Decoupling and Optimization of Differential-Algebraic Equations with Application in Flow Control

vorgelegt von
Dipl.-Math.Techn.

Jan Heiland

geb. in Friedrichshafen

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:

Vorsitzender: Prof. Dr. Rolf Möhring
Gutachter: Prof. Dr. Michael Hinze
Gutachter: Prof. Dr. Volker Mehrmann
Gutachter: Prof. Dr. Tomáš Roubíček

Tag der wissenschaftlichen Aussprache: 13. Februar 2014

Berlin 2014
D 83

Contents

1. Introduction

The notion of the state is used to describe all kinds of technical, physical, chemical, sociological, or mental phenomena. The state can mean a velocity, a temperature, a color, a noise, or just a state of mind. The state can vary with the time, with the location, or with other parameters.

For states that vary with time and location, one can use partial differential equations to model the underlying phenomena in mathematical terms [21, 22]. Then one can analyze the obtained equations, in order to understand or also to control the phenomena.

In view of employing computers for understanding or controlling, one approximates the continuous equations by discrete equations. This is called a discretization.

In practice, it is generally recognized that finite-dimensional or discrete formulations are well suited to deliver approximate solutions to continuous or infinite-dimensional problems. Thus one can use common numerical methods, optimized to solve the finite-dimensional equations, and to some extent forget about the actual aim of learning about the infinite-dimensional phenomena.

The design of discretizations that preserve certain properties of the continuous equations, like positivity of the solution, dissipassivity, symmetry, or energy of the system, has been proven very successful in providing efficient and reliable approximations.

It may happen that the discretization introduces properties that the continuous system does not have.

Consider the example of the Navier-Stokes Equation that model the state of an incompressible flow via its velocity v and its pressure p in a domain Ω and a time interval $(0, T)$. The describing equations are given via the system

$$\dot{v} + (v \cdot \nabla)v + \nabla p - \nu \Delta v = f, \tag{1.1a}$$

$$\operatorname{div} v = 0, \quad \text{in } \Omega \times (0, T), \tag{1.1b}$$

and

$$v|_{t=0} = \alpha \quad \text{and} \quad \mathrm{v}|_{\partial\Omega} = \gamma, \tag{1.1c}$$

consisting of the momentum equation with a viscosity parameter ν, the constraint that the flow is divergence-free, and a condition on the initial state of the velocity plus values for the velocity at the boundary $\partial\Omega$.

Discretizing the spatial component in (1.1), i.e. approximating $v(t)$ and $p(t)$ via finite-dimensional vectors $v_k(t)$ and $p_k(t)$, a discrete approximation to (1.1) is typically given as

$$M\dot{v}_k - A(v_k) - J_k^{\mathsf{T}} p_k = f_k, \tag{1.2a}$$

$$J_k v_k = 0, \quad \text{in } (0, T) \tag{1.2b}$$

and

$$v_k(0) = \alpha_k. \tag{1.2c}$$

Here, k is a parameter describing the order of approximation, and the matrices J_k and J_k^{T} represent the differential operators div and ∇ in finite dimensions.

For all common discretizations, the so called mass matrix M is invertible, so that from Equations (1.2a-c) one can infer, that if $J_k \alpha_k = 0$, then a solution v_k fulfills $J_k \dot{v}_k = 0$ for all time. Numerical methods [50, 144], that are state of the art, use the property that (1.2) implicitly defines the *Pressure Poisson Equation*

$$-J_k M^{-1} J_k^{\mathsf{T}} p_k = J_k M^{-1} f + J_k M^{-1} A(v_k)_k, \tag{1.3}$$

to decouple the computations of v_k and p_k.

However, the basic assumption that $J_k \dot{v}_k = 0$ will not necessarily transfer to the continuous case of (1.1), since in the general formulation $\dot{v}$ has to be assumed of low regularity and $\operatorname{div} \dot{v}$ is not defined. The proof whether or not there a solution possesses a higher regularity is a key in the quest for a solution to the *millennium problem* addressing the existence and uniqueness of solutions to the Navier-Stokes Equation [103]. In our considerations, the use of (1.3) is legitimate in finite dimensions but it might not be a proper approximation in the limit case, where the discretization is arbitrarily close to the continuous equation, cf. also the remarks in [50, p. 642].

Other examples relate to system theoretic properties like controllability and the question when their presence in the discrete equations is transferred to the limit case, see [105, 108] and see [77] for an illustrating example where this is not the case.

The above example of the *Pressure Poisson Equation* points to the general issue about the conditions under which a transformation of the discrete approximation commutes with a discrete approximation of a related transformation of the continuous problem. In this thesis on decoupling and optimization of differential-algebraic equations, we consider two such transformations, namely

(a) formulation of the optimal control problem as a system of optimality conditions and
(b) decoupling of the differential and algebraic equations

and their interaction with numerical approximations. We will investigate when do these transformations, whose finite-dimensional counterparts are well understood, also apply in the continuous setting and whether they commute with discretizations.

As for optimal control, the question of whether to optimize or to discretize first has been investigated in all fields of application, cf. [79] and see [33, 55, 146, 147] for examples in optimal flow control. As for infinite-dimensional differential-algebraic equations, the interaction of decoupling and discretization has attained attention recently [6, 7, 41].

Understanding the transformations in infinite dimensions may lead to discretizations that come with properties that are desired and compliant with those of the continuous equation. As an example consider the skew-symmetrization of the convective term in weak formulations of the Navier-Stokes Equation such that also the discretized convection is skew-symmetric, [68, Ch. 3]. Also, the direct application of algorithms for model reduction [63, 138, 123], update formulas in optimization [52, 72, 78, 95], or decoupling of differential and algebraic parts [6, 7] to infinite dimensional systems have been proven successful for numerical approximation.

The idea of interchanging transformations like discretization, optimization and decoupling is also reflected in the second part of the thesis dealing with the optimal control of finite-dimensional of differential-algebraic equations. Instead of first formulating a decoupling and then stating optimality conditions, we set up formal optimality conditions for the original system. Thus, a possibly necessary decoupling can be applied with respect to efficient numerical approximation rather than being used for theoretical considerations.

We use the particular structure of the equations to show that the solution of the original system implicitly leads to the solution of an equivalent system for which differential and algebraic parts are partially decoupled and optimality conditions are well understood.

Staying with the original differential-algebraic equations and the native variables is motivated by practical considerations. In computations, the decoupling, for example the restriction to the space of divergence-free functions in the Navier-Stokes setting, is not advisable in general or simply unfeasible, cf. the discussion in [7].

Also, if the optimality conditions are in the form of the state equations, then one can resort to the same solvers.

The thesis is structured as follows. We start with an introduction of notions related to differential-algebraic equations and of the functional analysis framework needed to treat infinite-dimensional equations. In Section 3, we define a class of infinite-dimensional differential-algebraic equations and derive a decoupling of differential and algebraic parts. We demonstrate how the abstract concepts apply to the Navier-Stokes Equation. In the following section, Section 4, we formulate the spatially semi-discretized approximation and state convergence of Galerkin schemes. In Section 5, we introduce basic concepts of optimization in Banach spaces. Section 6 is on optimal control of the infinite-dimensional equations. In order to extend known results to differential-algebraic equations, we recall basic and classical results on optimization in Banach spaces. Then we formulate necessary optimality conditions by introducing a formal adjoint equation. We derive necessary conditions for existence of solutions and prove convergence of Galerkin approximations. Also, we discuss how semi-discretization and formulation of optimality conditions are related. In Section 7 we discuss possible linearization approaches to the nonlinear optimality conditions. In Section 8, we consider, finite-dimensional approximations and, in particular, the linear-quadratic optimal control problem. We will give necessary and sufficient conditions for existence of optimal solutions and provide a solution representation via a differential-algebraic Riccati decoupling. In the section on numerical routines, we derive algorithms for efficient solution of the linear-quadratic optimality system.

We conclude the thesis by a summary of the presented work and an outlook. We discuss which questions have been answered and name remaining and related questions and problems for future research.

2. Preliminary Notions and Notations

We introduce some basic concepts concerning differential-algebraic equations and constrained optimal control problems formulated in finite-dimensional and infinite-dimensional spaces.

2.1. **A class of Semi-explicit Semi-linear DAEs.**

We start with defining a prototype of the DAEs that are considered throughout this work. We will define the tractability index of these equations in finite-dimensional cases and discuss its equivalence to other indices and extension to infinite-dimensional state spaces.

For $T > 0$, for a time parameter $t \in (0, T)$, and for variables $v(t)$ and $p(t)$, we will consider equations of type

$$\begin{bmatrix} M(t) & 0 \\ 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} v \\ p \end{bmatrix} (t) - \begin{bmatrix} A(t, v(t)) & J_1^T(t) \\ J_2(t) & 0 \end{bmatrix} \begin{bmatrix} v(t) \\ p(t) \end{bmatrix} = \begin{bmatrix} f(t) \\ g(t) \end{bmatrix}, \tag{2.1}$$

where the time derivative is interpreted either in the classical or in a generalized sense. If the state space is continuous, i.e. $v(t)$ and $p(t)$ are located in infinite-dimensional spaces, then we will refer to (2.1) as abstract differential-algebraic equation. Throughout this work, the time parameter t will always be continuous, i.e. we will only consider semi-discretizations. Nevertheless, we will refer to the semi-discretized equations as finite-dimensional or discrete equations.

To introduce basic concepts for the analysis differential-algebraic equations, we state the following finite-dimensional setup:

Problem 2.1. *Consider Equation* (2.1)*, with* $\frac{d}{dt}v(t)$*,* $v(t)$*, and* $f(t) \in \mathbb{R}^{n_v}$*, with* $p(t)$ *and* $g(t) \in \mathbb{R}^{n_p}$*,* n_v*,* $n_p \in \mathbb{N}$ *and* $n_v > n_p$*, and with* $M(t)$*,* $A(t, v(t))$*,* $J_2(t)$*, and* $J_1^T(t)$ *being matrices of suitable sizes with entries in* $\mathcal{C}(0, T; \mathbb{R})$*.*

We will always assume that $M(\cdot)$ is pointwise invertible and refer to (2.1) as a semi-explicit, semi-linear DAE. The *semi*s mean that, at least in the time derivative of the variables, the equation are explicit and linear. Additionally we assume that $J_1(t)$ and $J_2(t)$ are linear functions.

To the first equation in (2.1) we will refer as the *differential part* and to the second as the *algebraic part* or *algebraic constraint*. Note, that there are other, so-called *hidden constraints*, which constrain the motion of v and p but which are not apparent in (2.1), see [90, p. 177] and [102, p. 323, 426] for examples. Also, there is the notion of the *inherent ODE* that is obtained by formulating (2.1) as an ODE on a manifold that is prescribed via all algebraic constraints, cf. [102, p. 323].

2.2. **The Index of the DAEs.**

We will use the *tractability index* to quantify the considered DAEs, since, in finite dimensions, it can be directly checked for the semi-explicit semi-linear DAEs under consideration.

We introduce a general formulation of a linear DAE and some notions to define the *tractability index* of a DAE.

Omitting the time dependencies, we consider a general linear DAE of the form

$$\mathcal{E}_A \tfrac{d}{dt}(\mathcal{E}_D x) - \mathcal{A}x = q, \quad \text{on } (0, T), \tag{2.2}$$

for a state $x(t) \in \mathbb{R}^{n_x}$ and with $\mathcal{E}_A^T$, $\mathcal{E}_D \in \mathcal{C}(0, T; \mathbb{R}^{n_x, n_d})$, $\mathcal{A} \in \mathcal{C}(0, T; \mathbb{R}^{n_x, n_x})$, and $q \in \mathcal{C}(0, T; \mathbb{R}^{n_x})$.

If $\mathcal{E}_A$ and $\mathcal{E}_D$ are invertible, then (2.2) can be rewritten as an ordinary differential equation. If this is not the case, then (2.2) comprises both differential and algebraic equations and is, thus, referred to as a DAE.

Definition 2.2 (Cf. [114], Def. 2.1)**.** The DAE (2.2) has a *properly stated leading term*, if

$$\mathbb{R}^{n_x} = \operatorname{im} \mathcal{E}_A(t) \oplus \ker \mathcal{E}_D(t), \tag{2.3}$$

for all $t \in (0, T)$.

Definition 2.3 (Cf. [114], Eqn. (2.2))**.** Consider Equation (2.2). Given

$$\mathcal{E}_0 := \mathcal{E}_A \mathcal{E}_D \quad \text{and} \quad \mathcal{A}_0 := \mathcal{A},$$

for $i = 0, 1, 2, \cdots$, define the sequences of subspaces and matrices via

$$\begin{aligned} N_i :=& \ker \mathcal{E}_i, \\ S_i :=& \{x \in \mathbb{R}^{n_x} : \mathcal{A}_i x \in \operatorname{im} \mathcal{E}_i\}, \\ \mathscr{Q}_i :=& \mathcal{P}_{[N_i|\cdot]} \text{ (projector onto } \ker \mathcal{E}_i), \\ \mathscr{P}_i :=& I - \mathscr{Q}_i, \end{aligned}$$

and

$$\mathcal{E}_{i+1} := \mathcal{E}_i + \mathcal{A}_i \mathscr{Q}_i \quad \text{and} \quad \mathcal{A}_{i+1} := \mathcal{A}_i \mathscr{P}_i.$$

The definition of the spaces and matrices hold pointwise for $t \in (0, T)$.

We can now define the tractability index :

Definition 2.4 ([114], Def. 2.2)**.** Consider equation (2.2) and assume that the matrix coefficients are continuous and that $\mathcal{E}_A$ and $\mathcal{E}_D$ fulfill Definition (2.2). Consider the sequences of operators and subspaces defined in Definition 2.3. Then, the differential-algebraic equation (2.2) has

(a) *tractability index* $i_\mu = 1$, if $N_0 \cap S_0$ has a constant dimension $d_0 > 0$ and $\dim(N_1 \cap S_1) = 0$, for all $t \in (0, T)$.

(b) *tractability index* $i_\mu = 2$, if $\dim(N_0 \cap S_0) = d_0$ and $\dim(N_1 \cap S_1) = d_1$, with d_0, $d_1 > 0$, and $\dim(N_2 \cap S_2) = 0$, for all $t \in (0, T)$,

or,

(c) *tractability index* $i_\mu = \mu$, if $\dim(N_j \cap S_j) = d_j > 0$, for $j = 0, 1, \cdots, \mu$, and $\dim(N_\mu \cap S_\mu) = 0$.

Remark 2.5. For nonlinear DAEs, one defines the *tractability index* as the tractability index that, if it exists, is obtained via Definition 2.4, for all linearizations of the state equations about all states x^* in a neighborhood of a solution, cf. [115, Def. 2.3].

It will turn out, that in the semi-linear case of the form 2.1 the nonlinearity does not interfere with definition of the operators and subspaces, so that we can directly apply the linear theory.

Proposition 2.6. *Consider the setup of Problem 2.1. If M is pointwise invertible and if $J_2 M^{-1} J_1^T$ is pointwise invertible, then the DAE* (2.1) *is of tractability index* $i_\mu = 2$.

Proof. Having factorized, the leading matrix $\begin{bmatrix} M(t) & 0 \\ 0 & 0 \end{bmatrix}$, as $\mathcal{E}_A := \begin{bmatrix} M(t) \\ 0 \end{bmatrix}$ and $\mathcal{E}_D := \begin{bmatrix} I & 0 \end{bmatrix}$, and using that M is invertible, we find that (2.1) has a *properly stated leading term*, cf. Definition 2.4.

Furthermore, using the invertibility of $J_2 M^{-1} J_1^T$, we can give explicit representation of the matrix sequences defined in Definition 2.3 in its realization for the

DAE (2.1):

$$\begin{aligned}
\mathcal{E}_0 &:= \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{A}_0 := \begin{bmatrix} A(v) & J_1^T \\ J_2 & 0 \end{bmatrix}, \\
\mathscr{Q}_0 &:= \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \text{ (projector onto } \ker \mathcal{E}_0), \\
\mathscr{P}_0 &:= I - \mathscr{Q}_0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \\
\mathcal{E}_1 &:= \mathcal{E}_0 + \mathcal{A}_0 \mathscr{Q}_0 = \begin{bmatrix} M & J_1^T \\ 0 & 0 \end{bmatrix}, \quad \mathcal{A}_1 := \mathcal{A}_0 \mathscr{P}_0 = \begin{bmatrix} A(v) & 0 \\ J_2 & 0 \end{bmatrix}, \\
\mathscr{Q}_1 &:= \begin{bmatrix} M^{-1} J_1^T (J_2 M^{-1} J_1^T)^{-1} J_2 & 0 \\ -(J_2 M^{-1} J_1^T)^{-1} J_2 & 0 \end{bmatrix} =: \begin{bmatrix} \mathcal{Q} & 0 \\ -\mathcal{Q}^- & 0 \end{bmatrix} \text{ (projector onto } \ker \mathcal{E}_1), \\
\mathscr{P}_1 &:= I - \mathscr{Q}_1 =: \begin{bmatrix} \mathcal{P} & 0 \\ \mathcal{Q}^- & I \end{bmatrix}, \\
\mathcal{E}_2 &:= \mathcal{E}_1 + \mathcal{A}_1 \mathscr{Q}_1 = \begin{bmatrix} M + A\mathcal{Q} & J_1^T \\ J_2 & 0 \end{bmatrix}, \quad \mathcal{A}_2 := \mathcal{A}_1 \mathscr{P}_1.
\end{aligned}$$

To check for $i_\mu = 2$, see Definition 2.4(b), we now have to check the dimensions of the spaces $N_0 \cap S_0$ and $N_1 \cap S_1$, as defined in Definition 2.3. In what follows, we will arbitrarily switch between a space, e.g. N_0, and a matrix, e.g. $N_0 = \begin{bmatrix} 0 \\ I \end{bmatrix}$, if the columns of $\begin{bmatrix} 0 \\ I \end{bmatrix}$ span N_0. From the assumption that $J_2 M^{-1} J_1^T$ is invertible, for all $t \in (0, T)$, we have that J_2 and J_1 have full rank n_p.

With $N_0 = \begin{bmatrix} 0 \\ I \end{bmatrix}$ and $S_0 = \begin{bmatrix} \ker J_2 \\ I \end{bmatrix}$, we have that $N_0 \cap S_0$ has dimension n_p for all $t \in (0, T)$.

With $N_1 = \begin{bmatrix} \mathcal{Q} \\ -\mathcal{Q}^- \end{bmatrix}$, with $S_1 = \begin{bmatrix} \ker J_2 \\ I \end{bmatrix}$, and with the observation that $\mathcal{Q} \cap \ker J_2 = \{0\}$, we find that $N_1 \cap S_1 = \begin{bmatrix} 0 \\ -\mathcal{Q}^- \end{bmatrix}$ which has a rank of $n_v - n_p > 0$ for all $t \in (0, T)$.

Finally, we find that

$$\mathcal{E}_2^{-1} = \begin{bmatrix} \mathcal{P} M^{-1} & [I - \mathcal{P} M^{-1} A] M^{-1} J_1^T S^{-1} \\ \mathcal{Q}^- M^{-1} & -[I + \mathcal{Q}^- M^{-1} A M^{-1} J_1^T] S^{-1} \end{bmatrix}$$

is a inverse to $\mathcal{E}_2$ what means that $\dim(\ker \mathcal{E}_2 \cap S_2)$ is zero for all $t \in (0, T)$. □

Remark 2.7. The definition of the tractability index is applicable because A is assumed of the form $v \mapsto A(v)[v]$. For a general nonlinear A, the tractability index is defined for a linearization about states in the neighborhood of a solution, cf. Remark 2.5. Thus, provided the coefficients are sufficiently smooth, in the considered semi-linear setup, linearization will give an equation of type (2.1) but with $A(v)$ replaced by a term linear in v. Thus, the definition of the index via Definition also applies in the case that $A\colon v \mapsto A(v)$.

Before commenting on the index for abstract systems, we mention the commonly used *differentiation index* and its relation to the *tractability index*, see [116] for a general overview.

Definition 2.8 ([116] Def. 1)**.** Write (2.1) as $F(t, x, \dot{x}) = 0$ on $(0, T)$, assume that it has a solution x, and, formally, define

$$F_l(t, x, \dot{x}, \cdots, x^{(l+1)}) = \begin{bmatrix} F(t,x,\dot{x}) \\ \frac{d}{dt}F(t,x,\dot{x}) \\ \vdots \\ (\frac{d}{dt})^l F(t,x,\dot{x}) \end{bmatrix}.$$

The smallest integer i_ν, if it exists, such that the solution x is uniquely defined via $F_l(t, x, \dot{x}) = 0$ for any initial value that fulfills all (hidden) constraints, is called the *differentiation index.*

Remark 2.9. The definition of the *differentiation index* i_ν for (2.1) requires a certain smoothness of the coefficients. If, in the case of Problem 2.1, the *differentiation index* is defined, then it coincides with the *tractability index*, cf. [13, Rem. 2.3].

To fit also abstract systems of type (2.1), the *tractability index* has been generalized in [101, 145], see in particular [115] for a general discussion of index concepts for abstract DAEs. However, the basic assumption on properness of the leading term, cf. Definition (2.2), cannot be transferred to the evolution setting, that we will consider in Section 3 and that is discussed in [115, 145]. The problem here is that, the state space X is assumed strictly smaller than its dual X' where the equations are posed in. Then, in the generic case, cf. [145, Ch. 4], one has that $\frac{d}{dt}\colon \big((0,T) \to X\big) \to \big((0,T) \to X'\big)$ and the factorization of the leading term as $\mathcal{E}_A \frac{d}{dt} \mathcal{E}_D$ with $\mathcal{E}_A \colon X' \to X'$ as the dual operator of $\mathcal{E}_D \colon X \to X$.

Another index concept, the *perturbation index*, cf. [58], has been generalized to the evolution setting in [115, Def. 2.4].

We will quantify the equations of Section 3 by means of the index that is obtained after applying stable semi discretizations. In this sense, we will consider abstract DAEs of *tractability index* 2, while in [145] and [115] the index-1 case was investigated.

For completeness, we mention the *Kronecker index* for abstract DAEs as it was considered in [131], that can be defined in the linear time invariant case.

2.3. **The Index of DAEs with Inputs.** The definitions of the *tractability index* and the *differentiation index* apply for uncontrolled systems as (2.1). In this section, we will remark on index concepts for systems that include inputs, as, e.g.,

$$\begin{bmatrix} M(t) & 0 \\ 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} v \\ p \end{bmatrix}(t) - \begin{bmatrix} A(t, v(t)) & J_1^T(t) \\ J_2(t) & 0 \end{bmatrix} \begin{bmatrix} v(t) \\ p(t) \end{bmatrix} - \begin{bmatrix} B_1(t) \\ B_2(t) \end{bmatrix} u(t) = \begin{bmatrix} f(t) \\ g(t) \end{bmatrix}, \quad (2.4)$$

where u is located, e.g., in $\mathcal{U} := \mathcal{C}(0, T; \mathbb{R}^{n_u})$, $n_u \in \mathbb{N}$, and $\begin{bmatrix} B_1(t) \\ B_2(t) \end{bmatrix}$ is a linear operator that maps $u \in \mathcal{U}$ into the space where the equation is posed in.

For such controlled systems (2.4), the relation of inputs u and variables (v, p) has to be included in the definition of an index. As an illustrating example consider the system

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad x_1(0) = 0. \quad (2.5)$$

and note that the definition of the input can change the system's differential-algebraic structure. With the assignment $u_2 := \dot{x}_2$, System (2.5) can be interpreted as an ODE for x_1 and x_2. Assigning $u_2 := x_1$, the system can be reformulated to give only algebraic equations for x_1 and x_2.

A general approach to this issue bases on the *behavior formulation*, cf. [128], which considers the controlled DAE (2.4) as an underdetermined system $\mathcal{E}_z \dot{z}$ −

$\mathcal{A}_z(z) = 0$ in the augmented variable $z := [x, u]$. For the behavior formulation, one can define the *strangeness index* [26, 90] that generalizes the *differentiation index* to under- (and overdetermined) systems.

For the behavior system, one can identify *free components* of z that are then defined as the new controls, cf. [27, 90, 91]. This approach is natural, since if the chosen controls are not free variables in the behavior formulation, then the problem is ill-posed.

If it comes to applications, however, one cannot freely redefine the controls and variables since they are prescribed by the physical setup. In this case one can only hope that the problem is well-posed, in the sense that, the system can be reformulated as a system of *strangeness index* $i_s = 0$ with the current input u as free components, cf. [27]. In this case, one can apply the general algorithms provided in [27, 90, 91] to reformulate the considered system as a *strangeness-free* system, for which necessary and sufficient optimality conditions are well understood [27, 91, 92].

The definition of the *tractability index* has not been generalized to controlled systems like (2.4). We will not investigate the index for the controlled system. In view of solving associated optimal control problems numerically, we will, however, determine the *tractability index* for some particular formal optimality systems.

2.4. **Functional Analysis Framework.**

In view of abstract formulations we introduce basic definitions and notions of operators on Banach spaces. Then, in view of modelling states of dynamical systems, we introduce spaces of functions in spatial variables. To account for time evolution, we introduce the concept and functional analytical framework of abstract functions, i.e. functions of time that take on values in function spaces. If not stated otherwise, the basic notation and propositions of this section are taken from [133, Ch. 1, 7]. A detailed introduction into the functional analytical preliminaries for evolution equations can be found, e.g., in [45].

Banach Spaces and Operators

Let V, W be real or complex Banach spaces. A linear map $A\colon V \to W$ is called a *bounded* or *continuous* linear operator if there exists a constant c such that $\|Av\|_W \leq c\|v\|_V$. We denote the set of bounded linear operators by $\mathcal{L}(V, W)$, which – equipped with the norm

$$\|A\|_{\mathcal{L}(V,W)} = \sup_{v \in V \setminus \{0\}} \frac{\|Av\|_W}{\|v\|_V}$$

– is a Banach space.

Dual Spaces and Dual Operators

For a normed vector space V, its dual space V' is defined as $V' := \mathcal{L}(V, \mathbb{R})$, which is a Banach space with the dual norm

$$\|v'\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{\langle v', v\rangle_{V' \times V}}{\|v\|_V},$$

where $\langle v', v\rangle_{V' \times V} := v'(v)$ stands for the dual form. To emphasise its linearity, we will refer to the dual form as the dual product. For $A \in \mathcal{L}(V, W)$ there exists a unique dual operator $A' \in \mathcal{L}(W', V')$, satisfying

$$\langle w', Av\rangle_{W' \times W} = \langle A'w', v\rangle_{V' \times V} \quad \text{and} \quad \|A\|_{\mathcal{L}(V,W)} = \|A'\|_{\mathcal{L}(W',V')}$$

for all $w' \in W'$ and all $v' \in V'$. If A is defined on a domain of definition $D(A)$ that is closed and dense in V, then the *dual operator* $A'\colon D(A') \subset W' \to V'$ is defined on

$$D(A') := \{w' \in W' : \text{ there exists } v' \in V' : \big\langle w', Av\big\rangle = \big\langle v', v\big\rangle \text{ for all } v \in D(A)\}.$$

A normed vector space V is called *reflexive*, if there exists an isometric isomorphism between V and $(V')' =: V''$. If this is the case we will identify V'' with V to make use of the induced symmetry of the dual product:

$$\langle v', v \rangle_{V',V} = v'(v) = v''(v') = \langle v'', v' \rangle_{V'',V} = \langle v, v' \rangle_{V,V'}.$$

Operators on Hilbert Spaces. Let now H be a Hilbert space with scalar product $(\cdot,\cdot)_H$ and the induced norm $\|\cdot\|_H$. Then the *Riesz Representation Theorem* [45, Thm. 6.1] states, that for any $y' \in H'$ there exists a unique $y \in H$ such that

$$\langle y', h \rangle_{H',H} = (h, y)_H \quad \text{for all } h \in H \quad \text{and} \quad \|y'\|_{H'} = \|y\|_H.$$

From this, one can deduce [57, Cor. 6.3.6] the existence of the *Riesz isomorphism* $j'_H \in \mathcal{L}(H, H')$ with $j'_H y = y'$, ${j'_H}^{-1} y' = y$ and $\|j'_H\|_{\mathcal{L}(H,H')} = \|{j'_H}^{-1}\|_{\mathcal{L}(H',H)} = 1$. This implies that H' is also a Hilbert space with the scalar product $(h', y')_{H'} := ({j'_H}^{-1} h, {j'_H}^{-1} y)_H$. Since every Hilbert space is reflexive we can always identify H with H'', which yields $j'_H = j_H^{-1}$, $j_{H'} = j'_H$. If $A \in \mathcal{L}(H, W)$ and $W = W''$ is a Hilbert space as well, then $A'' = A$.

Complements, Annihilators and Projections We summarize some results of [83, Ch. III.4]. A bounded linear operator $\mathcal{P}_{[]}\colon V \to V$ with $\mathcal{P}_{[]}^2 = \mathcal{P}_{[]}$ is called a *projection*. A (closed) subspace $V_s \subset V$ is called *complemented* if there exists a second subspace $V_r \subset V$ such that $V = V_s \oplus V_r$, i.e., for any $v \in V$, there are uniquely defined v_s, v_r in V_s, V_r, such that $v = v_s + v_r$. In this case one can define the operator $\mathcal{P}_{[V_s|V_r]}\colon V \to V\colon v \mapsto v_s$. The operator $\mathcal{P}_{[V_s|V_r]}$ is a projection since it is linear, bounded, and $\mathcal{P}_{[V_s|V_r]}^2 = \mathcal{P}_{[V_s|V_r]}$ as is its complement $\mathcal{P}_{[V_r|V_s]} := I - \mathcal{P}_{[V_s|V_r]}\colon v \mapsto v_r$. Conversely, if a projection $\mathcal{P}_{[]}$ is in $\mathcal{L}(V, V)$, then $V = \ker \mathcal{P}_{[]} \oplus \operatorname{im} \mathcal{P}_{[]}$. This relation is reflected in the notation $\mathcal{P}_{[V_s|V_r]} = \mathcal{P}_{[\ker \mathcal{P}_{[]} \,|\, \operatorname{im} \mathcal{P}_{[]}]}$ which we will use throughout this work.

If the image, say V_k, of a projection is of interest, rather than its kernel, we use a placeholder $\cdot$ for the kernel and simply write $\mathcal{P}_{[V_k|\cdot]}$.

Because of noncomplemented subspaces in Banach spaces [121], given a subspace, there exists not necessarily a bounded projection onto it. Only if a Banach space is isomorphic to a Hilbert space, then every subspace is complemented [80].

In a Hilbert space, for any subspace $H_s \subset H$, one has the *orthogonal complement* $H_{s\perp}$ defined as

$$H_{s\perp} := \{v \in H : (v, w) = 0, \text{ for all } w \in H_s\}.$$

With $H = H_s \oplus H_{s\perp}$, we define the *orthogonal projector* $\mathcal{P}_{[H_s]} := \mathcal{P}_{[H_s|H_{s\perp}]}$. Regardless of the choice of the complement, the projections onto H_s are equivalent in the sense, that for any H_r being a complement to H_s, one has

$$\tfrac{1}{\|\mathcal{P}_{[H_s|H_r]}\|} \|\mathcal{P}_{[H_s|H_r]}\| \leq \|\mathcal{P}_{[H_s]}\| \leq \|\mathcal{P}_{[H_s|H_r]}\|.$$

This follows from the norm of a projection being ≥ 1 [46, Thm. 7.13] and equality holding for orthogonal projections and from $\mathcal{P}_{[H_s|H_r]} = \mathcal{P}_{[H_s|H_r]} \mathcal{P}_{[H_s]}$.

The orthogonal complement is a special case of the *annihilator* ${V_s}^0 \subset V'$ that can be defined for every subspace V_s of a Banach space V via

$$ {V_s}^0 := \{v' \in V' : \langle v', v \rangle_{V,V} = 0, \text{ for all } v \in V_s\}. \tag{2.6}$$

Convergence and Compactness A sequence $\{v_n\}_{n\in\mathbb{N}}$ in a Banach space V is called *strongly* or *norm convergent*, if there exists a $v \in V$ such that $\|v_n - v\|_V \to 0$ as $n \to \infty$. Then we write $v_n \to v$. It is called *weakly convergent* if $\langle f, v_n - v \rangle \to 0$, as $n \to \infty$, for all $f \in V'$. Weak convergence of $\{v_n\}_{n\in\mathbb{N}}$ to $v \in V$ is denoted by $v_n \rightharpoonup v$. By a corollary of the *Banach-Steinhaus Theorem* every weakly convergent sequence is bounded [133, Cor. 1.4].

Since the norm is convex, every Banach space is a locally convex space for which one can define the following notion [133, p. 2]. A subset U of a Banach space V is called closed, if every *Cauchy sequence* in U has its limit in U. The *closure* of U in the norm of V, that we denote by $\overline{U}^{\|\cdot\|_V}$, is the smallest closed subset of V that contains U. We say that U is *dense* in V, if $\overline{U}^{\|\cdot\|_V} = V$. We say that V is *separable* if there exists a dense subset of V that is at most countable.

A subset U of a Banach space V is called *(sequentially) compact*, if every sequence in U contains a norm convergent subsequence. In Banach spaces the definitions of compactness and sequential compactness coincide, cf. [133, p. 7]. A subset U is called *precompact*, if the closure of U with respect to $\|\cdot\|_V$ is compact.

A subset U is called *weakly (sequentially) compact* if every bounded sequence has a weakly convergent subsequence. By the *Eberlein-Šmulian Theorem* [156], in Banach spaces, weak sequential compactness is equivalent to *weak compactness* as it can be defined for more general spaces. As an important conclusion, one has, that in reflexive Banach spaces, bounded sets are weakly compact, so that every bounded sequence has a weakly convergent subsequence, cf. [161, Thm. 21.D].

Spaces of Integrable Functions The following definitions use notions as *measurability* and *integrability* in the sense of *Lebesgue* from measure theory. See e.g. [45, Ch. II] for the basic definitions.

Let $\Omega \in \mathbb{R}^d$, $d \in \mathbb{N}$, be a domain with a *regular boundary* $\partial\Omega$ in the sense of *Calderón* [45, Def. II.1.17]. By Lemma I.1.27 in [45], this means that Ω and $\partial\Omega$ fulfill the following assumption:

Assumption 2.10. *The subset $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ is a domain, i.e. it is open, simply connected, and bounded, and there are constants R, $L > 0$ such that for all $x_0 \in \partial\Omega$ there exists a neighborhood $U(x_0)$, that is the image of*

$$U = \left\{y = [y_1, \cdots, y_d] \in \mathbb{R}^d : \sqrt{y_1^2 + \cdots + y_{d-1}^2} < R,\ |y_d| < 2LR\right\}$$

possibly after translation or rotation, under the condition that

(a) $x_0 = S_{x_0}(0)$,

(b) there exists a Lipschitz-continuous function $f_{x_0} : \mathbb{R}^{d-1} \to \mathbb{R}$ with Lipschitz-constant L, such that

$$\partial\Omega \cap U_{x_0} = \left\{[y_1, \cdots, y_{d-1}, f_{x_0}(y_1, \cdots, y_{d-1})] : \sqrt{y_1^2 + \cdots + y_{d-1}^2} < R\right\},$$

(c) and that

$$\Omega \cap U_{x_0} = \left\{y \in U_{x_0} : \sqrt{y_1^2 + \cdots + y_{d-1}^2} < R, f_{x_0}(y_1, \cdots, y_{d-1}) < y_d < 2LR\right\}.$$

Convex and bounded sets of $\mathbb{R}^d$ have regular boundaries [45, Rem. II.1.11].

By $L^p(\Omega)$, $1 \le p < \infty$, we denote the space of all measurable functions $u \in \big(\Omega \to \mathbb{R}\big)$ with $\int_\Omega |u(\omega)|^p \, \mathrm{d}\omega < \infty$. With the norm

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u(\omega)|^p \, \mathrm{d}\omega\right)^{1/p},$$

the space $L^p(\Omega)$ is a Banach space. For $l \in \mathbb{N}$ and with the norm $\|u\|_{[L^p(\Omega)]^l} = \left(\int_\Omega \left(\sum_{i=1}^l |u_i(\omega)|^2\right)^{p/2} \mathrm{d}\omega\right)^{1/p}$, $[L^p(\Omega)]^l := \{(u_1, \cdots, u_l) : u_i \in L^p(\Omega), i = 1, \cdots, l\}$ is a Banach space.

With the scalar product

$$\big(u, v\big)_{[L^2(\Omega)]^l} := \sum_{i=1}^l \int_\Omega u_i v_i \, \mathrm{d}\omega,$$

the space $[L^2(\Omega)]^l$ is a Hilbert space.

By $L^1_{\mathrm{loc}}(\Omega)$, we denote the space of *locally integrable functions*, i.e. all $u \in (\Omega \to \mathbb{R})$ with $\int_K |u(\omega)| \,\mathrm{d}\omega < \infty$ for all compact $K \subset \Omega$.

For $1 < p < \infty$ and $l \in \mathbb{N}$, the space $[L^p(\Omega)]^l$ is separable and reflexive. In the case $p = 1$, it is separable [133, Lems. 1.15, 1.16].

Lemma 2.11 (Hölder's inequality)**.** *Consider given* conjugated exponents *p and p', i.e. $\frac{1}{p} + \frac{1}{p'} = 1$. If $1 < p < \infty$, or $p = 1$ and $p' = \infty$, then for $u \in L^p(\Omega)$ and $v \in L^{p'}(\Omega)$ one has $uv \in L^1(\Omega)$ and $|\int_\Omega uv \,\mathrm{d}\omega| \le \|u\|_{L^p(\Omega)} \|v\|_{L^{p'}(\Omega)}$.*

For $1 \le p < \infty$ one can identify the dual space $\big[[L^p(\Omega)]^l\big]'$ as $[L^{p'}(\Omega)]^l$, where p' is the conjugated exponent to p.

Remark 2.12. The elements of the spaces of integrable functions, defined in this and in the following sections, are the equivalence classes of functions that take on the same values on Ω except from subsets of zero measure, see e.g. [45, Ch. II.2] for a definition. In particular, two functions u and w belong to the same equivalence class, meaning they are considered the same element of the function space, if $u = v$ almost everywhere (a.e.) on Ω.

Generalized or Weak Derivatives Let $C_0^\infty(\Omega)$ be the set of infinitely often differentiable functions with compact support in Ω. For any $u \in L^1_{\mathrm{loc}}(\Omega)$ we have that

$$\langle u, \phi \rangle := \int_\Omega u(\omega)\phi(\omega) \,\mathrm{d}\omega < \infty,$$

for all $\phi \in C_0^\infty(\Omega)$. If there exists $g \in L^1_{\mathrm{loc}}(\Omega)$ such that

$$\langle g, \phi \rangle = (-1)^{|\alpha|} \langle u, \partial^\alpha \phi \rangle \quad \text{for all } \phi \in C_0^\infty(\Omega), \tag{2.7}$$

then g is called the derivative of u in the *weak sense*, with respect to the *multiindex* $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$. We will write $g = \partial^\alpha u$. Here ∂^α is short for $\frac{\partial^{|\alpha|} u}{\partial^{\alpha_1}\omega_1 \cdots \partial^{\alpha_d}\omega_d}$ and $|\alpha| := \sum_{i=1}^d \alpha_i$, and ω_i is the i-th coordinate in $\Omega \in \mathbb{R}^d$, $i = 1, \cdots, d$.

Since $L^p(\Omega) \subset L^1_{\mathrm{loc}}(\Omega)$ for $1 \le p < \infty$, relation (2.7) defines a derivative of any $u \in L^p(\Omega)$ if it exists.

Let $\nabla u := (\partial^{e_1} u, \cdots, \partial^{e_d} u)$ denote the d-tuple of the first order derivatives, where e_i is the i-th canonical unit vector, $i = 1, \cdots, d$.

For a d-tuple $u = (u_1, \dots, u_d) \in [L^1_{\mathrm{loc}}(\Omega)]^d$, we define $\operatorname{div} u := \sum_{i=1}^d \partial^{e_i} u_i$, if this sum is in $L^1_{\mathrm{loc}}(\Omega)$.

Sobolev and Bochner Spaces For $1 \le p < \infty$ and $k \in \mathbb{N}$ define the *Sobolev space*

$$W^{k,p}(\Omega) := \{u \in L^p(\Omega) : \partial^\alpha u \in L^p(\Omega), \text{ for all } \alpha \in \mathbb{N}^d, |\alpha| \le k\}$$

which is a Banach space, if one considers the norm

$$\|u\|_{W^{k,p}(\Omega)} := \left[\int_\Omega \Big(\sum_{\alpha \in \mathbb{N}^d : |\alpha| \le k} |\partial^\alpha u(\omega)|^2 \Big)^{p/2} \,\mathrm{d}\omega \right]^{1/p}.$$

If $p = 2$, then the norm is induced by the scalar product

$$\big(u, v\big)_{W^{k,2}} := \sum_{\alpha \in \mathbb{N}^d : |\alpha| \le k} \big(\partial^\alpha u, \partial^\alpha v\big)_{L^2(\Omega)}.$$

In particular, the space $W^{k,2}(\Omega)$ is a Hilbert space.

By $W_0^{k,p}(\Omega)$, we denote the closure of $C_0^\infty(\Omega)$ in the norm of $W^{k,p}(\Omega)$ [45, Def. II.1.16]. If Ω is bounded, then

$$\|u\|_{W_0^{k,p}(\Omega)} := \left[\int_\Omega \Big(\sum_{\alpha \in \mathbb{N}^d : |\alpha| = k} |\partial^\alpha u(\omega)|^2 \Big)^{p/2} \,\mathrm{d}\omega \right]^{1/p}.$$

defines a norm that is equivalent to $\|\cdot\|_{W_0^{k,p}(\Omega)}$ on $W_0^{k,p}(\Omega)$. In the case that $p = 2$, this norm is induced by

$$\big(u, v\big)_{W^{k,2}} := \sum_{\alpha\in\mathbb{N}^d:|\alpha|=k} \big(\partial^\alpha u, \partial^\alpha v\big)_{L^2(\Omega)}.$$

Thus, $W_0^{k,2}(\Omega)$ is a Hilbert space.

The definition of vector valued Sobolev spaces $[W^{k,p}(\Omega)]^l$ and $[W_0^{k,p}(\Omega)]^l$ is done by analogy with $[L^p(\Omega)]^l$.

