finite element methods for second-order hyperbolic equations. *Comput. Methods Appl. Mech. Engrg.*, 84(3):327–348, 1990.

[JPD93] M. Jahnke, K. Popp, and B. Dirr. Approximate analysis of flexible parts in multibody systems using the finite element method. In W. Schiehlen, editor, *Advanced Multibody System Dynamics*, pages 237–256. Kluwer Academic Publishers, Stuttgart, 1993.

[KPSG85] R. J. Kee, L. R. Petzold, M. D. Smooke, and J. F. Grcar. Implicit methods in combustion and chemical kinetics modeling. In J. U. Brackbill and B. I. Cohen, editors, *Multiple Time Scales*, pages 113–144. Academic Press, Orlando, 1985.

[KS95] R. Kouhia and R. Stenberg. A linear nonconforming finite element method for nearly incompressible elasticity and Stokes flow. *Comput. Methods Appl. Mech. Engrg*, 124:195–212, 1995.

[KM04] P. Kunkel and V. Mehrmann. Index reduction for differential-algebraic equations by minimal extension. *Z. Angew. Math. Mech. (ZAMM)*, 84(9):579–597, 2004.

[KM06] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations: Analysis and Numerical Solution*. European Mathematical Society (EMS), Zürich, 2006.

[LMT01] R. Lamour, R. März, and C. Tischendorf. PDAEs and further mixed systems as abstract differential algebraic systems. Preprint 2001–11, Humboldt-Universität zu Berlin, Germany, 2001.

[LMT13] R. Lamour, R. März, and C. Tischendorf. *Differential-algebraic equations: a projector based analysis*. Springer-Verlag, Heidelberg, 2013.

[LM72] J.-L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications I*. Springer-Verlag, New York, 1972.

[LS65] J.-L. Lions and W. A. Strauss. Some non-linear evolution equations. *Bull. Soc. Math. France*, 93:43–96, 1965.

[Lip04] M. K. Lipinski. *A posteriori Fehlerschätzer für Sattelpunktsformulierungen nicht-homogener Randwertprobleme*. PhD thesis, Ruhr Universität Bochum, 2004.

[LORW12] A. Logg, K. B. Ølgaard, M. Rognes, and G. N. Wells. FFC: the FEniCS form compiler. In A. Logg, K.-A. Mardal, and G. Wells, editors, *Automated Solution of Differential Equations by the Finite Element Method*, pages 227–238. Springer-Verlag, Berlin, 2012.

[LP86] P. Lötstedt and L. R. Petzold. Numerical solution of nonlinear differential equations with algebraic constraints. I. Convergence results for backward differentiation formulas. *Math. Comp.*, 46(174):491–516, 1986.

[LO93] C. Lubich and A. Ostermann. Runge-Kutta methods for parabolic equations and convolution quadrature. *Math. Comp.*, 60(201):105–131, 1993.

[LO95] C. Lubich and A. Ostermann. Runge-Kutta approximation of quasi-linear parabolic equations. *Math. Comp.*, 64(210):601–627, 1995.

[LSEL99] W. Lucht, K. Strehmel, and C. Eichler-Liebenow. Indexes and special discretization methods for linear partial differential algebraic equations. *BIT*, 39(3):484–512, 1999.

[LS09] C. Lunk and B. Simeon. The reverse method of lines in flexible multibody dynamics. In *Multibody dynamics*, volume 12 of *Comput. Methods Appl. Sci.*, pages 95–118. Springer-Verlag, Berlin, 2009.

[Mat12] M. Matthes. *Numerical Analysis of Nonlinear Partial Differential-Algebraic Equations: A Coupled and an Abstract Systems Approach*. PhD thesis, Universität zu Köln, 2012.

[MS06] G. Matthies and F. Schieweck. A multigrid method for incompressible flow problems using quasi divergence free functions. *SIAM J. Sci. Comput.*, 28(1):141–171, 2006.

[MS93] S. E. Mattsson and G. Söderlind. Index reduction in differential-algebraic equations using dummy derivatives. *SIAM J. Sci. Comput.*, 14(3):677–692, 1993.