For $l \in N$ and $1 \leq p < \infty$, the space $[W^{k,p}(\Omega)]^l$ is separable, for $1 < p < \infty$, it is reflexive [133, Ch. 1.2.3].

For functions in $[0, T] \to V$, where V itself is an infinite-dimensional Banach space one can define the corresponding function spaces resorting to the notion of *Bochner integrability*, see [45, Ch. IV] for a thorough or, e.g., [133, Ch. 1.1.5] for a brief introduction.

For $1 \leq p < \infty$ the *Bochner space* $L^p(0, T; V)$ is the space of Bochner integrable functions $u \in ([0, T] \to V)$ satisfying $\big(\int_{[0,T]}\|u(t)\|_V^p \,\mathrm{d}t\big)^{1/p} < \infty$. The space $L^\infty(0, T; V)$ is the space of functions $v \in \big((0, T) \to V\big)$ with $v(t) \leq C < \infty$, for almost all $t \in (0, T)$. This space is complete with the norm $\|v\|_{L^\infty(0,T;V)} := \operatorname{ess\,sup}_{t\in(0,T)}\|v(t)\|_V$, defined as the infimum of the suprema of $v(t)$ taken over all subsets of $(0, T)$ that have a nonzero measure.

Theorem 2.13 ([45], Thm. IV.1.14 and Rem. IV.1.10, pp. 131)**.** *Let $1 < p < \infty$ and $T > 0$. If W is reflexive or separable, then for any $l \in \big(L^p(0, T; W)\big)'$ there exists a unique representation via*

$$l[w] = \int_0^T \langle v_l(s), w(s)\rangle_{W',W} \,\mathrm{d}s, \quad \textit{for all } w \in L^p(0, T; W) \tag{2.8}$$

with $v_l \in L^{p'}(0, T; W')$, $\frac{1}{p} + \frac{1}{p'} = 1$. The mapping $l \mapsto v$ is linear and it holds that $\|l\|_{(L^p(0,T;W))'} = \|v\|_{L^{p'}(0,T;W')}$.

Remark 2.14 ([45], Rem. IV.1.11)*.* A direct consequence is that if W is reflexive, then $L^p(0, T; W)$ is reflexive as well.

For locally Bochner integrable functions u, a candidate weak derivative is defined by analogy with the definition on page 14 for Lebesgue integrable functions, see, e.g., [40, Def. 8.1.1].

Definition 2.15 (cf. [133], Ch. 7.1)**.** A function $v \in L^1_{\text{loc}}(0, T; W)$ is the weak time derivative $\dot{u}$ of $u \in L^1_{\text{loc}}(0, T; W)$, if

$$\int_0^T v(t)\phi(t) \,\mathrm{d}t = -\int_0^T u(t)\dot{\phi}(t) \,\mathrm{d}t,$$

for all $\phi \in \mathcal{C}_0^\infty(0, T)$.

We define the Sobolev-Bochner spaces

$$\mathcal{W}^{1;p,q}(0, T; V; W) := \{v \in L^p(0, T; V) : \dot{v} \in L^q(0, T; W)\}.$$

If $p = q = 2$, we write

$$\mathcal{W}(0, T; V; W) := \{v \in L^2(0, T; V) : \dot{v} \in L^2(0, T; W)\}.$$

For a space V, we define $\mathcal{V} := L^2(0, T; V)$. Similarly we define $\mathcal{V}' = L^2(0, T; V')$, $\mathcal{H} := L^2(0, T; H)$, $\mathcal{H}_{\text{df}} := L^2(0, T; H_{\text{df}})$, $\mathcal{Q} := L^2(0, T; Q)$ and others.

We will always assume that the assumptions of Theorem 2.13 hold, so that we can write the dual product in, e.g., $\mathcal{V}' \times \mathcal{V}$ as

$$\left\langle f, v \right\rangle_{\mathcal{V}',\mathcal{V}} := \int_0^T \left\langle f(t), v(t) \right\rangle_{V',V} \,\mathrm{d}t,$$

for $f \in L^2(0,T;V')$ and $v \in L^2(0,T;V)$.

Embeddings and Gelfand Triples If there exists a continuous *embedding operator* $i\colon X \to Y$, then we call X continuously embedded in Y, we write $X \hookrightarrow Y$ and we make use of $\|x\|_Y := \|ix\|_Y \leq c_i \|x\|_X$. If i is compact, then the embedding is called compact and is denoted by $X \overset{c}{\hookrightarrow} Y$. These definitions are given, e.g., in [133, p. 9] or [40], from where we have borrowed the notation for the embeddings.

If there exists an isometric isomorphism $i\colon X \to Y$, i.e. i is invertible and $\|ix\|_Y = \|x\|_X$, we write $X \cong Y$ and, tacitly omitting i, sometimes directly identify $X = Y$.

Given a reflexive Banach space V and a Hilbert space H, where V is dense in H and $V \hookrightarrow H$. Then by [45, Rem. I.5.14] one has $H' \hookrightarrow V'$, and, having identified $H = H'$ via the Riesz isomorphism, the triple embedding $V \hookrightarrow H \hookrightarrow V'$ – often referred to as *Evolution* or *Gelfand triple*[40, p. 83].

Within the Gelfand triple, we will always omit the injection i and treat, e.g., $v \in V$ as an element of H, that is we tacitly identify V with $i(V)$ to get the algebraic inclusion $V \subset H$.

Assuming $V \subset H$, the dual product $\left\langle \cdot, \cdot \right\rangle_{V',V}$ is the continuous extension of $\left\langle \cdot, \cdot \right\rangle_{H',H} = \left(j\cdot, \cdot\right)_{H,H}$ onto $V' \times V$. In particular, if $f \in H' \subset V'$, then $\left\langle f, v \right\rangle_{V',V} = \left(jf, v\right)_{H,H}$ for all $v \in V \subset H$, where $j\colon H' \to H$ is the Riesz isomorphism [133, Ch. 7.2].

Note that the dual product in the Gelfand triple is different from the dual product in V. In particular, in the case when V is a Hilbert space, one cannot use, e.g., the Riesz isomorphism to identify V and V', see [25, Ch. 5.2] for a concrete example. In fact, if V is a Hilbert space itself, if $V \subset H$, and if V' is defined with respect to the scalar product in H, then there exists a homeomorphism from V into V' only if $V \subset H$ is a closed subset. This can be seen as follows. Let $i\colon V \to H$ be the embedding operator and thus bounded and injective. Then $i'\colon H' \to V'$ is the embedding operator of H' into V'. Thus, $i'j'i\colon V \to V'$ is bounded and injective, where j is the *Riesz isomorphism* in H. Assume there exists an homeomorphism $h\colon V \to V'$. Then $i'j'i$ must be surjective, as one can use the injections to identify $V \cong i'j'i(V) \subset V' \overset{h^{-1}}{=} V$. Then, $i'\colon H' \to V'$ must be surjective, which by the *Closed Range Theorem* [83, Thm. IV.5.13] can only be the case, if the range of $i\colon V \to H$ is closed.

Recall that we consider a domain $\Omega \in \mathbb{R}^d$ with a regular boundary $\partial\Omega$ fulfilling Assumption 2.10.

The *Sobolev Embedding Theorems*, see, e.g., [45, Thm. II.1.2], state that for $1 \leq p < \infty$,

$$[W^{p,k}(\Omega)]^d \hookrightarrow [W^{q,l}(\Omega)]^d,$$

if $0 \leq k < l$ and $\frac{1}{p} - \frac{k-l}{d} \leq \frac{1}{q} < 1$.

If $1 \leq p < \infty$ and $k \geq 1$, then $W^{k,p}(\Omega) \hookrightarrow W^{k-1,p}(\Omega)$ is compact, i.e. $W^{k,p}(\Omega) \overset{c}{\hookrightarrow} W^{k-1,p}(\Omega)$ [45, Lem. II.1.28].

Since C_0^∞ is dense in $W_0^{k,p}(\Omega)$ by definition and dense in $L^q(\Omega)$ by [25, Cor. 4.23], we have that $W_0^{1,p}(\Omega)$ is dense in $L^q(\Omega)$, $k \geq 1$, $1 \leq p, q < \infty$. Thus, we can define the Gelfand triple

$$W_0^{1,2}(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow W^{-1,2}(\Omega) := (W_0^{1,2}(\Omega))',$$

where the embeddings are even compact.

For Bochner spaces we have the embedding $\mathcal{W}^{1;p,p'}(0,T;V,V') \hookrightarrow \mathcal{C}([0,T],H)$, for $1 \leq p, p' \leq \infty$ such that $\frac{1}{p}+\frac{1}{p'} = 1$, [133, Lem. 7.3], provided that $V \hookrightarrow H \hookrightarrow V'$ is a Gelfand triple. We will also make use of the *Aubin-Lions Lemma*:

Lemma 2.16 ([133] Lem. 7.7)**.** *If* V_1*,* V_2*,* V_3 *are Banach spaces and* $V_1 \overset{c}{\hookrightarrow} V_2 \hookrightarrow V_3$*,* $1 < p < \infty$*, and* $1 \leq q \leq \infty$*, then*

$$\mathcal{W}^{1;p,q}(0,T;V_1,V_3) \overset{c}{\hookrightarrow} L^p(0,T;V_2).$$

Operators on Spaces of Abstract Functions To treat abstract formulations, the following definitions and results, adapted from [133, Ch. 1.3], are needed. Let $A : \Omega \times V \to W$. Then A is a *Carathéodory* mapping if

$$\begin{aligned} A(\omega,\cdot)\colon V \to W \quad &\text{is continuous for } \omega \text{, a.e. in } \Omega \text{ and} \\ A(\cdot,v)\colon \Omega \to W \quad &\text{is measurable for all } v. \end{aligned}$$

If, in addition, $\|A(\omega,v)\|_W \leq \gamma(\omega) + c\|v\|_V^{p/p_0}$ for some $\gamma \in L^{p_0}(\Omega;\mathbb{R})$, then the *Nemyckij* map $\mathcal{N}_A \in (\Omega \to W)$ of a function $u \in L^p(\Omega;V)$, defined as

$$[\mathcal{N}_A(u)](x) = A(x,u(x)),$$

is in $L^{p_0}(\Omega;W)$, where $1 \leq p < \infty$ and $1 \leq p_0 \leq \infty$.

The same arguments are valid if one replaces Ω by $(0,T) \subset \mathbb{R}$. In particular, if $A\colon V \to W$ is linear and bounded, then $[\mathcal{N}_A(v)](\cdot) := A(v(\cdot))$ is in $L^p(0,T;W)$ for $v \in L^p(0,T;V)$. We will not distinguish notationally between $\mathcal{N}_A\colon L^p(0,T;V) \to L^{p_0}(0,T;W)$, as a mapping between abstract functions, and $A\colon V \to W$.

Nonlinear Operators and Fréchet Derivative

For an operator $A\colon V \to W\colon v \mapsto A(v)$ one defines the following properties [162, Def. 27.14]: Let $\{v_n\}_{n\in\mathbb{N}} \subset V$ be arbitrary. Then we say that A is

(a) *continuous*, if it holds that if $v_n \to v$, then $A(v_n) \to A(v)$,
(b) *strongly continuous*, if it holds that if $v_n \rightharpoonup v$, then $A(v_n) \to A(v)$,
(c) *weakly continuous*, if it holds that if $v_n \rightharpoonup v$, then $A(v_n) \rightharpoonup A(v)$,
(d) *demicontinuous*, if it holds that if $v_n \to v$, then $A(v_n) \rightharpoonup A(v)$,
(e) or, in the case that $W = V'$, *hemicontinuous*, if $t \mapsto, \langle A(u+tv),w\rangle_{V',V}$ is continuous on $[0,T]$ for all u, v, $w \in V$,

as $n \to \infty$,

Note that what is defined as *strongly continuous* here, is often referred to as *completely* or *totally continuous*.

Let V and W be Banach spaces. A mapping $A\colon V \to W$ is called *compact*, if it maps bounded sets into *precompact* sets [133, p. 7].

One has that if A is strongly continuous, then A is compact. The opposite direction holds if A is also linear, cf. [161, Prop. 26.2]. We call $A\colon V \to W$ *bounded* if it maps bounded sets in V into bounded sets in W [133, p. 5].

We now consider a reflexive and separable Banach space V and $A\colon V \to V'$. We call A *monotone* if for any v, $u \in V$, it holds that $\langle A(u) - A(v), u - v\rangle_{V',V} \geq 0$.

In view of semi-linear parabolic problems, cf. [133, Ch. 8.6], we introduce the notion of *pseudomonotonicity*:

Definition 2.17. [133, Def. 2.1] The operator $A\colon V \to V'$ is called *pseudomonotone* if A is bounded and if for given $\{u_k\}_{k\in\mathbb{N}} \subset V$, $u_k \rightharpoonup u \in V$,

$$\limsup_{k\to\infty} \langle A(u_k), u_k - u\rangle_{V',V} \geq 0$$

implies that for any $v \in V$,

$$\langle A(u), u - v\rangle_{V',V} \leq \liminf_{k\to\infty} \langle A(u_k), u_k - v\rangle_{V',V}.$$

The prototype example of a pseudomonotone operator is the sum of a monotone and hemicontinuous operator and a strongly continuous operator [162, p. 581].

For the sections on optimal control we will use the Fréchet derivatives.

Definition 2.18 ([159], Ch. 40.1)**.** An operator $A\colon V \to W$ is called *Fréchet differentiable* at $v_0 \in V$, if there exists an open subset $V_0 \subset V$ and a linear bounded Operator $D(A, v_0)\colon V_0 \to W$ and if for any direction $h \in V_0$ the limit

$$\lim_{\|h\|_V \to 0} \frac{A(v_0 + h) - A(v_0) - D(A, v_0)h}{\|h\|_V}$$

exists.

If it exists, we denote the *Fréchet derivative* of A at $v_0 \in V$ as $A_{;v}(v_0) := D(A, v_0)$. Similarly we denote partial derivatives for $A\colon V \times U \to W$ as for example $A_{;u}(v_0, u_0)\colon U \to W$.

By definition, the Fréchet derivative of a linear map is the map itself, i.e. $A_{;v}(v)[d_v] = Ad_v$.

We will use the *chain rule* for differentiation, see, e.g. [149, Thm. 2.20]. If $A\colon V \to W$ is Fréchet differentiable at v and $B\colon W \to U$ is is Fréchet differentiable at $A(v)$, then

$$(B \circ A(v))_{;v}[d_v] = B_{;w}(A(v))\big[A_{;v}(v)[d_v]\big].$$

As in the above examples, in ambiguous cases like $A(v)[d_v]$, we will use the square brackets in to point out that the action of an operator $A(v)$ onto d_v is linear.

3. Decoupling of Semi-linear Semi-explicit Index-2 ADAEs

If put into an abstract setting, i.e. considering v and p as functions of $t \in (0,T)$ taking on values in a function space, Equation (1.1) is a differential-algebraic equation (DAE), since v has to fulfill both a differential (1.1a) and an algebraic relation (1.1b). The interaction of such differential and algebraic equations is quantified by various index concepts, cf. Section 2.1. Some of these concepts have been generalized to abstract DAEs (ADAEs), see [115] for an overview. In our setting we speak of index-2 systems, as stable and conforming spatial discretizations of the considered ADAEs lead to DAEs of *tractability index* 2.

In finite dimensions, there are several ways to reduce the index by constructing equivalent systems of lower index [90]. Apart from being applied for numerical solution schemes, this – possibly partial – decoupling of the algebraic and the differential parts of the equation is used to investigate solvability. From the separated algebraic part one can read off consistency conditions, e.g. for the initial values. For the differential part one can use standard theory for abstract ODEs to establish solvability conditions. Again, some approaches have been generalized to the abstract setting [6, 101, 131, 145].

In this section we address the question, when an optimal control problem, constrained by a class of semi-explicit semi-linear ADAEs, is solvable. The "semis" mean that, as it is the case in (1.1), the time derivative appears explicitly and linearly. As the considered optimality system contains the constraints, this calls for results on solvability of the ADAE.

The considered abstract setting contains weak formulations of the Navier-Stokes Equation for which existence of solutions has been investigated since long, see [100, 143] for the fundamental notions and results. We will investigate these abstract DAEs from a DAE perspective that is well understood in finite dimensions [102]. A DAE point of view has been taken in [41] to analyse linearized Navier-Stokes Equation. Our approach differs from [41] and the classical works in so far as it does not eliminate the constraints by restricting the analysis to the subspace of an inherent ordinary differential equation (ODE) for the differential parts of the solution. The focus on the formulation in the native variables v and p, often referred to as saddle point formulation, makes our work comparable to [54], where a space-time variational formulation of the Navier-Stokes Equation has been investigated.

We introduce a decoupling of the ADAE that exploits the saddle point structure in a similar way as in the finite-dimensional setup, see Section 8.1. The presented generalization bases on additional regularity, such that the abstract equations are posed in a Hilbert space, which is densely embedded in the Banach space of the problem formulation. This additionally assumed regularity enables the splitting of the equations by projections. We will give reasoning, why this assumption is not very restrictive for the Navier-Stokes Equation.

3.1. **Semi-explicit Semi-linear ADAEs of Index 2.**

We investigate a class of abstract differential-algebraic equations, where the differential variable v takes on values in a Banach space V and the algebraic variable p in a Hilbert space Q_H. We assume that V is densely and continuously embedded in a Hilbert space H and consider the dynamical equation posed on the Gelfand triple $V \hookrightarrow H \hookrightarrow V'$. This setup, where the dynamical equation is posed on a Gelfand triple, we will refer to as *evolution setting*, which is in line with the notion for abstract differential equations without algebraic constraints [40, 106, 133].

We refer to the equations as index-2 equations, since a stable semi-discretization leads to a tractability index $i_\mu = 2$ of the resulting finite-dimensional DAE, cf. Section 2.1.

Problem 3.1. *Let V be a separable and reflexive Banach space, densely and continuously embedded in a Hilbert space H. Consider the Gelfand triple $V \hookrightarrow H \hookrightarrow V'$ and another Hilbert space Q_H. For given $f \in L^2(0,T;V')$, $g \in L^2(0,T;Q_H')$ and $\alpha \in H$ find $v \in L^2(0,T;V)$ and $p \in L^2(0,T;Q_H)$ that fulfill*

$$\dot{v}(t) - A(t,v(t)) - J_1'p(t) = f(t) \quad \text{in } V', \text{ a.e. in } (0,T), \tag{3.1a}$$

$$-J_2v(t) = g(t) \quad \text{in } Q_H', \text{ a.e. in } (0,T), \tag{3.1b}$$

$$v(0) = \alpha \quad \text{in } H, \tag{3.1c}$$

with an operator $A(t,\cdot)\colon V \to V'$ and bounded linear operators $J_1'\colon Q_H \to V'$ and $J_2\colon V \to Q_H'$.

We will frequently consider the symmetric version, where $J_1 = J_2$:

Problem 3.1(Sym). *Let the spaces and operators be as in Problem 3.1. For given $f \in L^2(0,T;V')$, $g \in L^2(0,T;Q_H')$ and $\alpha \in H$ find $v \in L^2(0,T;V)$ and $p \in L^2(0,T;Q_H)$ that fulfill*

$$\dot{v}(t) - A(t,v(t)) - J_2'p(t) = f(t) \quad \text{in } V', \text{ a.e. in } (0,T), \tag{3.2a}$$

$$-J_2v(t) = g(t) \quad \text{in } Q_H', \text{ a.e. in } (0,T), \tag{3.2b}$$

$$v(0) = \alpha \quad \text{in } H. \tag{3.2c}$$

A special case of Problem 3.1(Sym) is given by a weak formulation of the incompressible Navier-Stokes Equation:

Problem 3.1(NSE). *Let Ω be a regular domain in $\mathbb{R}^d$, $d \in \{2,3\}$ and consider the Gelfand triple*

$$V := [W_0^{1,2}(\Omega)]^d \hookrightarrow [L^2(\Omega)]^d \hookrightarrow [W^{-1,2}(\Omega)]^d =: V'$$

and the factor space $Q_H := L^2(\Omega)/\mathbb{R}$. Let $\operatorname{div} = -\nabla'\colon [W_0^{1,2}(\Omega)]^d \to (L^2(\Omega)/\mathbb{R})'$ be defined via

$$-\left\langle \nabla' v, q \right\rangle_{Q_H',Q_H} = \sum_{i=1}^{d} \int_\Omega (\partial^{e_i} v_i) q \, \mathrm{d}\omega \tag{3.3}$$

and $A\colon [W_0^{1,2}(\Omega)]^d \to [W^{-1,2}(\Omega)]^d$ via

$$\left\langle A(v), w \right\rangle_{V',V} = \left\langle (v \otimes \nabla)v, w \right\rangle_{V',V} + \sum_{i=1}^{d} \left(\nabla v_i, \nabla w_i\right)_{[L^2(\Omega)]^d}, \tag{3.4}$$

where

$$\left\langle (u \otimes \nabla)v, w \right\rangle_{V',V} := \sum_{i=1}^{d}\sum_{j=1}^{d} \int_\Omega u_i (\partial^{e_i} v_j) w_j \, \mathrm{d}\omega, \tag{3.5}$$

cf. [143, Ch. II.1.2]. For $f \in L^2(0,T;[W^{-1,2}(\Omega)]^d$ and $\alpha \in [L^2(\Omega)]^d$, find $v \in L^2(0,T;W_0^{1,2}(\Omega))$ and $p \in L^2(0,T;L^2(\Omega)/\mathbb{R})$ such that

$$\dot{v}(t) - A(v(t)) - \nabla p(t) = f(t) \quad \text{in } [W^{-1,2}(\Omega)]^d, \text{ a.e. in } (0,T), \tag{3.6a}$$

$$\operatorname{div} v(t) = 0 \quad \text{in } (L^2(\Omega)\mathbb{R})', \text{ a.e. in } (0,T), \tag{3.6b}$$

$$v(0) = \alpha \quad \text{in } [L^2(\Omega)]^d. \tag{3.6c}$$

Most results will be derived for the general case of Problem 3.1. For the symmetric case, some assumptions can be relaxed. We will prove convergence of Galerkin schemes only for the symmetric problem. Applicability of the results will be demonstrated by confirming that for the Navier-Stokes Equation as in Problem 3.1(NSE) the taken assumptions are valid.

Equating the left and right hand sides in (3.1a,b) *a.e.* on the time interval refers to equality in $L^1_{\mathrm{loc}}(0,T)$ pointwise in the corresponding dual product, cf. the *Fundamental Lemma of Variational Calculus* as given, e.g., in [40, Thm. 8.1.3].

With the time derivative understood in the weak sense, we say that v and p fulfill (3.1a) if for all $w \in V$ and for all $\phi \in C_0^\infty(0,T)$,

$$\int_0^T -\dot{\phi}(t)\big(v(t),w\big)_H - \phi(t)\langle A(t,v(t)),w\rangle_{V',V} \tag{3.7a}$$

$$-\phi(t)\langle J_1'p(t),w\rangle_{V',V}\,\mathrm{d}t = \int_0^T \phi(t)\langle f(t),w\rangle_{V',V}\,\mathrm{d}t$$

As for algebraic constraints, we say that, e.g., (3.1b) holds if, for all $q \in Q$ and for all $\phi \in \mathcal{C}_0^\infty(0,T)$,

$$\int_0^T \phi(t)\big\langle -J_2v(t) - g(t), q\big\rangle_{Q_H',Q_H}\,\mathrm{d}t = 0. \tag{3.7b}$$

We will also frequently write, e.g., $J_2v = g$ in $\mathcal{Q}_H'$ or $\dot{v} = f$ in $\mathcal{V}'$, what is short for, e.g., (3.1b) meaning (3.7b).

We will consider solutions v with $\dot{v} \in L^2(0,T;V')$ or $\dot{v} \in L^q(0,T;H')$, with $q \geq 1$. Then, the initial value is well-defined in the norm of H, since $\mathcal{W}^{1;2,2}(0,T;V;V')$ and $\mathcal{W}^{1;p,q}(0,T;V;H')$ are continuously embedded in $C([0,T];H)$, for $p,q \geq 1$, see [133, Lems. 7.1, 7.3].

Remark 3.2. Because of the continuous embedding $V \hookrightarrow H \hookrightarrow V'$ a given function $v \in L^2(0,T;V)$ is also in $L^2(0,T;V')$. Since $(0,T)$ is bounded, we have $L^2(0,T;V') \subset L^1(0,T;V')$. Thus, by [40, Thm. 8.1.5], the function $\dot{v} \in L^2(0,T;V')$ is the weak derivative of v as defined in Definition 2.15 if, and only if, for all $w \in V''$, the function $t \mapsto \big\langle \dot{v}(t),w\big\rangle_{V',V}$ is the weak derivative of the real valued function $t \mapsto \big\langle v(t),w\big\rangle_{V',V''}$. Since V is assumed reflexive, i.e. $V'' = V$, and $\big\langle \cdot,\cdot\big\rangle_{V',V}$ is the continuous extension of $\big(\cdot,\cdot\big)_H$, in the considered setup, the definition of the weak derivative in (3.7bb) is equivalent to the formal definition of Definition 2.15.

We will often omit the time dependencies and also make use of a block operator formulation [41] and write (3.1a-b) as

$$\begin{bmatrix} I_{V'\leftarrow V'} & 0 \\ 0 & 0 \end{bmatrix} \frac{d}{dt}\begin{bmatrix} v \\ p \end{bmatrix} - \begin{bmatrix} A(\cdot) & J_1' \\ J_2 & 0 \end{bmatrix}\begin{bmatrix} v \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \quad \text{in } \mathcal{V}' \times \mathcal{Q_H}', \tag{3.8a}$$

$$v(0) = \alpha \quad \text{in } H. \tag{3.8b}$$

3.2. **Decoupling of the Equations.** In this section we present a decoupling of the abstract semi-explicit index-2 DAE (3.1) into the algebraic and the differential part using the approach of the *operator chain*, see [102] and [51] for the finite-dimensional formulations and [101] for abstract DAEs.

Via the decoupling we will establish necessary conditions for existence and uniqueness of solutions also in view of optimal control formulations.

Unlike in a Hilbert space, in general, a closed subspace of a Banach space does not induce a decomposition the Banach space, cf. the survey paper [136] or [121] which addresses this issue for function spaces. Equivalently, given a subspace, there is no guarantee for the existence of a bounded projection that maps onto this subspace. Therefore, we will assume that the differential equation is posed in the Hilbert space H', where one always can define bounded projections via the orthogonal complement.

In a PDE setting, the requirement that the equations are posed in H' rather than in V' implies a higher regularity of the solutions. We will comment on this with respect to the Navier-Stokes Equation in Section 3.5.

In the evolution setting, the solution space is not isomorphic to or the same as the space where the equations are posed in. From this argument and from the fact that the splitting projection for the equations may not be bounded in the Banach space of the solution, we confer that the solution must be split in a separate way. To get a well defined separation of the solution space, we will assume that the kernel of J_2 splits the solution space. For the Navier-Stokes Equation with states typically located in a Hilbert space, this assumption is obsolete.

We will split the solutions in order to investigate their existence. In practise, however, splitting up a solution space may be infeasible or unstable, cf. [7]. One should rather split or transform the equations into parts that define the solution components. This is the motivation of the twofold approach of moving the equations to a Hilbert space, where the inner product enables the explicit definition of projections, while sticking to a technical decomposition of the solutions in a Banach space.

To obtain a decoupling one has to ensure that the split equations enable the computation of the solution components. In the evolution setting, one has to take into account that the solution space is only dense and continuously embedded in the space where the equations are posed in.

Therefore, we will use assumptions on the linear operators J_1 and J_2 – accounting for the differential-algebraic coupling – that formalize a particular property of differential operators. Namely, differentiation acts decremental with respect to the degree of smoothness of a function. For example, the divergence operator is bounded if defined as $-\nabla' := \operatorname{div}\colon [W_0^{1,2}(\Omega)]^d \to L^2(\Omega)$ as it is bounded if defined as $-\overline{\nabla}' := \operatorname{div}\colon [L^2(\Omega)]^d \to W^{-1,2}(\Omega)$. Then, if defined in this way, we have that $\nabla'([W_0^{1,2}(\Omega)]^d) \hookrightarrow \overline{\nabla}'([L^2(\Omega))$.

For general operators, we need to explicitly assume these properties to state the following proposition.

Proposition 3.3. *Consider Problem 3.1. Assume there is a separable and reflexive Banach space Q, densely and continuously embedded in Q_H, such that J_1, $J_2\colon V \subset H \to Q'$ are bounded, i.e. there is a constant c such that $\|J_i v\|_{Q'} \leq c\|v\|_H$ for all $v \in V$, $i = 1, 2$. Then we can define unique extensions $\bar{J}_2$, $\bar{J}_1 \in \mathcal{L}(H, Q')$, via $\bar{J}_2 v = J_2 v \in Q'$ and $\bar{J}_1 v = J_1 v \in Q'$ for all $v \in V \subset H$.*

Proof. We define $\bar{J}_1$, $\bar{J}_2\colon H \to Q'$ as the closures of the given J_1, $J_2 \in \mathcal{L}(V, Q'_H)$, cf. [83, P. 166]. Let $i \in \{1, 2\}$ and $J_i \in \mathcal{L}(V \subset H, Q'_H)$. Since $V \hookrightarrow H$, for any $v \in H$, there is a sequence $\{v_n\}_{n=1}^\infty \subset V$ that converges to v in the norm of H. Since $Q'_H \hookrightarrow Q'$ and $J_i : H \subset V \to Q'$ is bounded, the sequence $\{J_i v_n\}_{n=1}^\infty$ converges to a $q' \in Q'$. Since q' does not depend on the particular choice of $\{v_n\}_{n=1}^\infty$, the relation $q' = \bar{J}_i v := \lim_{n\to\infty} J v_n$ well defines the extension of $J_i \in \mathcal{L}(V, Q'_H)$ to $\bar{J}_i \in \mathcal{L}(H, Q')$. □

The following lemma links the duals of the extensions to the dual operators.

Lemma 3.4. *Let $V \hookrightarrow H$ and $Q \hookrightarrow Q_H$. If $\bar{J}_1\colon H \to Q'$ is the closure of $J_1\colon V \subset H \to Q'_H \subset Q'$, then $J'_1\colon Q_H \to V'$ is the closure of $\bar{J}_1{}'\colon Q \subset Q_H \to H' \subset V'$. The same holds for $\bar{J}_2$.*

Proof. Since for any $q \in Q_H$, there is a sequence $\{q_n\}_{n=1}^\infty \subset Q$ converging to q in the norm of Q_H, we can define the closure of $\bar{J}_1{}'$ via its action on $v \in V$: $\langle \bar{J}_1{}' q, v\rangle_{V',V} = \lim_{n\to\infty}(\bar{J}_1{}' q_n, v)_{H',H}$. Using that $\bar{J}_1 v = J_1 v$, for all $v \in V$, and that the dual products are the continuous extensions of the scalar products, we derive that

$$\lim_{n\to\infty}(\bar{J}_1{}' q_n, v)_{H',H} = \lim_{n\to\infty}(q_n, \bar{J}_1 v)_{Q_H,Q'_H} = (q, J_1 v)_{Q_H,Q'_H} = \langle J'_1 q, v\rangle_{V',V}.$$

□

The space Q defines a Gelfand triple $Q \hookrightarrow Q_H \cong Q'_H \hookrightarrow Q$ that is in line with $V \hookrightarrow H \hookrightarrow V'$ with respect to J_1' and J_2, as illustrated in Figure 1.

$$\begin{tikzcd}
V \arrow[r, phantom, "\hookrightarrow"] \arrow[drr, dotted, "J_2"'] & H \arrow[r, phantom, "\cong"] \arrow[drr, "\bar{J}_2"] & H' \arrow[r, phantom, "\hookrightarrow"] & V' \\
Q \arrow[r, phantom, "\hookrightarrow"] \arrow[urr, "{\bar{J}_1}'"] & Q_H \arrow[r, phantom, "\cong"] \arrow[urr, dotted, "J_1'"'] & Q'_H \arrow[r, phantom, "\hookrightarrow"] & Q'
\end{tikzcd}$$

FIGURE 1. Illustration of the arrangements of the considered function spaces and the actions of the operators J_1' and J_2.

The scheme of Figure 1 can be interpreted also in a PDE sense. Namely, additional regularity of a function is preserved under the action of the (differential) operators J_1 and J_2: If $q \in Q_H$ has additional regularity, then $J_1'q$ has more regularity than a general function in V'. Conversely, for $v \in H$, $J_2 v$ has to be defined in a space larger than Q'_H.

Note that boundedness of $J_1,\ J_2 \colon V \subset H \to Q'$ must be established for the particular choice of Q.

For the considered Problem 3.1 we assume this boundedness of J_1, $J_2 \colon V \subset H \to Q'$ and, in view of decoupling, a complementing property of image and kernel of ${\bar{J}_1}'$ and $\bar{J}_2$:

Assumption 3.5. *Consider $J_1,\ J_2 \colon V \to Q'_H$ from Problem 3.1. Assume that there is a Banach space $Q \subset Q_H$ densely and continuously embedded, such that J_1, $J_2 \colon V \subset H \to Q'$ are bounded, so that we can define the extensions $\bar{J}_1,\ \bar{J}_2 \colon H \to Q'$. We assume that:*

(a) ${\bar{J}_1}',\ {\bar{J}_2}' \in \mathcal{L}(Q, H')$ are homeomorphisms onto their range and

(b) $H = \ker \bar{J}_2 \oplus j(\operatorname{im} {\bar{J}_1}')$,

where $j \colon H' \to H$ is the Riesz isomorphism.

The following definitions and relations will be helpful for the decoupling of the abstract DAE (3.8).

Lemma 3.6. *Let H be a Hilbert space and Q a Banach space. Let $\bar{J}_1,\ \bar{J}_2 \colon H \to Q'$ fulfill Assumption 3.5, i.e. their duals are homeomorphisms onto their range and $H = j(\operatorname{im} {\bar{J}_1}') \oplus \ker \bar{J}_2$. Then*

$$S := \bar{J}_2 j {\bar{J}_1}' \colon Q \to Q'$$

is invertible and with

$$L := {\bar{J}_1}' S^{-1} \bar{J}_2 \colon H \to H', \tag{3.9}$$

one has

$$H_{\mathrm{df}} := \ker \bar{J}_2 = \operatorname{im}[I_H - jL], \tag{3.10a}$$

$$H_{\mathrm{c}} := \operatorname{im} jL = \operatorname{im} j{\bar{J}_1}', \tag{3.10b}$$

$$H = H_{\mathrm{df}} \oplus H_{\mathrm{c}}, \tag{3.10c}$$

$$H'_{\mathrm{c}} := \operatorname{im} Lj = j'(H_{\mathrm{c}}), \tag{3.10d}$$

$$H'_{\mathrm{df}} := \operatorname{im}[I_{H'} - Lj] = j'(H_{\mathrm{df}}), \tag{3.10e}$$

$$H' = H'_{\mathrm{df}} \oplus H'_{\mathrm{c}}. \tag{3.10f}$$

Proof. Since ${\bar{J}_2}'$ is an homeomorphism onto its range, it has a closed range and the *Closed Range Theorem*, see e.g. [83, Thm. IV.5.13] applies. In particular $\bar{J}_2$ is surjective, since it holds that $\operatorname{im} \bar{J}_2 = (\ker {\bar{J}_2}')^0 = Q'$ and since ${\bar{J}_2}'$ is injective. Consider the *factor space* $H/\ker \bar{J}_2$, which is the space of equivalence classes $[h]$, defined via: h_1, h_2 belong to a class $[h]$ if $h_1 - h_2 \in \ker \bar{J}_2$. Consider $\tilde{\bar{J}}_2 \colon H/\ker \bar{J}_2 \to Q'$, defined as $\tilde{\bar{J}}_2[h] = \bar{J}_2 h_0$, for h_0 being a member of $[h]$. Since for h_1, $h_2 \in [h]$ we have that $\bar{J}_2 h_1 - \bar{J}_2 h_2 = \bar{J}_2(h_1 - h_2) = 0$, so that this map is well defined. By definition, this map is injective, by surjectivity of $\bar{J}_2$ it is also surjective, so that, by the *Open Mapping Theorem* [67, Thms. 39.2/4], it has a bounded inverse $\tilde{\bar{J}}_2^{-1} \colon Q \to H/\ker \bar{J}_2$, with $\tilde{\bar{J}}_2^{-1} \tilde{\bar{J}}_2[h] = [h]$ for all $[h] \in H/\ker \bar{J}_2$.

Consider the linear map $i_1 \colon j(\operatorname{im} {\bar{J}_1}') \to H/\ker \bar{J}_2 \colon h \mapsto [h]$, which is bounded since $\|i_1(h)\|_{H/\ker \bar{J}_2} = \min_{c \in \ker \bar{J}_2} \|h + c\|_H \leq \|h\|_H$. Since $H = \ker \bar{J}_2 \oplus j(\operatorname{im} {\bar{J}_1}')$, we obtain that i_1 is injective and surjective: take any $[h] \in H/\ker \bar{J}_2$ and any representative $h_0 \in [h]$. Then $h_0 = h_1 + h_2$ with unique $h_1 \in j(\operatorname{im} {\bar{J}_1}')$ and $h_2 \in \ker \bar{J}_2$. Thus $[h]$ is the unique image of $h_1 \in j(\operatorname{im} {\bar{J}_1}')$ under i_1. Thus, by the *Open Mapping Theorem* [67, Thms. 39.2/4], i_1 has a bounded inverse.

Since for all $h \in j(\operatorname{im} {\bar{J}_1}')$, $\tilde{\bar{J}}_2 i_1 h = \bar{J}_2 h$ and $\tilde{\bar{J}}_2 i_1$ is invertible, we find that $\bar{J}_2 \colon j(\operatorname{im} {\bar{J}_1}') \to Q'$ has a bounded inverse. Then, with j being invertible and J_1' having a left inverse by assumption, we conclude that $S := \bar{J}_2 j {\bar{J}_1}'$ is invertible.

Thus, we can define $L := {\bar{J}_1}' S^{-1} \bar{J}_2 \colon H \to H'$ and prove the assertions in (3.10):

(a) If $v \in \ker \bar{J}_2$, then $v = [I - jL]v \in \operatorname{im}[I - jL]$. Conversely, if $v \in \operatorname{im}[I - jL]$, then $\bar{J}_2 v = 0$.
(b) $\operatorname{im} jL = \operatorname{im} j {\bar{J}_1}' S^{-1} \bar{J}_2 = \operatorname{im} j {\bar{J}_1}'$ since, S^{-1} and $\bar{J}_2$ are surjective.
(c) Follows by assumption and from the definition of the spaces H_{df} and H_{c}.
(d) Since j is bijective, we find that $j'(H_{\mathrm{c}}) \overset{\text{b.)}}{=} \operatorname{im}(j'jL) = \operatorname{im} L = \operatorname{im}(Lj)$.
(e) Follows with the arguments of d.) and $j'j = I$.
(f) Since $(Lj)^2 = Lj$, $Lj \colon H' \to H'$ is a projector and H' decomposes into $\ker Lj$ and $\operatorname{im} Lj$.

□

The following corollary of Lemma 3.6 will be used to rephrase Assumption 3.5 in the discrete setting.

Corollary 3.7. *Consider the setup of Assumption 3.5. If part (a) holds, then part (b) holds if, and only if, there exists a constant* γ*, such that* $\|\bar{J}_2 h\|_{Q'} \geq \gamma \|h\|_H$*, for all* $h \in j(\operatorname{im} {\bar{J}_1}')$.

Proof. In the proof of Lemma 3.6 we have established that under the conditions of Assumption 3.5 the operator $J_2 \colon j(\operatorname{im} {\bar{J}_1}') \to Q'$ has a bounded inverse, which is equivalent to $\|\bar{J}_2 h\|_{Q'} \geq \gamma \|h\|_H$, for all $h \in j(\operatorname{im} {\bar{J}_1}')$, cf. [3, Thm. 2.5]. Conversely, if $\bar{J}_2$ on $j(\operatorname{im} {\bar{J}_1}')$ has a bounded right inverse ${\bar{J}_2}^-$, then $\ker \bar{J}_2$ and $j(\operatorname{im} {\bar{J}_1}')$ split the space H, since ${\bar{J}_2}^- \bar{J}_2$ is a bounded projector having $\ker \bar{J}_2$ as its kernel and $j(\operatorname{im} {\bar{J}_1}')$ as its image. □

We assume additional regularity of the problem:

Assumption 3.8. *Consider Problem 3.1, assume that Assumption 3.5 holds, and assume that* g *and* α *are sufficiently smooth. For more regular data* $f \in L^2(0, T; H')$ *(rather than in* $L^2(0, T; V')$*) any corresponding solution* (v, p) *of Problem 3.1 fulfills* $p(t) \in Q$ *and* $A(t, v(t)) \in H'$*, for almost all* $t \in (0, T)$.

Remark 3.9. As a consequence of Assumption 3.8 one has $J_1'p(t) \in H'$ and, thus, $\dot{v}(t) \in \mathcal{H}'$, for almost all $t \in (0,T)$. This means that the solution part v must lie in $\mathcal{W}(0,T;V;H')$. For the considered abstract setup, we will specify the necessary and sufficient smoothness of g in space and time in Lemma 3.28. The necessary smoothness of α in space depends, in particular, on the operator A. We will give sufficient conditions for α when considering the Navier-Stokes Equation in Section 3.5.

We can now decouple the equations (3.8a) into a system of four equations in H', V', Q_H', and Q', respectively.

Lemma 3.10. *Consider the setup of Problem 3.1 and assume that Assumption 3.5 holds, i.e. there is $Q \hookrightarrow Q_H$ such that the extensions $\bar{J}_1$, $\bar{J}_2 \colon H \to Q'$ of J_1, J_2 are bounded. Assume that Assumption 3.8 holds, i.e. for $f \in L^2(0,T;H')$, the equation (3.1a) is posed in H' rather than in V'.*

If $J_1' \colon Q_H \to V'$ is injective and if $f \in L^2(0,T;H')$, then system (3.8a) *is equivalent to*

$$\begin{bmatrix} \mathcal{P}_{[H_{\mathrm{df}}'|H_{\mathrm{c}}']}\dot{v} \\ 0 \\ \bar{J}_2 j\dot{v} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathcal{P}_{[H_{\mathrm{df}}'|H_{\mathrm{c}}']}A(\cdot) & 0 \\ J_1' j_{Q_H} J_2 & 0 \\ \bar{J}_2 j A(\cdot) & S \\ J_2 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} = \mathcal{E}_2^{-1} \begin{bmatrix} f \\ g \end{bmatrix}, \tag{3.11}$$

where $j_{Q_H} \colon Q_H' \to Q_H$ and $j \colon H' \to H$ are the Riesz isomorphisms, H_{c}', H_{df}', and $S = \bar{J}_2 j \bar{J}_1{}'$ are as defined in Lemma 3.6, $\mathcal{P}_{[H_{\mathrm{df}}'|H_{\mathrm{c}}']} := I_{H'} - \bar{J}_1 S^{-1} \bar{J}_2 j \colon H' \to H'$ is the projector that realizes part (3.10f), and

$$\mathcal{E}_2^{-1} := \begin{bmatrix} \mathcal{P}_{[H_{\mathrm{df}}'|H_{\mathrm{c}}']} & 0 \\ 0 & J_1' j_{Q_H} \\ \bar{J}_2 j & 0 \\ 0 & I_{Q_H'} \end{bmatrix} : H' \times Q_H' \to H' \times V' \times Q' \times Q_H'. \tag{3.12}$$

Proof. As a consequence of Assumption 3.8, one has that $\dot{v}(t) \in H'$, for almost all $t \in (0,T)$ and, thus, that $\mathcal{P}_{[H_{\mathrm{df}}'|H_{\mathrm{c}}']}\dot{v}$ and $j\dot{v}$ well-defined. System (3.11) is obtained from (3.8) by using that for $p(t) \in Q$, $J_1'p(t) = \bar{J}_1{}'p(t)$, and by applying $\mathcal{E}_2^{-1}$ from the left. We show that $\mathcal{E}_2^{-1}(v', q') = 0$ in $H' \times V' \times Q' \times Q_H'$ if, and only if, $(v', q') = 0$ in $H' \times Q'$, i.e. both systems have the same solution set. By linearity, one has $\mathcal{E}_2^{-1}(0,0) = 0$. Let now $(v', q') \in H' \times Q'$ such that $\mathcal{E}_2^{-1}(v', q') = 0$. Using the decomposition $v' = v_{\mathrm{df}}' + v_{\mathrm{c}}'$ with $\mathcal{P}_{[H_{\mathrm{df}}'|H_{\mathrm{c}}']}v' = v_{\mathrm{df}}'$ and $\bar{J}_2 j v' = \bar{J}_2 j v_{\mathrm{c}}'$, we find that

$$\mathcal{E}_2^{-1} \begin{bmatrix} v_{\mathrm{df}}' + v_{\mathrm{c}}' \\ q' \end{bmatrix} = \begin{bmatrix} v_{\mathrm{df}}' \\ J_1' j_{Q_H} q' \\ \bar{J}_2 j v_{\mathrm{c}} \\ q' \end{bmatrix} = 0. \tag{3.13}$$

And $(v', q') = 0$ follows from $\bar{J}_2$ being injective on H_{c}'. □

Remark 3.11. The definition of $\mathcal{E}_2^{-1}$ is motivated by the elaborations of the finite-dimensional case where the decoupling operator chain can be explicitly derived [59]. Here $\mathcal{E}_2^{-1}$ is defined not resorting to $A(v)$. In the finite-dimensional linear case the incorporation of A led to a complete decoupling of the solution components. The analogous approach is not possible here, since the solution space $\mathcal{V}$ does not coincide with $\mathcal{H}'$, where the equations are posed.

Remark 3.12. Requiring injectivity of J_1' is not an additional restriction, since it is particularly necessary for the splitting of the solution space, cf. Assumption 3.13 just below.

3.3. **Decomposition of the Solution.** In this section, we define a splitting of the solution state space V that matches the splitting of the equations as given in Lemma 3.10.

In general, $\mathcal{P}_{[H_{\mathrm{df}}|H_{\mathrm{c}}]}\colon V \to V$ is not bounded. An example is the *Leray projector* that is bounded in $[L^2(\Omega)]^d$ but not in $[W_0^{1,2}(\Omega)]^d$, cf. [54, Rem. 4.1].

As discussed in Section 2.4, even if the kernel of J_2 is closed in V, the existence of a complement in V is not guaranteed, cf. [136]. For this reason, we will assume that $J_2\colon V \to Q_H'$ has a right inverse J_2^- so that $J_2^- J_2$, $I_V - J_2^- J_2 \in \mathcal{L}(V, V)$ decompose V into the kernel of J_2 and a remainder.

Assumption 3.13. *Consider Problem 3.1. The operator $J_1' \in \mathcal{L}(Q_H, V')$ is an homeomorphism onto its range and $J_2 \in \mathcal{L}(V, Q_H')$ has a bounded right inverse.*

Remark 3.14. As long as V is a general Banach space, Assumption 3.13 is stronger than Assumption 3.5(a). By the *Closed Range Theorem*, see, e.g., [83, Thm. IV.5.13], if J_2 has a right inverse, then it is surjective and, thus, closed and $J_2'\colon Q_H \to V'$ is a homeomorphism onto its range. Conversely, J_2' being an homeomorphism onto its range, only makes J_2 surjective but not necessarily right-invertible via a linear bounded operator.

The following arguments rephrase the results given in [67, Ch. 37]: Surjectivity of J_2 gives rise to an isomorphism $I_{J_2}\colon V/\ker J_2 \to Q_H'$, but there is no general way to identify $V/\ker J_2$ with a subspace of V. However, if there is a subspace V_g such that $V = \ker J_2 \oplus V_g$, then $V/\ker J_2$ is isomorphic to V_g and one can define a right inverse of J_2 by means of I_{J_2}.

In particular, if V itself is a Hilbert space, then every subspace has a complement and, thus, surjectivity implies existence of a bounded linear right inverse and the Assumptions 3.5(a) and 3.13 are equivalent.

We summarize some notions related to right inverses.

Remark 3.15. A mapping $J\colon V \to Q_H'$ has a right inverse, if there exists a mapping $J^-\colon Q_H' \to V$ such that $JJ^-q = q$ for all $q \in Q_H'$. As can be derived from the arguments in Remark 3.14, a right inverse is only unique if J is bijective. Also, even if J is linear and bounded, a right inverse need not be linear or bounded. If J has a right inverse, then $J|_{J^-(Q_H')}$ is injective.

Remark 3.16. An operator $J_2'\colon Q_H \to V'$ is an homeomorphism onto its range, if, and only if, it has a left inverse $J_2'^-$ and there is a $\gamma > 0$ such that $\|J_2'^- v\| \leq \frac{1}{\gamma}\|v\|$ for all $v \in \operatorname{im} J_2'$, implying $\|J_2' q\| \geq \gamma \|q\|$ for all $q \in Q_H$, cf. [3, Thm. 2.5]. Thus, if V is a Hilbert space, cf. Remark 3.14, Assumption 3.13 is equivalent [49, Lem. I.4.1] to the so-called *inf-sup* or *LBB condition*

$$\inf_{0 \neq q \in Q_H} \sup_{0 \neq v \in V} \frac{\langle J_2' q, v\rangle_{V',V}}{\|q\|_{Q_H}\|v\|_V} \geq \gamma > 0, \tag{3.14}$$

which is frequently used in literature.

Lemma 3.17. *Let Assumption 3.13 hold. Then $V = V_{\mathrm{df}} \oplus V_{\mathrm{c}}$, where $V_{\mathrm{df}} := \ker J_2$ and V_{c} is the image of Q_H' under J_2^- denoting the right inverse of J_2.*

Proof. Since J_2 and J_2^- are bounded, $J_2^- J_2\colon V \to V$ is a projection having V_{df} as its kernel and V_{c} as its image. Thus, V_{df} and V_{c} split V, cf. Section 2.4. □

3.4. **Decoupling of the System.** In this section we will show how the decoupling of the equations and of the solution gives a decoupled system. We will need several technical lemmas to characterize the relation of the split solution and equation spaces. In particular, we need to establish that V_{df}, defined in Lemma 3.17, is dense in H_{df}, defined in Lemma 3.6, so that $V_{\mathrm{df}} \hookrightarrow H_{\mathrm{df}} \cong j'(H_{\mathrm{df}}) \hookrightarrow (V_{\mathrm{df}})'$ gives a Gelfand triple to define the underlying abstract differential equation. In view of the algebraic part, we will show that for $v_{\mathrm{c}} := J_2^- g$ sufficiently smooth, one has $\dot{v}_{\mathrm{c}}(t) = j'\bar{J_2}^- \dot{g}(t) \in H'$, for almost all $t \in (0,T)$.

Lemma 3.18. *Let $V \subset H$ and $Q'_H \subset Q'$ be densely and continuously embedded, let $J_2 \colon V \to Q'_H$ be surjective, let $J_2 \colon V \subset H \to Q'$ be bounded, and let $\bar{J}_2 \colon H \to Q'$ be the closure of J_2. Then $V_{\mathrm{df}} := \ker J_2$ is dense in $H_{\mathrm{df}} := \ker \bar{J}_2$.*

Proof. We will use that for any Banach space W and subsets $W_1 \subset W_2 \subset W$ it follows that W_1 is dense in W_2 if, and only if, the annihilators coincide, i.e. ${W_1}^0 = {W_2}^0$, cf. [49, Cor. I.2.5]. By the continuity of the closure, we immediately have $V_{\mathrm{df}} \subset H_{\mathrm{df}}$ and, thus, ${H_{\mathrm{df}}}^0 \subset {V_{\mathrm{df}}}^0$. It remains to establish that ${V_{\mathrm{df}}}^0 \subset {H_{\mathrm{df}}}^0$, i.e., that any $l \in H'$ that vanishes on $V_{\mathrm{df}} \subset H$ also vanishes on H_{df}.

Let $l \in {V_{\mathrm{df}}}^0 = \{v' \in H' : \langle v', h\rangle_{H',H} = 0 \text{ for all } h \in V_{\mathrm{df}} \subset H\}$. Since the dual product in V is the extension of the product in H, we have

$$ {V_{\mathrm{df}}}^0 \subset (V_{\mathrm{df}})^0 := \{v' \in V' : \langle v', v_{\mathrm{df}}\rangle_{V',V} = 0, \text{ for all } v_{\mathrm{df}} \in V_{\mathrm{df}} \subset V\}, $$

and, thus, $l \in (V_{\mathrm{df}})^0$. Since $J_2 \colon V \subset H \to Q'_H$ has a closed range, it is closed, and, by the *Closed Range Theorem*, see e.g. [83, Thm. IV.5.13], one has $(V_{\mathrm{df}})^0 = (\ker J_2)^0 = \operatorname{im} J'_2$. Thus, there exists a $q \in Q_H$, such that $l = J'_2 q$. Take any $h \in H_{\mathrm{df}}$ and a sequence $\{v_n\}_{n\in\mathbb{N}} \subset V$ converging to h in H. Then

$$ \begin{aligned} \langle l, h\rangle_{H',H} = \lim_{n\to\infty} \langle l, v_n\rangle_{V',V} &= \lim_{n\to\infty} \langle J'_2 q, v_n\rangle_{V',V} \\ &= \lim_{n\to\infty} \langle q, J_2 v_n\rangle_{Q_H,Q'_H} = \langle q, \bar{J}_2 h\rangle_{Q,Q'} = 0, \end{aligned} $$

since $h \in \ker \bar{J}_2$. Thus $l \in {H_{\mathrm{df}}}^0$. □

Since with Assumption 3.5, the considered spaces and operators fulfill the assumptions of Lemma 3.18, for $\overline{V_{\mathrm{df}}}^{\|\cdot\|_H}$ denoting the closure of V_{df} in H, we can conclude that

$$ \overline{V_{\mathrm{df}}}^{\|\cdot\|_H} = H_{\mathrm{df}}, \text{ i.e. } V_{\mathrm{df}} \text{ is dense in } H_{\mathrm{df}}. \tag{3.15} $$

Now, we show that $(H_{\mathrm{df}})' \cong H'_{\mathrm{df}} := j'(H_{\mathrm{df}})$, where $(H_{\mathrm{df}})'$ is the dual space of the subspace $H_{\mathrm{df}} \subset H$. For later reference, we formulate the next result for general Banach spaces and state the needed and well-known Hilbert space result, see the proof of [49, Cor. I.2.4], as a corollary.

Lemma 3.19. *For a Banach space V, assume that $V = V_1 \oplus V_2$. Then the dual space of V_1 is isometrically isomorphic to the annihilator of V_2, i.e. $(V_1)' \cong {V_2}^0$.*

Proof. For any $f \in (V_1)'$, we can define a $\tilde{f} \in V'$ via

$$ \langle \tilde{f}, v\rangle := \langle f, \mathcal{P}_{[V_1|V_2]} v\rangle, \quad \text{for all } v \in V. \tag{3.16} $$

As for $w \in V_2$, we have $\mathcal{P}_{[V_1|V_2]} w = 0$, we find that $\tilde{f} \in {V_2}^0$. Since

$$ \|\tilde{f}\|_{V'} = \sup_{V \ni v \neq 0} \frac{\langle \tilde{f}, v\rangle}{\|v\|} = \sup_{V_1 \ni v_{\mathrm{df}} \neq 0} \frac{\langle f, v_{\mathrm{df}}\rangle}{\|v_{\mathrm{df}}\|_V} = \|f\|_{(V_1)'}, $$

relation (3.16) defines an isometric linear mapping $i_1 \colon (V_1)' \to {V_2}^0$. Conversely, as any $\tilde{f} \in (V_2)^0$ vanishes on V_2, it can be associated with $f \in (V_1)'$ and we have the injection $i_2 \colon \tilde{f} \mapsto f := \tilde{f}\big|_{V_1}$. Since i_1 is the inverse of i_2 and vice versa, we have that $i_1 \colon (V_1)' \to {V_2}^0$ is an isometric isomorphism and, thus, $(V_1)' \cong {V_2}^0$. □

In a Hilbert space, one can use the orthogonal complement and Riesz isomorphism to characterize the annihilator:

Lemma 3.20. *Let H_{df} be a subspace of a Hilbert space H and $H_{\mathrm{df}\perp}$ be its orthogonal complement. Then,*

$$j'(H_{\mathrm{df}}) = (H_{\mathrm{df}\perp})^0.$$

Proof. Since $H_{\mathrm{df}} := \ker \bar{J}_2$ is closed and H is a Hilbert space and therefore reflexive, we can use the identity $(H_{\mathrm{df}\perp})_\perp = H_{\mathrm{df}}$, cf. [83, p. 252]. For $v' \in (H_{\mathrm{df}\perp})^0$, one has that $\langle v', w\rangle_{H',H} = 0$, for all $w \in H_{\mathrm{df}\perp}$ if, and only if, $(jv', w)_H = 0$, for all $w \in H_{\mathrm{df}\perp}$. This is the definition of jv' being in $(H_{\mathrm{df}\perp})_\perp = H_{\mathrm{df}}$. Application of j then gives $v' \in (H_{\mathrm{df}\perp})^0$ if, and only if, $v' \in j'(H_{\mathrm{df}})$. □

Combining Lemmata 3.19 and 3.20 we can state the following:

Corollary 3.21. *If H_{df} is a closed subspace of the Hilbert space H, then the spaces $(H_{\mathrm{df}})'$ and $j'(H_{\mathrm{df}}) =: H'_{\mathrm{df}}$ are isometrically isomorphic.*

Remark 3.22. We now can show that if $\bar{J}_1 = \bar{J}_2$, then part (a) of Assumption 3.5 implies (b). By the *Closed Range Theorem*, see e.g. [83, Thm. IV.5.13], we have that $\operatorname{im} \bar{J}_1{}' = (\ker \bar{J}_2)^0$ and by Lemma 3.20 that

$$(\ker \bar{J}_2)^0 = (H_{\mathrm{df}})^0 = j'(H_{\mathrm{df}\perp})$$

and thus $H = H_{\mathrm{df}} \oplus H_{\mathrm{df}\perp} = \ker \bar{J}_2 \oplus j(\operatorname{im} \bar{J}_2{}')$. Note, that $H_{\mathrm{df}} := \ker \bar{J}_2$ and Lemma 3.20 are established not recurring to the invertibility of $S := J_2 j J_1'$ as defined in Lemma 3.6, and, hence, not presupposing Assumption 3.5 (b).

For further reference and to recall the overall setting, we collect the preceding assumptions and their implications relevant for the splitting of the spaces in one assumption:

Assumption 3.23. *Consider Problem 3.1, posed on the Gelfand triple $V \hookrightarrow H = H' \hookrightarrow V'$ and Hilbert space Q_H.*

(a) (Assumption 3.5): There is a Banach space $Q \hookrightarrow Q_H$, such that the extensions $\bar{J}_1$, $\bar{J}_2 \colon H \to Q'$ of J_1, $J_2 \colon V \to Q'_H$ are bounded and such that $H = H_{\mathrm{df}} \oplus H_{\mathrm{c}}$, where $H_{\mathrm{df}} = \ker \bar{J}_2$ and $H_{\mathrm{c}} = j(\operatorname{im} \bar{J}_1{}')$.

(b) (Assumption 3.13, Lemma 3.17, Lemma 3.18): The operator $J_2 \colon V \to Q'_H$ has a bounded right inverse, so that the solution space V decomposes into $V = V_{\mathrm{df}} \oplus V_{\mathrm{c}}$, with $V_{\mathrm{df}} = \ker J_2$ and V_{df} is dense in H_{df}. The operator $J_1 \colon Q_H \to V'$ is an isomorphism onto its range.

Next we will show, that if $v(t) \in V_{\mathrm{df}}$, i.e. $v(t)$ is in the kernel of J_2, then $\dot{v}(t)$ can be found in the Riesz representation of the kernel of $\bar{J}_2$.

Lemma 3.24. *Consider Problem 3.1 and let Assumption 3.23 hold. Let $v \in \mathcal{W}(0, T; V; H')$. Then $v \in L^2(0, T; V_{\mathrm{df}})$ if, and only if, $\dot{v} \in L^2(0, T; H'_{\mathrm{df}})$.*

To prove this, we use the following lemma:

Lemma 3.25. *Let Assumption 3.23 hold and let $\overline{V_{\mathrm{c}}}^{\|\cdot\|_H}$ be the closure of V_{c} in H. Then, $H_{\mathrm{df}} \cap \overline{V_{\mathrm{c}}}^{\|\cdot\|_H} = \{0\}$.*

Proof. We will show that if $v_{\mathrm{c}} \in \overline{V_{\mathrm{c}}}^{\|\cdot\|_H}$ and $v_{\mathrm{c}} \neq 0$, then $\|\bar{J}_2 v_{\mathrm{c}}\|_{Q'} > 0$. We will use a generic constant $\tilde{c} > 0$ that may change in every step. Given $v_g \in \overline{V_{\mathrm{c}}}$, there is a sequence $\{v_{g,n}\}_{n\in\mathbb{N}} \subset V_{\mathrm{c}}$ with $v_{g,n} \to v_{\mathrm{c}}$ in H, as $n \to \infty$. Since on V_{c}, J_2 has a bounded inverse, we obtain $\|J_2 v_{g,n}\|_{Q'_H} \geq \tilde{c}\|v_{g,n}\|_V \geq \tilde{c}\|v_{g,n}\|_H$, for all $n \in \mathbb{N}$, making use of the embedding $V \hookrightarrow H$. With $v_{g,n} \to v_g$ and $\|v_{\mathrm{c}}\|_H = \tilde{c}$, there is an $N \in \mathbb{N}$ such that $\|J_2 v_{g,n}\|_{Q'_H} \geq \tilde{c}$, for all $n > N$, meaning that

$$\sup_{0\neq q\in Q_H} \frac{\left\langle q, J_2 v_{g,n}\right\rangle_{Q_H, Q'_H}}{\|q\|_{Q_H}} \geq \tilde{c}.$$

Since Q is dense in Q_H, one also has

$$\sup_{0\neq q\in Q} \frac{\big\langle q, J_2 v_{g,n}\big\rangle_{Q_H,Q'_H}}{\|q\|_{Q_H}} \geq \tilde{c}, \tag{3.17}$$

for all $n > N$. Thus,

$$\sup_{0\neq q\in Q} \frac{\big\langle q, \bar{J}_2 v_g\big\rangle_{Q,Q'}}{\|q\|_{Q_H}} = \lim_{n\to\infty} \sup_{0\neq q\in Q} \frac{\big\langle q, J_2 v_{g,n}\big\rangle_{Q_H,Q'_H}}{\|q\|_{Q_H}} \geq \tilde{c} > 0.$$

Since there exists a $q_\varepsilon \in Q$, such that $\big\langle q_\varepsilon, \bar{J}_2 v_g\big\rangle_{Q,Q'} = (\tilde{c}-\varepsilon)\|q_\varepsilon\|_{Q_H} > 0$ and, thus,

$$\frac{\big\langle q_\varepsilon, \bar{J}_2 v_g\big\rangle_{Q,Q'}}{\|q_\varepsilon\|_Q} = (\tilde{c}-\varepsilon)\frac{\|q_\varepsilon\|_{Q_H}}{\|q_\varepsilon\|_Q} > 0,$$

we have

$$\|\bar{J}_2 v_g\|_{Q'} = \sup_{0\neq q\in Q} \frac{\big\langle q, \bar{J}_2 v_g\big\rangle_{Q,Q'}}{\|q\|_Q} \geq (\tilde{c}-\varepsilon) > 0. \qquad \square$$

Proof of Lemma 3.24. For a $v \in \mathcal{W}(0,T;V;H') \cap L^2(0,T;H_{\mathrm{df}})$, by definition of the orthogonal complement, we have $\big(v(t), w\big)_H = 0$, for all $w \in (H_{\mathrm{df}})_\perp$ and for almost all $t \in (0,T)$. The latter is equivalent to

$$\int_0^T (v(t), w)\dot{\phi}(t)\ \mathrm{d}t = -\int_0^T \langle \dot{v}(t), w\rangle \phi(t)\ \mathrm{d}t = 0$$

for all $w \in (H_{\mathrm{df}})_\perp$ and for all $\phi \in \mathcal{C}_0^\infty$, or equivalent to $\langle \dot{v}(t), w\rangle_{H',H} = 0$ or $\dot{v}(t) \in (H_{\mathrm{df}})^0$, which, by Lemma 3.20 is the case, if, and only if, $\dot{v}(t) \in H'_{\mathrm{df}}$. Thus, since $\dot{v} \in L^2(0,T;H')$ by assumption, we conclude that $\dot{v} \in L^2(0,T;H'_{\mathrm{df}})$.

Conversely, we can take a $v \in \mathcal{W}(0,T;V,H')$, with $\dot{v} \in L^2(0,T;Hk')$, and go backwards through the arguments above to find that $v \in L^2(0,T;H_{\mathrm{df}})$.

By assumption, $v(t)$ is also in V and, in line with Lemma 3.17, it can be split into $v(t) = v_{\mathrm{df}}(t) + v_{\mathrm{c}}(t)$. Because of $\overline{V_{\mathrm{df}}}^{\|\cdot\|_H} = H_{\mathrm{df}}$, one has $v_{\mathrm{df}}(t) \in H_{\mathrm{df}}$ and, thus, also $v_{\mathrm{c}}(t) \in H_{\mathrm{df}}$. Since $V_g \cap H_{\mathrm{df}} = \{0\}$, cf. Lemma 3.25, and since the embedding $V \hookrightarrow H$ is injective, we have $v_{\mathrm{c}}(t) = 0 \in V$ and thus $v(t) = v_{\mathrm{df}}(t) \in V_{\mathrm{df}}$. □

Since the subspace V_{df} can be replaced by any complementable subspace of V, we conclude that for $v \in \mathcal{W}(0,T;V;H')$ it holds that

$$v(t) \in V_{\mathrm{c}},\ \text{if, and only if, } \dot{v}(t) \in j'(\overline{V_{\mathrm{c}}}^{\|\cdot\|_H}) \tag{3.18}$$

for almost all $t \in (0,T)$. However, because H_g is a particular choice while V_g is defined via a right inverse J_2^-, that is not uniquely defined, in general, $\overline{V_{\mathrm{c}}}^{\|\cdot\|_H} \neq H_g$.

We will define v_{c} via the right inverse of J_2 applied to the inhomogeneity g in the algebraic constraint of the state equations (3.1). The following lemmas will be used to link the derivative of the inhomogeneity $\dot{g}(t) \in Q'$ with $\dot{v}_{\mathrm{c}}(t) \in H'$.

Lemma 3.26. *Consider Problem 3.1 and assume that Assumptions 3.23 hold and consider the right inverse J_2^- of $J_2\colon V \to Q'_H$ inverse with $J_2^-(Q'_H) = V_{\mathrm{c}}$. Then $\bar{J}_2\colon \overline{V_{\mathrm{c}}}^{\|\cdot\|_H} \to Q'$ has a right inverse $\bar{J}_2{}^-\colon \bar{J}_2(\overline{V_{\mathrm{c}}}^{\|\cdot\|_H}) \to \overline{V_{\mathrm{c}}}^{\|\cdot\|_H}$ and the restriction $\bar{J}_2{}^-\big|_{Q'_H\subset Q'}\colon Q'_H \subset Q' \to V_{\mathrm{c}} \subset H$ coincides with J_2^-.*

Proof. By Assumption 3.23(b) and by Lemma 3.17, the considered right inverse J_2^- exists. By Lemma 3.25, we have that the map $\bar{J}_2\big|_{\overline{V_{\mathrm{c}}}^{\|\cdot\|_H}}$ into Q' is injective and thus invertible on its range. Since for $v \in V \subset H$, it holds that $\bar{J}_2 v = J_2 v \in Q'_H \subset Q'$, on $\bar{J}_2(V_g) = Q'_H \subset Q'$, the right inverse $\bar{J}_2{}^-$ of $\bar{J}_2$, coincides with J_2^- considered as a map from $Q'_H \subset Q'$ into $V_{\mathrm{c}} \subset H$. □

Remark 3.27. Since $\bar{J}_2{}^-$ as defined in the proof of Lemma 3.26 is closed but not necessarily bounded, cf. Remark 3.15, it does not simply extend to a map from Q' onto $\overline{V_\mathrm{c}}^{\|\cdot\|_H}$. That is why we will require $\dot{g}(t)$ to be in $\bar{J}_2(\overline{V_\mathrm{c}}^{\|\cdot\|_H})$ which is dense in Q'.

Lemma 3.28. *Let $g \in L^2(0,T;Q'_H)$ and $v_g := J_2^- g$. Then $\dot{v}_\mathrm{c} \in L^2(0,T;H')$, if, and only if, $\dot{g} \in \mathcal{Q}'_-$, where*

$$\mathcal{Q}'_- := \{f \in \mathcal{Q}' : \textit{ there exists a } w \in L^2(0,T;\overline{V_\mathrm{c}}^{\|\cdot\|_H}), \textit{ such that } \bar{J}_2 w = f\}.$$

Proof. Assume that $\dot{v}_\mathrm{c} \in L^2(0,T;H')$. Then v_c is in $\mathcal{W}(0,T;V:H')$ and by Lemma 3.25 and (3.18) we have that $j\dot{v}_\mathrm{c} \in L^2(0,T;\overline{V_\mathrm{c}}^{\|\cdot\|_H})$. Since $\bar{J}_2$ is bounded, we have $\bar{J}_2 j \dot{v}_\mathrm{c} \in Q'_- \subset L^2(0,T;\bar{J}_2(\overline{V_\mathrm{c}}^{\|\cdot\|_H})) \subset \mathcal{Q}'$. We show that $\bar{J}_2 j \frac{d}{dt}(J_2^- g) = \dot{g}$ in $\mathcal{Q}'$. For all $\phi \in C_0^\infty(0,T)$ and for all $q \in Q$, we have that

$$\begin{aligned}
\big\langle \bar{J}_2 j \tfrac{d}{dt}(J_2^- g), \phi q\big\rangle_{\mathcal{Q}',\mathcal{Q}} &= \big\langle j\tfrac{d}{dt}(J_2^- g), \phi \bar{J}_2{}' q\big\rangle_{\mathcal{H},\mathcal{H}'} = \big\langle \tfrac{d}{dt}(J_2^- g), \phi j \bar{J}_2{}' q\big\rangle_{\mathcal{H}',\mathcal{H}} \\
&= -\big(J_2^- g, \dot{\phi}\bar{J}_2{}' q\big)_{\mathcal{H},\mathcal{H}} = -\big\langle J_2^- g, \dot{\phi}\bar{J}_2{}' q\big\rangle_{\mathcal{H},\mathcal{H}'} \\
&= -\big\langle \bar{J}_2 J_2^- g, \dot{\phi} q\big\rangle_{\mathcal{Q}',\mathcal{Q}} = -\big(g, \dot{\phi} q\big)_{\mathcal{Q}_\mathcal{H},\mathcal{Q}_\mathcal{H}} \\
&= \big\langle \dot{g}, \phi q\big\rangle_{\mathcal{Q}',\mathcal{Q}}.
\end{aligned}$$

To prove the converse direction we show that if $\dot{g} \in Q'_-$, then the unique $w \in L^2(0,T;\overline{V_\mathrm{c}}^{\|\cdot\|_H})$ that fulfills $\bar{J}_2 w = \dot{g}$ is the *Riesz representation* of the derivative of v_c in $L^2(0,T;H)$. That is we need to establish that given $\phi \in C_0^\infty(0,T)$, we have

$$\big\langle j' w, \phi v\big\rangle_{\mathcal{H}',\mathcal{H}} = \big(v_\mathrm{c}, \dot{\phi} v\big)_{\mathcal{H},\mathcal{H}}$$

for all $v \in j\bar{J}_2{}'(Q)$. Taking v from the range of $j\bar{J}_2{}'$ is enough because it can be identified with a dense subset in $\overline{V_\mathrm{c}}^{\|\cdot\|_H}$, and because the inner product with vectors from the complement is zero by definition. Denseness follows by injectivity of $\bar{J}_2 \colon \overline{V_\mathrm{c}}^{\|\cdot\|_H} \to Q'$ that makes the range of its adjoint dense in $(\overline{V_\mathrm{c}}^{\|\cdot\|_H})'$ and by Lemma 3.20 stating that $(\overline{V_\mathrm{c}}^{\|\cdot\|_H})'$ can be identified with $j'(\overline{V_\mathrm{c}}^{\|\cdot\|_H})$. Given $v \in j\bar{J}_2{}'(Q)$ and the corresponding $q \in Q$ that fulfills $j\bar{J}_2{}' q = v$, we compute that

$$\begin{aligned}
\big\langle j' w, \phi v\big\rangle_{\mathcal{H}',\mathcal{H}} &= \big\langle w, \phi \bar{J}_2{}' q\big\rangle_{\mathcal{H},\mathcal{H}'} = \big\langle \bar{J}_2 w, \phi q\big\rangle_{\mathcal{Q},\mathcal{Q}} \\
&= \big\langle \dot{g}, \phi q\big\rangle_{\mathcal{Q}',\mathcal{Q}} = -\big(g, \dot{\phi} q\big)_{\mathcal{Q},\mathcal{Q}} \\
&= -\big(J_2 J_2^- g, \dot{\phi} q\big)_{\mathcal{Q},\mathcal{Q}} = -\big\langle J_2^- g, \dot{\phi} J_2' q\big\rangle_{\mathcal{V},\mathcal{V}'} \\
&= -\big\langle v_\mathrm{c}, \dot{\phi}\bar{J}_2{}' q\big\rangle_{\mathcal{H},\mathcal{H}'} = -\big(v_\mathrm{c}, \dot{\phi} v\big)_{\mathcal{H},\mathcal{H}'},
\end{aligned}$$

where we have used that for $q \in Q$, $J_2 q \in V'$ equals $\bar{J}_2 q \in H' \subset V'$, cf. Lemma 3.4. □

Remark 3.29. We have that $\mathcal{Q}'_- \subset L^2(0,T;\bar{J}_2(\overline{V_\mathrm{c}}^{\|\cdot\|_H})) \subset \mathcal{Q}'$ and that all spaces collapse into $\mathcal{Q}'$ if, and only if, $\bar{J}_2{}^-$ is bounded.

Having established the necessary properties and relations of the splittings of solution and equation spaces, we now can separate the differential and algebraic components of the abstract differential algebraic equation (3.1). First, we state the equations that, under certain conditions, the algebraic and differential parts of any solution to (3.1) need to fulfill. Secondly, from the separated components, we can read off necessary and sufficient smoothness and consistency conditions of the problem. Finally, we will use the underlying abstract differential equation to state existence and uniqueness of solutions to (3.1).

Lemma 3.30. *Consider the setup of Problem 3.1 and let Assumption 3.23 hold. Let $f \in L^2(0,T;H')$ and Assumptions 3.8 hold, i.e. a solution (v,p) to System (3.1a,b) is in $\mathcal{W}(0,T;V;H') \times L^2(0,T;Q)$. Then any solution (v,p) of System (3.1a,b) has the representation $(v_{\mathrm{df}} + v_{\mathrm{c}}, p)$, where $v_{\mathrm{c}} \in L^2(0,T;V_{\mathrm{c}})$ satisfies*

$$v_{\mathrm{c}}(t) = -J_2^- g(t), \tag{3.19a}$$

where J_2^- is the right inverse for J_2 as defined in Lemma 3.17, where v_{df} is a solution to

$$\mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}\dot{v}_{\mathrm{df}}(t) - \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}A(v_{\mathrm{df}}(t) - J_2^- g(t)) = \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[f(t) - \tfrac{d}{dt}(J_2^- g)(t)], \tag{3.19b}$$

and where $p \in \mathcal{Q}$ satisfies

$$p(t) = S^{-1}\big[\bar{J}_2 j A(v_{\mathrm{df}}(t) - J_2^- g(t)) + \bar{J}_2 j f(t) + \dot{g}(t)\big], \tag{3.19c}$$

with $S = J_2 j J_1'$ and $\mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}$ defined in Lemma 3.6. The equalities hold for a.a. $t \in (0,T)$.

Proof. By Assumption 3.23, J_1' is injective and by Lemma 3.10, System (3.8a) is equivalent to (3.11). By Lemma 3.17 we can write $v = v_{\mathrm{df}} + v_{\mathrm{c}}$ with unique $v_{\mathrm{df}} \in \mathcal{V}_{\mathrm{df}}$ and $v_{\mathrm{c}} \in \mathcal{V}_{\mathrm{c}}$. Then the equations from the decoupling given in (3.11), with parts in H', V', Q'_H, and Q', cf. (3.12), read as follows. With $J_2 v_{\mathrm{df}} = 0$ the part in $\mathcal{Q}'_H$ is given via

$$-J_2 v = -J_2 v_{\mathrm{c}} = g \in \mathcal{Q}'_H.$$

By means of the right inverse J_2^- one obtains (3.19a). Since J_1' is injective, the equation $-J_1' j_{Q_H} J_2 v_{\mathrm{c}} = J_1' j_{Q_H} g$ in $\mathcal{V}'$ is redundant.

The equation part in $\mathcal{H}'_{\mathrm{df}}$ is given via

$$\mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}\dot{v} - \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}A(v) = \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[f + \tfrac{d}{dt}(J_2^- fp)],$$

which is (3.19b), since $v = v_{\mathrm{df}} - J_2^- g$.

Finally, the part in $\mathcal{Q}'$ is given via

$$\bar{J}_2 j \dot{v} - \bar{J}_2 j A(v) - Sp = \bar{J}_2 j f. \tag{3.20}$$

By Lemma 3.24, it holds that $\bar{J}_2 j \dot{v}_{\mathrm{df}} = 0$, by Lemma 3.28 that $\bar{J}_2 j J_2^- \dot{g} = \dot{g}$. Thus, the invertibility of S, as defined in Lemma 3.6, gives the necessary condition (3.19c) for the algebraic variable p. □

Corollary 3.31. *Under the assumptions of Lemma 3.30, and with $f \in L^2(0,T;H')$, for the existence of a solution $(v,p) \in \mathcal{W}(0,T;V;H') \times L^2(0,T;Q)$ to Problem 3.1 it is necessary, that $g \in W^{1,2}(0,T;Q'_H;Q'_-)$ and $\alpha = \alpha_{\mathrm{df}} - \bar{J}_2^{\,\bar{-}} g(0)$, with $\alpha_{\mathrm{df}} \in H_{\mathrm{df}}$, i.e. $-\bar{J}_2\alpha = g(0)$.*

Proof. By Lemmas 3.10 and 3.30, solutions $(v,p) \in \mathcal{W}(0,T;V;H') \times \mathcal{Q}$ of (3.1) fulfill Equations (3.19). According to (3.19c) $\dot{g} \in \mathcal{Q}'$ is necessary for the existence of $p \in \mathcal{Q}$. With $g \in \mathcal{W}(0,T;Q'_H;Q'_-) \hookrightarrow \mathcal{C}(0,T;Q')$ also $g(0) \in Q'$ is well defined. We have $\mathcal{W}(0,T;V_{\mathrm{df}},H'_{\mathrm{df}}) \hookrightarrow \mathcal{C}([0,T],H_{\mathrm{df}})$, so that, for $v_{\mathrm{df}} \in \mathcal{V}_{\mathrm{df}}$ to solve (3.19b), $v_{\mathrm{df}}(0)$ must be in H_{df}.

Furthermore, since $v_{\mathrm{c}} \in \mathcal{W}(0,T;V_{\mathrm{c}},j'(\overline{V_{\mathrm{c}}}^{\|\cdot\|_H})) \hookrightarrow \mathcal{C}(0,T;\overline{V_{\mathrm{c}}}^{\|\cdot\|_H})$, a value for $v_{\mathrm{c}}(0) \in H$ must be well defined. Since $J_2 v_{\mathrm{c}} = -g \in \mathcal{C}([0,T];Q')$, we have that $J_2 v_{\mathrm{c}}(t) \to -g(0)$ in $\mathcal{Q}'$, as $t \to 0$. Then for $t_n := \frac{1}{n}$, the sequence $(\bar{J}_2 v_{\mathrm{c}})_n := \bar{J}_2 v_{\mathrm{c}}(t_n) \to g(0)$, as $n \to \infty$. Also, because of the continuity of v_{c}, the sequence $\bar{J}_2^{\,\bar{-}}(\bar{J}_2 v_{\mathrm{c}}(t_n)) = v_{\mathrm{c}}(t_n)$ converges to a function h in H. Since $\bar{J}_2^{\,\bar{-}}$, as the inverse of a bounded injective operator, is closed, this limit is in the range of J_2^- and equals $h = v_{\mathrm{c}}(0) = -\bar{J}_2^{\,\bar{-}} g(0)$.

Note, that this gives $-\bar{J}_2\alpha = -J_2[v_{\mathrm{df}}(0)+v_{\mathrm{c}}(0)] = g(0)$, i.e., the initial condition must fulfill the algebraic constraints. □

Remark 3.32. Unlike v, that can be shown to be a continuous function in time, the variable p is only in $L^2(0,T;Q)$ and the point value $p(0)$ has no meaning. Thus, in the current setting, there is no consistency restriction imposed by (3.19c). However, if one requires continuity in time of p or, equivalently, of $\bar{J}_2 jA(v)$, then (3.19c) demands that $f(0)$, $\dot{g}(0)$, and α are consistent such that

$$p(0) = -S^{-1}[\bar{J}_2 jA(\alpha) + \bar{J}_2 jf(0) + \dot{g}(0)],$$

which is in line with the conditions that were established for the Navier-Stokes case [68, Sec. 2].

Remark 3.33. If the case of the Navier-Stokes Equations as defined Problem 3.1(NSE), Equation 3.20 gives the Pressure Poisson Equation. Thus, the assumptions of Lemma 3.30 give sufficient conditions for the existence of the Pressure Poisson Equation in infinite dimensions, cf. the discussion in the introduction of this thesis.

Remark 3.34. Since the consistency of the initial value as specified in Corollary 3.31 is necessary for any solution to (3.1), it cannot depend on a particular choice of J_2^- in Lemma 3.17. Thus, an initial condition α is genuine consistent or not, but the parts α_{df} and $-\bar{J}_2{}^- g(0)$ may depend on the choice of J_2^-.

The preceding remark gives rise to the following definition:

Definition 3.35. Consider Problem 3.1, assume that Assumptions 3.5 and 3.13 hold, and that $g \in \mathcal{W}(0,T;Q'_H,Q')$. An initial condition $\alpha \in H$ is called *consistent*, if there is a right inverse $\bar{J}_2{}^-$ to $\bar{J}_2$ such that $\alpha = \alpha_{\mathrm{df}} - \bar{J}_2{}^- g(0)$, with $\alpha_{\mathrm{df}} \in \ker \bar{J}_2$.

Corollary 3.36. *An initial condition is consistent in the sense of Definition 3.35 if, and only if,* $-\bar{J}_2\alpha = g(0)$.

Proof. If α is consistent, then $-\bar{J}_2\alpha = g(0)$. If $-\bar{J}_2\alpha = g(0)$, then for any right inverse $\bar{J}_2{}^-$ one has that $\alpha = (\alpha + \bar{J}_2{}^- g(0)) - \bar{J}_2{}^- g(0)$ with $\alpha + \bar{J}_2{}^- g(0) \in \ker \bar{J}_2$. □

Theorem 3.37. *Consider Problem 3.1 and let Assumption 3.23 hold. Let* $f \in L^2(0,T;H')$ *and let Assumption 3.8 hold. Let* $g \in L^2(0,T;Q'_H)$ *and let* $\alpha \in H$. *Then the ADAE* (3.1) *has a (unique) solution* $(v,p) \in \mathcal{W}(0,T;V;H') \times \mathcal{Q}$ *if, and only if,* $\dot{g} \in L^2(0,T;Q') \cap \mathcal{Q}'_-$, $\alpha = \alpha_{\mathrm{df}} - \bar{J}_2{}^- g(0)$, *with* $\alpha_{\mathrm{df}} \in H_{\mathrm{df}}$, *and*

$$\dot{w}(t) - \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[A(w(t) - J_2^- g(t))] = \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[f(t) + \tfrac{d}{dt}(J_2^- g)(t)] \quad \text{in } V', \tag{3.21a}$$

$$w(0) = \alpha_{\mathrm{df}} \quad \text{in } H, \tag{3.21b}$$

for almost all $t \in (0,T)$, *has a (unique) solution* $w \in \mathcal{W}(0,T;V_{\mathrm{df}};H'_{\mathrm{df}})$.

Proof. Assume that the ADAE (3.1) has a solution (v,p). Then by Corollary 3.31 the conditions on g and α are necessarily fulfilled and by Lemma 3.30 $\mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}v$ solves (3.21). While the solution components v_{c} and p are fixed via (3.19a,c), all solutions of (3.21) also solve (3.19b). Thus (v,p) can only be unique, if (3.21) has a unique solution.