[Meh13] V. Mehrmann. Index concepts for differential-algebraic equations. In T. Chan, W.J. Cook, E. Hairer, J. Hastad, A. Iserles, H.P. Langtangen, C. Le Bris, P.L. Lions, C. Lubich, A.J. Majda, J. McLaughlin, R.M. Nieminen, J. Oden, P. Souganidis, and A. Tveito, editors, *Encyclopedia of Applied and Computational Mathematics*. Springer-Verlag, Berlin, 2013.

[Mos06] M. S. Moslehian. A survey on the complemented subspace problem. *Trends in Mathematics*, 9(1):91–98, 2006.

[NS11] M. Neumüller and O. Steinbach. Refinement of flexible space-time finite element meshes and discontinuous Galerkin methods. *Comput. Vis. Sci.*, 14(5):189–205, 2011.

[New59] N. M. Newmark. A method of computation for structural dynamics. *Proceedings of A.S.C.E.*, 3, 1959.

[Ost93] A. Ostermann. A class of half-explicit Runge-Kutta methods for differential-algebraic systems of index 3. *Appl. Numer. Math.*, 13(1-3):165–179, 1993.

[Ost91] G.-P. Ostermeyer. On Baumgarte stabilization for differential algebraic equations. In E. Haug and R. Deyo, editors, *Real-Time Integration Methods for Mechanical System Simulation*, volume 69 of *NATO ASI Series*, pages 193–207. Springer-Verlag, Heidelberg, 1991.

[PW60] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal.*, 5:286–292, 1960.

[Paz83] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*, volume 44 of *Applied Math. Sciences*. Springer-Verlag, New York, 1983.

[Pet82] L. R. Petzold. Differential–algebraic equations are not ODEs. *SIAM J. Sci. Statist. Comput.*, 3(3):367–384, 1982.

[RA05] J. Rang and L. Angermann. Perturbation index of linear partial differential-algebraic equations. *Appl. Numer. Math.*, 53(2-4):437–456, 2005.

[RT92] R. Rannacher and S. Turek. Simple nonconforming quadrilateral Stokes element. *Numer. Meth. Part. D. E.*, 8(2):97–111, 1992.

[RR04] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*. Springer-Verlag, New York, second edition, 2004.

[Ria08] R. Riaza. *Differential-algebraic systems*. World Scientific Publishing Co. Pte. Ltd., Hackensack, 2008.

[RS88] R. E. Roberson and R. Schwertassek. *Dynamics of multibody systems*. Springer-Verlag, Berlin, 1988.

[Rot30] E. Rothe. Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Randwertaufgaben. *Math. Ann.*, 102(1):650–670, 1930.

[Rou05] T. Roubíček. *Nonlinear Partial Differential Equations with Applications*. Birkhäuser Verlag, Basel, 2005.

[Ruž04] M. Ružička. *Nichtlineare Funktionalanalysis: Eine Einführung*. Springer-Verlag, London, 2004.

[Sad10] M. H. Sadd. *Elasticity: Theory, Applications, and Numerics*. Elsevier Science, Amsterdam, 2010.

[SB98] M. Schemann and F. A. Bornemann. An adaptive Rothe method for the wave equation. *Computing and Visualization in Science*, 1(3):137–144, 1998.

[SHD11] R. Seifried, A. Held, and F. Dietmann. Analysis of feed-forward control design approaches for flexible multibody systems. *Journal of System Design and Dynamics*, 5(3):429–440, 2011.

[Sha97] A. A. Shabana. Flexible multibody dynamics: review of past and recent developments. *Multibody Syst. Dyn.*, 1(2):189–222, 1997.

[She95] J. Shen. On error estimates of the penalty method for unsteady Navier-Stokes equations. *SIAM J. Numer. Anal.*, 32(2):386–403, 1995.

[Sim96] B. Simeon. Modelling a flexible slider crank mechanism by a mixed system of DAEs and PDEs. *Math. Comp. Model. Dyn.*, 2:1–18, 1996.