Conversely, with consistent g and α and v_{df} solving (3.21), by Lemma 3.30 $(v_{\mathrm{df}} + v_{\mathrm{c}}, p)$ solves (3.1). Since, under the assumptions made, any solution in $\mathcal{W}(0,T;V;H') \times \mathcal{Q}$ to the ADAE (3.1) has a representation as given in Lemma 3.30, uniqueness is given if

$$\begin{aligned} \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}\dot{w} - \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}A(w - J_2^- g) = \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[f + \tfrac{d}{dt}(J_2^- g)], \\ w(0) = \alpha_{\mathrm{df}}, \end{aligned} \tag{3.22}$$

determines a unique $w \in \mathcal{W}(0,T;V_{\mathrm{df}},V')$. We consider solutions in $\mathcal{W}(0,T;V;H')$, for which Equations (3.22) and (3.21) are the same, cf. Lemma 3.24, and, thus, unique solvability of (3.21) implies unique solvability of the ADAE (3.1). □

Equation (3.21) is formulated on $V \hookrightarrow H \hookrightarrow V'$ but – under Assumption 3.8 – defines solutions in $\mathcal{V}_{\mathrm{df}}$ with derivatives in $\mathcal{H}'_{\mathrm{df}} \cong j'(\mathcal{H}_{\mathrm{df}})$. Thus, we can consider (3.21) on the Gelfand triple $V_{\mathrm{df}} \hookrightarrow H_{\mathrm{df}} \hookrightarrow (V_{\mathrm{df}})'$, cf. Lemma 3.18 stating the denseness of the embeddings, and use standard results to establish sufficient conditions for the existence of solutions in terms of properties of the operator

$$A_0(\cdot) := A(\cdot - J_2^- g) \colon \mathcal{V}_{\mathrm{df}} \to (\mathcal{V}_{\mathrm{df}})'. \tag{3.23}$$

For later reference for spatial semi-discretization via Galerkin schemes, we state an existence result that bases on possible semi-coerciveness of A_0. This fits the weak formulation of the Navier-Stokes Equation. Other results that prove existence of solution under different conditions are discussed at the end of this section in Remark 3.39.

Theorem 3.38 ([133], Thm. 8.27)**.** *Let $V_{\mathrm{df}} \hookrightarrow H_{\mathrm{df}} \hookrightarrow (V_{\mathrm{df}})'$ be a Gelfand triple and let $-A_0 \colon [0,T] \times V_{\mathrm{df}} \to (V_{\mathrm{df}})'$ be a Carathéodory mapping, be* semi-coercive, *i.e. there is $c_0 > 0$, $c_1 \in L^{p'}(0,T)$, and $c_2 \in L^1(0,T)$ such that for all $v \in V_{\mathrm{df}}$ it holds that*

$$-\left\langle A_0(t,v), v \right\rangle_{(V_{\mathrm{df}})',V_{\mathrm{df}}} \geq c_0 |v|_V^p - c_1(t)|v|_V - c_2(t)\|v\|_{H_{\mathrm{df}}}^2, \tag{3.24}$$

and satisfy the following growth condition*: there is $\gamma \in L^{p'}([0,T])$ and $c \colon \mathbb{R} \to \mathbb{R}$ increasing, so that*

$$\|A_0(t,v)\|_{(V_{\mathrm{df}})'} \leq c(\|v\|_H)(\gamma(t) + \|v\|_V^{p-1}),$$

for all $v \in V_{\mathrm{df}}$ and almost all $t \in (0,T)$, and for a $1 \leq p < \infty$ and p' such that $\frac{1}{p} + \frac{1}{p'} = 1$. Let $-A_0(t,\cdot) \colon V_{\mathrm{df}} \to (V_{\mathrm{df}})'$ be pseudomonotone, see Definition 2.17, for almost all $t \in (0,T]$.

Then, there is a solution $u \in \mathcal{W}^{1,p,p'}(0,T;V_{\mathrm{df}};(V_{\mathrm{df}})')$ to

$$\dot{u}(t) - A_0(t,u(t)) = f_{\mathrm{df}}(t), \quad \text{for a.a. } t \in (0,T), \quad u(0) = a_0, \tag{3.25}$$

for any $f_{\mathrm{df}} \in L^{p'}(0,T;(V_{\mathrm{df}})')$ and $a_0 \in H_{\mathrm{df}}$.

If A_0 fulfills Assumption 3.8 on $V_{\mathrm{df}} \hookrightarrow H_{\mathrm{df}}$, then a solution to (3.25), with $f_{\mathrm{df}} = \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]} f \in H'_{\mathrm{df}} \subset (H_{\mathrm{df}})'$ and $p = p' = 2$, also solves (3.21), as by Corollary 3.21, a function $w \in \mathcal{W}(0,T;V_{\mathrm{df}};(H_{\mathrm{df}})')$ can be identified with a function $\tilde{w} \in \mathcal{W}(0,T;V_{\mathrm{df}};H'_{\mathrm{df}})$.

Remark 3.39. Theorem 3.38 establishes existence of solutions to the nonlinear abstract ODE (3.25) via semi-coerciveness and bounded growth of the involved operator A_0. Existence of solutions can also be proved if, among other conditions, A_0 is a sum of a monotone and a strongly continuous operator [40, Thm. 8.4.2]. Unique solvability is proved for A_0 monotone, coercive, and hemicontinuous [162, Thm. 30A] or radially continuous [45, Thm. VI.1.1].

Remark 3.40. Except from Theorem 3.38, and the statements that an initial condition for v must be well-defined, see Section 3.1, all results in this Section 3 are independent of the degree of Bochner integrability of the functions. All results hold for candidate solutions $v \in \mathcal{W}^{1;p,q}(0,T;V;H')$, for a $1 \leq p,q < \infty$, a right-hand side $f \in L^q(0,T;H')$, and Equation (3.1a) posed in $L^q(0,T;V')$. In order to apply Theorem 3.38, one can consider solutions in $\mathcal{W}^{1;p,p'}(0,T;V,V')$ with $1 \leq p \leq \infty$, p' such that $\frac{1}{p} + \frac{1}{p'} = 1$, and suitable right hand sides.

3.5. **Application to the Navier-Stokes Equation.** In this section we show the applicability of the results of Section 3 to the weak formulation of the Navier-Stokes Equation

$$\dot{v} + (v \otimes \nabla)v - \operatorname{div}(\nu \nabla v) + \operatorname{grad} p = f \quad \text{in } (0,T) \times \Omega, \tag{3.26a}$$

$$\operatorname{div} v = 0 \quad \text{in } (0,T] \times \Omega, \tag{3.26b}$$

$$v\big|_{(0,T)\times\partial\Omega} = 0 \quad \text{and} \quad v\big|_{\{0\}\times\Omega} = \alpha, \tag{3.26c}$$

modelling the evolution of the velocities $v \in \big((0,T] \times \Omega \to \mathbb{R}^d\big)$, $d \in \{2,3\}$, and the pressure $p \in \big((0,T] \times \Omega \to \mathbb{R}\big)$ of a flow in a domain $\Omega \subset \mathbb{R}^d$ for a time interval $(0,T]$.

We will assume that Ω and its boundary are regular, i.e. they fulfill Assumption 2.10.

The standard weak formulation, as derived e.g. in [49, 103, 143], formulates (3.26) as Problem 3.1(NSE). In particular, we have that

$$V := [W_0^{1,2}(\Omega)]^d := \big\{v \in [W^{1,2}(\Omega)]^d : v\big|_{\partial\Omega} = 0\big\},$$
$$H := [L^2(\Omega)]^d,$$

and

$$Q_H := L^2(\Omega)/\mathbb{R}.$$

Because of $v \in V$ being zero at the boundary, we can define $J_2 := -\operatorname{div}\colon V \to Q_H'$ and J_1' – representing grad in (3.26) – via $J_1' = J_2'$ using

$$\big\langle \bar{J_1}' q, w \big\rangle := -\int_\Omega q \cdot \operatorname{div} w \,\mathrm{d}\omega, \quad \text{for } q \in Q_H \text{ and } w \in V. \tag{3.27}$$

We will use the short notation $J_1' = -J_2' =: \nabla$, with ∇ being the formal vector of partial derivatives as introduced in Section 2.4. Note, that the formal notation ∇ has to be interpreted with respect to the domains of definition.

To apply the decoupling of Lemma 3.30, we introduce $Q := W^{1,2}(\Omega)/\mathbb{R}$ and show that with this choice $Q \hookrightarrow Q_H$ and that $J_2 = -\nabla'$ and $J_1' = \nabla$ fulfill Assumptions 3.5 and 3.13. To establish also Assumption 3.8, we give sufficient conditions for the additional regularity of solutions of weak formulations of (3.26).

The quotient space $W^{1,2}(\Omega)/\mathbb{R}$ is the space of equivalence classes $[q] \subset W^{1,2}(\Omega)$ defined as q_1, $q_2 \in [q]$, if $q_1 - q_2 \in \mathbb{R}$. Considering the norm $\|[q]\|_{W^{1,2}(\Omega)/\mathbb{R}} := \inf_{q\in[q]} \|q\|_{W^{1,2}(\Omega)}$, it is a Banach space, see e.g. [122, Ch. 1.1.7]. Writing $[q] = q_0 + \mathbb{R}$ for some $q_0 \in [q]$, we have that $\|[q]\|_{W^{1,2}(\Omega)/\mathbb{R}} = \inf_{c\in\mathbb{R}} \|q_0 + c\|_{W^{1,2}(\Omega)}$.

Proposition 3.41. *The space $W^{1,2}(\Omega)/\mathbb{R}$ is continuously and densely embedded in $L^2(\Omega)/\mathbb{R}$ and $L^2(\Omega)/\mathbb{R}$ is a Hilbert space.*

Proof. The norm in $L^2(\Omega)/\mathbb{R}$ is defined as $\|[q]\|_{L^2(\Omega)/\mathbb{R}} := \inf_{q\in[q]} \|q\|_{L^2(\Omega)}$. Since $W^{1,2}(\Omega) \hookrightarrow L^2(\Omega)$, we have that the injection $I\colon W^{1,2}(\Omega)/\mathbb{R} \to L^2(\Omega)/\mathbb{R}$ is continuous. Next, we show that $W^{1,2}(\Omega)/\mathbb{R}$ is dense in $L^2(\Omega)$. Viewing $\mathbb{R}$ as a closed subspace of $L^2(\Omega)$ one finds that the Hilbert space $\mathbb{R}_\perp = \{q \in L^2(\Omega) : \int_\Omega q \,\mathrm{d}\omega = 0\}$ is rendered isometrically isomorph to $L^2(\Omega)/\mathbb{R}$ by the isomorphism $T\colon \mathbb{R}_\perp \to L^2(\Omega)/\mathbb{R}\colon q \mapsto q + \mathbb{R}$. Thus, any $[q] \in L^2(\Omega)/\mathbb{R}$ can be identified with a unique $L^2(\Omega) \ni q_0 = T^{-1}[q]$. Since $W^{1,2}(\Omega)$ is dense in $L^2(\Omega)$, there is a sequence $\{q_{0,n}\}_{n\in\mathbb{N}} \in W^{1,2}(\Omega)$ that converges to q_0 in $L^2(\Omega)$. Then, the sequence $\{[q_{0,n}]\}_{n\in\mathbb{N}}$ approaches $[q] \in L^2(\Omega)/\mathbb{R}$, since $\inf_{c\in\mathbb{R}} \|q_{0,n} - q_0 + c\|_{L^2(\Omega)} \leq \|q_{0,n} - q_0\|_{L^2(\Omega)} \to 0$ as $n \to \infty$. Furthermore, $L^2(\Omega)/\mathbb{R}$ can be equipped with the scalar product $\big(T^{-1}\cdot, T^{-1}\cdot\big)_{L^2(\Omega)}$, which is taking the $L^2(\Omega)$ scalar-product of the unique representatives of the equivalence classes that have a zero mean. □

We will write $q \in W^{1,2}(\Omega)/\mathbb{R}$ instead of $[q]$.

So, we can define the Gelfand triple

$$Q := W^{1,2}(\Omega)/\mathbb{R} \hookrightarrow Q_H := L^2(\Omega)/\mathbb{R} \hookrightarrow (W^{1,2}(\Omega)/\mathbb{R})' = Q' \tag{3.28}$$

to weakly formulate the divergence constraints (3.26). For the weak formulation of the differential part (3.26a), we use the standard Gelfand triple

$$V := [W_0^{1,2}(\Omega)]^d \hookrightarrow H := [L^2(\Omega)]^d \hookrightarrow [W_0^{1,2}(\Omega)']^d = V'.$$

With the assumption of a regular boundary of Ω, we have that $W^{1,2}(\Omega) \overset{c}{\hookrightarrow} L^2(\Omega)$ and that there exist c_1, $c_2 > 0$, such that

$$c_1\|q\|_{W^{1,2}(\Omega)/\mathbb{R}} \leq \|\nabla q\|_{[L^2(\Omega)]^d} \leq c_2\|q\|_{W^{1,2}(\Omega)/\mathbb{R}}, \quad \text{for all } q \in W^{1,2}(\Omega)/\mathbb{R}, \tag{3.29}$$

i.e. on $W^{1,2}(\Omega)/\mathbb{R}$, the norms $\|\cdot\|_{W^{1,2}(\Omega)/\mathbb{R}}$ and $\|\nabla\cdot\|_{[L^2(\Omega)]^d}$ are equivalent [122, Thm. II.7.2].

This implies that $\nabla' \colon [W_0^{1,2}(\Omega)]^d \subset [L^2(\Omega)]^d \to (W^{1,2}(\Omega)/\mathbb{R})'$ is bounded. Having this, we can establish Assumption 3.5(a), stating that

Proposition 3.42. *The closure of* $\nabla \colon [W_0^{1,2}(\Omega)]^d \subset [L^2(\Omega)]^d \to (W^{1,2}(\Omega)/\mathbb{R})'$,

$$\overline{\nabla}' \colon [L^2(\Omega)]^d \to (W^{1,2}(\Omega)/\mathbb{R})' \tag{3.30}$$

is surjective, i.e. its dual $\overline{\nabla} \colon W^{1,2}(\Omega)/\mathbb{R} \to [L^2(\Omega)]^d$ *is a homeomorphism onto its range.*

Proof. By (3.29) one has that $\overline{\nabla}$ is bounded from below, what makes it injective. By linearity and boundedness from below, it follows that the preimage of any Cauchy sequence in the image of $\overline{\nabla}$ is a convergent sequence in $W^{1,2}(\Omega)/\mathbb{R}$. Thus, by continuity, we have that the range of $\overline{\nabla}$ also contains the limits of the sequences and, thus, is closed. □

Remark 3.43. With the choice $\bar{J}_1 = -\bar{J}_2 = \overline{\nabla}$, we get that also part (b) of Assumption 3.5 holds, cf. Remark 3.22.

Since the definition of weak derivatives given in Section 2.4 requires the derivative to be in $L^1_{\mathrm{loc}}(\Omega)$, in this setup, we cannot simply write $\overline{\nabla}' = -\operatorname{div}$. However, for $v \in [W_0^{1,2}(\Omega)]^d$ and $q \in W^{1,2}(\Omega)/\mathbb{R}$, by *Green's Formula* [122, Thm. III.1.1], one has that

$$\int_\Omega q \operatorname{div} v \, \mathrm{d}\omega = \int_{\partial\Omega} qv \cdot \vec{n} \, \mathrm{d}\sigma - \int_\Omega \overline{\nabla} q \cdot v \, \mathrm{d}\omega. \tag{3.31}$$

This implies that $\overline{\nabla}'v = -\operatorname{div} v$, for $v \in [W_0^{1,2}(\Omega)]^d$, as $v \cdot \vec{n} = 0 \in L^2(\partial\Omega)$, where $\vec{n}$ is the outer normal vector of $\partial\Omega$.

By continuity arguments, Green's formula (3.31) can be extended to functions in $W^{\mathrm{div},2}(\Omega) := \{v \in [L^2(\Omega)]^d : \operatorname{div} v \in L^2(\Omega)\}$. Also, one can continuously extend the domain of definition of the mapping $v \to v \cdot \vec{n} \in L^2(\partial\Omega)|_{\partial\Omega}$ to $W^{\mathrm{div},2}(\Omega)$, see [49, Thms. I.2.5-6].

Thus, the space $H_{\mathrm{df}} := \{v \in [L^2(\Omega)]^d : \operatorname{div} v = 0 \text{ and } \vec{n} \cdot v|_{\partial\Omega} = 0\}$ is well defined. Its orthogonal complement is given by $H_{\mathrm{c}} := j \operatorname{im} J' = \operatorname{im} \overline{\nabla}$, with $\overline{\nabla}$ as defined in (3.30), cf., e.g., [49, Thm. I.2.7]. Since $H_{\mathrm{df}} = \ker J' = \ker \overline{\nabla}'$, cf. [54, Lem. 4.4], these spaces set up the splitting of $H = H_{\mathrm{df}} \oplus H_{\mathrm{c}}$ as established by Lemma 3.6.

We can call on a classical result on *strong solutions* of the Navier-Stokes Equation:

Proposition 3.44 ([140], Lems. 25.1,2)**.** *Consider the setup of Problem 3.1(NSE). Let* V_{df} *be the kernel of* $J_2 = \nabla'$ *and* H_{df} *be the kernel of* $\bar{J}_2 = \overline{\nabla}'$*. If* $f_{\mathrm{df}} \in$

$L^2(0,T;H_{\mathrm{df}})$ and $\alpha \in V_{\mathrm{df}}$, then there exists $0 < T_c \leq T$, and a unique $v \in L^2(0,T_c;V_{\mathrm{df}} \cap W^{2,2}(\Omega)) \cap \mathcal{C}([0,T_c];V_{\mathrm{df}})$ with $\dot{v} \in L^2(0,T_c;V_{\mathrm{df}})$ that fulfills

$$\dot{v}(t) - A(t,v(t)) = f_{\mathrm{df}}(t) \quad \text{in } (V_{\mathrm{df}})',\ \text{a.e. in } (0,T),$$
$$v(0) = \alpha \quad \text{in } H.$$

If the spatial dimension $d = 2$, then $T_c = T$.

Application to our particular setup gives sufficient conditions for Assumption 3.8.

Proposition 3.45. *Consider Problem 3.1(NSE). Then*

(a) Assumption 3.13 holds.

(b) If, in addition, Q is chosen as $W^{1,2}(\Omega)/\mathbb{R} \subset Q_H$, then Assumption 3.5 is fulfilled.

Additionally,

(c) if T is sufficiently small, $f \in L^2(0,T;H')$ and $\alpha \in V_{\mathrm{df}} := \ker J_2$ are sufficient for Assumption 3.8 to hold.

Proof. The existence of the splitting of the variable space, i.e. Assumption 3.13, follows from $\nabla'\colon W_0^{1,2}(\Omega) \to L^2(\Omega)/\mathbb{R}$ fulfilling the *inf-sup* condition [103, Prop. 4.5] and $W_0^{1,2}(\Omega)$ being a Hilbert space, cf. Remarks 3.16 and 3.14. Part (b) follows from Proposition 3.42 and Remark 3.43. Part (c) follows from Proposition 3.44 as follows. Given $f \in L^2(0,T;H')$, by (b) we can consider the components $f_{\mathrm{df}} \in L^2(0,T;H'_{\mathrm{df}})$, that defines a $v_{\mathrm{df}} \in L^2(0,T;V_{\mathrm{df}} \cap W^{2,2}(\Omega))$, and the remainder $f_{\mathrm{c}} = f - f_{\mathrm{df}}$. Since with $v_{\mathrm{df}}(t) \in V \cap W^{2,2}(\Omega)$, it holds that $A(v(t))$ is in H' for almost all $t \in (0,T)$, we can similarly split $A(v(t))$ into a part in H'_{df} and the remainder $A_{\mathrm{c}}(v(t))$. By (a), for almost all $t \in (0,T)$, there exists a unique $p(t) \in Q$, so that $-J_1'p(t) = A_{\mathrm{c}}(v(t)) + f_{\mathrm{c}}(t) \in H'$. There can be no other solution, since by homogeneity of (3.6b), any solution v to must lie in the kernel of J_2 and is thus uniquely defined by Proposition 3.44. □

Thus, if the assumptions of Proposition 3.45 hold, then the decoupling of Lemma 3.30 and the solvability results of Theorem 3.37 apply for the weak formulation of the Navier-Stokes Equation as given in Problem 3.1(NSE).

By now, we have only addressed space regularity of the problem. If the decoupling is possible, the necessary time regularity of the data, that give e.g. $v \in \mathcal{W}(0,T;V;H')$, can be read off from Equations (3.19). In particular for the differential variables, time regularity depends on the nonlinear operator A, specifically, on A_0 as defined in (3.23). In the case of $d = 3$ in the Navier-Stokes Equation, one has

$$A_0\colon L^2(0,T;V_{\mathrm{df}}) \cap L^\infty(0,T;H_{\mathrm{df}}) \to L^{4/3}(0,T;(V_{\mathrm{df}})'),$$

see e.g. [143, Thm. III.3.3]. Accordingly, recalling (3.19), $\dot{v}$ and p in (3.26) are not square integrable in time. This is different from $d = 2$, where A_0 maps into $L^2(0,T;V')$, see [143, Lem. III.3.4]. In the linear case this regularity follows from the boundedness of $A\colon V \to V'$.

Accordingly, the formulation of the optimal control problem and particularly the adjoint ADAE in three spatial dimensions will require $\dot{\lambda}_v \in L^4(0,T;H')$, see Remarks 3.40 and 6.19.

3.6. **Application to the Maxwell Equation.** As a further application example, in this section, we briefly illustrate that the methods developed in the beginning of the current section also apply to a class of equations that model electrodynamical phenomena.

In a time interval $(0, T)$ and on a bounded domain $\Omega \in \mathbb{R}^d$, $d \in \{2, 3\}$, we consider the *Maxwell Equation*

$$\epsilon \dot{E} + \sigma E - \operatorname{curl} H = j, \tag{3.32a}$$

$$\mu \dot{H} + \operatorname{curl} E = 0, \tag{3.32b}$$

$$\operatorname{div}(\epsilon E) = \rho, \tag{3.32c}$$

$$\operatorname{div}(\mu H) = 0, \tag{3.32d}$$

describing the coupled time evolution of an *electric field density* E and the coupled *magnetic field intensity* H for a given current j and a given *charge density* ρ. The material parameters ϵ, μ, and σ represent the *dielectric constant*, the *magnetic permeability*, and the *conductivity*, respectively. See [124] for the basic principles of Maxwell equations.

We complete equations (3.32) by initial conditions

$$E\big|_{\{0\}\times\Omega} = E_0 \quad \text{and} \quad H\big|_{\{0\}\times\Omega} = H_0 \tag{3.33}$$

and boundary conditions that model a perfectly conducting boundary, cf. [120, Ch. 1],

$$n \times E\big|_{(0,T)\times\partial\Omega} = 0 \quad \text{and} \quad n \cdot H\big|_{(0,T)\times\partial\Omega} = 0, \tag{3.34}$$

where n is the outer normal to $\partial\Omega$.

Typically, the applied *charge density* fulfills the *charge conservation law*

$$\dot{\rho} + \operatorname{div}(\sigma E - j) = 0 \quad \text{in } (0, T) \times \Omega.$$

In this case, using the identity $\operatorname{div} \operatorname{curl} = 0$ that holds for smooth vector fields [120, App. B], one can show that for consistent initial values, the algebraic relations (3.32c) and (3.32d) are a-priori fulfilled for all time, cf. [74]. For this reason, in the analytical and numerical analysis of the Maxwell Equation, the constraint for the divergence of the magnetic and the electric field is typically omitted [74, 104, 119, 120].

However, in recent publications [30, 158] the enforcement of the divergence constraint by means of an *Lagrange multiplier* has been proven beneficial for the numerical approximation. We will show, how the results of Section 3 apply to the part of (3.32) that constitutes the magnetic field H and comment on a possible treatment of the part that defines E.

We introduce the following spaces

$$W^{\mathrm{curl},2}(\Omega) := \{v \in [L^2(\Omega)]^d : \operatorname{curl} v \in [L^2(\Omega)]^d\} \quad \text{and}$$
$$W^{\mathrm{div},2}(\Omega) := \{v \in [L^2(\Omega)]^d : \operatorname{div} v \in [L^2(\Omega)]^d\}$$

that are complete with the norms $\|\cdot\|_{W^{\mathrm{curl},2}(\Omega)} := \|\cdot\|_{[L^2(\Omega)]^d} + \|\operatorname{curl} \cdot\|_{[L^2(\Omega)]^d}$ and $\|\cdot\|_{W^{\mathrm{div},2}(\Omega)} := \|\cdot\|_{[L^2(\Omega)]^d} + \|\operatorname{div} \cdot\|_{L^2(\Omega)}$, cf. [120, Ch. 3.5]. Since the trace operator $v \mapsto n \cdot v|_{\partial\Omega} \colon W^{\mathrm{div},2}(\Omega) \to H^{-1/2}(\partial\Omega)$ is well defined, one can use

$$W_0^{\mathrm{div},2}(\Omega) = \{v \in [L^2(\Omega)]^d : \operatorname{div} v \in [L^2(\Omega)]^d, \; n \cdot v|_{\partial\Omega} = 0\}$$

and

$$W_0^{\mathrm{curl},2}(\Omega) = \{v \in [L^2(\Omega)]^d : \operatorname{div} v \in [L^2(\Omega)]^d, \; n \times v|_{\partial\Omega} = 0\}$$

to define the space where the magnetic and electric fields are sought in:

$$V_t := W^{\mathrm{curl},2}(\Omega) \cap W_0^{\mathrm{div},2}(\Omega) \quad \text{and} \quad V_n := W^{\mathrm{curl},2}(\Omega) \cap W_0^{\mathrm{curl},2}(\Omega), \tag{3.35}$$

cf. [120, Ch. 3.8]. It holds that $V_t \hookrightarrow [L^2(\Omega)]^d$, see [120, Cor. 3.49]. With $H := [L^2(\Omega)]^d$, we can consider the Gelfand triple

$$V_t \hookrightarrow [L^2(\Omega)]^d := H \cong H' \hookrightarrow (V_t)'.$$

For the modelling of the divergence constraint (3.32d) we use the same evolution triple 3.28 as for the Navier-Stokes Equation

$$Q := W^{1,2}(\Omega)/\mathbb{R} \hookrightarrow Q_H := L^2(\Omega)/\mathbb{R} \hookrightarrow (W^{1,2}(\Omega)/\mathbb{R})' = Q'. \tag{3.36}$$

Because of the zero normal component of functions in V_t, we can define the divergence $\nabla' \colon V_t \to Q_H$ similarly to $\nabla' \colon V \to Q_H$, cf. (3.27). By definition of V_t, for $v \in V_t$, we have $\nabla' v \in L^2(\Omega)$, and, thus, also $\nabla' v \in L^2(\Omega)/\mathbb{R}$ is well-defined. With this definition of the operator $\nabla' \colon V_t \to Q_H$ and V_t as defined in (3.35), we state the following weak formulation for the time-dependent magnetic field h:

Problem 3.46. *Let $T > 0$ and $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, a domain with a regular boundary. For given $e \in L^2(0, T; V_n)$ and $h_0 \in V_t$, find $h \in L^2(0, T; V_t)$ and $\phi \in L^2(0, T; Q)$, such that*

$$\begin{aligned} \dot{h}(t) + \operatorname{curl} e(t) - \nabla\phi(t) = 0 \quad & \textit{in } (V_t)', \textit{ a.e. in } (0, T), & \text{(3.37a)} \\ \nabla' h = 0 \quad & \textit{in } (L^2(\Omega)/\mathbb{R})', \textit{ a.e. in } (0, T), & \text{(3.37b)} \\ h(0) = h_0 \quad & \textit{in } V_t, & \text{(3.37c)} \end{aligned}$$

holds in the sense described in Section 3.1.

The magnetic permeability μ is a material constant. Thus, it is assumed to be piecewise constant across subdomains [120, Ch. 4.2]. Accordingly, for a single domain, one can set $\mu \equiv 1$ as we did in (3.37a).

We will show that in the continuous case, in (3.37a), the multiplier ϕ will be zero, so that (3.37) is a weak formulation of (3.32b)-(3.32d).

We show that the spaces and operators used in Problem 3.46 fulfill Assumptions 3.5 and 3.13, i.e. the divergence operator and its extension allow for splitting of the solution and the equations. In a second step, we show that Assumption 3.8 holds, i.e. Equation (3.37a) is posed in H' rather then in $(V_t)' \supset H'$. In summary, we prove that we can apply Theorem 3.37 to split the equation (3.37).

Proposition 3.47. *Consider V_t as defined in (3.35) for a smooth domain Ω. The divergence operator $\nabla' \colon V_t \to (L^2(\Omega)/\mathbb{R})'$ has a bounded right inverse.*

Proof. Since $\nabla' \colon [W_0^{1,2}(\Omega)]^d \to (L^2(\Omega)/\mathbb{R})'$ fulfills an *inf-sup* condition [103, Prop. 4.5], it fulfills the *inf-sup condition* if one extends the considered space to $V_t \supset [W_0^{1,2}(\Omega)]^d$. Since the kernel of $\nabla' \colon V_t \to (L^2(\Omega)/\mathbb{R})'$ is closed, the operator is bounded and the claim follows from Remark 3.16. □

Since V_t is a Hilbert space, $\nabla' \colon V_t \to Q_H$ fulfills Assumption 3.13. The extension $\overline{\nabla} \colon H := [L^2(\Omega)]^d \to (W^{1,2}(\Omega)/\mathbb{R})'$ is the same operator as the one considered in Proposition 3.42 and, thus, also Assumption 3.5 is fulfilled.

In particular, we can split $H = [L^2(\Omega)]^d$ and there is projection $\mathcal{P}_{[]} \colon H \to H$ onto the kernel of $\overline{\nabla}'$ with $\mathcal{P}_{[]} j \nabla = 0$ as defined by virtue of Lemma 3.6. For $e \in L^2(0, T; V_n)$ and $\phi \in L^2(0, T; Q_H)$, we obtain that a solution h to (3.37) will be in the kernel of ∇' and thus of $\overline{\nabla}'$ for almost all $t \in (0, T)$.

Since curl maps V_n into the kernel of ∇, cf. [120, Eq. 3.59], we find that for any $w \in V_t$, for any $\phi \in \mathcal{C}_0^\infty(0, T)$, and for any sequence $\{q_n\}_{n \in \mathbb{N}} \subset L^2(0, T; Q)$, with

$q_n \to \pi$ in $L^2(0,T;Q_H)$, by the symmetry of $\mathcal{P}_{[]}$ it holds that

$$\int_0^T \big(v(t), \nabla\pi(t)\big)_H \phi(t)\ \mathrm{d}t =$$
$$\lim_{n\to\infty} \int_0^T \big(v(t), \nabla q_n(t)\big)_H \phi(t)\ \mathrm{d}t =$$
$$\int_0^T \big([I-\mathcal{P}_{[]}]\operatorname{curl} e(t), w\big)_H \phi(t)\ \mathrm{d}t - \int_0^T \big([I-\mathcal{P}_{[]}]h(t), w\big)_H \dot\phi(t)\ \mathrm{d}t = 0.$$

Thus, at a solution, one has that $\nabla\pi = 0$. Consequently, Problem 3.46 is indeed a consistent weak formulation of (3.32b) and (3.32c).

Another immediate consequence of $\nabla\pi = 0$ is that a solution h has its time derivative $\dot h \in L^2(0,T;H)$ so that Assumption 3.8 is fulfilled. By virtue of Proposition 3.47 and Proposition 3.42, the formulation of Problem 3.46 meets all assumptions for the application of Theorem 3.37.

Remark 3.48. Since in Problem 3.46, eventually $\nabla\pi$ will be zero, in the first place, Theorem 3.37 recovers the known fact that the divergence free constraint (3.32d), and in the same way (3.32c), is redundant in theory. Nevertheless, the splitting of the infinite dimensional problem, gives a base for the convergence analysis of spatial discretization where the solution is not divergence free a priori and where the constraint is enforced in a discrete sense.

Remark 3.49. For the treatment of the equations for the electric field e one has to consider the tangential boundary conditions on $n \times e|_{\partial\Omega}$ in the definition of the divergence and its dual. The definition of $\overline{\nabla}'\colon [L^2(\Omega)]^d \to (W_0^{1,2}(\Omega))'$ is possible via Green's formula, cf. (3.31), and gives an operator that fulfills an *inf-sup* condition, cf. [158, Eqn. 3.13] which is readily extended to $[L^2(\Omega)]^d$. By Remark 3.16, in the considered setting, the *inf-sup* condition is equivalent to Assumption 3.13.

However, the definition of ∇' on the space V_t is unclear. Following the derivation for the Navier-Stokes Equation in Section 3.5 and for the magnetic field in the current section, the domain is to be assumed as a subset of $L^2(\Omega)$. Then, however, a definition via Green's formula (3.31) is not possible since it does not respect the tangential values of e. This is not a problem for $\overline{\nabla}$ that maps into $(W_0^{1,2}(\Omega))'$ where a zero trace of the elements of the dual space is well defined.

4. Semi Discretization of the State Equations

In this section we investigate finite-dimensional approximations to the abstract DAEs as defined in Problem 3.1. The approximations are semi-discrete as the time parameter $t \in (0, T)$ remains continuous. Having in mind finite element schemes, we tend to talk of spatial semi-discretizations here, although the used Galerkin approximations are more general.

The approach of discretizing the spatial variable and solving sequences of DAEs in the time variable is commonly referred to as *method of lines.* The other approach - to discretize the time and to solve a sequence of unsteady problems in space - is often called *Rothe's method*, see [133] for general results and [39] for results on the Navier-Stokes Equation.

We will consider separate discretizations of V and Q via sequences of finite-dimensional subspaces $\{V_k\}_{k\in\mathbb{N}}$ and $\{Q_k\}_{k\in\mathbb{N}}$. This so called *mixed Galerkin* approach leads to a standard Galerkin scheme for $V \times Q$ but an *external approximation* to $V_{\mathrm{df}} = \ker J_2$, i.e. $V_{k\mathrm{df}} \subsetneq V_{\mathrm{df}}$, where $V_{k\mathrm{df}}$ is the finite-dimensional approximation to V_{df}. An internal approximation of V_{df}, such that a solution v_k to the semi-discrete equations will always fulfill $J_2 v_k = 0$, via spaces that a priori fulfill the algebraic constraint, is in general impractical [50, p. 267] and intrinsically unstable in the approximation of the algebraic variables [7].

We start by formulating a decoupling of the discrete equations that – in the same way as for the continuous equations – defines separate equations for the differential $v_{k\mathrm{df}}$ and algebraic variables $v_{k\mathrm{c}}$ and p_k.

To show convergence of the discrete approximations to a solution of the continuous problem, we will restrict our analysis to the symmetric case of Problem 3.1, where $J_1 = J_2$. We will apply the theory of external approximation schemes, see [142] for an introduction, in combination with results on standard abstract evolution equations [133] to show that the discrete solutions $v_{k\mathrm{df}}$ converge to solutions v_{df} of the continuous equation (3.21), when $k \to \infty$. We will use the uniform boundedness of the decoupling operators to conclude that also the associated $v_{k\mathrm{c}}$ and p_k converge to v_{c} and p solving the continuous equations (3.19a,c). Finally, from Theorem 3.37 and a discrete counterpart, we will infer that the sequence of solutions (v_k, p_k) converges to (v, p) solving the continuous problem (3.1).

We point out that the external approximation is a nonconforming scheme as defined, e.g., in [53]. In most cases, and in particular in the realm of computational fluid dynamics [150], nonconforming relates to the case where the trial functions are not part of the solution space in terms of regularity as, for example, in the approximation of smooth solutions via discontinuous Galerkin schemes [66]. This is different to our case, where the discrete solutions are not included in the actual solution space, because they do not fulfill an algebraic constraint. Thus, we can rely on the theory of external approximations to establish existence of converging sequences and we can use the norms of the solution space to measure the convergence.

4.1. **Galerkin Approximation.** We start by defining a general Galerkin scheme and related notions and state the semi-discretized system. Having assumed the existence of a Schauder basis of the approximated space, we formalize the approximation properties via projectors.

We define the *abstract Galerkin approximation* property:

Definition 4.1. A sequence of finite-dimensional subspaces $\{W_k\}_{k\in\mathbb{N}}$ of a Banach space W fulfills the *abstract Galerkin approximation* property, if

$$W_0 \subset W_1 \subset \cdots \subset W \quad \text{and} \quad \bigcup_{k\in\mathbb{N}} W_k \text{ is dense in } W. \tag{4.1}$$

In view of approximating solutions in W in a subspace W_k, we will refer to a sequence $\{W_k\}_{k\in\mathbb{N}}$ as *approximation scheme*. If the sequence fulfills (4.1), one has

$$\lim_{k\to\infty} \inf_{v_k\in W_k} \|v - v_k\|_W = 0, \tag{4.2}$$

for all $v \in W$, and we will call it a *Galerkin scheme*, cf. [40, Def. 4.1].

We give sufficient conditions for the existence of these approximation schemes and accompanying projectors that project from W onto the subspace W_k.

Lemma 4.2 ([40], Lem. 4.1.2)**.** *If W is a separable Banach space, then there exists a sequence $\{W_k\}_{k\in\mathbb{N}}$ that has the abstract Galerkin approximation property* (4.1)*.*

In view of establishing the existence of a uniformly bounded projection onto the subspaces, we state a result that holds for spaces W with a *Schauder basis.*

Definition 4.3 (cf. [46], Def. 7.18)**.** A normed space W has a Schauder basis, if there exists a countable subset $\{\phi_i\}_{i\in\mathbb{N}} \subset W$, such that for every element $w \in W$ there exists a unique sequence of scalars $\{s_i\}_{i\in\mathbb{N}}$ with

$$w = \sum_{i=0}^{\infty} s_i\phi_i. \tag{4.3}$$

Lemma 4.4. *Let W be a Banach space with a Schauder basis $\{\phi_i\}_{i\in\mathbb{N}}$. For $k \in \mathbb{N}$, define $W_k := \operatorname{span}\{\phi_0, \cdots, \phi_k\}$ and $\mathcal{P}_{[W_k|\cdot]} \colon W \to W$ via*

$$\mathcal{P}_{[W_k|\cdot]} \colon w \mapsto \sum_{i=0}^{k} s_i\phi_i \tag{4.4}$$

making use of the representation (4.3)*. Then*

(a) *$\{W_{k'}\}_{k'\in\mathbb{N}}$ fulfill* (4.1)*,*
(b) *$\mathcal{P}_k \colon W \to W$ is a projection with $\mathcal{P}_kW = W_k$ and $\|\mathcal{P}_{[W_k|\cdot]}\|_{\mathcal{L}(W,W)}$ is bounded independently of k, and*
(c) *for all $w \in W$, $\|w - \mathcal{P}_{k'}w\|_W \to 0$ as $k' \to \infty$.*

Proof. The approximation property (4.1) of $\{W_k\}_{k\in\mathbb{N}}$ follows from the definition of W_k as the span of the truncated Schauder basis. The projection property and uniform boundedness of $\mathcal{P}_{[W_k|\cdot]}$ is proven in [46, Thm. 7.19]. By (4.3) the series $\sum_{i=0}^{\infty} s_i\phi_i$ converges in the norm of W so that $0 = \lim_{k'\to\infty}\|w - \sum_{i=0}^{k} s_i\phi_i\|_W = \lim_{k'\to\infty}\|w - \mathcal{P}_{k'}w\|_W$. □

Since V and Q, as introduced in Problem 3.1 and Assumption 3.5, are separable, we can assume the existence of sequences of finite-dimensional subspaces $\{V_k\}_{k\in\mathbb{N}}$ and $\{Q_k\}_{k\in\mathbb{N}}$ that fulfill (4.1). There are separable Banach spaces, that do not have a Schauder basis, but all commonly used Banach spaces, as the L^p-spaces, $1 < p < \infty$, do have such a basis, see [25, p. 146] and the references provided there. Thus, the following assumption on the approximation schemes of V and Q is not a severe restriction.

Assumption 4.5. *The approximation schemes $\{V_k\}_{k\in\mathbb{N}}$, $\{Q_k\}_{k\in\mathbb{N}}$ that are used to approximate the Banach spaces V, Q, are chosen such, that for every $k \in \mathbb{N}$, there exists a projector $\mathcal{P}_{[V_k|\cdot]} \colon V \to V$ with $\mathcal{P}_{[V_k|\cdot]}V = V_k$ bounded independently of k and $\|v - \mathcal{P}_{k'}v\|_V \to 0$, as $k' \to \infty$.*

If W is a Hilbert space, then every subspace W_k comes with an orthogonal projection which has always norm 1. In the general separable Banach space, one can prove that there exists a projection that is at least bounded for every k:

Proposition 4.6. *Given a d_k dimensional subspace W_k of a Banach space W. Then there exists a bounded projection $\mathcal{P}_{[W_k|\cdot]}\colon W \to W$ with $\mathcal{P}_{[W_k|\cdot]}W = W_k$.*

Proof. The space W_k is d_k-dimensional. Thus, we can take a basis of normalized vectors $e_1, \cdots, e_{d_k}$ and define functionals $f_j\colon W_k \to \mathbb{R}$ via $f_j(e_i) = \delta_{ij}$, $i, j = 1, \cdots, d_k$. The functionals f_j have norm $\|f_j\|_{W_k'} = 1$ and, by the Hahn-Banach Theorem as stated in [83, Thm. III.1.21], extend to functionals in W' of the same norm. Thus, we can define the projection $\mathcal{P}_{[W_k|\cdot]}\colon W \to W$ via $\mathcal{P}_{[W_k|\cdot]}v = \sum_{i=1}^{d_k} f_i(v)e_i$ that is bounded by the dimension d_k, as can be seen from $\|\mathcal{P}_{[W_k|\cdot]}\|_W \leq \sum_{i=1}^{d_k} \|f_i\|_{W'}\|v\|_W\|e_i\|_W \leq d_k\|v\|_W$. □

We will refer to k as the discretization parameter. A fixed k that realizes a finite-dimensional approximation to Problem 3.1 we will call discretization level.

In finite dimensions all norms are equivalent. Thus, at one discretization level, one may identify, e.g., H_k and V_k. However, in view of asymptotic results for $k \to \infty$, we will always distinguish between H_k and V_k and Q_k and Q_{Hk}. In particular, if we assume an f_k in H_k for all $k \in \mathbb{N}$, this will mean that f_k is bounded in H independently of k.

Noting that $H_k := \overline{V_k}^{\|\cdot\|_H}$ is a closed subspace of H, we find $(H_k)' \cong j'(H_k) =: H_k'$, cf. Corollary 3.21. In the same way, we define $Q_{Hk} := \overline{Q_k}^{\|\cdot\|_{Q_H}}$ and identify $(Q_{Hk})' \cong Q'_{Hk} := j_q'(Q_{Hk}) \hookrightarrow (Q_k)'$. Thus, we can define the discrete Gelfand triples

$$V_k \hookrightarrow H_k \cong H_k' \hookrightarrow (V_k)' \quad \text{and} \quad Q_k \hookrightarrow Q_{Hk} \cong Q'_{Hk} \hookrightarrow (Q_k)', \tag{4.5}$$

with embedding operators that are bounded independently of k.

As in the infinite-dimensional case, there will be necessary conditions for the consistency of the initial values in the discrete approximations. We will characterize them in the end of this section.

We refer to infinite-dimensional formulation given in Problem 3.1 as *continuous problem* and define the *discrete problem* as follows:

Problem 4.7. *Consider the setup of the continuous Problem 3.1 and let $Q \subset Q_H$ be as in Assumption 3.5. Let the approximation schemes $\{V_{k'}\}_{k'\in\mathbb{N}}$, $\{Q_{k'}\}_{k'\in\mathbb{N}}$ to V, Q fulfill* (4.1)*. Let $k \in \mathbb{N}$ and consider the discrete Gelfand triples* (4.5)*. Let the finite-dimensional approximations $f_k \in L^2(0,T;(V_k)')$ and $g_k \in L^2(0,T;Q'_{Hk})$ be the restriction of $f(t)$ and $g(t)$ to V_k and Q_{Hk}, respectively, and let $\alpha_k \in H_k$ be an approximation of $\alpha \in H$.*

Find $v_k \in \mathcal{W}(0,T;V_k,(V_k)')$ and $p_k \in L^2(0,T;Q_k)$ that fulfill

$$\dot{v}_k(t) - A_k(t, v_k(t)) - J'_{1k}p_k(t) = f_k(t) \quad \textit{in } (V_k)',\ \textit{a.e. in } (0,T), \tag{4.6a}$$

$$-J_{2k}v_k(t) = g_k(t) \quad \textit{in } Q'_{Hk},\ \textit{a.e. in } (0,T), \tag{4.6b}$$

$$v_k(0) = \alpha_k \quad \textit{in } H_k, \tag{4.6c}$$

with $J_{2k}\colon V_k \to Q'_{Hk}$ defined via

$$\langle J_{2k}v_k, q_k\rangle := \langle J_2v_k, q_k\rangle \quad \textit{for } v_k \in V_k \textit{ and for all } q_k \in Q_{Hk},$$

with $J'_{1k}\colon Q_{Hk} \to (V_k)'$ defined via

$$\langle J'_{1k}q_k, v_k\rangle := \langle J'_1q_k, v_k\rangle \quad \textit{for } q_k \in Q_{Hk} \textit{ and for all } v_k \in V_k,$$

and with $A_k(t,\cdot)\colon V_k \to (V_k)'\colon v_k \mapsto A(t,v_k)\big|_{V_k}$.

4.2. **Decomposition of the Discrete Solutions.** The existence of a decomposition of the continuous solution space was ensured by Assumption 3.13. This property of V and Q in the interplay with J_2 and J_1 is not necessarily taken over by their discrete approximations. One can construct simple examples but there are also commonly used Galerkin schemes, like the so called $Q_1 - P_0$ scheme for the Navier-Stokes Equation, that fail to fulfill the *LBB condition* [38, 50]. Recall that, in particular cases, the *LBB condition* (3.14) is equivalent to 3.13, cf. Remark 3.16.

Also, to ensure stability of the approximation when k goes to infinity, we need bounds on the decomposing projections that are independent of k. Thus, we will call on the following assumption.

Assumption 4.8. *Consider Problem 3.1 and its finite-dimensional approximation defined in Problem 4.7 with the operators* $J_{2k} \colon V_k \to Q'_{Hk}$ *and* $J'_{1k} \colon Q_H \to (V_k)'$.

- *(a) For all* $k \in \mathbb{N}$, J_{2k} *has a bounded right inverse and* J'_{1k} *is a homeomorphism onto its range, i.e., it has a bounded left inverse*
- *(b) The norms of both inverses are bounded independently of* k.

Remark 4.9. Assumption 4.8(b), i.e. uniformity of the bounds with respect to k, is posed in view of asymptotic results for $k \to \infty$. To establish existence of solutions of Problem 4.7 for a fixed k, Assumption 4.8(a) is sufficient.

We denote the right inverse of J_{2k} by $J_{2k}^- \colon Q'_{Hk} \to V_k$. If we assume existence of J_{2k}^- (Assumption 4.8), then there exists a decomposition

$$V_k = V_{k\mathrm{df}} \oplus V_{k\mathrm{c}}, \tag{4.7}$$

where $V_{k\mathrm{df}} := \ker J_{2k}$ and $V_{k\mathrm{c}}$ is the image of Q'_{Hk} under J_{2k}^-, cf. Lemma 3.17.

Remark 4.10. If for all $k \in \mathbb{N}$, there exists a projection $\mathcal{P}_{[V_k|\cdot]} \colon V \to V$ onto V_k that is bounded independently of k, cf. Assumption 4.5, then, with Assumption 4.8, we have that $[I - J_{2k}^- J_{2k}]\mathcal{P}_{[V_k|\cdot]}$ is a projection from V onto $V_{k\mathrm{df}}$ that is uniformly bounded in k. If, in addition, $J_1 = J_2$, then Assumption 4.8 is equivalent to the *discrete uniform LBB condition* commonly used in the literature [49, Lem. II.1.1].

Remark 4.11. If the discrete operators J_{1k}, J_{2k} fulfill Assumption 4.8, then their continuous representation J_1, J_2 fulfill Assumption 3.13. Consider, for example, $J_{1k}' \colon Q_{Hk} \to (V_k)'$ being bounded from below by a constant c independent of k. Since $\{Q_k\}_{k\in\mathbb{N}}$ fulfills (4.1) and since Q is dense in Q_H, for every $q \in Q_H$ there exists a sequence of $q_k \in Q_k$ so that $q_k \to q$ in Q_H. Then by

$$\|J_1' q\|_{V'} = \lim_{k\to\infty} \|J_{1k}' q_k\|_{V'} \geq \lim_{k\to\infty} c\|q_k\|_{Q_H} = c\|q\|_{Q_H},$$

we find that also J_1 is bounded from below and thus an homeomorphism onto its range. The existence of a bounded right inverse to J_2 is established within the proof of Lemma 4.26.

In view of decoupling the equations, as it was the case for the decoupling of the solution space, the properties (Assumption 3.5) that were the base for the decoupling of the continuous system are not inherited by the discrete formulation. They have to be established for the chosen discretization, cf. [49, Ch. II.1.4]. In finite dimension surjectivity and injectivity of J_{1k} and $\bar{J}_{1k}$ carry over to their extensions. For the decoupling we will assume that $\ker \bar{J}_{2k}$ and $j(\operatorname{im} \bar{J}_{1k}')$ split the space. This gives a discrete version of Assumption 3.5(b). Furthermore, to ensure a stable approximation of the algebraic variables, we assume the splitting projection to be bounded uniformly in k.

Proposition 4.12. *Consider Problem 3.1 and its discrete approximation Problem 4.7, and let Assumption 4.8 hold. Then the discrete representations* $\bar{J}_{1k}' \colon Q_k \to H_k'$

and $\bar{J}_{2k}\colon H_k \to (Q_k)'$ *of the extensions* $\bar{J}_1$, $\bar{J}_2\colon H \to Q'$ *introduced in Assumption 3.5, defined via*

$$\big\langle \bar{J}_{1k}' q_k, h_k \big\rangle_{H',H} := \big\langle \bar{J}_{1k}' q_k, h_k \big\rangle_{H',H}, \qquad \textit{for } q_k \in Q_k \textit{ and for all } h_k \in H_k, \tag{4.8a}$$

and

$$\big\langle \bar{J}_{2k} h_k, q_k \big\rangle_{Q',Q} := \big\langle \bar{J}_2 h_k, q_k \big\rangle_{Q',Q}, \qquad \textit{for } h_k \in H_k \textit{ and for all } q_k \in Q_k, \tag{4.8b}$$

is a homeomorphism onto its range and surjective, respectively. In particular, there exists a constant $\gamma_1(k)$*, such that*

$$\|\bar{J}_{1k}' q_k\|_{H'} \geq \gamma_1(k) \|q_k\|_Q, \qquad \textit{for all } q_k \in Q_k. \tag{4.9}$$

Proof. By Assumption 3.13 the operators J_{1k}, $J_{2k}\colon V_k \to Q_{Hk}$, as defined in Problem 4.7, have the stated properties. Since the considered finite-dimensional spaces are isomorphic these properties are transferred to the extension $\bar{J}_{1k}$, $\bar{J}_{2k}$. In particular, there exists c such that $\|J_{1k} q_k\|_{(V_k)'} \geq c \|q_k\|_{Q_{Hk}}$, for all $q_k \in Q_{Hk}$, cf. Remark 3.16, and with the constants $c_1(k)$, $c_2(k)$ from the norm equivalence of $(V_k)'$ and H_k' and Q_{Hk} and Q_k, respectively, we have that

$$\|\bar{J}_{1k} q_k\|_{H_k'} \geq c_1(k) \|J_{1k} q_k\|_{V_k'} \geq c_1(k) c \|q_k\|_{Q_{Hk}} \geq c_1(k) c_2(k) c \|q_k\|_{Q_k}.$$

Thus, we arrive at (4.9) having defined $\gamma_1(k) := c_1(k) c_2(k) c$ and having used that the norm of the discrete spaces use the same norm as the continuous spaces. □

Assumption 4.13. *Consider Problem 3.1, let Assumption 3.5 hold, and consider* $\bar{J}_{1k}$*,* $\bar{J}_{2k}$ *as defined in* (4.8).

(a) There exists a constant $\gamma_2(k)$*, such that*

$$\|\bar{J}_{2k} h_k\|_{Q'} \geq \gamma_2(k) \|h_k\|_H,$$

for all $h_k \in j(\operatorname{im} \bar{J}_{1k}')$.

(b) There is a constant γ *such that* $\gamma_1(k)$, $\gamma_2(k) \geq \gamma > 0$ *for all* $k \in \mathbb{N}$*, where* $\gamma_1(k)$ *is the constant defined in* (4.9).

Remark 4.14. With Proposition 4.12, Assumption 4.13(a) is the discrete version of Assumption 3.5, cf. Corollary 3.7. Assumption 4.13(b) is posed in view of establishing asymptotic results as $k \to \infty$.

Remark 4.15. If $\bar{J}_{1k} = \bar{J}_{2k}$, then Assumption 4.13 is equivalent to a *discrete uniform LBB condition*, i.e. there exists a γ such that

$$\inf_{0 \neq q_k \in Q_k} \sup_{0 \neq h_k \in H_k} \frac{\langle \bar{J}_{2k}' q_k, h_k \rangle_{H',H}}{\|q_k\|_Q \|h_k\|_H} \geq \gamma > 0,$$

cf. [49, Lem. I.4.1].

Remark 4.16. By Assumption 4.13(a), we have that

$$\|\bar{J}_{2k} j \bar{J}_{1k}' q_k\|_{Q'} \geq \gamma_2 \|\bar{J}_{1k}' q_k\|_{H'} \geq \gamma_2 \gamma_1 \|q_k\|_Q,$$

for all $q_k \in Q_k$, and, thus, that $S_k := \bar{J}_{2k} j \bar{J}_{1k}'\colon Q_k \to (Q_k)'$ is invertible with $\|S_k^{-1}\|_{\mathcal{L}((Q_k)', Q_k)} \leq \frac{1}{\gamma_1 \gamma_2}$. If also Assumption 4.13(b) holds, then S_k^{-1} is bounded independently of k.

Lemma 4.17. *Consider Problem 4.7, let* $\bar{J}_{1k}'$*,* $\bar{J}_{2k}$ *fulfill Assumption 4.13, and define*

$$L_k := \bar{J}_{1k}' S_k^{-1} \bar{J}_{2k}\colon H_k \to H_k'. \tag{4.10}$$

Then one has

$$\begin{aligned}
H_{k\mathrm{df}} &:= \ker \bar{J}_{2k} = \operatorname{im}[I_{H_k} - jL_k] && \text{(a)}\\
H_{k\mathrm{c}} &:= \operatorname{im} jL_k = \operatorname{im} j\bar{J}_{1k}', && \text{(b)}\\
H_k &= H_{k\mathrm{df}} \oplus H_{k\mathrm{c}}, && \text{(c)}\\
H'_{k\mathrm{c}} &:= \operatorname{im} L_k j = j'(H_{k\mathrm{c}}), && \text{(d)}\\
H'_{k\mathrm{df}} &:= \operatorname{im}[I_{H'_k} - L_k j] = j'(H_{k\mathrm{df}}), && \text{(e)}\\
H'_k &= H'_{k\mathrm{df}} \oplus H'_{k\mathrm{c}}, && \text{(f)}
\end{aligned}$$

where $j\colon H' \to H$ is the Riesz isomorphism.

Proof. The same arguments as in the continuous case apply, see Lemma 3.6. □

Remark 4.18. Since a finite-dimensional subspaces $W_k \subset W$ of a Banach space W is closed, its dual $(W_k)'$ can be identified with a subspace of W', cf. Lemma 3.19. Thus, we find that the embedding operators in the discrete triples $V_k \hookrightarrow H_k \hookrightarrow (V_k)'$ and $Q_k \hookrightarrow Q_{H\,k} \hookrightarrow (Q_k)'$ are bounded independently of k.

Remark 4.19. Because for any $k \in \mathbb{N}$, we can identify H'_k with $(V_k)'$, Assumption 3.8 has no meaning in finite dimensions.

By Remark 4.19, Assumptions 4.8 and 4.13 are sufficient for the application of Lemma 3.30 in order to decouple the discrete system (4.6a,b) as follows: Every solution $(v_k, p_k) \in \mathcal{W}(0,T;V_k;(V_k)') \times L^2(0,T;Q_k)$ of System (4.6a,b) can be written as $(v_{k\mathrm{df}} + v_{k\mathrm{c}}, p_k)$, where $v_{k\mathrm{c}} \in L^2(0,T;V_{k\mathrm{c}})$ satisfies

$$v_{k\mathrm{c}}(t) = -J_{2k}^{-} g_k(t), \tag{4.11a}$$

where $v_{k\mathrm{df}} \in \mathcal{W}(0,T;V_{k\mathrm{df}};H'_{k\mathrm{df}})$ is a solution to

$$\begin{aligned}
&\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}\dot{v}_{k\mathrm{df}}(t) -\\
&\qquad \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}A_k\bigl(t, v_{k\mathrm{df}}(t) - J_{2k}^{-} g_k(t)\bigr) = \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}[f_k(t) - \tfrac{d}{dt}(J_{2k}^{-} g_k)(t)],
\end{aligned} \tag{4.11b}$$

and where $p_k \in L^2(0,T;Q_k)$ satisfies

$$p_k(t) + S_k^{-1}\bigl[\bar{J}_{2k} j A_k\bigl(t, v_{k\mathrm{df}}(t) - J_{2k}^{-} g_k(t)\bigr) = \bar{J}_{2k} j f_k(t) + \bar{J}_{2k} j \tfrac{d}{dt}(J_{2k}^{-} g_k(t))\bigr]. \tag{4.11c}$$

Equations (4.11a-c) hold for almost all $t \in (0,T)$ as specified in Lemma 3.30.

Repeating the arguments of Corollary 3.31 and noting that on the discrete level, $\bar{J}_{2k}^{-} : Q'_H \subset Q'_k \to V_{k\mathrm{c}} \subset H_k$ is bounded, so that $\dot{g}_k$ can be assumed in $L^2(0,T;Q'_k)$, cf. Remark 3.29, we obtain a discrete analogue to Theorem 3.37.

Theorem 4.20. *Consider Problem 4.7. Let Assumptions 4.8(a) and 4.13(a) hold, let $f \in L^2(0,T;V')$, $g \in L^2(0,T;Q')$, and $\alpha_k \in H_k$. Then, the semi-discretized ADAE (4.6) has a (unique) solution $(v_k, p_k) \in \mathcal{W}(0,T;V_k;H'_k) \times \mathcal{Q}_k$ if and only if $\dot{g} \in L^2(0,T;Q')$, $\alpha_k = \alpha_{k\mathrm{df}} - \bar{J}_{2k}^{-}(0)g_k(0)$, with $\alpha_{k\mathrm{df}} \in H_{k\mathrm{df}}$, and*

$$\dot{w}_k - \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}A_k(w_k - J_{2k}^{-} g_k) = \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}[f_k - \tfrac{d}{dt}(J_{2k}^{-} g_k)], \quad \textit{in } \mathcal{V}_{k\mathrm{df}}, \tag{4.12a}$$

$$w_k(0) = \alpha_{k\mathrm{df}} \quad \textit{in } H, \tag{4.12b}$$

has a (unique) solution $w_k \in \mathcal{W}(0,T;V_{k\mathrm{df}};H'_{k\mathrm{df}})$.

Proof. The proof is analogous to the continuous case for Theorem 3.37. Note, that the restriction g_k of g, has the same time regularity as g, since by assumption there is a bounded projection onto Q_{Hk}. □

4.3. **Convergence of the (External) Approximation Scheme.** As mentioned in the beginning of this section, the Galerkin scheme that defines the semi-discrete equations in Problem 4.7 leads to an external approximation of the space V_{df}. In view of proving convergence of the solutions $v_{k\mathrm{df}}$ of the discrete inherent differential equation (4.12) to a solution v_{df} of the continuous inherent differential equation (3.21), we need to establish stability and convergence properties of the external scheme.

We will prove that any mixed Galerkin scheme, that fulfills Assumption 4.8 and, thus, by Remark 4.10 any discrete LBB stable mixed finite element scheme, defines a stable external approximation to the subspace V_{df} of the differential variable v_{df}.

We follow the notation of [143].

Definition 4.21. [143, Def. I.3.2] An *external approximation* of a normed space W is a set consisting of

(a) a normed space F and a linear, bounded, and injective *synchronization operator* $\bar{\omega}\colon W \to F$ and

(b) a family of triples $\{W_k, p_k, r_k\}$, in which, for each k,
  - W_k is a normed space,
  - p_k is a linear continuous mapping of W_k into F, and
  - r_k is a (possibly nonlinear) mapping of W into W_k.

As an example, consider a Galerkin scheme $\{V_k\}_{k\in\mathbb{N}}$ to a Banach space V. Then, with $\bar{\omega} := I\colon V \to V$ being the identity, with $p_k := I|_{V_{\mathrm{df}}}\colon V_{\mathrm{df}} \to V$, and with $R_k := \mathcal{P}_{[V_k|\cdot]}$ defined as the projection of V onto V_k, we find that a general Galerkin scheme $\{V_k\}_{k\in\mathbb{N}}$ is included in the definition of an external scheme.

Definition 4.22. cf. [143, Def. I.3.4] An external approximation of the space W as defined in Definition 4.21 is said to be *stable*, if the norm of the prolongation operator $p_k \in \mathcal{L}(W_k, F)$ is bounded independently of k.

Definition 4.23. [143, Def. I.3.6] An external approximation of the space W as defined in Definition 4.21 is said to be *convergent*

(a) if for all $u \in W$,
$$\lim_{k\to\infty} p_k r_k u = \bar{\omega} u \quad \text{in } F,$$

(b) and if for each sequence $\{v_{k'}\}$ of elements in $W_{k'}$, such that $p_{k'}$ converges to some element ϕ weakly in F, there exists a $u \in W$ such that $\phi = \bar{\omega} u$.

We show that the spaces $V_{k\mathrm{df}}$, $k \in \mathbb{N}$, containing the solutions of the discrete inherent differential equation (4.11b) define an external approximation to V_{df}.

Lemma 4.24. *Consider the setup of Problem 4.7 and $\mathcal{P}_{[V_k|\cdot]}$ from Proposition 4.6. If Assumption 4.8 holds, then there exists a splitting $V_k = V_{k\mathrm{df}} \oplus V_{k\mathrm{c}}$, for all $k \in \mathbb{N}$, and the choice of*

$$F := V, \tag{4.13a}$$
$$\bar{\omega} := I\colon V \to V, \tag{4.13b}$$
$$r_k := \mathcal{P}_{[V_{k\mathrm{df}}|V_{k\mathrm{c}}]}\mathcal{P}_{[V_k|\cdot]}\big|_{V_{\mathrm{df}}}\colon V_{\mathrm{df}} \to V_{k\mathrm{df}}, \tag{4.13c}$$
$$p_k := I\colon V_{k\mathrm{df}} \to V \tag{4.13d}$$

defines an external approximation scheme to V_{df} as illustrated in Figure 2.

$$\begin{tikzcd}
V_{\mathrm{df}} \arrow[rr, "\bar{\omega}"] \arrow[dr, "r_k"] & & V \\
& V_{k\mathrm{df}} \arrow[ur, "p_k"'] &
\end{tikzcd}$$

FIGURE 2. Illustration of the external approximation of V_{df} via $\{V_{k\mathrm{df}}\}_{k\in\mathbb{N}}$.

Proof. Assumption 4.8 provides the splitting $V_k = V_{k\mathrm{df}} \oplus V_{k\mathrm{c}}$ and the projector $\mathcal{P}_{[V_{k\mathrm{df}}|V_{k\mathrm{c}}]} \colon V_k \to V_k$. Proposition 4.6 states the existence of $\mathcal{P}_{[V_k|\cdot]} \colon V \to V_k$ and Assumption 3.13 implies that $\mathcal{P}_{[V_k|\cdot]}\big|_{V_{\mathrm{df}}}$ is well defined. Thus, we can define $r_k := \mathcal{P}_{[V_{k\mathrm{df}}|V_{k\mathrm{c}}]}\mathcal{P}_{[V_k|\cdot]}\big|_{V_{\mathrm{df}}}$. Since $\bar{\omega}$ and p_k are the identity operators, they are linear, injective, and bounded. Thus, the proposed scheme fulfills Definition 4.21. □

To prove that the external approximation scheme defined in Lemma 4.24 is convergent, we use that J_{2k} has a right inverse bounded uniformly in k.

Lemma 4.25. *Consider the setup of Problem 4.7, assume that J_{2k} fulfills Assumption 4.8, and that $\{V_k\}_{k\in\mathbb{N}}$ fulfills Assumption 4.5. Then the external approximation scheme* (4.13) *is stable and convergent.*

Proof. As $p_k := I$ is bounded independently of k, the external approximation (4.13) is stable. To establish convergence, we need to show that parts (a) and (b) of Definition 4.21 hold.

Ad (a):

With $\bar{\omega}$ and p_k chosen as the identities in V, we have to show that for $v \in V_{\mathrm{df}}$, $\|v - r_k v\|_V \to 0$ as $k \to \infty$. Let $v \in V_{\mathrm{df}}$ and consider $\mathcal{P}_{[V_k|\cdot]}v \in V_k$. With $V_k = V_{k\mathrm{df}} \oplus V_{k\mathrm{c}}$, we have

$$\mathcal{P}_{[V_k|\cdot]}v = v_{k\mathrm{df}} + v_{k\mathrm{c}} \quad \text{and} \quad r_k v = \mathcal{P}_{[V_k|\cdot]}v - v_{k\mathrm{c}},$$

with $v_{k\mathrm{df}} \in V_{k\mathrm{df}}$ and $v_{k\mathrm{c}} \in V_{k\mathrm{c}}$.

Since $J_{2k}v = 0$ and $J_{2k}v_{k\mathrm{df}} = 0$, we have $J_{2k}[-v_{k\mathrm{c}}] = J_{2k}[v - \mathcal{P}_{[V_k|\cdot]}v]$. By Assumption 4.8, J_{2k} has a uniformly bounded right inverse. This implies that $J_{2k}|_{V_{k\mathrm{c}}}$ is uniformly bounded from below by a constant γ. Thus,

$$\|v_{k\mathrm{c}}\|_V \le \frac{1}{\gamma}\|J_{2k}v_{k\mathrm{c}}\|_{Q'_{Hk}} = \frac{1}{\gamma}\|J_{2k}[v - \mathcal{P}_{[V_k|\cdot]}v]\|_{Q'_{Hk}} \le \frac{\|J_2\|_{\mathcal{L}(V,Q'_H)}}{\gamma}\|v - \mathcal{P}_{[V_k|\cdot]}v\|_V,$$

where we have used that on $V_{k\mathrm{c}}$ the norm of J_{2k} is majorized by the norm of J_2 since $Q'_{Hk} \subset Q'_H$. Thus, we obtain that

$$\|v - r_k v\|_V \le \|v - \mathcal{P}_{[V_k|\cdot]}v\|_V + \|v_{k\mathrm{c}}\|_V \le (1 + \frac{\|J_2\|_{\mathcal{L}(V,Q'_H)}}{\gamma})\|v - \mathcal{P}_{[V_k|\cdot]}v\|_V \to 0,$$

as $k \to \infty$, using the approximation property of $\mathcal{P}_{[V_k|\cdot]}$ established in Assumption 4.5.

Ad (b):

Let $v_{k\mathrm{df}} \in V_{k\mathrm{df}}$, for $k \in \mathbb{N}$, and $v_{k\mathrm{df}} \rightharpoonup \phi \in F$. We need to show that there exists a $v_{\mathrm{df}} \in V_{\mathrm{df}}$, such that $\phi = \bar{\omega}v_{\mathrm{df}} = v_{\mathrm{df}}$. Since $\bar{\omega}$ is the identity, there exists $v \in V$, such that $\phi = v$. We show that $v \in V_{\mathrm{df}}$, i.e. $J_2 v = 0$. Since $\langle v_{k\mathrm{df}} - \phi, f\rangle_{V,V'} \to 0$, as $k \to \infty$, for all $f \in V'$, this convergence is in particular valid for $f \in \operatorname{im} J'_2 \subset V'$. Since J'_2 is an homeomorphism onto its range, cf. Remark 4.11, we find that for all $q \in Q_H$ we have

$$\langle J_2[v_{k\mathrm{df}} - \phi], q\rangle_{Q'_H,Q_H} = \langle v_{k\mathrm{df}} - \phi, J'_2 q\rangle_{V,V'} \to 0,$$

as $k \to \infty$. We show that $\lim_{k\to\infty}\langle J_2 v_{k\mathrm{df}}, q\rangle_{Q'_H,Q_H} = 0$ for all $q \in Q_H$, to get that $J_2 v = J_2\phi = 0 \in Q'_H$. As Q is dense in Q_H, the scheme $\{Q_k\}_{k\in\mathbb{N}}$ has property (4.1) also with respect to Q_H. Thus for any $q \in Q_H$ there exists a sequence $\{q'_k\}_{k'\in\mathbb{N}}$, with $q_{k'} \in Q_{k'}$, that converges to q in Q_H. By definition of V_{df}, we have $\langle J_2 v_{k\mathrm{df}}, q_{k'}\rangle_{Q'_H,Q_H} = 0$ for all $k' \le k$. Thus we obtain for given $q \in Q_H$ that

$$|\langle J_2 v_{k\mathrm{df}}, q\rangle_{Q'_H,Q_H}| = |\langle J_2 v_{k\mathrm{df}}, q - q_k\rangle_{Q'_H,Q_H}| \le \|J_2\|_{\mathcal{L}(V,Q'_H)}\|v_{k\mathrm{df}}\|_V\|q-q_k\|_{Q_H} \to 0$$

with $k \to \infty$, since $\{v_{k\mathrm{df}}\}_{k\in\mathbb{N}}$ as a weakly convergent sequence is bounded.

□

4.4. **Convergence of the Galerkin Approximations.** In this section we will establish conditions for the convergence of solutions (v_k, p_k) of (4.6) to (v, p) solving (3.1).

We will separate the Galerkin approximation v_k into the differential $v_{k\mathrm{df}}$ and algebraic $v_{k\mathrm{c}}$ part, see Theorem 4.20 for the discrete case.

To prove convergence of $v_{k\mathrm{c}} := -J_{2k}^{-} g_k$, we will establish a continuity in the choice of J_{2k}^{-}, which is both possible and necessary, since the right inverse is not unique.

Then, assuming that $J_1 = J_2$, we prove weak convergence of the approximations for the differential parts as defined in Theorems 3.37 and 4.20. Having formulated pseudomonotonicity and semi-coerciveness conditions for the shifted nonlinear operator as it results from the decoupling procedure, we will extend standard results on (internal) Galerkin approximations to the considered setup. If the continuous problem is uniquely solvable, we can drop the limiting factor, that the convergence holds only for subsequences.

As a direct consequence of the stability assumptions, we will also obtain convergence of the approximations of the algebraic variable p_k.

We start by establishing convergence in $v_{k\mathrm{c}}$:

Lemma 4.26. *Consider Problem 4.7 and let Assumption 4.8 hold. If the sequence $\{g_k\}_{k\in\mathbb{N}} \subset \mathcal{W}(0,T;Q_H;Q')$ converges to g strongly in $L^2(0,T;Q_H)$, then we can choose J_{2k}^{-}, $k \in \mathbb{N}$, and J_2^{-}, such that*

$$-J_{2k}^{-} g_k =: v_{k\mathrm{c}} \to v_{\mathrm{c}} := J_2^{-} g$$

in the norm of $L^2(0,T;V)$. If, in addition, $\frac{d}{dt}(J_2^{-}g) \in L^2(0,T;H')$, then

$$-\tfrac{d}{dt}(J_{2k}^{-} g_k) = \dot v_{k\mathrm{c}} \rightharpoonup \dot v_{\mathrm{c}} = \tfrac{d}{dt}(J_2^{-} g)$$

in $L^2(0,T;H')$.

Proof. We start by proving that it is possible to choose J_{2k}^{-} such that they converge to a J_2^{-}, where J_2^{-} is a right inverse of J_2 and such that $V_{k\mathrm{c}} \subset V_{k+1\mathrm{c}} \subset \cdots \subset V_{\mathrm{c}}$. Then the choice of J_2^{-} as a limit of the J_{2k}^{-} will give the convergence of $v_{k\mathrm{c}}$ to v_{c}.

We will use induction and we assume that $\dim Q_k = k$. The case that the growth in dimension of Q_k is not incremental can be covered by doing the corresponding number of induction steps at once.

The proof is constructive – start with any $k \in \mathbb{N}$.

By Assumption 4.8, there exists a right inverse $J_{2k}^{-} : Q'_{Hk} \to V_{k\mathrm{c}}$ to J_{2k} bounded by a constant γ independent of k. Consider the next discretization level $k+1$. Again, by assumption there exists a right inverse $\tilde J_{2k+1}^{-} : Q'_{Hk+1} \to V_{k+1\mathrm{c}}$ of J_{2k+1} bounded by the same constant. We have that $J_{2k+1}(V_{k\mathrm{c}}) = Q'_{Hk} \subset Q'_{Hk+1}$ is a k-dimensional subspace. Thus, there exists a $q \in Q'_{Hk+1}$ such that $Q'_{Hk+1} = J_{2k+1}(V_{k\mathrm{c}}) \oplus \mathrm{span}\{q\}$. Let $w := \tilde J_{2k+1}^{-} q \in V_{k+1}$. This w exists since $Q'_{Hk+1} = J_{2k+1}(V_{k+1})$. For $\tilde q \in Q'_{Hk+1}$, written as $\tilde q = q_k + \alpha q$, with $q_k \in Q'_{Hk}$ and $\alpha \in \mathbb{R}$, define the alternative right inverse ${J_2}^-_{k+1}$ via ${J_2}^-_{k+1}\tilde{q} = {J_2}^-_k q_k + \alpha w$. Since ${J_2}^-_k$ and $\tilde{J}{}_{2}{}^-_{k+1}$ are bounded by γ, also ${J_2}^-_{k+1}$ is bounded by γ. By construction, we have that $V_{k\mathrm{c}} \subset {J_2}^-_{k+1}(V_{k+1_\mathrm{c}})$. Set $V_{k+1_\mathrm{c}} := {J_2}^-_{k+1}({Q'_H}_{k+1})$ and proceed to $k+2$.

With this procedure we can construct the sequences of ${J_2}^-_k$ and $V_{k\mathrm{c}}$, such that

$$V_{k\mathrm{c}} \subset V_{k+1_\mathrm{c}} \subset \cdots . \tag{4.14}$$

Next, we show that these ${J_2}^-_k$ converge to a right inverse of $J_2 \colon V \to Q'_H$, so that we can define V_c as the superspace of all $V_{k\mathrm{c}}$.

Define $J_2^- : Q'_H \to V$ as follows. Let $q \in Q'_H$, then there exists $\{q_k\}_{k\in\mathbb{N}} \subset Q'_H$, with $q_k \in {Q'_H}_k$ and $q_k \to q$, as $k \to \infty$, to give

$$J_2^- q = \lim_{k\to\infty} {J_2}^-_k q_k \tag{4.15}$$

Since ${J_2}^-_k$ is uniformly bounded, this limit exists and J_2^- is bounded.

Now, we show that $J_2 {J_2}^-_k q_k \to q \in Q'_H$. Consider the orthogonal projection $\mathcal{P}_{[{Q'_H}_k]} \colon Q'_H \to Q'_H$ onto ${Q'_H}_k$ and that ${J_2}_k w_k = J_2 w_k$ on ${Q_H}_k$, for any $w_k \in V_k$. Then

$$\begin{aligned}
\lim_{k\to\infty} \|J_2 {J_2}^-_k q_k\|^2_{Q'_H} &= \lim_{k\to\infty} \|\mathcal{P}_{[{Q'_H}_k]} J_2 {J_2}^-_k q_k\|^2_{Q'_H} + \lim_{k\to\infty} \|\mathcal{P}_{[{Q'_H}_{k\perp}]} J_2 {J_2}^-_k q_k\|^2_{Q'_H} \\
&= \lim_{k\to\infty} \|{J_2}_k {J_2}^-_k q_k\|^2_{Q'_H} \\
&= \|q\|^2_{Q'_H}
\end{aligned} \tag{4.16}$$

since J_2, ${J_2}^-_k$, and q_k are uniformly bounded and since $\mathcal{P}_{[{Q'_H}_{k\perp}]} \to 0$, as $k \to \infty$.

Thus, we have that J_2^-, as defined in (4.15), is a bounded right inverse to $J_2 \colon V \to Q'_H$ and, by Lemma 3.17, we can define the splitting $V = V_\mathrm{df} \oplus V_\mathrm{c}$ with $V_\mathrm{c} = \operatorname{im} J_2^- \supset V_{k\mathrm{c}}$, for all $k \in \mathbb{N}$.

Since $J_2 \colon V \to Q'_H$, $J_2^- : Q'_H \to V_\mathrm{c}$, and ${J_2}^-_k : {Q'_H}_k \to V_{k\mathrm{c}}$ are linear and bounded, their extension to mappings on abstract functions $J_2 \colon L^2(0,T;V) \to \mathcal{Q}_\mathcal{H}$ and $J_2^- : \mathcal{Q}'_\mathcal{H} \to L^2(0,T;V)_\mathrm{c}$, and ${J_2}^-_k : \mathcal{Q}'_{\mathcal{H}k} \to L^2(0,T;V_{k\mathrm{c}})$ are bounded by the same constants. Using the same arguments as in (4.16) and that $g_k \to g$ in $\mathcal{Q}'_\mathcal{H}$, as $k \to \infty$, we obtain that $J_2 v_{k\mathrm{c}} = -J_2 {J_2}^-_k g_k \to -g = J_2 v_\mathrm{c}$ in $\mathcal{Q}'_\mathcal{H}$. By boundedness of J_2^- we obtain that $\mathcal{P}_{[V_\mathrm{c}|V_\mathrm{df}]} v_{k\mathrm{c}} \to \mathcal{P}_{[V_\mathrm{c}|V_\mathrm{df}]} v_\mathrm{c}$ in $L^2(0,T;V)_\mathrm{c}$. Since by construction, for every k, $V_{k\mathrm{c}} \subset V$, we also have that $v_{k\mathrm{c}} \to v_\mathrm{c}$ in $L^2(0,T;V)$, as $k \to \infty$.

Having established the convergence of $v_{k\mathrm{c}} \to v_\mathrm{c}$ and having assumed that $\dot{v}_\mathrm{c} \in L^2(0,T;H')$, we will now prove weak convergence in the derivatives. By Lemma 3.28 we have that g has a time derivative $\dot{g} \in L^2(0,T;Q')$. That is why also its restriction has a time derivative $\dot{g}_k \in L^2(0,T;(Q_k)')$. Because of the equivalence of the norms in finite-dimensional spaces, we have that $\frac{d}{dt}({J_2}^-_k g_k) \in L^2(0,T;H'_k) \subset L^2(0,T;H)$ for any $k \in \mathbb{N}$. Thus, for all $v \in H$ and all $\phi \in \mathcal{C}_0^\infty(0,T)$, we have, by the definition of the weak derivative as in (3.7bb), that

$$\left\langle \tfrac{d}{dt}({J_2}^-_k g_k)\phi, v\right\rangle_{\mathcal{H}',\mathcal{H}} = -\big({J_2}^-_k g_k \dot{\phi}, v\big)_\mathcal{H} \to -\big(J_2^- g\dot{\phi}, v\big)_\mathcal{H} = \left\langle \tfrac{d}{dt}(J_2^- g)\phi, v\right\rangle_{\mathcal{H}',\mathcal{H}} ,$$

as $k \in \mathbb{N}$, since $-{J_2}^-_k g_k \to -J_2^- g$ in $L^2(0,T;V)$ as proven in the first part of the lemma. As this relation holds for all $\phi \in \mathcal{C}_0^\infty(0,T)$, by the *Fundamental Lemma of Calculus of Variations* as given, e.g., in [40, Lem. 3.1.5], the pointwise convergence of $\dot{v}_{k\mathrm{c}}$ to $\dot{v}_\mathrm{c}$ is shown, i.e. for all $w \in H$,

$$\left\langle \dot{v}_{k\mathrm{c}}(t), w\right\rangle_{H',H} \to \left\langle \dot{v}_\mathrm{c}(t), w\right\rangle_{H',H} \tag{4.17}$$

for almost all $t \in (0,T)$. Since, by definition, the space of *simple functions* in $\big((0,T) \to H\big)$, namely the space of functions that are piecewise constant on measurable subsets of $(0,T)$, is dense in $L^2(0,T;H)$, see [133, Def. IV.1.5] and since $v_\mathrm{c} \in L^2(0,T;H')$, we conclude from (4.17) that $\dot{v}_{k\mathrm{c}} \rightharpoonup v_\mathrm{c}$ in $L^2(0,T;H')$. □

Remark 4.27. The particular choice of the right inverse to J_2 is needed to prove the individual convergence of the separated parts. This should be only a theoretical concern. In practise, at a discretization level k, the use of $J_{2_k}^-$ is in general not advisable [7]. Nevertheless, the solutions $v_k = v_{k\mathrm{df}} + v_{k\mathrm{c}}$ converge to v independent of the choice of $J_{2_k}^-$ at a particular level. Thus, if required, the choice of $J_{2_k}^-$ can be done with respect to, e.g., stability or efficiency issues.

Having, in theory, established convergence in the part $v_{k\mathrm{c}}$, we will now investigate the behavior of the differential parts of the solution. The nonlinearity $A_k(t, \cdot + v_{k\mathrm{c}}(t))$ in the differential equation (4.11b) appears as a shifted operator, with a shift $v_{k\mathrm{c}}$ that varies for every $k \in \mathbb{N}$. We will assume properties of the nonlinearity in the Galerkin approximations that are uniform with respect to these shifts.

Assumption 4.28. *Consider Problem 3.1(Sym) and its finite-dimensional approximation defined in Problem 4.7. Assume that Assumption 4.8 holds, such that we can define the sequence* $\{v_{k\mathrm{c}}\}_{k\in\mathbb{N}}$*, where* $v_{k\mathrm{c}} := -J_{2_k}^- g_k \in L^2(0,T;V_{k\mathrm{c}})$*, that converges to* $v_\mathrm{c} := -J_2^- g$*, cf.* (4.11a) *and Lemma 4.26. For* $k \in \mathbb{N}$ *define*

$$A_{0_k} \colon L^2(0,T;V) \to \big((0,T) \to V'\big) \colon v \mapsto \mathcal{N}_A(v + v_{k\mathrm{c}}), \tag{4.18}$$

via the Nemyckij map of $A\colon (0,T) \times V \to V'$.

We will add the shift to the explicit dependence of A *on* t *and write* $A_{0_k}\colon (0,T)\times V \to V'$ *and* $A_{0_k}(t, v(t)) \in V'$.

We assume that $A_{0_k}\colon L^2(0,T;V) \to L^2(0,T;V')$ *is bounded and that it has the following properties:*

(a) Bounded Growth: *There is* $\gamma \in L^1(0,T)$ *and* $\beta\colon \mathbb{R} \to \mathbb{R}$*, increasing, such that for all* $v \in V$ *with* $A_{0_k}(t,v) \in H'$*:*

$$\|A_{0_k}(t,v)\|_{H'} \le \beta(\|v\|_H)(\gamma(t) + \|v\|_H), \tag{4.19}$$

(b) $v_{k\mathrm{c}}$-uniform Semi-Coerciveness: *There is* $c_0 > 0$*,* $c_1 \in L^2(0,T)$*, and* $c_2 \in L^1(0,T)$*, such that for all* $v \in V$*:*

$$\big\langle A_{0_k}(t,v), v\big\rangle_{V',V} \ge c_0\|v\|_V^2 - c_1(t)\|v\|_V - c_2(t)\|v\|_H^2, \tag{4.20}$$

and

(c) $v_{k\mathrm{c}}$-uniform Pseudomonotonicity, cf. [133, Def. 2.1]: *The operator* A_{0_k} *is bounded and, for given* $\{u_k\}_{k\in\mathbb{N}} \subset L^2(0,T;V)$*, with* $u_k \rightharpoonup u \in L^2(0,T;V)$*, if*

$$\limsup_{k\to\infty} \big\langle A_{0_k}(u_k), u_k - u\big\rangle_{\mathcal{V}',\mathcal{V}} \ge 0$$

it follows that for any $w \in L^2(0,T;V)$*,*

$$\big\langle A(u + v_\mathrm{c}), u - w\big\rangle_{\mathcal{V}',\mathcal{V}} \le \liminf_{k\to\infty} \big\langle A_{0_k}(u_k), u_k - w\big\rangle_{\mathcal{V}',\mathcal{V}}. \tag{4.21}$$

The assumed bounds in (a) and (b) are independent of $k \in \mathbb{N}$ *and hold for almost all* $t \in (0,T)$.