[Sim00] B. Simeon. *Numerische Simulation Gekoppelter Systeme von Partiellen und Differential-algebraischen Gleichungen der Mehrkörperdynamik*. VDI Verlag, Düsseldorf, 2000.

[Sim06] B. Simeon. On Lagrange multipliers in flexible multibody dynamics. *Comput. Method. Appl. M.*, 195(50–51):6993–7005, 2006.

[Sim13] B. Simeon. *Computational flexible multibody dynamics. A differential-algebraic approach*. Differential-Algebraic Equations Forum. Springer-Verlag, Berlin, 2013.

[SH90] J. L. Sohn and J. C. Heinrich. A Poisson equation formulation for pressure calculations in penalty finite element models for viscous incompressible flows. *Int. J. Numer. Meth. Eng.*, 30(2):349–361, 1990.

[Ste08] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements*. Springer-Verlag, New York, 2008.

[Ste06] A. Steinbrecher. *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems*. PhD thesis, Technische Universität Berlin, 2006.

[Tar06] L. Tartar. *An Introduction to Navier-Stokes Equation and Oceanography*. Springer-Verlag, Berlin, 2006.

[Tar07] L. Tartar. *An Introduction to Sobolev Spaces and Interpolation Spaces*. Springer-Verlag, Berlin, 2007.

[TH73] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Internat. J. Comput. & Fluids*, 1(1):73–100, 1973.

[Tem77] R. Temam. *Navier-Stokes Equations. Theory and Numerical Analysis*. North-Holland, Amsterdam, 1977.

[Tis96] C. Tischendorf. *Solution of index-2 differential algebraic equations and its application in circuit simulation*. PhD thesis, Humboldt-Universität zu Berlin, 1996.

[Tis03] C. Tischendorf. *Coupled systems of differential algebraic and partial differential equations in circuit and device simulation. Modeling and numerical analysis*. Habilitationsschrift, Humboldt-Universität zu Berlin, 2003.

[Trö09] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*. Vieweg+Teubner Verlag, Wiesbaden, 2009.

[Tur99] S. Turek. *Efficient Solvers for Incompressible Flow Problems. An Algorithmic and Computational Approach*. Springer-Verlag, Berlin, 1999.

[Ver96] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley-Teubner, Stuttgart, 1996.

[Wei96] J. Weickert. Navier-Stokes equations as a differential-algebraic system. Preprint SFB393/96-08, Technische Universität Chemnitz-Zwickau, 1996.

[Wei97] J. Weickert. *Applications of the Theory of Differential-Algebraic Equations to Partial Differential Equations of Fluid Dynamics*. PhD thesis, Technische Universität Chemnitz-Zwickau, Chemnitz, 1997.

[Wil98] E. L. Wilson. *Three Dimensional Static and Dynamic Analysis of Structures: A Physical Approach with Emphasis on Earthquake Engineering*. Computers and Structures Inc., Berkeley, 1998.

[Wlo87] J. Wloka. *Partial Differential Equations*. Cambridge University Press, Cambridge, 1987.

[Woh99] B. I. Wohlmuth. Hierarchical a posteriori error estimators for mortar finite element methods with Lagrange multipliers. *SIAM J. Numer. Anal.*, 36(5):1636–1658 (electronic), 1999.

[YPR98] J. Yen, L. R. Petzold, and S. Raha. A time integration algorithm for flexible mechanism dynamics: The DAE α-method. *Comput. Method. Appl. M.*, 158(3–4):341–355, 1998.

[Yos80] K. Yosida. *Functional analysis*. Springer-Verlag, Berlin, sixth edition, 1980.

[Zei86] E. Zeidler. *Nonlinear Functional Analysis and its Applications I: Fixed–Point Theorems*. Springer-Verlag, New York, 1986.

[Zei90a] E. Zeidler. *Nonlinear Functional Analysis and its Applications IIa: Linear Monotone Operators*. Springer-Verlag, New York, 1990.

[Zei90b] E. Zeidler. *Nonlinear Functional Analysis and its Applications IIb: Nonlinear Monotone Operators*. Springer-Verlag, New York, 1990.

Index