Note that the notions of *bounded growth, semi-coerciveness*, and *pseudomonotonicity* are adjusted to the particular setup and, thus, are different from the definitions in [133]. For the v_k*-uniform pseudomonotonicity* we have the following sufficient condition:

Lemma 4.29. *If* $A\colon L^2(0,T;V) \to L^2(0,T;V')$ *is pseudomonotone and weakly continuous then* A_{0_k}*, as defined in* (4.18)*, is* v_k*-uniform pseudomonotone as defined in* (4.21)*.*

Proof. Consider sequences and limits $u_k \rightharpoonup u$ and $v_k \to v$ in $L^2(0,T;V)$. We write $A_{0_k}(u_k)$ as $A(u_k+v_k)$. For a constant shift, i.e. $v_k = v$, for all $k \in \mathbb{N}$, a similar result was proven in [133, Lem. 2.11]. By weak continuity of A we have that

$$\begin{aligned}\limsup_{k\to\infty}\big\langle A(u_k+v_k), u_k-u\big\rangle_{\mathcal{V}',\mathcal{V}} &= \limsup_{k\to\infty}\big\langle A(u_k+v_k), u_k+v_k-u-v_k\big\rangle_{\mathcal{V}',\mathcal{V}}\\ &= \limsup_{k\to\infty}\big\langle A(u_k+v_k), u_k+v_k-u-v\big\rangle_{\mathcal{V}',\mathcal{V}}.\end{aligned}$$

In fact, since the sequence $A(u_k+v_k) \rightharpoonup A(u+v)$ in $L^2(0,T;V')$ and $v_k \to v$ in $L^2(0,T;V)$, by [161, Prop. 21.23(k)] the limit of $\big\langle A(u_k+v_k), u+v_k\big\rangle_{\mathcal{V}',\mathcal{V}}$, as $k\to\infty$, exists and thus $\limsup_{k\to\infty}\big\langle A(u_k+v_k), u+v_k\big\rangle_{\mathcal{V}',\mathcal{V}} = \big\langle A(u+v), u+v\big\rangle_{\mathcal{V}',\mathcal{V}} = \limsup_{k\to\infty}\big\langle A(u_k+v_k), u+v\big\rangle_{\mathcal{V}',\mathcal{V}}$.

This means that if $\limsup_{k\to\infty}\big\langle A_{0_k}(u_k), u_k-u\big\rangle_{\mathcal{V}',\mathcal{V}} \geq 0$, then it holds that also $\limsup_{k\to\infty}\big\langle A(u_k+v_k), u_k+v_k-u-v\big\rangle_{\mathcal{V}',\mathcal{V}} \geq 0$. Thus, by the pseudomonotonicity of A, we find that for any $w \in L^2(0,T;V)$,

$$\big\langle A(u+v), u+v-w\big\rangle_{\mathcal{V}',\mathcal{V}} \leq \liminf_{k\to\infty}\big\langle A(u_k+v_k), u_k+v_k-w\big\rangle_{\mathcal{V}',\mathcal{V}},$$

or

$$\begin{aligned}\big\langle A_0(u), u-w\big\rangle_{\mathcal{V}',\mathcal{V}} \leq{}& \liminf_{k\to\infty}\big\langle A_{0_k}(u_k), u_k-w\big\rangle_{\mathcal{V}',\mathcal{V}}+\\ &+\lim_{k\to\infty}\big\langle A(u_k+v_k), v_k\big\rangle_{\mathcal{V}',\mathcal{V}} - \big\langle A(u+v), v\big\rangle_{\mathcal{V}',\mathcal{V}}\ .\end{aligned}$$

Since $\lim_{k\to\infty}\big\langle A(u_k+v_k), v_k\big\rangle_{\mathcal{V}',\mathcal{V}} - \big\langle A(u+v), v\big\rangle_{\mathcal{V}',\mathcal{V}} = 0$, we can conclude that A_{0_k} is v_k*-uniform* pseudomonotone. □

Corollary 4.30. *If* $A\colon L^2(0,T;V) \to L^2(0,T;V')$ *is strongly continuous, then* A_{0_k} *as defined in* (4.18) *is* v_k*-uniform pseudomonotone as defined in* (4.21)*.*

Proof. Since every strongly convergence sequence is weakly convergent, strong continuity implies weak continuity. Also, strong continuity of A implies pseudomonotonicity of A, so that the assumptions of Lemma 4.29 are fulfilled. □

Remark 4.31. We will show convergence of the approximations from the discrete differential equations from Theorem 4.20 only for the symmetric problem, where $J_1 = J_2$, and, thus, $J_{1k} = J_{2k}$. The general case is not covered here since we have not established pointwise convergence of the projections $\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} \to \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}$, as $k\to\infty$.

In particular, in the proof of Theorem 4.36 we call on the convergence of

$$\big\langle \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}f_k, u_k\big\rangle_{\mathcal{H}',\mathcal{H}} \to \big\langle \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}f, u\big\rangle_{\mathcal{H}',\mathcal{H}}, \quad \text{as } k\to\infty, \tag{4.22}$$

given $\{u_k\}_{k\in\mathbb{N}} \subset L^2(0,T;V)$, $u_k \in V_{k\mathrm{df}}$ for all $k \in \mathbb{N}$, and $u_k \rightharpoonup u$ in $L^2(0,T;V)$.

As a consequence of the *Closed Range Theorem*, see e.g. [83, Thm. IV.5.13], the dual of $\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}\colon H'_k \to H'_k$ is given as $\mathcal{P}'_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} = \mathcal{P}_{[(H'_{k\mathrm{c}})^0|(H'_{k\mathrm{df}})^0]}$, see [83, p. 156]. If $(H'_{k\mathrm{c}})^0 = H_{k\mathrm{df}}$ and $(H'_{\mathrm{c}})^0 = H_{\mathrm{df}}$, then we obtain the convergence in (4.22) via

$$\begin{aligned}\lim_{k\to\infty}\big\langle \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}f_k, u_k\big\rangle_{\mathcal{H}',\mathcal{H}} = \lim_{k\to\infty}\big\langle f_k, u_k\big\rangle_{\mathcal{H}',\mathcal{H}} &= \big\langle f, u\big\rangle_{\mathcal{H}',\mathcal{H}}\\ &= \big\langle f, \mathcal{P}_{[(H'_{\mathrm{c}})^0|(H'_{\mathrm{df}})^0]}u\big\rangle_{\mathcal{H}',\mathcal{H}}\\ &= \big\langle \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}f, u\big\rangle_{\mathcal{H}',\mathcal{H}},\end{aligned}$$

since by Lemma 3.18 it holds that $V_{k\mathrm{df}} \subset H_{k\mathrm{df}}$ and $V_{\mathrm{df}} \subset H_{\mathrm{df}}$, i.e. u_k and u are already in the range of the dual projections.

The preceding remark motivates the following lemma:

Lemma 4.32. *Consider the symmetric problem 3.1(Sym), where $J_1 = J_2$, and its finite-dimensional approximation defined in Problem 4.7. If Assumptions 4.13 and Assumption 3.5 hold, such that $H = H_{\mathrm{df}} \oplus H_{\mathrm{c}}$ and $H_k = H_{k\mathrm{df}} \oplus H_{k\mathrm{c}}$ as defined in Lemma 3.6 and Lemma 4.17, then*

$$(H_{k\mathrm{c}}{}')^0 = H_{k\mathrm{df}}, \text{ for all } k \in \mathbb{N}, \text{ and } (H_{\mathrm{c}}')^0 = H_{\mathrm{df}}.$$

Proof. Since all considered spaces are closed subspaces of the reflexive space H, by the *Closed Range Theorem*, see e.g. [83, Thm. IV.5.13], and by Lemma 3.6 with $J_1 = J_2$ we have that

$$H_{\mathrm{df}} = ((H_{\mathrm{df}})^0)^0 = ((\ker J_2)^0)^0 = (\operatorname{im} J_2')^0 = (H_{\mathrm{c}}')^0.$$

Analogous arguments hold in particular for the finite-dimensional case. □

Before we show the convergence of the discrete solutions, we state convergence in the right hand sides that is, in particular, a necessary condition for the convergence of the algebraic parts $v_{k\mathrm{c}}$, cf. Lemma (4.14).

Proposition 4.33. *Consider Problem 4.7. For the restrictions f_k and g_k of $f \in L^2(0,T;H')$ and $g \in L^2(0,T;Q_H')$ onto H_k and Q_{Hk}, respectively, we have that $f_k \to f \in L^2(0,T;H')$ and $g_k \to g \in L^2(0,T;Q_H')$, as $k \to \infty$.*

Proof. Since V, Q are dense in H, Q_H, the schemes $\{V_k\}_{k\in\mathbb{N}}$, $\{Q_k\}_{k\in\mathbb{N}}$ have the *abstract Galerkin approximation* property (4.1) also for the Hilbert spaces H, Q_H. The strong convergence then follows from the existence of orthogonal and, thus, uniformly bounded projections onto the finite-dimensional subspaces. □

As mentioned in Remark 4.19, Assumption 3.8 stating that $A(t, v(t))$ is in H' rather than in V' does not make sense on finite-dimensional spaces. Instead, we assume smoothness in the equations uniformly in k:

Assumption 4.34. *If f is in $L^2(0,T;H')$ and if $g \in L^2(0,T;Q')$ and $\alpha_k \in H_k$ are sufficiently smooth, then any solution (v_k, p_k) to (4.6) is such that $\|A(t, v_k(t))\|_{H'} \leq c_1$ and $\|p_k(t)\|_Q \leq c_2$ almost everywhere on $(0,T)$ and with constants c_1, c_2 independent of k.*

It will turn out, that if g fulfills the necessary and sufficient smoothness conditions as specified in Lemma 3.28, then its finite-dimensional approximation is sufficiently smooth, see Lemma 4.26. Again, the necessary smoothness of the initial values depends on the particular choice of the nonlinearity A in Problem 3.1.

Before we can prove convergence, we need to establish a priori estimates on the sequences of the discrete solutions to Problem 4.7.

Lemma 4.35 (C.f. [133], Lem. 8.23)**.** *Consider Problem 4.7 and let the assumptions of Theorem 4.20 hold. Let $A\colon (0,T) \times V \to V'$ be a Carathéodory mapping and let Assumption 4.28(a-b) hold for $-A_{0_k}$, with A_{0_k} as defined in (4.18).*

If $\{\alpha_{k\mathrm{df}}\}_{k\in\mathbb{N}}$ is bounded in H, then

(a) there exists a constant $C_1 > 0$ independent of k and a solution v_k to (4.12), satisfying

$$\|v_k\|_{L^\infty(0,T;H)} \leq C_1 \quad \text{and} \quad \|v_k\|_{L^2(0,T;V)} \leq C_1, \tag{4.23}$$

(b) and there exists a constant $C_2 > 0$ independent of k so that

$$\|\dot{v}_k\|_{L^2(0,T;H')} \leq C_2. \tag{4.24}$$

Proof. Having assumed *semi-coerciveness* as defined in (3.24), we obtain existence of $v_k \in L^2(0,T;V) \cap L^\infty(0,T;H)$ and estimate (4.23a) from [133, Lem. 8.23].

To prove the uniform bound on $\dot{v}_k$, we use the additional regularity that we have introduced with Assumption 4.34.

Because of boundedness of convergent sequences we have that $\tilde{f}_k := f_k + \frac{d}{dt}(J_{2k}^- g_k)$ is uniformly bounded in $L^2(0,T;H')$ by a constant c_4, see Proposition 4.33 and Lemma 4.26 for convergence of f_k and $\frac{d}{dt}(J_{2k}^- g_k)$, as $k \to \infty$.

By Remark 4.16 and Lemma 4.17, we have that $\|\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}\|_{\mathcal{L}(H',H')} = \|I_{H'_k} - L_k j\|_{\mathcal{L}(H',H')} \leq 1 + c_5$, with c_5 independent of k.

Since $A\colon (0,T) \times V \to V'$ is a Carathéodory mapping, $A_{0_k}(\cdot, v)\colon (0,T) \to V'$ is measurable for all $v \in V$ and $k \in \mathbb{N}$. Since the dual product in V is the extension of the dual product in H, $A_{0_k}(\cdot, w_k)\colon (0,T) \to H'$ is measurable, provided that $A_{0_k}(t, w_k) \in H'$ for almost all $t \in (0,T)$. If we take a v_k solving (4.12), by Assumption 4.34, we have that $A_{0_k}(t, v_k(t))$ is in H', for almost all $t \in (0,T)$. Since v_k is in $L^\infty(0,T,H)$, the growth condition (4.19) implies that the *Nemyckij* map $t \mapsto A_{0_k}(t, w_k(t)) \in H'$ is measurable, cf. Section 2.4. Thus, we can estimate

$$\begin{aligned}\int_0^T \|\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} A_{0_k}(t, v_k(t))\|_{H'}^2 \,\mathrm{d}t &\leq (1+c_5)^2 \int_0^T \|A_{0_k}(t, v_k(t))\|_{H'}^2 \,\mathrm{d}t \\ &\leq (1+c_5)^2 \int_0^T \big[\beta(\|v_k(t)\|_H)\big(\gamma(t) + \|v_k\|_H\big)\big]^2 \,\mathrm{d}t \\ &\leq c_6^2 := (1+c_5)^2 \beta^2(C_1)\big[\|\gamma\|_{L^2(0,T)} + C_1\big]^2, \end{aligned} \tag{4.25}$$

using (4.23) and (4.19).

With this, we find that

$$\|\dot{v}_k\|_{\mathcal{H}'} \leq \|\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} A_{0_k}(v_k)\|_{\mathcal{H}'} + \|\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} \tilde{f}_k\|_{\mathcal{H}'} \leq C_2 := c_6 + (1+c_5)c_4.$$

□

Having established that the sequences of the discrete solutions and their derivatives are uniformly bounded, we can now prove convergence to a solution of the continuous differential equation (3.21).

Theorem 4.36 (Cf. [133], Thm. 8.27)**.** *Consider the setup of Problem 4.7 for the symmetric case defined in Problem 3.1(Sym), with $J_1 = J_2$ and with the approximation scheme $\{V_k\}_{k\in\mathbb{N}}$ to V fulfilling Assumption 4.5. Let Assumptions 4.8 and 4.13 hold, let $f \in L^2(0,T;V')$, $g \in L^2(0,T;Q')$, and $\alpha_k \in H_k$. Let the sequences of inverses to J_{2k} be chosen as defined in Lemma 4.26, let $\frac{d}{dt}(J_2^- g) \in L^2(0,T;H')$, and let the Carathéodory mapping $-A\colon (0,T) \times V \to V'$ satisfy Assumption 4.28.*

If for the sequence of initial values to the discrete problem converges to the initial value of the continuous problem, i.e. $\alpha_{k\mathrm{df}} \to \alpha_{\mathrm{df}}$ in H, then there exists a $v \in L^2(0,T;V)$ solving the continuous equation (3.21) *and $v_k \rightharpoonup v$ in $L^2(0,T;V)$ and $\dot{v}_k \rightharpoonup \dot{v}$ in $L^2(0,T;H')$, as $k \to \infty$, (possibly in terms of subsequences), where, for $k \in \mathbb{N}$, v_k is a solution to the discrete equations* (4.12)*.*

Proof. The proof follows the lines of the proof for [133, Thm. 8.27] but with modifications that in particular account for the external approximation of the variable v.

Since by Lemma 4.35 the sequence $\{v_k\}_{k\in\mathbb{N}}$ is bounded, and since with V reflexive, also $L^2(0,T;V)$ is reflexive, there exists a subsequence $\{v_{k'}\}_{k'\in\mathbb{N}} \subset \{v_k\}_{k\in\mathbb{N}}$ that converges weakly to v in $L^2(0,T;V)$. By the same arguments, there exists a subsequence $\{\dot{v}_{k''}\}_{k''\in\mathbb{N}} \subset \{\dot{v}_k\}_{k\in\mathbb{N}}$ and a $\nu \in L^2(0,T;H')$ so that $\dot{v}_{k''} \rightharpoonup \nu$ in $L^2(0,T;H')$. In what follows, we will assume that the subscript k always labels the subsequence that is currently under consideration.

We conclude that there exists a subsequence $\{v_k\}_{k\in\mathbb{N}}$ with limit v, so that $\dot{v} = \nu$ in $L^2(0,T;H')$, since for all $w \in V$ and all $\phi \in \mathcal{C}_0^\infty$, we have

$$\begin{aligned}\int_0^T \big\langle \nu(t), w\big\rangle_{V',V} \phi(t)\ \mathrm{d}t &= \lim_{k\to\infty} \int_0^T \big\langle \dot{v}_k(t), w\big\rangle_{H',H} \phi(t)\ \mathrm{d}t \\ &= -\lim_{k\to\infty} \int_0^T \big(v_k(t), w\big)_H \dot{\phi}(t)\ \mathrm{d}t \\ &= -\int_0^T \big(v(t), w\big)_H \dot{\phi}(t)\ \mathrm{d}t.\end{aligned}$$

So far, we have established existence of $v \in \mathcal{W}(0,T;V;H')$, that is the weak limit of a subsequence of solutions $v_k \in V_{k\mathrm{df}}$ to the discrete equations (4.12). We will show that v is in $\mathcal{W}(0,T;V_{\mathrm{df}};H'_{\mathrm{df}})$ and that it solves (3.21).

Since the external approximation scheme $\{V_{k\mathrm{df}}\}_{k\in\mathbb{N}}$ to V_{df} is stable and since v is the weak limit of functions v_k with $v_k(t) \in V_{k\mathrm{df}}$, for almost all $t \in (0,T)$, we have that $v(t) \in V_{\mathrm{df}}$ for almost all $t \in (0,T)$, see Lemma 4.25, and, thus, $v \in L^2(0,T;V_{\mathrm{df}})$.

Take a $u \in \mathcal{W}(0,T;V_{\mathrm{df}};H'_{\mathrm{df}})$ and $\{u_k\}_{k\in\mathbb{N}} \subset L^2(0,T;V_{k\mathrm{df}})$ with $u_k \to u$ in $L^2(0,T;V)$ as $k\to\infty$. Such a sequence exists, see [143, Lem. III.5.10], since $\{V_{k\mathrm{df}}\}_{k\in\mathbb{N}}$ is a stable and convergent external approximation scheme for V_{df}, see Lemma 4.25.

By the assumption on $-A$, we can consider $-A_{0_k}$, with A_{0_k} as defined in (4.18), satisfying the semi-coerciveness condition (4.20), growth condition (4.19), and the condition for $v_{k\mathrm{c}}$-uniform pseudomonotonicity (4.21).

Since $v_{k\mathrm{df}}$ is a solution to (4.12), we have for any $w \in L^2(0,T;V_{k\mathrm{df}})$:

$$\begin{aligned}\int_0^T \big\langle \dot{v}_k(t) - \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} A_{0_k}(v_k(t)), w(t)\big\rangle_{H',H}\ \mathrm{d}t& \\ = \int_0^T \big\langle \mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]} \tilde{f}_k(t), w(t)\big\rangle_{H',H}\ \mathrm{d}t&,\end{aligned} \tag{4.26}$$

where $\tilde{f}_k := f_k + \frac{d}{dt}(J_{2k}^{-} g_k)$, which, by Lemma 4.26 is in $L^2(0,T;H')$ for all k.

With the assumption that $J_1 = J_2$, in (4.26), the projection $\mathcal{P}_{[H'_{k\mathrm{df}}|H'_{k\mathrm{c}}]}$ has no effect, cf. Remark 4.31. Thus, putting $w = u_k - v_k$, we can write (4.26) as

$$-\big\langle A_{0_k}(v_k), u_k - v_k\big\rangle_{\mathcal{V}',\mathcal{V}} = \big\langle \tilde{f}_k, u_k - v_k\big\rangle_{\mathcal{V}',\mathcal{V}} - \big\langle \dot{v}_k, u_k - v_k\big\rangle_{\mathcal{V}',\mathcal{V}} =: I_k^{(1)} - I_k^{(2)}. \tag{4.27}$$

Note that for functionals in $H' \subset V'$ and elements of V the dual product in H and V coincide.

With $v_k \rightharpoonup v$ and $u_k \rightharpoonup u$ weakly in $L^2(0,T;V)$, by [161, Prop. 21.23(j)] we have $I_k^{(1)} = \big\langle \tilde{f}_k, u_k - v_k\big\rangle_{\mathcal{V},\mathcal{V}'} \to \big\langle \tilde{f}, u - v\big\rangle_{\mathcal{V}',\mathcal{V}}$ as $k\to\infty$, since $\tilde{f}_k \to \tilde{f} := f + \frac{d}{dt}(J_2^{-} g)$, see Proposition 4.33 and Lemma 4.26. Since the limit $u - v$ is in $L^2(0,T;V_{\mathrm{df}})$, by Remark 4.31 and Lemma 4.32, we have that

$$\begin{aligned}\lim_{k\to\infty} I_k^{(1)} = \lim_{k\to\infty} \big\langle \tilde{f}_k, u_k - v_k\big\rangle_{\mathcal{V}',\mathcal{V}} = \big\langle \tilde{f}, u - v\big\rangle_{\mathcal{V}',\mathcal{V}} &= \big\langle \tilde{f}, \mathcal{P}'_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[u - v]\big\rangle_{\mathcal{V}',\mathcal{V}} \\ &= \big\langle \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]} \tilde{f}, u - v\big\rangle_{\mathcal{V}',\mathcal{V}}.\end{aligned} \tag{4.28}$$

With $v_k \in \mathcal{W}(0,T;V_k,H'_k)$, we have that $\big\langle \dot{v}_k, v_k\big\rangle_{\mathcal{V}',\mathcal{V}} = \frac{1}{2}(\|v_k(T)\|_H^2 - \|v_k(0)\|_H^2)$, see [133, Rem. 7.5]. Since, by Lemma 4.35 and by assumption, v_k, $\dot{v}_k$, and $v_k(0) = \alpha_{k\mathrm{df}}$, are uniformly bounded, $v_k(T)$ is uniformly bounded in H and we can conclude that there exists a $\zeta \in H$, such that $v_k(T) \rightharpoonup \zeta$ in H. We will show that $\zeta = v(T)$ and that $v_k(0) \to v(0)$.

Let $\psi \in C^\infty([0,T])$, and for any $w \in V$ consider $\psi w \in \mathcal{W}(0,T;V;V')$. Then, by [133, Lem. 7.3] and by the reflexivity of $\mathcal{V} = L^2(0,T;V)$, we have

$$\begin{aligned}\big(\zeta, \psi(T)w\big)_H - \big(v(0), \psi(0)w\big)_H &= \lim_{k\to\infty}\big[\big(v_k(T), \psi(T)w\big)_H - \big(v_k(0), \psi(0)w\big)_H\big] \\ &= \lim_{k\to\infty}\big[\big\langle \dot v_k, \psi w\big\rangle_{\mathcal{V}',\mathcal{V}} + \big\langle v_k, \dot\psi w\big\rangle_{\mathcal{V},\mathcal{V}'}\big] \\ &= \big\langle v, \psi w\big\rangle_{\mathcal{V}',\mathcal{V}} - \big\langle v, \dot\psi\big\rangle_{\mathcal{V},\mathcal{V}'} \\ &= \big(v(T), \psi(T)w\big)_H - \big(v(0), \psi(0)w\big)_H.\end{aligned}$$

Having chosen ψ, such that $\psi(0) = 0$ and $\psi(T) = 1$, we conclude that for any $w \in V$, it holds that $\big(v(T), w\big)_H = \big(\zeta, w\big)_H$ and, thus, $v(T) = \zeta$ since V is dense in H. Choosing ψ, such that $\psi(0) = 1$ and $\psi(T) = 0$, it also follows that $v(0) = \alpha_{\mathrm{df}}$ in H. By assumption, we have that $v_k(0) = \alpha_{k\mathrm{df}} \to \alpha_{\mathrm{df}}$, so that we can conclude that $v_k(0) \to v(0)$ in H.

To proceed, we recall two fundamental facts for Banach spaces. The norm is *weakly lower semicontinuous*, i.e. if

$$u_k(T) \rightharpoonup u(T), \qquad \text{then} \qquad \|u(T)\|_H \le \liminf_{k\to\infty}\|u_k(T)\|_H, \tag{4.29}$$

see [149, Thm 2.12]. Secondly, if

$$\dot v_k \rightharpoonup \dot v \text{ in } L^2(0,T;V') \text{ and } u_k \to u \text{ in } L^2(0,T;V), \text{ then } \big\langle \dot v_k, u_k\big\rangle_{\mathcal{V}',\mathcal{V}} \to \big\langle \dot v, u\big\rangle_{\mathcal{V}',\mathcal{V}}, \tag{4.30}$$

as $k \to \infty$, see [161, Prop. 21.23(k)].

Consider $I_k^{(2)} = \big\langle \dot v_k, u_k - v_k\big\rangle_{\mathcal{V}',\mathcal{V}}$ from (4.27). By (4.29) and (4.30) it follows that

$$\begin{aligned}\limsup_{k\to\infty} I_k^{(2)} &= \lim_{k\to\infty}\big\langle \dot v_k, u_k\big\rangle_{\mathcal{V}',\mathcal{V}} - \frac{1}{2}\liminf_{k\to\infty}\|v_k(T)\|_H^2 + \frac{1}{2}\lim_{k\to\infty}\|v_k(0)\|_H^2 \\ &\le \big\langle \dot v, u\big\rangle_{\mathcal{V}',\mathcal{V}} - \frac{1}{2}\|v(T)\|_H^2 + \frac{1}{2}\|v(0)\|_H^2 = \big\langle \dot v, u - v\big\rangle_{\mathcal{V}',\mathcal{V}}.\end{aligned} \tag{4.31}$$

By (4.25) we have that $A_{0_k}(v_k)$ is uniformly bounded in $L^2(0,T;V')$, so that $\big\langle A_{0_k}(v_k), u_k - u\big\rangle_{\mathcal{V}',\mathcal{V}} \to 0$, as $k \to \infty$. Thus, using also (4.31), we can estimate

$$\begin{aligned}-\limsup_{k\to\infty}\big\langle A_{0_k}(v_k), v_k - u\big\rangle_{\mathcal{V}',V} &= -\limsup_{k\to\infty}\big\langle A_{0_k}(v_k), v_k - u_k\big\rangle_{\mathcal{V}',V} \\ &\qquad - \lim_{k\to\infty}\big\langle A_{0_k}(v_k), u_k - u\big\rangle_{\mathcal{V}',V} \\ &= \lim_{k\to\infty}\big\langle \tilde f_k, v_k - u\big\rangle_{\mathcal{V}',V} - \limsup_{k\to\infty}\big\langle \dot v_k, v_k - u\big\rangle_{\mathcal{V}',V} \\ &\le \big\langle \tilde f - \dot v, v - u\big\rangle_{\mathcal{V}',\mathcal{V}}.\end{aligned} \tag{4.32}$$

Setting $u := v$ in (4.32), we find that $-\limsup_{k\to\infty}\big\langle A_{0_k}(v_k), v_k - v\big\rangle_{\mathcal{V}',\mathcal{V}} \le 0$, so that the $v_{k\mathrm{c}}$-*uniform* pseudomonotonicity of A_{0_k} implies that

$$\liminf_{k\to\infty} -\big\langle A_{0_k}(v_k), v_k - u\big\rangle_{\mathcal{V}',\mathcal{V}} \ge -\big\langle A_0(v), v - u\big\rangle_{\mathcal{V}',\mathcal{V}}. \tag{4.33}$$

Combining (4.32) and (4.33), we obtain that $-\big\langle A_0(v), v-u\big\rangle_{\mathcal{V}',\mathcal{V}} \le \big\langle f - \dot v, v - u\big\rangle_{\mathcal{V}',\mathcal{V}}$. Since $u \in L^2(0,T;V_{\mathrm{df}})$ is arbitrary, we get equality. In the same way as in (4.28) we can put back the projectors to get

$$\dot v - \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}A_0(v) = \mathcal{P}_{[H'_{\mathrm{df}}|H'_{\mathrm{c}}]}[f + \tfrac{d}{dt}(J_{2k}^- g)] \text{ in } L^2(0,T;V').$$

□

We now show how the convergence of the partial solutions $v_{k\mathrm{c}} \to v_{\mathrm{c}}$ and $v_{k\mathrm{df}} \to v_{\mathrm{df}}$ in $L^2(0,T;V)$ makes the remaining part p_k of the solutions to (4.6) converge to p as the part of a solution to the continuous problem problem as defined in (3.19c).

Lemma 4.37. *Consider Problem 3.1 and its finite-dimensional approximation defined in Problem 4.7. Assume that the approximation schemes* $\{V_k\}_{k\in\mathbb{N}}$ *and* $\{Q_k\}_{k\in\mathbb{N}}$ *fulfill Assumption 4.5. Let the assumptions of Theorem hold 4.20 and let the nonlinear operator* $A\colon L^2(0,T;V) \to L^2(0,T;V')$ *be bounded and weakly continuous.*

If for $\{v_k\}_{k\in\mathbb{N}}$*, with* $A(v_k) \in L^2(0,T;H')$*,* $v_k = v_{k\mathrm{c}} + v_{\mathrm{df}}$*, for all* $k \in \mathbb{N}$*, there exists* v_{df}*,* $v_{\mathrm{c}} \in L^2(0,T;V)$*, such that* $v_{k\mathrm{c}} \to v_{\mathrm{c}}$ *and* $v_{k\mathrm{df}} \rightharpoonup v_{\mathrm{df}}$ *in* $L^2(0,T;V)$*, then* $p_k \rightharpoonup p$*, where* $p_k \in L^2(0,T;Q_k)$ *is defined in* (4.11c) *and* $p \in L^2(0,T;Q)$ *is defined in* (3.19c).

Proof. By (3.19c) and with $v_{k\mathrm{c}} := -J_{2_k}^{-} g_k$, for $k \in \mathbb{N}$, we have that

$$p_k = -S_k^{-1}\bigl[\bar{J}_{2k} j A_k\bigl(v_{k\mathrm{df}} + v_{k\mathrm{c}}\bigr) + \bar{J}_{2k} j f_k + \dot{g}_k\bigr].$$

By Proposition 4.33 and Lemma 4.26, we have that $f_k \to f$ in $L^2(0,T;H')$ and $\dot{g}_k \to \dot{g}$ in $\mathcal{Q}'$. By the weak continuity of A we have that $A(v_k) \rightharpoonup A(v)$, and by the existence of uniformly bounded projections onto H_k also that $A_k(v_k) \rightharpoonup A(v)$ in $L^2(0,T;H')$. By Assumption 4.5 for $\{Q_k\}_{k\in\mathbb{N}}$, there exists a uniformly bounded projection $\mathcal{P}_{[Q_k]}\colon Q \to Q$ onto Q_k. Thus, for all $q \in Q$, one has that

$$\begin{aligned}\bigl\langle \bar{J}_{2k} j A(v_k), q\bigr\rangle_{Q',Q} =& \bigl\langle jA(v_k), \bar{J}_2{}'\mathcal{P}_{[Q_k]} q\bigr\rangle_{\mathcal{H},\mathcal{H}'} \\ &\to \bigl\langle jA(v), \bar{J}_2{}' q\bigr\rangle_{\mathcal{H},\mathcal{H}'} = \bigl\langle \bar{J}_2 j A(v), q\bigr\rangle_{\mathcal{Q}',\mathcal{Q}},\end{aligned}$$

as $k \to \infty$, i.e., recalling that Q is reflexive, $\bar{J}_{2k} j A_k(v_k) \rightharpoonup \bar{J}_2 j A(v)$ in Q'. Since strong convergence is preserved by $\bar{J}_{2k}$ we conclude that

$$-S_k p_k = \bar{J}_{2k}[jA_k(v_k) + jf_k] + \dot{g}_k \rightharpoonup \bar{J}_2[jA(v) + jf] + \dot{g} = -Sp \text{ in } Q', \tag{4.34}$$

as $k \to \infty$. Since S_k^{-1} is uniformly bounded, see Remark 4.16, from (4.34) we have that the sequence $\{p_k\}_{k\in\mathbb{N}} \subset \mathcal{Q}$ is bounded. We show that every weakly convergent subsequence converges to p, to conclude that the sequence $\{p_k\}_{k\in\mathbb{N}}$ itself converges weakly to p [161, Prop. 21.23(i)].

Let $\{p_{k'}\}_{k'\in\mathbb{N}} \subset \{p_k\}_{k\in\mathbb{N}}$ be a convergent subsequence, i.e. there exists a $p' \in \mathcal{Q}$ with $p_{k'} \rightharpoonup p'$, as $k' \to \infty$. By (4.34), we have that $S_{k'} p_{k'} \rightharpoonup Sp$, as $k' \to \infty$. Let $q \in Q$, and consider

$$\lim_{k'\to\infty} \bigl\langle S_{k'} p_{k'}, q\bigr\rangle_{\mathcal{Q}',\mathcal{Q}} = \bigl\langle Sp, q\bigr\rangle_{\mathcal{Q}',\mathcal{Q}} = \bigl\langle p, \bar{J}_1 j \bar{J}_2{}' q\bigr\rangle_{\mathcal{Q},\mathcal{Q}'}. \tag{4.35}$$

On the other hand we have that

$$\bigl\langle S_{k'} p_k, q\bigr\rangle_{\mathcal{Q}',\mathcal{Q}} = \bigl\langle p_{k'}, \bar{J}_1 \mathcal{P}_{[H_{k'}]} j \bar{J}_2{}' \mathcal{P}_{[Q_{k'}]} q\bigr\rangle_{\mathcal{Q},\mathcal{Q}'} \to \bigl\langle p', \bar{J}_1 j \bar{J}_2{}' q\bigr\rangle_{\mathcal{Q},\mathcal{Q}'}, \tag{4.36}$$

using [161, Prop. 21.23(j)] and, in particular, $\bar{J}_1 \mathcal{P}_{[H_{k'}]} j \bar{J}_2{}' \mathcal{P}_{[Q_{k'}]} q \to \bar{J}_1 j \bar{J}_2{}' q$, because the involved projections fulfill Assumption 4.5. Since $\bar{J}_1 j \bar{J}_2{}'\colon Q \to Q'$ is bijective, see Lemma 6.12, combining Equations (4.35) and (4.36) yields $\bigl\langle p', \tilde{q}\bigr\rangle_{\mathcal{Q},\mathcal{Q}'} = \bigl\langle p, \tilde{q}\bigr\rangle_{\mathcal{Q},\mathcal{Q}'}$ for all $\tilde{q} \in \mathcal{Q}'$, i.e. $p' = p \in \mathcal{Q}$.

Thus we have that $p_k \rightharpoonup p$.

□

Remark 4.38. Convergence of the finite-dimensional approximations to a solution of the abstract DAE (Problem 3.1) was established on the base of several assumptions that address the following aspects of the problem statement:

(a) *Space Regularity*: We assume additional smoothness of the right hand side $f \in L^2(0,T;H')$ and that the corresponding solution of Problem 3.1 is such that $A(t, v(t))$ and $J_1' p(t)$ are in H', for almost all $t \in (0,T)$ (Assumption 3.8). Since at a fixed discretization level k, $(V_k)'$ and H_k' are isomorphic, in finite dimension, instead, we assume uniform boundedness of $A_k(v_k(t))$

in $L^2(0,T;H')$ and of $p_k(t)$ in Q. (Assumption 4.34). To formalize this, we assume existence of $Q \subset Q_H$, densely and continuously embedded in Q_H, such that, e.g., $J_1'(Q) \in H'$ and such that J_1, $J_2 \colon H \to Q'$ are bounded (Assumption 3.5).

(b) *Decoupling of Equation and Solution Spaces*: We assume that the operators J_1, $J_2 \colon V \to Q_H'$ that account for the coupling of differential and algebraic equations in (3.1) allow for a decoupling as follows: The kernel of J_2 splits the solution space V into a differential and an algebraic part and J_1' can be inverted to compute p separately (Assumption 3.13). To split the equations conformingly, we assume that H' decomposes into the image of $\bar{J_1}'$ and a part that is identified with $\ker \bar{J}_2$ (Assumption 3.5).

(c) *Consistency and Regularity of the Data*: If the setup allows for a splitting, existence of solutions to Problem 3.1 implies existence of the separated solution components and, thus, defines necessary conditions for solvability. These are smoothness of g in time, i.e. $\dot{g} \in \mathcal{Q}_-$ as specified in Lemma 3.28, and consistency of the initial value as defined in Definition 3.35.

(d) *Stable Approximation Schemes*: Consider the discrete setup of Problem 4.7. We assume there are approximation schemes $\{Q_k\}_{k\in\mathbb{N}}$ and $\{V_k\}_{k\in\mathbb{N}}$ that fulfill Assumption 4.5, i.e. they come with a sequence of associated uniformly bounded projections. Considering J_{1k} and J_{2k}, we assume that $\{Q_k\}_{k\in\mathbb{N}}$ and $\{V_k\}_{k\in\mathbb{N}}$ allow for splittings of the equations and solutions as in the continuous case and uniformly with respect to the discretization parameter k (Assumption 4.8 and 4.13). Then, in particular, the underlying *external approximation scheme* to $\ker J_2$ is stable and convergent, cf. Lemma 4.25. Without further restrictions, we can assume that J_{2k}^- and J_2^- are chosen such that $v_{k\mathrm{c}} := -J_{2k}^- g_k \to J_2^- g =: v_\mathrm{c}$ in $L^2(0,T;V)$, as $k \to \infty$, cf. Lemma 4.26.

(e) *Consistent and Convergent Approximation of the Initial Value*: We assume that for every k, the initial value α_k of the discrete problem writes as $\alpha_k = \alpha_{k\mathrm{df}} - \bar{J}_{2k}^- g_k(0)$, with $\alpha_{k\mathrm{df}} \to \alpha_\mathrm{df}$ in H.

(f) *Symmetry*: To show convergence of the discrete solutions to a solution of the continuous problem, we have assumed that $J_1 = J_2$, cf. Remark 4.31.

(g) *Continuity, Coerciveness, Monotonicity and Boundedness of the Nonlinearity*: Given the sequence $\{v_{k\mathrm{c}}\}_{k\in\mathbb{N}}$ converging to v_c, we define the shifted nonlinearity A_{0_k} as in (4.18) and assume that it has particular boundedness, coerciveness and pseudo-monotonicity properties uniformly with respect to $v_{k\mathrm{c}}$ (Assumption 4.28). This was sufficient to prove weak convergence of subsequences of v_k to v. To prove convergence in p, we additionally assumed weak continuity of the nonlinearity A, cf. Lemma 4.37.

We summarize the findings of this section in the following theorem:

Theorem 4.39. *Consider Problem 3.1(Sym), where $J_1 = J_2$, and the corresponding finite-dimensional approximation defined in Problem 4.7. Let all assumptions listed in Remark 4.38 hold. In particular, assume right hand sides $f \in L^2(0,T;H')$ and $g \in W^{1,2}(0,T;Q,Q_-)$ and an initial value α and approximations α_k that are consistent as specified in Definition 3.35 and convergent as defined in Remark 4.38(e).*

For $k \in \mathbb{N}$ let $(v_k, p_k) \in \mathcal{W}(0,T;V_k,H_k') \times L^2(0,T;Q_k)$ be a solution of the discrete problem. Then there exists $(v,p) \in \mathcal{W}(0,T;V,H') \times L^2(0,T;Q)$ such that $v_k \rightharpoonup v$ in $\mathcal{W}(0,T;V,H')$ and $p_k \rightharpoonup p$ in $L^2(0,T;Q)$ and (v,p) solves the continuous problem. The convergence is in terms of subsequences.

If the continuous problem has only one solution, then the whole sequence $\{(v_k, p_k)\}_{k\in\mathbb{N}}$ *converges.*

If $V \hookrightarrow H$ *is compact, then* $v_k \to v$ *in* $L^2(0, T; H)$*.*

Proof. Under the given conditions, by Theorem 4.20, for all $k \in \mathbb{N}$, we can decompose any solution of the discrete approximations: A solution (v_k, p_k) to Problem 4.7 writes $(v_{k\mathrm{df}} + v_{k\mathrm{c}}, p_k)$, with $v_{k\mathrm{c}}$ specified in (4.11a), $v_{k\mathrm{df}}$ solving the differential equation (4.12), and p_k defined via (4.11c).

Let k be the subscript of a convergent subsequence. By Lemma 4.26, we have that $v_{k\mathrm{c}} \to -J_2^- g =: v_\mathrm{c}$ in $L^2(0, T; V)$. By Theorem 4.36 we have that $v_{k\mathrm{df}} \rightharpoonup v_\mathrm{df}$ in $\mathcal{W}(0, T; V, H')$, where v_df solves (3.21). With $\dot v_{k\mathrm{df}} \in L^2(0, T; H')$ for all $k \in \mathbb{N}$ and Assumption 4.34, we have that $\dot v_{k\mathrm{c}} \in L^2(0, T; H')$, for all $k \in \mathbb{N}$ and thus, by Lemma 4.26, that $v_{k\mathrm{c}} \rightharpoonup v_\mathrm{c}$ in $\mathcal{W}(0, T; V; H')$. By Lemma 4.37, we have that $p_k \rightharpoonup p$ in $L^2(0, T; Q)$, where p fulfills (3.19c). Since with g being sufficiently smooth also g_k is sufficiently smooth and since we have assumed consistency of α_k, by Theorem 3.37 we have that, for all k, $(v_k, p_k) := (v_{k\mathrm{df}} + v_{k\mathrm{c}}, p_k)$ solves Problem 4.7.

By Lemma 3.30 and Theorem 3.37 a solution (v, p) to Problem 3.1 has the representation as $(v_\mathrm{df} + v_\mathrm{c}, p)$, where v_c is defined via (3.19a), v_df is the solution of the abstract differential equation (3.21), and p fulfills (3.19c). Thus, the limit of a convergent subsequence $\{(v_k, p_k)\}_{k\in\mathbb{N}} = \{(v_{k\mathrm{df}} + v_{k\mathrm{c}}, p_k)\}_{k\in\mathbb{N}}$ of solutions to the discrete problem is a solution to Problem 3.1.

If Problem 3.1 has only one solution, then every convergent subsequence of $\{(v_k, p_k)\}_{k\in\mathbb{N}}$ converges to the same limit. This implies that the sequence itself is convergent [161, Prop. 21.23(i)].

If $V \overset{c}{\hookrightarrow} H$, then, by Lemma 2.16, $\mathcal{W}(0, T; V; H') \overset{c}{\hookrightarrow} L^2(0, T; H)$, and thus the weak convergence of $v_k \to v$ in $\mathcal{W}(0, T; V; H')$ implies strong convergence of $v_k \to v$ in $L^2(0, T; H)$.

□

4.5. **Initial Conditions.** Under the assumptions of Theorem 4.39, for the existence of a solutions to the sequence of Problems 4.7, it is necessary that the initial value α_k to the discrete equations 4.6 is consistent in the sense of Definition 3.35 for the discrete spaces and for every $k \in \mathbb{N}$.

In general, the canonical projections of a consistent initial condition onto a discrete subspace does not give a consistent initial condition for the projected equations, see [60], for a canonical example for the finite element approximation of Navier Stokes Equation. Thus, for the proper choice of the initial value α_k, a particular approximation is necessary that takes into account the finite-dimensional realizations J_{2k} of the operator J_2, cf. Corollary 3.36. In addition, in view of convergence of the solution, any strategy has to ensure the $\alpha_k \to \alpha$ in H, as $k \to \infty$.

In theory, under the assumptions of Remark 4.38(a,b,d), this does not pose a problem. With the arguments of Section 4.3, one can show that $\{V_k\}_{k\in\mathbb{N}}$ and $\{Q_k\}_{k\in\mathbb{N}}$ also define a stable and convergent approximation scheme for H_df via the sequence $H_{k\mathrm{df}} := \ker \bar J_{2k}$. Let $\overline{r_k}$ be the corresponding restriction operator. Let $\alpha = \alpha_\mathrm{df} - \bar J_2{}^{-} g(0)$ be a consistent initial value. Then we have by Definition 4.23(a) that $\overline{\alpha_{k\mathrm{df}}} := \overline{r_k}\alpha_\mathrm{df} \to \alpha_\mathrm{df}$ in H and by Lemma 4.26 and by $v_{k\mathrm{c}}(0) = -\bar J_{2k}{}^{-} g(0)$, cf. the proof of Corollary 3.31, that $-\bar J_{2k}{}^{-} g_k(0) \to -\bar J_2{}^{-} g(0)$ in H. Thus, the sequence of initial values $\alpha_k := \overline{\alpha_{k\mathrm{df}}} - \bar J_{2k}{}^{-} g_k(0)$ is consistent and convergent.

However, this approach for the retrieval of consistent and convergent initial values is not feasible in practice, since the right inverses of J_2 and J_{2k} are not available in general.

To ensure consistency, at a given discretization level, with α_k being an arbitrary approximation to α, one can use regularization techniques to modify α_k such that it becomes consistent, see [42], [90, Ch. 6.1], and [102, Ch. 4.3]. However, this modification may destroy the convergence properties of the parts of $\alpha_k = \alpha_{k\mathrm{df}} - \bar{J}_{2k}^{-} g_k(0)$ that were needed in the proof of convergence in Theorem 4.39.

In the case where $g = 0$, convergent sequences of consistent initial values can be defined via $\alpha_k := \mathcal{P}_{k_{\mathrm{df}}} \mathcal{P}_{[H_k]}$, where $\mathcal{P}_{k_{\mathrm{df}}}$ is an uniformly bounded projector from H into H that maps H_k onto $H_{k\mathrm{df}}$, and $\mathcal{P}_{[H_k]}$ is the orthogonal projection in H onto H_k. By [49, Lem. II.1.1] such a projector $\mathcal{P}_{k\mathrm{df}}$ exists. If $J_2 = J_1$, then $\mathcal{P}_{k_{\mathrm{df}}}$ can be chosen as the orthogonal projector defined in Lemma 4.17.

5. Constrained Optimization and Optimal Control

In view of optimal control of systems, we recall classical results that extend the theory of Lagrangian multipliers for elimination of side constraints in finite dimensional optimization problems, cf. [44], to infinite-dimensional problems. Particularly, we will refer to results that take into account the specific structure of dynamical systems.

We define a general constrained optimization problem.

Problem 5.1. *Let Z and W be metric spaces and let W be complete. Consider the minimization problem*

$$\mathcal{J}(z) \to \min_{z \in Z}, \quad \textit{subject to} \quad \mathcal{G}(z) = 0, \tag{5.1}$$

for a functional $\mathcal{J} \colon Z \to \mathbb{R}$ and a mapping $\mathcal{G} \colon Z \to W$. Let $N_{\mathcal{G}}$ denote the subset of Z on which the constraints are fulfilled, i.e.

$$N_{\mathcal{G}} := \{ z \in Z : \mathcal{G}(z) = 0 \}.$$

In order to state optimality conditions, we will call on the Fréchet derivative, see Definition 2.18.

5.1. **Multipliers and First Order Necessary Optimality Conditions.**

Assuming Fréchet-differentiability of $\mathcal{J}$ and $\mathcal{G}$ locally in the region of interest, one has the classical result due to Ljusternik:

Theorem 5.2 ([109], p. 398)**.** *Consider Problem 5.1. If z_0 is a stationary and an absolutely regular point for $\mathcal{J}$ in $N_{\mathcal{G}}$, then there exists a linear functional l on $\mathcal{G}_{;z}(z_0)[z]$ such that*

$$\mathcal{J}_{;z}(z_0)[z] = l(\mathcal{G}_{;z}(z_0)[z]). \tag{5.2}$$

In (5.2) and in what follows the subscript $;z$ to, e.g., $\mathcal{J}_{;z}$ denotes the possibly partial Fréchet derivative of $\mathcal{J}$ with respect to z as it was introduced along with Definition 2.18.

Since the differentiation commutes with the application of bounded linear functionals, (5.2) can be written as

$$\mathcal{L}_{;z}(z_0)[z] = 0, \quad \text{where} \quad \mathcal{L}(z) := \mathcal{J}(z) - l(\mathcal{G}(z)). \tag{5.3}$$

Here, z_0 is a stationary point for $\mathcal{J}$ in $N_{\mathcal{G}}$, if

$$|\mathcal{J}(z') - \mathcal{J}(z_0)| = \mathrm{o}(\|z' - z_0\|_Z) \quad \text{for all } z' \in N_{\mathcal{G}}. \tag{5.4}$$

A point $z_0 \in N_{\mathcal{G}}$ is called regular if $\operatorname{im} \mathcal{G}_{;z}(z_0) = W$.

Let $T_{z_0} := \ker \mathcal{G}_{;z}(z_0)$ denote the *tangent space* to $N_{\mathcal{G}}$ in z_0. One has that in a regular point z_0 the set $N_{\mathcal{G}}$ is asymptotically close to T_{z_0} in the sense that for all $z' = z_0 + h' \in N_{\mathcal{G}}$ there exists $h \in T_{z_0}$ such that

$$\|h - h'\|_Z = \mathrm{o}(\|h\|_Z), \tag{5.5}$$

see [109, p. 393] for a proof. If also the converse is true, i.e. for all $h \in T_{z_0}$ in a neighborhood of the origin there exists a point $z_0 + h'$, such that $\|h - h'\|_Z = \mathrm{o}(\|h\|_Z)$, then z_0 is called absolutely regular.

Regularity of a point z_0 is often referred to as surjectivity of $d\mathcal{G}_{;z}(z_0)$. For the absolute regularity there are sufficient conditions that have been used to formulate more specific versions of Theorem 5.2.

If there exists a complement space $T_c \subset Z$ such that $Z = T_{z_0} \oplus T_c$ and if the mapping $z_0 \mapsto \mathcal{J}_{;z}(z_0)$ is continuous in the norm of $\mathcal{B}(Z, W)$ in a neighborhood of z_0, then, in terms of [109, p. 396], if z_0 is regular, then z_0 is absolutely regular.

Accordingly, cf. [159, Thm. 43D], the functional l from (5.2) exists, if $\mathcal{G}$ is a *submersion* at z_0, i.e. z_0 is regular, $\mathcal{G}$ is locally continuously Fréchet-differentiable and T_{z_0} splits W.

The necessary splitting of $Z = T_{z_0} \oplus T_c$ is guaranteed if $\dim T_{z_0} < \infty$, if $\dim Z/T_{z_0} < \infty$ or if Z is a Hilbert space, cf. [159, Ex. 43.16].

If $\dim W = n < \infty$, then the constraints in (5.1) can be written as $g_i(z) = 0 \in \mathbb{R}$, $i = 1, \cdots, n$ and a variant of Theorem 5.2 states that if z_0 is a stationary point of $\mathcal{J}$, then there exist $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$ such that

$$\mathcal{J}_{;z}(z_0) = \lambda_i g_{i;z}(z_0), \tag{5.6}$$

cf. [65, Thm. 9.1] and [109, p. 399].

If Z and W are Banach spaces, then the local continuity of the Fréchet differentiation with respect to z makes a regular point an absolutely regular point, see [109, p. 396] for a proof. This is used in the following variant of Theorem 5.2:

Theorem 5.3 ([96], Thm. 5.4.2)**.** *Consider Problem 5.1 and let Z and W be Banach spaces. Suppose that $\mathcal{J}$ and $\mathcal{G}$ are continuously Fréchet-differentiable on an open set $\mathcal{O} \subset X$ and that $z_0 \in \mathcal{O}$ is a regular point of the constraints $\mathcal{G}$. If $\mathcal{J}$ has a local extremum under the constraint $\mathcal{G}(z_0) = 0$ at the regular point z_0, then there exists $l \in W'$ such that the Lagrangian*

$$\mathcal{J}(z) - l\mathcal{G}(z)$$

is stationary at z_0, i.e. $\mathcal{J}_{;z}(z_0) - l\mathcal{G}_{;z}(z_0) = 0$.

One can weaken the assumption on regularity of z_0 in $N_{\mathcal{G}}$ and require only that the range of $G_{;z}(z_0)$ is closed in W. Then one has that, if $\mathcal{J}$ takes on a local extremum at z_0, then there exists $\lambda_0 \in \mathbb{R}$ and $l \in W'$ such that

$$\lambda_0 \mathcal{J}_{;z}(z_0) - l\mathcal{G}_{;z}(z_0) = 0,$$

cf. [96, Thm. 5.4.3] or [159, Prp. 43.19].

5.2. **Relation to Optimal Control of Systems and the Adjoint State.** In optimal control of systems the variable z is given as $z = (x, u)$, where $x \in X$ denotes the state of the system and $u \in U$ is an input or a parameter. Then the optimization problem is written as:

Problem 5.4. *Let X, U, and W be Banach spaces and let $\mathcal{J} \colon X \times U \to \mathbb{R}$ be a functional. Consider*

$$\mathcal{J}(x, u) \to \min_{(x,u)}, \quad \textit{subject to} \quad \mathcal{G}(x, u) = 0, \tag{5.7}$$

where $\mathcal{G} \colon X \times U \to W$ represents the relation of states and inputs.

In PDE constrained optimization, $\mathcal{G}$ represents a partial differential equations. We will consider cases, where $\mathcal{G}$ stands for an abstract differential algebraic equation.

Application of a variant of Theorem 5.2 would give that if (x_0, u_0) is a stationary point of $\mathcal{J}$ in (5.7), then there exists $l \in W'$ such that $\mathcal{J}(x, u) - l\mathcal{G}(x, u)$ is stationary at (x_0, u_0), i.e.

$$\mathcal{J}_{;x}(x_0, u_0) + \mathcal{J}_{;u}(x_0, u_0) - l\mathcal{G}_{;x}(x_0, u_0) - l\mathcal{G}_{;u}(x_0, u_0) = 0. \tag{5.8}$$

In the frequent case that u is a free variable while $x = x(u)$ is well defined in terms of u via the relation $\mathcal{G}(x, u) = 0$, then the functional l can be interpreted as the *adjoint state* defined as the solution $\lambda \in W'$ of the *adjoint equation*

$$\lambda \mathcal{G}_{;x}(x, u) + \mathcal{J}_{;x}(x, u) = 0. \tag{5.9}$$

This can be deduced from the following theorem.

Theorem 5.5 ([96], Thm. 5.5.1)**.** *Consider Problem 5.4 with X, U, and W being normed vector spaces, with the cost functional $\mathcal{J}\colon X\times U\to\mathbb{R}$, and with $\mathcal{G}\colon X\times U\to W$ such that it well defines a mapping $u\mapsto x(u)$ via $\mathcal{G}(x(u),u)=0$. Assume that $\mathcal{G}$ and $\mathcal{J}$ are Fréchet-differentiable with respect to x at $x=x(u)$ and continuously on $X\times U$ Gateaux differentiable with respect to u and that $u\mapsto x(u)$ is Lipschitz continuous.*

If there exists a solution $\lambda\in W'$ to (5.9) at $x=x(u)$, then $\hat{\mathcal{J}}(u):=\mathcal{J}(x(u),u)$ is Gateaux differentiable at u and

$$\hat{\mathcal{J}}_{;u}(u)=\mathcal{J}_{;u}(x,u)+\lambda\mathcal{G}_{;u}(x,u). \tag{5.10}$$

Thus, if $\mathcal{J}$, $\mathcal{G}$ and u_0 meet the conditions of Theorem 5.5, u_0 is a stationary point of $\hat{\mathcal{J}}(u)$ and the adjoint state $\lambda\in W'$ exists, then one has that

$$\begin{aligned} 0=&\hat{\mathcal{J}}_{;u}(u_0)=\mathcal{J}_{;u}(x(u_0),u_0)+\lambda\mathcal{G}_{;u}(x(u_0),u_0)\\ &=\mathcal{J}_{;u}(x(u_0),u_0)+\lambda\mathcal{G}_{;u}(x(u_0),u_0)+\mathcal{J}_{;x}(x(u_0),u_0)+\lambda\mathcal{G}_{;x}(x(u_0),u_0),\end{aligned}$$

i.e. the adjoint state λ serves as the functional $-l$ in (5.8). The following theorem establishes sufficient conditions for the applicability of Theorem 5.5:

Theorem 5.6 ([160], Thm. 4.B)**.** *Consider Problem 5.4. Assume that $\mathcal{G}\colon D\subseteq X\times U\to W$ is defined on an open neighborhood of D of (x_0,u_0), with $\mathcal{G}(x_0,u_0)=0$, that $\mathcal{G}_{;x}$ exists as a partial Fréchet derivative on D and $\mathcal{G}_{;x}(x_0,u_0)\colon X\to W$ is bijective, and that $\mathcal{G}$ and $\mathcal{G}_{;x}$ are continuous on D, then the following statements hold:*

- *(a) There exist r_0, $r>0$ such that for every $u\in U$ satisfying $\|u-u_0\|\le r_0$, there exists exactly one $x(u)\in X$ for which $\|x(u)-x_0\|\le r$ and $\mathcal{G}(x(u),u)=0$.*
- *(b) If $\mathcal{G}$ is continuous in a neighborhood of (x_0,u_0), then $u\mapsto x$ is continuous in a neighborhood of u_0.*
- *(c) If $\mathcal{G}$ is m-times continuously Fréchet-differentiable on a neighborhood of (x_0,u_0), $1\le m\le\infty$, then so is $u\mapsto x$ on a neighborhood of u_0.*

In particular, if $\mathcal{G}$ is Fréchet-differentiable, then the input to state map $u\mapsto x$ is locally Lipschitz continuous. If the Fréchet differential of $\mathcal{G}$ is uniformly bounded, then $u\mapsto x$ is globally Lipschitz continuous, cf. [160, Problem 4.1b].

For dynamical systems that are formulated with a time variable t, we consider the optimization problem (5.1) but with $\mathcal{G}\colon Z\to\mathcal{W}$, where, for $T>0$ and $1<p<\infty$, $\mathcal{W}:=L^p(0,T;W)$, is a Bochner space. In this case, one can call on Theorem 2.13 to determine a representation of $l\in\mathcal{W}'$ in a Bochner space.

5.3. **First Order Sufficient Optimality Conditions.** In this section we state conditions that are sufficient for optimality for general optimization problems and their formulation for optimal control problems.

Theorem 5.7 ([159], Thm. 43D(2))**.** *Let Z and W be Banach spaces and $\mathcal{J}\colon Z\to\mathbb{R}$ and $\mathcal{G}\colon Z\to W$ be n-times continuously Fréchet-differentiable in an open neighborhood of z_0, where n is an even integer, $n\ge 2$. Let $\mathcal{G}$ be a submersion at z_0. Then z_0 is a local solution to (5.1), if there exists $c>0$ and $l\in W'$ such that*

$$\begin{aligned} &\mathcal{J}_{;z^r}(z_0)[k]^r-l\mathcal{G}_{;z^r}(z_0)[k]^r=0, &&r=1,\cdots,n-1\\ &\mathcal{J}_{;z^n}(z_0)[h]^n-l\mathcal{G}_{;z^n}(z_0)[h]^n\ge c\|h\|^n,\end{aligned}$$

for all $k\in Z$ and $h\in T_{z_0}$.

In the most frequent case when $n=2$, one has to check existence of the functional l and $c>0$ with

$$\mathcal{J}_{;z}(z_0)-l\mathcal{G}_{;z}(z_0)=0 \quad\text{and}\quad J_{;zz}(z_0)[h,h]-l\mathcal{G}_{;zz}[h,h]\ge c\|h\|^2,$$

to conclude that z_0 is a local minimum.

Remark 5.8. With the assumptions of Theorem 5.7 the set $N_{\mathcal{G}} := \{z \in Z : \mathcal{G}(z) = 0\}$ is a $\mathcal{C}^1$-manifold, cf. [159, Thm. 43.C(3)]. If $Z = X \times U$ and $\mathcal{G}(x,u) = 0$ defines a Fréchet-differentiable mapping $u \mapsto x(u)$, then $N_{\mathcal{G}} = \{(x(u), u) : u \in U\}$, with the tangent space at $(x(u_0), u_0)$ given via

$$T_{(x(u_0),u_0)} = \{(x_{;u}(u_0)h_u, h_u) : h_u \in U\}.$$

In fact, for any smooth curve $\gamma\colon t \mapsto (x(u(t)), u(t)) \in N_{\mathcal{G}}$ with parameter t in a neighborhood of $0 \in \mathbb{R}$ and with $\gamma(0) = (x(u_0), u_0)$ the tangential vector $h = \gamma'(0) = (x_{;u}(u_0)u'(0), u'(0))$ is in $T_{(x(u_0),u_0)}$. Conversely, any $h = (x_{;u}(u_0)h_u, h_u) \in T_{(x(u_0),u_0)}$ is a tangent vector with the curve $t \mapsto (x(u + th_u), u + th_u)$. See [159, Def. 43.8] for the definition of a manifold and the tangent space in Banach spaces.

Furthermore, in the setting of Theorem 5.7 one has

$$T_{(x(u_0),u_0)} = \ker \mathcal{G}_{;x,u}(x(u_0), u_0), \tag{5.11}$$

as the tangent space was defined for Theorem 5.2, cf. [159, Thm. 43.C(1)].

Thus, one can formulate sufficient conditions tailored to dynamical systems with state x and input u:

Theorem 5.9. *Let X, U and W be Banach spaces $\mathcal{G}\colon X \times U \to W$ such that it defines a Fréchet-differentiable mapping $u \mapsto x(u)$ via $\mathcal{G}(x(u), u) = 0$. Let $\mathcal{G}$ and $\mathcal{J}\colon X \times U \to \mathbb{R}$ be 2-times continuously Fréchet-differentiable in an open neighborhood of $(x_0, u_0) := (x(u_0), u_0)$. Let $\mathcal{G}$ be a submersion at (x_0, u_0). Then (x_0, u_0) is a local solution to (5.7), if there exists $c > 0$ and $l \in W'$ such that*

$$\mathcal{J}_{;(x,u)}(x_0, u_0) - l\mathcal{G}_{;(x,u)}(x_0, u_0) = 0 \quad \textit{and}$$
$$\big(\mathcal{J}_{;(x,u)^2}(x_0, u_0) - l\mathcal{G}_{;(x,u)^2}(x_0, u_0)\big)[h_x(h_u), h_u]^2 \geq c\|(h_x(h_u), h_u)\|^2$$

for all $h_u \in U$ and $h_x(h_u) := x_{;u}(u_0)h_u$.

Remark 5.10. As the conditions of Theorem 5.9 imply that the local solution is a strict extremum, there exists a neighborhood in which the local solution is the only solution to the optimization problem, cf. the proof of [159, Thm. 43D].

Remark 5.11. If, in addition to the assumptions of Theorem 5.9 the partial Fréchet derivative $\mathcal{G}_{;x}$ is invertible, the differential $x_{;u}(u_0)h_u$ of the input to state map $u \mapsto x(u)$ at u_0 can be established by resolving the total derivative of $\mathcal{G}(x(u), u) = 0$ with respect to u at (x_0, u_0) applied to h_u given via

$$\mathcal{G}_{;x}(x_0, u_0)x_{;u}h_u + \mathcal{G}_{;u}(x_0, u_0)h_0 = 0.$$

See Theorem 5.6 establishing the necessary smoothness of $u \mapsto x$ and see [139, Ch. 2.3] for a general discussion of the relation between $x_{;u}h_u$ and linearized state equations for finite dimensions, [149, Thm. 4.25] for a concrete formulation for some generic PDE cases, and [1] for the Navier-Stokes case.

For an overview of results regarding existence of optimal controls we refer to [28]. Here, we list results for constrained optimization problems with convex or related properties. Writing the minimization (5.1) as

$$\min_{z \in N_{\mathcal{G}}} \mathcal{J}(z) = \alpha, \tag{5.12}$$

one can state existence of a minimizer z_0 via the following theorem.

Theorem 5.12 ([159], Thm. 38.A, Cor. 38.8)**.** *For the functional $\mathcal{J}\colon Z \supset N_{\mathcal{G}} \to \bar{\mathbb{R}}$ with $N_{\mathcal{G}} \neq \varnothing$, there exists a solution to (5.12) provided the following conditions hold:*

(a) Z is a real reflexive Banach space.

(b) $N_{\mathcal{G}}$ is bounded, closed and convex.

(c) $\mathcal{J}$ is sequentially lower semicontinuous on $N_{\mathcal{G}}$.

The case that $N_{\mathcal{G}}$ is convex and closed but not bounded is treated in the following proposition that also lists alternative requirements on $\mathcal{J}$.

Proposition 5.13 ([159], Prop. 38.15, [96], Thm. 7.2.3)**.** *On real Banach space Z, a functional $\mathcal{J}\colon Z \supset N_{\mathcal{G}} \to \bar{\mathbb{R}}$, with $N_{\mathcal{G}} \neq \varnothing$ convex and closed and with $\mathcal{J}(z) \to \infty$ as $\|z\| \to \infty$, possesses a minimum if, in addition, $\mathcal{J}$ is convex and continuous or convex and Gateaux differentiable over $N_{\mathcal{G}}$.*

The conditions can be checked locally to state existence of local minima. Global convexity, however, leads to global results:

Proposition 5.14 ([112], Ch. 7.8)**.** *Consider Problem 5.1. If $\mathcal{J}$ and $N_{\mathcal{G}}$ are convex, then a local solution to* (5.12) *is a global solution.*

Uniqueness of a minimizer z_0 is given if, in addition, $\mathcal{J}$ is strictly convex on $N_{\mathcal{G}}$, cf. [159, Thm. 38.C]. Another property of convex optimization is that the necessary conditions $\mathcal{J}_{;z}(z_0) = 0$ for z_0 being a *free minimum*, i.e. a minimum in the interior of $N_{\mathcal{G}}$, is also sufficient, see [96, Thm. 6.2.1]. Thus, combining Theorem 5.5 and Proposition 5.13 one can derive the following theorem that in particular suits linear quadratic optimal control of systems:

Theorem 5.15. *Consider Problem 5.4 and let the assumptions of Theorem 5.5 hold. If $\hat{\mathcal{J}}(u) := \mathcal{J}(x(u), u)$ is (strictly) convex and $\hat{\mathcal{J}}(u) \to \infty$, as $\|u\| \to \infty$, and U is convex and closed, then* (5.7) *has a (unique) solution.*

Furthermore, if there exists a solution $u_0 \in U$ and $\lambda_0 \in W'$ to the adjoint equation

$$\lambda_0 \mathcal{G}_{;x}(x_0, u_0) + \mathcal{J}_{;x}(x_0, u_0) = 0$$

and to

$$\lambda_0 \mathcal{G}_{;u}(x_0, u_0) + \mathcal{J}_{;u}(x_0, u_0) = 0,$$

at $x_0 = x(u_0)$, then u_0 is a (the) global solution to (5.7)*.*

Remark 5.16. The basic assumption that the input to state map is one-to-one, cf. Theorem 5.5 an 5.6 is very restrictive as it is tied to the existence of unique solutions of the state equations. For the Navier-Stokes Equation in the standard weak formulation as considered in Section 3.5, uniqueness of solutions in three spatial dimensions is only proven for small right-hand sides, small initial values, and for short time intervals [143, Ch. 3.3]. Furthermore, the sufficient conditions, cf. [143, Eq. 3.115], are hard to check in practical applications. In [152], it has been shown that the conditions for unique solvability of optimal control problems are less restrictive and more immediate for a certain class of *non-Newtonian* flows.

6. Optimal Control Problem and the Adjoint Equation

In this section we discuss a general optimal control problem for the abstract differential-algebraic equations considered in Section 3. Optimal control problems constrained by partial differential equations or more general abstract equations have been widely investigated, see e.g. [149, 16, 43, 106, 23, 73]. The literature on analysis of infinite-dimensional control problems that explicitly treat abstract differential-equations as constraints is sparse, as, in theory, the differential-algebraic structure can be resolved in the input to state map. However, there is a vast amount of results on optimal control for the Navier-Stokes Equation, see [1, 56] for an overview and, in particular, [70] for existence and representation of adjoint states.

We define the optimal control problem:

Problem 6.1. *Consider the setup of the abstract DAE in Problem 3.1. Let U be a Banach space and let $\mathcal{U} \subset \big((0,T) \to U\big)$ be the space of input functions. Let $\mathcal{M}\colon H \to \mathbb{R}$ and $\mathcal{K}\colon (0,T)\times V\times U \to \mathbb{R}$ be Fréchet differentiable weighting functions and $\mathcal{K}(t,v,u) = \mathcal{K}_1(t,v) + \mathcal{K}_2(t,u)$ be such that the Nemyckij mappings associated with the partial derivatives, $\mathcal{K}_{;u}\colon \mathcal{U} \to \mathcal{U}'$ and $\mathcal{K}_{;v}\colon \mathcal{V} \to \mathcal{V}'$, are well defined.*

Consider the task of minimizing the cost functional

$$\mathcal{J}\colon \mathcal{V}\times\mathcal{U} \to \mathbb{R}\colon \mathcal{J}(v,u) = \mathcal{M}(v(T)) + \int_0^T \mathcal{K}(t,v(t),u(t))\,\mathrm{d}t, \tag{6.1}$$

subject to the constraints $\mathcal{G}(v,p,u) = 0$ defined via $\mathcal{G}(v,p,u) = 0$, if $(v,u) \in \mathcal{V}\times\mathcal{U}$ solve

$$\begin{aligned} \dot v(t) - A(t,v(t)) - J_1'p(t) - B_1u(t) &= f(t) \quad \text{in } V',\ \text{a.e. in } (0,T), && (6.2a)\\ -J_2v(t) &= g(t) \quad \text{in } Q_H',\ \text{a.e. in } (0,T), && (6.2b)\\ v(0) &= \alpha \quad \text{in } H, && (6.2c)\end{aligned}$$

for a $p \in L^2(0,T;Q_H')$, for given $f \in L^2(0,T;V')$, $g \in L^2(0,T;Q_H')$ and $\alpha \in H$, and where $B_1\colon \mathcal{U} \to L^2(0,T;V')$ is a bounded and injective input operator.

Remark 6.2. The assumption that the partial derivatives of $\mathcal{K}$ with respect to u and v are independent of v and u, respectively, i.e. $\mathcal{K}_{;v}(t,v(t),u(t)) = \mathcal{K}_{;v}(t,v(t))$ and $\mathcal{K}_{;u}(t,v(t),u(t)) = \mathcal{K}_{;u}(t,u(t))$ is no restriction, as long as we consider necessary conditions for solvability and since we allow for explicit time dependency of $\mathcal{K}$.

The terms coupling v and u in the cost functional are commonly referred to as *cross terms*. In the finite-dimensional case, we will illustrate, why the exclusion of cross terms in the theoretical considerations is not a restriction, see Section 8.5.

Remark 6.3. Requiring injectivity of B_1 is necessary to have a unique state (v,p) for any input u. This is not a restriction if one considers B_1 on $\mathcal{U}/\ker B_1$ that can be seen as a closed subspace of $\mathcal{U}$ since $\mathcal{U}$ is a Hilbert space and B_1 is bounded, cf. the discussion in Remark 3.14.

In view of establishing optimality conditions as defined in Theorem 5.5 we state the formal adjoint equation.

Problem 6.4. *Consider Problem 6.1. Let A be Fréchet-differentiable and let $w \in \mathcal{W}(0,T;V;V')$. Find $(\lambda,\mu) \in \mathcal{W}(0,T;V;V') \times L^2(0,T;Q_H)$ that solve*

$$\begin{aligned} -\dot\lambda(t) - A_{;v}(t,w(t))'\lambda(t) - J_2'\mu(t) + \mathcal{K}_{;v}(t,w(t)) &= 0 \quad \text{in } V', && (6.3a)\\ J_1\lambda(t) &= 0 \quad \text{in } Q_H', && (6.3b)\end{aligned}$$

for almost all $t \in (0,T)$, and

$$\lambda(T) = -j\mathcal{M}_{;v}(w(T)) \quad \text{in } H. \tag{6.3c}$$

The formal definition of (6.3) is motivated by the following lemma.

Lemma 6.5. *Consider Problem 6.1. If for a $w \in \mathcal{W}(0,T;V;V')$, there exists a solution $(\lambda,\mu) \in \mathcal{W}(0,T;V;V') \times L^2(0,T;Q_H)$ to* (6.3)*, then $\Lambda := (\lambda,\mu,\lambda(0))$ is in $\mathcal{V} \times \mathcal{Q}_H \times H = \big(\mathcal{V}' \times \mathcal{Q}_H' \times H\big)'$ and*

$$\Lambda\mathcal{G}_{;(v,p)}(w) + \mathcal{J}_{;(v,p)}(w) = 0 \qquad \text{in } \mathcal{V}' \times \mathcal{Q}_H'. \tag{6.4}$$

Proof. The representation of Λ in $\mathcal{V} \times \mathcal{Q}_H \times H$ being a functional on $\mathcal{V}' \times \mathcal{Q}_H \times H$ is well defined by the embedding of $\mathcal{W}(0,T;V,V') \hookrightarrow \mathcal{C}([0,T],H)$ and the reflexivity of the considered spaces.

The Fréchet derivative of $\mathcal{G}$ with respect to the states at (w,p_0) is given via

$$(h_v,h_p) \mapsto \begin{bmatrix} \dot{h}_v - A_{;v}(w)h_v - J_1'h_p \\ -J_2h_v \\ h_v(0) \end{bmatrix}. \tag{6.5}$$

Let $h_v \in \mathcal{W}(0,T;V,V')$ and $h_p \in L^2(0,T;Q_H)$. To interpret the derivative of $\mathcal{M}(v(T))$ applied to h_v we introduce the operator $T\colon \mathcal{W}(0,T) \to H\colon v \mapsto v(T)$ which is linear and well-defined since $\mathcal{W}(0,T) \hookrightarrow \mathcal{C}([0,T],H)$. Then $\mathcal{M}\colon H \to \mathbb{R}$ extends to $\mathcal{M} \circ T\colon \mathcal{W}(0,T) \to \mathbb{R}$ and by the chain rule we can obtain the representation of $\mathcal{M}_{;v}$ in H' via

$$\begin{aligned}\langle \mathcal{M}(T(v))_{;v}, h_v\rangle_{\mathcal{V}',\mathcal{V}} &= \mathcal{M}_{;T(v)}(T(v))\big[T_v[h_v]\big] \\ &= \langle \mathcal{M}_{;T(v)}T(v), T(h_v)\rangle_{H',H} = \langle \mathcal{M}_{;v(T)}v(T), h_v(T)\rangle_{H',H},\end{aligned}$$

since $T(h_v) = h_v(T) \in H$. Thus, omitting the arguments (w,p_0), we have

$$\begin{aligned}(\Lambda\mathcal{G}_{;(v,p)} + \mathcal{J}_{;(v,p)})[h_v,h_p] &= \langle \lambda, \dot{h}_v\rangle_{\mathcal{V},\mathcal{V}'} + \langle \lambda, -A_{;v}h_v - J_1'h_p\rangle_{\mathcal{V},\mathcal{V}'} \\ &\quad - \langle \mu, J_2h_v\rangle_{\mathcal{Q},\mathcal{Q}'} + \big(\lambda(0), h_v(0)\big)_H + \\ &\quad \big(j\mathcal{M}_{;v}(w(T)), h_v(T)\big)_H + \langle \mathcal{K}_{;v}, h_v\rangle_{\mathcal{V}',\mathcal{V}} \\ &= -\langle \dot{\lambda}, h_v\rangle_{\mathcal{V}',\mathcal{V}} + \langle \lambda, -A_{;v}h_v - J_1'h_p\rangle_{\mathcal{V},\mathcal{V}'} \\ &\quad - \langle \mu, J_2h_v\rangle_{\mathcal{Q},\mathcal{Q}'} + \langle \mathcal{K}_{;v}, h_v\rangle_{\mathcal{V}',\mathcal{V}} \\ &= \langle -\dot{\lambda} - A_{;v}'\lambda - J_2'\mu + \mathcal{K}_{;v}, h_v\rangle_{\mathcal{V}',\mathcal{V}} + \langle -J_1\lambda, h_p\rangle_{\mathcal{Q}',\mathcal{Q}} \\ &= 0,\end{aligned}$$

where we have used the assumption that (λ,μ) solves Problem (6.4) and, in particular, that λ admits the application of the formula of integration by parts, as given, e.g., in [45, p. 147] to obtain

$$\langle \lambda, \dot{h}_v\rangle_{\mathcal{V},\mathcal{V}'} = -\langle \dot{\lambda}, h_v\rangle_{\mathcal{V}',\mathcal{V}} + \big(\lambda(T), h_v(T)\big)_H - \big(\lambda(0), h_v(0)\big)_H. \tag{6.6}$$

□

Remark 6.6. Since $\mathcal{G}$ is linear in p and $\mathcal{J}$ does not depend on p, the algebraic variable p does only formally appear in the definitions of the Fréchet derivatives.

Since with A and $\mathcal{J}$ being Fréchet-differentiable, see Theorem 5.6, the smoothness assumptions in Theorem 5.5 are fulfilled and we can state the following corollary:

Corollary 6.7. *Consider Problem 6.1 and assume that the ADAE* (6.2) *has a unique solution for all inputs. Let $u_0 \in \mathcal{U}$ and let $(v(u_0), p(u_0))$ be the corresponding solution. If $(v(u_0), u_0)$ is a local minimum of* (6.1)*, then for (λ,μ) solving the formal adjoint equation* (6.3) *at $w = v(u_0)$ one has*

$$-\lambda B_1 + \mathcal{K}_{;u}(u_0) = 0 \qquad \text{in } \mathcal{U}'. \tag{6.7}$$

For further reference we state another immediate corollary:

Corollary 6.8. *Consider Problem 6.1 and assume that the ADAE* (6.2) *has a unique solution for all inputs u. Let $u_0 \in \mathcal{U}$ and let $(v(u_0), p(u_0))$ be the corresponding solution. If the adjoint equation* (6.3) *has a solution (λ, μ) at $w = v(u_0)$, then the mutual solvability of* (6.2), (6.3), *and* (6.7) *is necessary for optimality of $(v(u_0), u_0)$. This means that $(v, p, \lambda, \mu, u_0)$ is a solution to*

$$\dot{v}(t) - A(t, v(t)) - J_1' p(t) - B_1 u_0(t) = f(t) \quad \text{in } V', \tag{6.8a}$$

$$-J_2 v(t) = g(t) \quad \text{in } Q_H', \tag{6.8b}$$

$$-\dot{\lambda}(t) - A_{;v}(t, v(t))' \lambda(t) - J_2' \mu(t) + \mathcal{K}_{;v}(t, v(t)) = 0 \quad \text{in } V', \tag{6.8c}$$

$$J_1 \lambda(t) = 0 \quad \text{in } Q_H', \tag{6.8d}$$

for almost all $t \in (0, T)$, and

$$v(0) = \alpha \quad \text{in } H, \tag{6.8e}$$

$$\lambda(T) + j \mathcal{M}_{;v}(v(T)) = 0 \quad \text{in } H, \tag{6.8f}$$

as well as

$$-\lambda B_1 + \mathcal{K}_{;u}(u_0) = 0 \quad \text{in } \mathcal{U}'. \tag{6.8g}$$

Remark 6.9. The necessary optimality conditions (6.8) fail in the case that the optimal control problem has a solution while the adjoint ADAE is not solvable because of inconsistency or insufficient regularity of the data, cf. [91, 92]. Similarly, the sufficient conditions that we will state in Corollary 6.10 may fail, if the linearized ADAE with zero initial conditions does not have a solution.

We reformulate the sufficient second order conditions given in Theorem 5.9 for the optimal control problem defined in Problem 6.1.

Corollary 6.10. *Let A, $\mathcal{M}$, and $\mathcal{K}$ be two times Fréchet differentiable. Let $(v_0, u_0) \in \mathcal{V} \times \mathcal{U}$ solve* (6.2) *and assume that the Fréchet derivative with respect to state and control, cf.* (6.5), *of the constraints is surjective. Then (v_0, u_0) is locally optimal for* (6.1) *if there is (λ, μ) solving* (6.3) *and* (6.7) *at (v_0, u_0) and if there is $c > 0$ such that*

$$\begin{aligned}\big(\mathcal{M}_{;vv}(v_0(T)) h_v(T), h_v(T)\big)_H + \big\langle \mathcal{K}_{;(v,u)^2}(h_v, h_u), (h_v, h_u)\big\rangle_{\mathcal{V}' \times \mathcal{U}', \mathcal{V} \times \mathcal{U}} \\ - \big\langle \lambda A_{;vv} h_v, h_v \big\rangle_{\mathcal{V}', \mathcal{V}} \geq c \|(h_v, h_u)\|^2_{\mathcal{V} \times \mathcal{Q}}\end{aligned} \tag{6.9}$$

at (v_0, u_0) and for all $(h_v, h_u) \in T_{(v_0, u_0)} \subset \mathcal{V} \times \mathcal{U}$, i.e. all (h_v, h_0) solving the linearized about (v_0, u_0) state equations.

Remark 6.11. As p appears only linear, the above optimality conditions do not depend on the associated algebraic variable $p_0 := p(u_0)$. For formal considerations, we will also use the tangent space $T_{(v_0, p_0, u_0)} \subset \mathcal{V} \times \mathcal{Q} \times \mathcal{U}$ containing all (h_v, h_p, h_u) that solve the linearized state equations.

To investigate existence of solutions to the adjoint ADAE (6.3), we can use a similar approach as for the primal ADAE (3.1). In particular, Assumption 3.5 also works for the decoupling of the adjoint equation:

Lemma 6.12. *Consider Problem 3.1 and assume that Assumption 3.5 holds. Then,*

$$S_{ad} := \bar{J}_1 j \bar{J}_2{}' : Q \to Q'$$

is invertible and

$$
\begin{aligned}
\ker \bar{J}_1 &= H_{\mathrm{c}\perp}, && \textit{(a)}\\
\operatorname{im} j\bar{J}_2{}' &= H_{\mathrm{df}\perp}, && \textit{(b)}\\
H &= H_{\mathrm{c}\perp} \oplus H_{\mathrm{df}\perp}, && \textit{(c)}\\
\textit{and} \quad H' &= H_{\mathrm{c}}{}^0 \oplus H_{\mathrm{df}}{}^0, && \textit{(d)}
\end{aligned}
$$

where H_{c} and H_{df} are defined in Lemma 3.6.

Proof. By assumption, the operators $\bar{J}_1$ and $\bar{J}_2$ are surjective, and thus closed, so we can use the *Closed Range Theorem*, see e.g. [83, Thm. IV.5.13]. Using the identities $\left(H_g{}^0\right)^0 = H_g$ and $H_g{}^0 = \left(j(H_g)\right)_\perp$ that hold for subspaces of Hilbert spaces, we conclude with the arguments of Lemma 3.20 that

$$\ker \bar{J}_1 = \left(\operatorname{im} \bar{J}_1{}'\right)^0 = \left(H_{\mathrm{c}}'\right)^0 = \left(H_{\mathrm{c}\perp}{}^0\right)^0 = H_{\mathrm{c}\perp}$$

and that

$$j(\operatorname{im} \bar{J}_2{}') = j(\left(\ker \bar{J}_2\right)^0) = j(H_{\mathrm{df}}{}^0) = H_{\mathrm{df}\perp}.$$

This proves parts (a) and (b). Ad (c): $j(\operatorname{im} \bar{J}_1{}') \oplus \ker \bar{J}_2 = H_{\mathrm{c}\perp} \oplus H_{\mathrm{df}\perp} = (H_{\mathrm{c}} \cap H_{\mathrm{df}})_\perp = H$, cf. [46, Thm. 7.57], since $H = H_{\mathrm{c}} \oplus H_{\mathrm{df}}$. Part (d) follows by Lemma 3.20 and [83, Thm. IV.4.8].

Since by assumption $\bar{J}_1{}'$ and $\bar{J}_2{}'$ are homeomorphisms onto their range and since $H = j(\operatorname{im} \bar{J}_1{}') \oplus \ker \bar{J}_2$, the invertibility of S_{ad} follows by the same arguments that were used to show that $S := J_2 j J_1$ is invertible, see the proof of Lemma 3.6. □

As for the decoupling of the ADAE (3.1), we need additional regularity of the problem.

Assumption 6.13 (Cf. Assumption 3.8)**.** *Let Assumption 3.5 hold. For an inhomogeneity $\mathcal{K}_{;v}(v) \in \mathcal{H}'$ (rather than in $\mathcal{V}'$) and $v \in \mathcal{W}(0,T;V,V')$, such that $-j\mathcal{M}(v(T))$ is sufficiently smooth, any corresponding solution (λ, μ) to the formal adjoint ADAE* (6.3) *is such that $A_{;v}(w)\lambda \in \mathcal{H}'$ and such that $\mu \in \mathcal{Q}$.*

Remark 6.14. Since there is no inhomogeneity in (6.3b), there is no smoothness constraint as for the primal equations, cf. Assumption 3.8.

To decouple the solutions of the adjoint ADAE analogously to the solutions of the DAE, we need to mirror Assumption 3.13:

Assumption 6.15. *The operator $J_2' \in \mathcal{L}(Q_H, V')$ is a homeomorphism onto its range and $J_1 \in \mathcal{L}(V, Q_H')$ has a bounded right inverse.*

Remark 6.16. If Assumption 3.13 holds, then J_2' is a homeomorphism onto its range. However, Assumption 3.13 is not sufficient for a bounded right inverse to J_1, cf. Remark 3.14.

Theorem 6.17. *Consider Problem 3.1, let Assumptions 3.5 and 6.13 hold, and let $\mathcal{K}_{;v}(w) \in \mathcal{H}'$. Then the formal adjoint ADAE* (6.3) *has a (unique) solution $(\lambda, \mu) \in \mathcal{W}(0,T;V;H') \times L^2(0,T;Q)$ if, and only if, $j\mathcal{M}_{;v}(w(T))$ is in $\ker \bar{J}_1$ and*

$$-\mathcal{P}_{[H_{\mathrm{c}}{}^0|H_{\mathrm{df}}{}^0]}\dot{\lambda}_{\mathrm{df}'} - \mathcal{P}_{[H_{\mathrm{c}}{}^0|H_{\mathrm{df}}{}^0]}A_{;v}(w)'\lambda_{\mathrm{df}'} - \mathcal{P}_{[H_{\mathrm{c}}{}^0|H_{\mathrm{df}}{}^0]}\mathcal{K}_{;v}(w) = 0, \quad \textit{in } \mathcal{V}' \tag{6.10a}$$

$$\lambda_{\mathrm{df}'}(T) = -j\mathcal{M}_{;v}(w(T)), \tag{6.10b}$$

has a (unique) solution $\lambda_{\mathrm{df}'} \in \mathcal{W}(0,T;\ker J_1;\mathcal{P}_{[H_{\mathrm{c}}{}^0|H_{\mathrm{df}}{}^0]}H')$, where $\mathcal{P}_{[H_{\mathrm{c}}{}^0|H_{\mathrm{df}}{}^0]} := I_{H'} - \bar{J}_2{}'S_{ad}^{-1}\bar{J}_1 j$.

Proof. The formal adjoint ADAE (6.3) has the same structure as the ADAE (3.1) with $\bar{J}_1$ and $\bar{J}_2$ interchanged. By assumption and, in particular, by Lemma 6.12, all conditions of Theorem 3.37 are fulfilled. In particular, we can define the decoupling projector

$$\mathcal{P}_{[H_{\mathrm{c}}{}^0|H_{\mathrm{df}}{}^0]} := I_{H'} - \bar{J}_2{}' S_{ad}^{-1} \bar{J}_1 j \tag{6.11}$$

that has the image of $\bar{J}_2{}'$ as its kernel and $j'(\ker \bar{J}_1)$ as its range. □

Remark 6.18. In the optimal control setup, the condition $j\mathcal{M}_{;v}(w(T)) \in \ker \bar{J}_1$ is not restrictive. By Theorem 3.37, for a solution v, one has that $v(T) = \mathcal{P}_{[H_{\mathrm{df}}]} v(T) - j\bar{J}_1{}' S^{-1} g(T)$. Thus, the endpoint restriction in the cost functional (6.1), if constrained by the DAE under the conditions of Theorem 3.37, can also be written as $\mathcal{M}\big(\mathcal{P}_{[H_{\mathrm{df}}]} v(T) - j\bar{J}_1 S^{-1} g(T)\big)$, where $\mathcal{P}_{[H_{\mathrm{df}}]} := I_H - j\bar{J}_1{}' S^{-1} \bar{J}_2$. In this case, the end condition for λ in (6.3) reads

$$\begin{aligned} -jP_H{}'\mathcal{M}_{;v}(w(T)) &= j[I_{H'} - \bar{J}_2{}' S_{ad}^{-1} \bar{J}_1{}' j]\mathcal{M}_{;v}(w(T)) \\ &= [I_H - j\bar{J}_2{}' S_{ad}^{-1} \bar{J}_1] j\mathcal{M}_{;v}(w(T)), \end{aligned}$$

which is in the kernel of $\bar{J}_1$.

Remark 6.19. With the same arguments as for the state equations, one can replace the solution space $\mathcal{W}(0,T;V;V')$ for λ by $\mathcal{W}^{1;p',q'}(0,T;V;V')$, where p', q' are the conjugated exponents to p, q in the definition of the space $\mathcal{W}^{1;p,q}(0,T;V,V')$ that is used to formulate the state equations, see Remark 3.40.

As an example, we consider the linear-quadratic optimal control problem to minimize

$$\begin{aligned} \mathcal{J}(v,u) =& \frac{1}{2}\big(\mathcal{M}_1[v(T) - v^*(T)], v(T) - v^*(T)\big)_H + \\ &+ \frac{1}{2}\int_0^T \big(\mathcal{K}_1[v(t) - v^*(t)], v(t) - v^*(t)\big)_H + \big(\mathcal{R}u(t), u(t)\big)_U \, \mathrm{d}t, \end{aligned} \tag{6.12}$$

subject to (3.1) with linear A, with a target state $v^* \in L^2(0,T;H)$, and with self-adjoint and positive operators $\mathcal{M}_1$, $\mathcal{K}_1\colon H \to H$ and self-adjoint, positive and invertible $\mathcal{R}\colon U \to U$.

Assume, that for given data the unique solvability of the ADAE and its formal adjoint with respect to the quadratic cost functional has been established, e.g. via Theorems 3.37 and 6.17. Then, by Lemma 6.5, if (v,u) are optimal, then the system

$$\begin{aligned} \dot{v} - Av - J_1'p - B_1 u = f, &\quad v(0) = \alpha \\ -J_2 v = g,& \end{aligned} \tag{6.13a}$$

$$\begin{aligned} \dot{\lambda} - A'\lambda - J_2'\mu + \mathcal{K}_1 v = \mathcal{K}_1 v^*, &\quad \lambda(T) + \mathcal{M}_1 v(T) = 0, \\ -J_1\lambda = 0,& \end{aligned} \tag{6.13b}$$

$$-B_1'\lambda + \mathcal{R}u = 0, \tag{6.13c}$$

is solvable.

6.1. **Semi-discretization of the Adjoint Equation.** In this section we address the semi-discretization of the necessary optimality conditions (6.8). We start with formulating a semi-discretization for the formal adjoint equation (6.3) in the same way as for the state equations in Section 4. Then, we formulate the conditions under which the semi-discretized formal adjoint equation is the formal adjoint to the semi-discretized state equations with respect to the cost functional (6.1). This will in particular answer the question, when does the spatial discretization of the necessary optimality conditions coincide with the necessary optimality condition for the semi-discretized optimal control problem.

We define the semi-discrete approximations to the formal adjoint equation (6.3):

Problem 6.20. *Consider Problem 6.1 and let $Q \subset Q_H$ as in Assumption 3.5. Assume that A, $\mathcal{K}$, and $\mathcal{M}$ are Fréchet differentiable. Let the approximation schemes $\{V_{k'}\}_{k' \in \mathbb{N}}$, $\{Q_{k'}\}_{k' \in \mathbb{N}}$ to V, Q fulfill* (4.1)*. Let $k \in \mathbb{N}$ and consider the discrete Gelfand triples* (4.5)*. For given $w \in \mathcal{W}(0,T;V;V')$ and ω_k approximating the terminal value $-j\mathcal{M}_{;v}(w(T))$ in H_k, find $(\lambda_k, \mu_k) \in \mathcal{W}(0,T;V_k,(V_k)') \times L^2(0,T;{Q_H}'_k)$ that fulfill*

$$-\dot{\lambda}_k - A_{k;v}(t,w(t))'\lambda_k(t) - {J_2}'_k\mu_k(t) + \mathcal{K}_{k;v}(t,w(t)) = 0 \quad \textit{in } (V_k)', \tag{6.14a}$$

$$J_{1k}\lambda_k(t) = 0 \quad \textit{in } {Q_H}'_k, \tag{6.14b}$$

for almost all $t \in (0,T)$, and

$$\lambda_k(T) = \omega_k \quad \textit{in } H, \tag{6.14c}$$

with J_{2k} and J_{1k} as defined in (4.6) *and with $A_{k;v}$ and $\mathcal{K}_{k;v}$ denoting the restrictions of $A_{;v}$ and $\mathcal{K}_{;v}$ onto V_k.*

System (6.14) arises from necessary optimality conditions for an unknown function $w \in \mathcal{W}(0,T;V;V')$. In practice, at a discretization level k, w may not be accessible but a finite-dimensional approximation $w_k \in \mathcal{W}(0,T;V_k;V_k') \cap L^\infty(0,T;H)$, cf. Lemma 4.35. Thus, we formulate a further approximation to the formal adjoint equation (6.14).

Problem 6.20(w_k). *Consider the setup of Problem 6.20. Let $w_k \in L^\infty(0,T;H_k) \cap \mathcal{W}(0,T;V_k;(V_k)')$ and let ω_k approximate the terminal value $-j\mathcal{M}_{;v}(w_k(T))$ in H_k. Find $(\lambda_k,\mu_k) \in \mathcal{W}(0,T;V_k,(V_k)') \times L^2(0,T;{Q_H}'_k)$ that fulfill*

$$-\dot{\lambda}_k - A_{k;v}(t,w(t))'\lambda_k(t) - {J_2}'_k\mu_k(t) + \mathcal{K}_{k;v}(t,w_k(t)) = 0 \quad \textit{in } (V_k)', \tag{6.15a}$$

$$J_{1k}\lambda_k(t) = 0 \quad \textit{in } {Q_H}'_k, \tag{6.15b}$$

for almost all $t \in (0,T)$, and

$$\lambda_k(T) = \omega_k \quad \textit{in } H, \tag{6.15c}$$

with J_{2k} and J_{1k} as defined in (4.6) *and with $A_{k;v}$ and $\mathcal{K}_{k;v}$ denoting the restrictions of $A_{;v}$ and $\mathcal{K}_{;v}$ onto V_k.*

We will establish conditions for solvability of Problem 6.20 for $k \in \mathbb{N}$ via a decoupling as it was introduced for the semi-discrete state equations in Section 4.

Then we investigate convergence of the discrete solutions of (6.14) to solutions of the formal adjoint equations (6.3). This will also give a solvability result for the formal adjoint that is needed for the formulation of the necessary optimality conditions given in Corollary 6.8.

We start with a splitting of the solution space.

Proposition 6.21. *Consider Problem 6.20 with the operators $J_{1k} \colon V_k \to {Q_H}'_k$ and ${J_2}'_k \colon Q_H \to (V_k)'$. If Assumption 4.8(a) holds, then, for all $k \in \mathbb{N}$, J_{1k} has a right inverse ${J_1}_k^-$ and ${J_2}'_k$ is a homeomorphism onto its range. Furthermore, it holds that*

$$V_k = \ker J_{1k} \oplus \operatorname{im} {J_1}_k^-. \tag{6.16}$$

Proof. By Assumption 4.8(a) one has that J_{2k} has a right inverse and that J_{1k}' is a homeomorphism onto its range. The former implies that J_{2k}' is a homeomorphism onto its range and the latter, together with finite-dimensionality of V_k that makes every subspace complementable, implies that J_{1k} has a right inverse, cf. Remark 3.14. The projection $I - J_{1k}^{-}J_{1k}$ gives the splitting (6.16). □

In view of decoupling the equations (6.14), we state the following result:

Proposition 6.22. *Consider Problem 6.20, let Assumption 4.8(a) hold, let the extensions $\bar{J}_{1k}$, $\bar{J}_{2k}\colon H_k \to (Q_k)'$ be defined as in* (4.8)*, and let $k \in \mathbb{N}$. Then there is a constant $\gamma_2'(k)$, such that*

$$\|\bar{J}_{2k}'q_k\|_{H'} \geq \gamma_2'(k)\|q_k\|_Q, \qquad \textit{for all } q_k \in Q_k,$$

If, in addition, Assumption 4.13(a) holds, then there is a constant $\gamma_1'(k)$, such that

$$\|\bar{J}_{1k}h_k\|_{Q'} \geq \gamma_2(k)\|h_k\|_H, \tag{6.17}$$

for all $h_k \in j(\mathrm{im}\,\bar{J}_{2k}')$.

Proof. By Assumption 4.8 and Proposition 6.22, $\bar{J}_{1k}'$, $\bar{J}_{2k}'$ are homeomorphisms onto their range, cf. Proposition 4.12, from which also the existence of $\gamma_2'(k)$ is inferred. Given this, by Lemma 4.17, Assumption 4.13(a) implies that $H_k = \ker \bar{J}_{2k} \oplus \mathrm{im}\, j\bar{J}_{1k}'$. With the arguments of Lemma 6.12, this implies that $H_k = \ker \bar{J}_{1k} \oplus \mathrm{im}\, j\bar{J}_{2k}'$, which, by Corollary 3.7, implies existence of γ_1' as in (6.16). □

Based on Propositions 6.21 and 6.22, and since Assumption 3.8 has no meaning at a fixed discretization level k, we conclude that the conditions that were sufficient for the decoupling of the semi-discrete state equations defined in Problem 4.7, are also sufficient for the decoupling of the semi-discretizations of the formal adjoint equations as given in Problem 6.20.

As in the continuous case, cf. Lemma 6.12, if the conditions of Proposition 6.22 are fulfilled, then we have that

$$H_k = H_{k\mathrm{c}\perp} \oplus H_{k\mathrm{df}\perp} \quad \text{and} \quad H_k' = {H_{k\mathrm{c}}}^0 \oplus {H_{k\mathrm{df}}}^0,$$

where $H_{k\mathrm{c}}$ and $H_{k\mathrm{df}}$ are defined in Lemma 4.17 and with $H_{k\mathrm{c}\perp} = \ker \bar{J}_{1k}$ and ${H_{k\mathrm{df}}}^0 = j'(H_{k\mathrm{df}\perp}) = \mathrm{im}\,\bar{J}_{2k}'$. In particular, we can define the decoupling projector

$$\mathcal{P}_{[{H_{k\mathrm{c}}}^0|{H_{k\mathrm{df}}}^0]}\colon H_k' \to H_k', \tag{6.18}$$

with $\ker \mathcal{P}_{[{H_{k\mathrm{c}}}^0|{H_{k\mathrm{df}}}^0]} = \mathrm{im}\,\bar{J}_{2k}'$ and $\mathrm{im}\,\mathcal{P}_{[{H_{k\mathrm{c}}}^0|{H_{k\mathrm{df}}}^0]} = j(\ker \bar{J}_{1k})$.

Lemma 6.23. *Consider Problem 6.20 and let the conditions of Propositions 6.21 and 6.22 be fulfilled. Then any $(\lambda_k, \mu_k) \in \mathcal{W}(0,T;V_k,V_k') \times L^2(0,T;Q_{H'k})$ that fulfill (6.14a-b) also solve*

$$\begin{aligned}-\mathcal{P}_{[{H_{k\mathrm{c}}}^0|{H_{k\mathrm{df}}}^0]}\dot{\lambda}_k(t) - \mathcal{P}_{[{H_{k\mathrm{c}}}^0|{H_{k\mathrm{df}}}^0]}A_{k;v}(w(t))'\lambda_k(t)& \\ - \mathcal{P}_{[{H_{k\mathrm{c}}}^0|{H_{k\mathrm{df}}}^0]}\mathcal{K}_{k;v}(w(t)) = 0,& \quad \textit{in } V_k'\end{aligned}$$

and

$$\mu_k(t) = -S_{k\,ad}^{-1}\bar{J}_{1k}j[A_{k;v}(w(t))'\lambda_k(t) - \mathcal{K}_{k;v}(w(t))], \quad \textit{in } Q_k,$$

where $S_{k\,ad} := J_{1k}j\bar{J}_{2k}'\colon Q_k \to (Q_k)'$ and the equalities hold for almost all $t \in (0,T)$ as specified in Lemma 3.30.

Proof. This is a special case of Lemma 3.30 with a zero right hand side in the algebraic constraint. □

Application of Corollary 3.31 gives necessary solvability conditions for Problem 6.20.

Corollary 6.24. *Consider Problem 6.20 and let the conditions of Propositions 6.21 and 6.22 be fulfilled. For the existence of* $(\lambda_k, \mu_k) \in \mathcal{W}(0,T; V_k, (V_k)') \times L^2(0,T; Q_{H'_k})$ *solving* (6.14) *it is necessary, that* $\omega_k \in \ker \bar{J}_{1k}$.

Application of Theorem 3.37 gives the following necessary and sufficient conditions. Note that because of linearity of the equations, there can only be one solution.

Corollary 6.25. *Consider Problem 6.20 and let the conditions of Propositions 6.21 and 6.22 be fulfilled. Then, for given* $w \in \mathcal{W}(0,T;V;V')$*, the semi-discretized formal adjoint equation* (6.14) *has a solution, if, and only if,* $\omega_k \in \ker \bar{J}_{1k}$ *and*

$$-\dot{\lambda}_k(t) - \mathcal{P}_{[{H_k}_{\mathrm{c}}{}^0 | {H_k}_{\mathrm{df}}{}^0]} A_{k;v}(t, w(t))' \lambda_k(t) \tag{6.19a}$$

$$- \mathcal{P}_{[{H_k}_{\mathrm{c}}{}^0 | {H_k}_{\mathrm{df}}{}^0]} \mathcal{K}_{k;v}(w(t)) = 0 \quad \textit{in } (\ker J_{1k})', \tag{6.19b}$$

$$\lambda_k(T) = \omega_k \quad \textit{in } H, \tag{6.19c}$$

for almost all $t \in (0,T)$*, has a solution in* $\mathcal{W}(0,T; \ker J_{1k}; j'(\ker \bar{J}_{1k}))$*. If a solution exists, then it is unique.*

Remark 6.26. As the results above hold for a fixed discretization level k, one can replace w by any w_k, so that all derivations hold in particular for Problem 6.20(w_k).

6.1.1. *Convergence of the Discrete Solutions.* To establish convergence of the discrete solutions (λ_k, μ_k) to a solution of the continuous problem (6.3), we proceed in a similar way as for the state equations in Section 4. We will assume existence of a stable decoupling and prove convergence in the differential part, i.e. that the sequence of discrete solutions to (6.19) converges to a $\lambda \in \mathcal{W}(0,T;V;V')$ solving (6.10). By stability of the decoupling, we then can reconstruct convergent sequences of the algebraic variables μ_k.

The formal adjoint equation is linear in (λ, μ) so that one can employ the theory of Galerkin approximations for linear evolution equations, see, e.g., [161]. However, as discussed in Section 4, a standard semi-discretization of the ADAE (6.3) leads to an *external approximation* scheme to (6.19). Also, existence and uniqueness of solutions will depend on properties of the Fréchet derivative $A_{;v}$ of the nonlinearity A of the state equations (3.1). We will establish convergence for a generic case of A, for which the results on external approximations derived in Section 4.4 carry over to the approximation of the adjoint equation.

We start with formulating the counterpart to Assumption 4.28, that ensured uniform boundedness, semi-coerciveness, and pseudomonotonicity of the nonlinearity in the state equations, for the operator $A_{;v}(t, w(t))' \colon V \to V'$ from the formal adjoint (6.3). For the case that $w \in \mathcal{W}(0,T;V;V')$ is not directly available but a sequence $\{w_k\}_{k \in \mathbb{N}} \subset \mathcal{W}(0,T;V;V')$ that converges to w in $L^2(0,T;V)$, cf. Problem 6.20(w_k), in this section, we formulate the conditions uniformly with respect to such a converging sequence. In Remark 6.39 will discuss how the conditions can be relaxed for the case that $w_k \equiv w \in \mathcal{W}(0,T;V;V')$ as defined in Problem 6.20.

Assumption 6.27. *Consider Problem 6.4 and its finite-dimensional approximation defined in Problem 6.20. Given a* $\{w_k\}_{k \in \mathbb{N}} \subset L^2(0,T;V)$ *such that* $w^k \rightharpoonup w$ *in* $L^2(0,T;V)$*, as* $k \in \mathbb{N}$*. For* $k \in \mathbb{N}$ *consider*

$$A_{;v}(w_k)' \colon L^2(0,T;V) \to \big((0,T) \to V'\big) \colon v \mapsto \mathcal{N}_{A_{;v}(w_k)'} v, \tag{6.20}$$

defined via the Nemyckij map of $A_{;v}(w_k) \colon (0,T) \times V \to V' : (t,v) \mapsto A_{;v}(t, w_k(t))'v$.

We assume that $A_{;v}(w_k)' \colon L^2(0,T;V) \to L^2(0,T;V')$ *is bounded and that it has the following properties:*

(a) *Bounded Growth: There is $\gamma \in L^1(0,T)$ and $\beta\colon \mathbb{R} \to \mathbb{R}$, increasing, such that for all $v \in V$ with $A_{;v}(t, w_k(t))' \in H'$ for almost all $t \in (0,T)$ it holds that*

$$\|A_{;v}(t, w_k(t))'\|_{H'} \leq \beta(\|v\|_H)(\gamma(t) + \|v\|_H), \tag{6.21}$$

(b) *w_k-uniform Semi-Coerciveness: There is $c_0 > 0$, $c_1 \in L^2(0,T)$, and $c_2 \in L^1(0,T)$, such that for all $v \in V$:*

$$\big\langle A_{;v}(t, w_k(t))'v, v\big\rangle_{V',V} \geq c_0\|v\|_V^2 - c_1(t)\|v\|_V - c_2(t)\|v\|_H^2, \tag{6.22}$$

and

(c) *w_k-uniform Pseudomonotonicity, cf. [133, Def. 2.1]: For given $\{u_k\}_{k\in\mathbb{N}} \subset L^2(0,T;V)$, with $u_k \rightharpoonup u \in L^2(0,T;V)$, it holds that if*

$$\limsup_{k\to\infty} \big\langle A_{;v}(w_k)'u_k, u_k - u\big\rangle_{\mathcal{V}',\mathcal{V}} \geq 0$$

then it follows that for any $v \in L^2(0,T;V)$,

$$\big\langle A_{;v}(w)'v, u - v\big\rangle_{\mathcal{V}',\mathcal{V}} \leq \liminf_{k\to\infty} \big\langle A_{;v}(w_k)'u_k, u_k - v\big\rangle_{\mathcal{V}',\mathcal{V}}. \tag{6.23}$$

The assumed bounds in (a) and (b) are independent of $k \in \mathbb{N}$ and hold for almost all $t \in (0,T)$.

Remark 6.28. We will base our convergence analysis on v_k-pseudomonotonicity, since we can show that for a generic case, that includes the Navier-Stokes Equation as given in Problem 3.1(NSE), the v_k-pseudomonotonicity carries over from the state equations to the formal adjoint so that we can directly apply Theorem 4.39. However, considering the linearity of the equation one may also extend standard results, as given in [161], to the considered setup of external approximations and k-dependent operators.

Proposition 6.29. *Consider Problem 3.1 and assume that the Nemyckij map of $A\colon (0,T) \times V \to V'$ is given as the sum of a strongly continuous $A^c\colon L^2(0,T;V) \to L^2(0,T;V')$ and a linear, bounded, and positive part $A^m\colon L^2(0,T;V) \to L^2(0,T;V')$. Then,*

(a) *$A = A^m + A^c$ is pseudomonotone.*

If, in addition, A is Fréchet-differentiable, then

(b) *the linear operator $A_{;v}(w)'\colon L^2(0,T;V) \to L^2(0,T;V')$ is pseudomonotone, for all $w \in L^2(0,T;V)$.*

If, in addition, $w \mapsto A_{;v}(w)$ is continuous, then

(c) *$A_{;v}(w_k)'\colon L^2(0,T;V) \to L^2(0,T;V')$ is w_k-pseudomonotone, for any strongly convergent $\{w_k\}_{k\in\mathbb{N}} \subset L^2(0,T;V)$.*

Proof. Ad (a): Consider a sequence of $u_k \in \mathcal{V}$, such that $u_k \rightharpoonup u \in \mathcal{V}$. Then, because of $\mathcal{V}'' = \mathcal{V}$, for all $f \in \mathcal{V}'' = \mathcal{V}$, we have that $\big\langle A^m u_k - Au, f\big\rangle_{\mathcal{V}',\mathcal{V}} = \big\langle u_k - u, A^{m\prime} f\big\rangle_{\mathcal{V},\mathcal{V}'} \to 0$, which means that A^m is demicontinuous. Then, A as the sum of a monotone and demicontinuous and a strongly continuous operator, is pseudomonotone [162, Prop. 27.6].

Ad (b): Since A^m is linear, $A^m_{;v}(\cdot)' = A^{m\prime}$ and, thus, $A^m_{;v}(\cdot)'$ again is linear, bounded, and monotone. Since A^c is strongly continuous, it is compact and thus $A_{;v}(w)$ is compact, cf. [87, Lem. 4.4.1]. Since then $A_{;v}(w)'$ is compact [135] and, thus, as $A_{;v}(w)'$ is linear, it is also strongly convergent [162, Prop. 26.2]. Then the claim follows with the arguments of (a).

Ad (c): Since the mapping $A_{;v}(w) \mapsto A_{;v}(w)'$ is linear and preserves the norm and since the considered spaces are reflexive, we have that $L^2(0,T;V) \ni w \mapsto A_{;v}(w) \in \mathcal{L}(L^2(0,T;V), L^2(0,T;V'))$ is continuous. Let $\{w_k\}_{k\in\mathbb{N}} \subset L^2(0,T;V)$ be such that $w_k \to w$ in $L^2(0,T;V)$, then $\|A_{;v}(w_k)' - A_{;v}(w)\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \to 0$, as $k \to \infty$. We show that for $\{u_k\}_{k\in\mathbb{N}} \subset \mathcal{V}$, such that $u_k \rightharpoonup u \in \mathcal{V}$, it holds that

$$\limsup_{k\to\infty} \big\langle A_{;v}(w_k)' u_k, u_k - u \big\rangle_{\mathcal{V}',\mathcal{V}} = \limsup_{k\to\infty} \big\langle A_{;v}(w)' u_k, u_k - u \big\rangle_{\mathcal{V}',\mathcal{V}}$$

and that, for any $v \in L^2(0,T;V)$,

$$\liminf_{k\to\infty} \big\langle A_{;v}(w_k)' u_k, u_k - v \big\rangle_{\mathcal{V}',\mathcal{V}} = \liminf_{k\to\infty} \big\langle A_{;v}(w)' u_k, u_k - v \big\rangle_{\mathcal{V}',\mathcal{V}}.$$

If this is the case, then pseudomonotonicity of $A_{;v}(w)'$ (that was established in (b)) will imply w_k-pseudomonotonicity of $A_{;v}(w_k)'$, cf. the Proof of Lemma 4.29.

Since, for $k \in \mathbb{N}$,

$$\big\langle A_{;v}(w_k)' u_k, u_k - u \big\rangle_{\mathcal{V}',\mathcal{V}} = \big\langle [A_{;v}(w_k)' - A_{;v}(w)' + A_{;v}(w)'] u_k, u_k - u \big\rangle_{\mathcal{V}',\mathcal{V}}$$

and, with the strong convergence of $\{w_k\}_{k\in\mathbb{N}}$ and boundedness of $\{u_k\}_{k\in\mathbb{N}}$ we have that

$$\begin{aligned} &|\big\langle [A_{;v}(w_k)' - A_{;v}(w)'] u_k, u_k - u \big\rangle_{\mathcal{V}',\mathcal{V}}| \\ &\qquad\qquad \leq \|A_{;v}(w_k)' - A_{;v}(w)'\|_{\mathcal{L}(\mathcal{V},\mathcal{V}')} \|u_k\|_{\mathcal{V}} \|u_k - u\|_{\mathcal{V}} \to 0, \end{aligned}$$

as $k \to \infty$, the equality of the limits superior holds. Using the same arguments, also the equality of the limits inferior follows. □

Remark 6.30. In the optimal control setup, smoothness of the Fréchet derivative $A_{;v}$ is commonly assumed, cf. Theorem 5.9.

By Proposition, in certain cases, compactness and monotonicity properties of A are carried over to $A_{;v}(w)'$. To ensure a stable decoupling via smoothness properties of $A_{;v}(w)'$, we introduce another assumption. See Assumption 4.34 for its counterpart for the semi-discrete state equations and Assumption 6.13 for the infinite-dimensional version.

Assumption 6.31. *Consider Problem 6.20(w_k). Given $\{w_k\}_{k\in\mathbb{N}} \subset \mathcal{W}(0,T;V_k;(V_k'))$ that is uniformly bounded in $L^2(0,T;V') \cap L^\infty(0,T;H)$. Then $\mathcal{K}_{;v}(t, w_k(t))$ is bounded in $L^2(0,T;H')$ uniformly in k and ω_k is sufficiently smooth for all $k \in \mathbb{N}$ and any solution $(\lambda_k, \mu_k) \in \mathcal{W}(0,T;V_k;(V_k)') \times L^2(0,T;Q_{H}{}'_k)$ to (6.14) is such that $\|A_{;v}(w_k)'\lambda_k(t)\|_{H'}$ and $\|\mu_k(t)\|_Q$ are bounded almost everywhere on $(0,T)$ independent of k.*

Finally, we impose the same assumptions on the approximation schemes that were used to establish convergence of the semi-discrete state solutions, cf. Remark 4.38(d).

Assumption 6.32. *Consider Problem 6.20(w_k). We assume that $J_1 = J_2$, and that $\{V_k\}_{k\in\mathbb{N}}$ and $\{Q_k\}_{k\in\mathbb{N}}$ are such that they fulfill Assumption 4.5 and such that J_{2k} and $\bar{J}_{2k}$ allow for splittings of the equations and solutions as in the continuous case and uniformly with respect to the discretization parameter k, i.e. Assumption 4.8 and 4.13 hold.*

Remark 6.33. Consider the formal adjoint equation (6.3) and its finite-dimensional approximation defined in Problem 6.20(w_k) with given $\{w_k\}_{k\in\mathbb{N}} \subset \mathcal{W}(0,T;V_k;(V_k'))$ and a $w \in \mathcal{W}(0,T;V;V')$ such that $w_k \rightharpoonup w$ in $L^2(0,T;V) \cap L^\infty(0,T;H)$ and $w_k(T) \to w(T)$ in H. With the findings and assumptions of this section and with Assumptions 6.13, 6.15, and 3.5, that were imposed to decouple the formal adjoint equation, we have established the situation that was described in Remark 4.38. In particular, we have

(a) *Space Regularity*: As it follows by Assumption 3.5, 6.13, uniform with respect to k, and Assumption 6.31.

(b) *Decoupling of Equation and Solution Spaces*: As it follows by Assumptions 3.5, 6.15, and 6.32 that in particular include the assumption that $J_1 = J_2$.

(c) *Consistency of the Data*: Since Equations (6.3b) and (6.15b) are homogeneous, there is no additional smoothness constraint on the right hand sides. For solvability, however, we need assume consistency of the terminal value, i.e., considering that $J_1 = J_2$, that $-j\mathcal{M}_{;v}(w(T)) \in \ker \bar{J}_2$, cf. Theorem 6.17.

(d) *Stable Approximation Schemes*: As follows by Assumptions 6.32 and, in particular, by the assumed symmetry $J_1 = J_2$ that makes all results on the convergence and stability of the inherent *external approximation schemes* from Section 4.3 hold for the considered case. Note that because of the homogeneity of (6.3b) there is no need for a particular choice of the discrete right inverses as for the state equations, cf. Lemma 4.26.

(e) *Consistent and Convergent Approximation of the Initial Value*: As it is ensured by the arguments presented in Section 4.5 and the homogeneity of (6.3b) and (6.15b).

(f) *Symmetry*: To show convergence of the discrete solutions to a solution of the continuous problem, we have assumed that $J_1 = J_2$, cf. also Remark 4.31.

(g) *Continuity, Coerciveness, Monotonicity and Boundedness of $A_{;v}(w_k)'$ and $A_{;v}(w)$*: As it follows by Assumption 6.27. In view of applications, we will, in particular, consider cases for which Proposition 6.29 applies. Also, assuming that $w_k \mapsto A_{;v}(w_k)'$ is continuous and since $A_{;v}(w_k)'\lambda_k$ is assumed uniformly bounded in $\mathcal{H}'$ the assumptions of Lemma 4.37 are fulfilled.

Thus, applying Theorem 4.39, we can state the following theorem:

Theorem 6.34. *Consider Problem* (6.4) *and its finite-dimensional approximation defined in Problem 6.20(w_k), with $J_1 = J_2$. Given $\{w_k\}_{k\in\mathbb{N}} \subset \mathcal{W}(0,T;V_k;(V_k'))$ and a $w \in \mathcal{W}(0,T;V;V')$ such that $w_k \rightharpoonup w$ in $L^2(0,T;V') \cap L^\infty(0,T;H)$, and such that $\mathcal{K}_{;v}(\cdot, w_k(\cdot)) \to \mathcal{K}_{;v}(\cdot, w_k(\cdot))$ in $L^2(0,T;H')$. Let all assumptions listed in Remark 6.33 hold. In particular, let the terminal value $-j\mathcal{M}_{;v}(w(T))$ and its approximations $-j\mathcal{M}_{;v}(w_k(T))$ be consistent as specified in Definition 3.35 and convergent as defined in Remark 4.38(e).*

For $k \in \mathbb{N}$ let $(\lambda_k, \mu_k) \in \mathcal{W}(0,T;V_k,H_k') \times L^2(0,T;Q_k)$ be a solution of the discrete problem. Then there is $(\lambda,\mu) \in \mathcal{W}(0,T;V,H') \times L^2(0,T;Q)$ such that $\lambda_k \rightharpoonup \lambda$ in $\mathcal{W}(0,T;V,H')$ and $\mu_k \rightharpoonup \mu$ in $L^2(0,T;Q)$ and (λ,μ) solves the continuous problem. The convergence is in terms of subsequences.

If the continuous problem has only one solution, then the complete sequence $\{(\lambda_k,\mu_k)\}_{k\in\mathbb{N}}$ converges.

If $V \hookrightarrow H$ is compact, then $\lambda_k \to \lambda$ in $L^2(0,T;H)$.

Remark 6.35. Convergence of the Galerkin approximations defined in Problem 6.20 with $J_1 = J_2$, can be obtained using the above results and considering w instead of w_k at every discretization level. In this case the operator $A_{;v}$ is independent of k, and the required uniformity in k in Assumptions 6.27 and 6.31 can be omitted. Then, also continuity of $w \mapsto A_{;v}(w)$, used to establish convergence in μ_k can be dropped, as well as Proposition 6.29(c). The case that the optimal state w is available is rather theoretical, however, Theorem 6.34, in particular, addresses solvability of the formal adjoint equation (6.3) and thus gives sufficient conditions for the validity of the necessary optimality conditions (6.8).

6.2. **Optimal Control of the Semi-discrete Equations.**

The results of the previous section establish convergence of discrete approximations solving Problem 6.20 to the multipliers (λ,μ) that are part of the continuous optimality conditions

(6.8). In general applications, however, in (6.14), the linearization point w is not available. Also, in general, the discrete approximations (λ_k, μ_k) do not define necessary optimality conditions, cf. Corollary 6.8, for Problem 6.1 formulated for a discretization level k.

In this section, we formulate necessary conditions for optimality with respect to (6.1) of (v_k, u) solving the spatially discretized state equations (6.2).

We start with formulating the optimal control problem Problem 6.1 for the spatial discretization level k.

Problem 6.36. *Consider the setup of the abstract optimal control problem as defined in Problem 6.1. In particular, let $\mathcal{U}$ be the space of input functions, let $\mathcal{M}\colon H \to \mathbb{R}$ and $\mathcal{K}\colon (0,T) \times V \times U \to \mathbb{R}$ be Fréchet differentiable weighting functions, and let $\mathcal{K}(t,v,u) = \mathcal{K}_1(t,v) + \mathcal{K}_2(t,u)$ be such that the Nemyckij mappings associated with the partial derivatives, $\mathcal{K}_{;u}\colon \mathcal{U} \to \mathcal{U}'$ and $\mathcal{K}_{;v}\colon \mathcal{V} \to \mathcal{V}'$, are well defined.*

Consider the task of minimizing the cost functional

$$\mathcal{J}\colon \mathcal{V}_k \times \mathcal{U} \to \mathbb{R}\colon \mathcal{J}(v_k, u) = \mathcal{M}(v_k(T)) + \int_0^T \mathcal{K}(t, v_k(t), u(t))\, \mathrm{d}t, \tag{6.24}$$

subject to the constraints $\mathcal{G}_k(v_k, p_k, u) = 0$ defined via $\mathcal{G}_k(v_k, p_k, u) = 0$, if $(v_k, u) \in \mathcal{V}_k \times \mathcal{U}$ solve

$$\dot{v}_k(t) - A_k(t, v_k(t)) - J'_{1k} p_k(t) - B_{1k} u = f_k(t) \quad \textit{in } (V_k)',\ \textit{a.e. in } (0,T), \tag{6.25a}$$

$$-J_{2k} v_k(t) = g_k(t) \quad \textit{in } {Q_H}'_k,\ \textit{a.e. in } (0,T), \tag{6.25b}$$

$$v_k(0) = \alpha_k \quad \textit{in } H_k, \tag{6.25c}$$

for a $p_k \in L^2(0,T; {Q_H}'_k)$ and where the restrictions A_k, J'_{1k}, J_{2k}, f_k, g_k, α_k, and B_{1k} of given A, J'_1, J_2, f, g, α and B_1 are defined as in Problem 4.7.

As in the continuous case, in view of formulating necessary optimality conditions as in Theorem 5.5, we formally define an adjoint equation to (6.25) via an application of Lemma 6.5.

Corollary 6.37 (of Lemma 6.5)**.** *Consider Problem 6.36 and assume that A is Fréchet differentiable. If for some $w \in \mathcal{W}(0,T;V;V')$, the functions $(\lambda_k, \mu_k) \in \mathcal{W}(0,T;V_k;(V_k)') \times L^2(0,T;{Q_H}_k)$ solve*

$$-\dot{\lambda}_k(t) - A_{k;v}(t, w(t))'\lambda_k(t) - (J_{2k})'\mu_k(t) + \mathcal{K}_{;v}(t, w(t)) = 0 \quad \textit{in } (V_k)', \tag{6.26a}$$

$$J_{1k}\lambda_k(t) = 0 \quad \textit{in } {Q_H}'_k, \tag{6.26b}$$

for almost all $t \in (0,T)$, and

$$\lambda_k(T) = -j\mathcal{M}_{;v}(w(T)) \quad \textit{in } H_k, \tag{6.26c}$$

meaning that $\big(\lambda_k(T), h_k\big)_H = -\big(j\mathcal{M}_{;v}(w(T)), h_k\big)_H$ for all $h_k \in H_k$, then $\Lambda_k := (\lambda_k, \mu_k, \lambda_k(0))$ is in $\mathcal{V}_k \times {Q_H T}_k \times H_k = \big((\mathcal{V}_k)' \times {Q_H T}_k{}' \times H_k\big)'$ and

$$\Lambda_k \mathcal{G}_{k;(v,p)}(w) + \mathcal{J}_{;(v,p)}(w) = 0 \quad \textit{in } (\mathcal{V}_k)' \times {Q_H T}_k{}'. \tag{6.27}$$

Then, applying Corollary 6.8, we can state necessary optimality conditions for Problem 6.36:

Corollary 6.38. *Consider Problem 6.36 and assume that the DAE (6.25) has a unique solution for all inputs. Let $u_0 \in \mathcal{U}$ and let $(v_k(u_0), p_k(u_0))$ be the corresponding solution. If $(v_k(u_0), u_0) \in \mathcal{W}(0,T;V_k;(V_k)') \times L^2(0,T;{Q_H}_k)$ is a local minimum of (6.24), and if there is $(\lambda_k, \mu_k) \in \mathcal{W}(0,T;V_k;(V_k)') \times L^2(0,T;{Q_H}_k)$ solving the* *formal adjoint equation* (6.26) *at* $w = v_k(u_0)$*, then* $(v_k(u_0), p_k(u_0), \lambda_k, \mu_k, u_0)$ *is a solution to the system*

$$\dot{v}_k(t) - A_k(t, v_k(t)) - J_{1k}' p_k(t) - B_{1k} u_0(t) = f_k(t) \quad \text{in } (V_k)', \tag{6.28a}$$

$$-J_{2k} v_k(t) = g_k(t) \quad \text{in } Q_H{}_k', \tag{6.28b}$$

$$-\dot{\lambda}_k(t) - A_{k;v}(t, v_k(t))'\lambda_k(t) - J_{2k}'\mu_k(t) + \mathcal{K}_{;v}(t, v_k(t)) = 0 \quad \text{in } (V_k)', \tag{6.28c}$$

$$J_{1k}\lambda_k(t) = 0 \quad \text{in } Q_H{}_k', \tag{6.28d}$$

for almost all $t \in (0, T)$*, and*

$$v_k(0) = \alpha_k \quad \text{in } H, \tag{6.28e}$$

$$\lambda(T) + j\mathcal{M}_{;v}(v(T)) = 0 \quad \text{in } H_k, \tag{6.28f}$$

and

$$-\lambda_k B_{1k} + \mathcal{K}_{;u}(u_0)' = 0 \quad \text{in } \mathcal{U}'. \tag{6.28g}$$

Proof. This is an application of Corollary 6.8. Note that taking the dual commutes with restricting to finite-dimensional spaces, i.e. $J_{1k}' = (J_{1k})' \colon Q_{Hk} \to (V_k)'$. Also, considered as a map from $\mathcal{V}_k$ to $(\mathcal{V}_k)'$, it holds that $(A_k)_{;v}(v_k) = A_{;v}(v_k)$. □

Remark 6.39. To ensure that any $u \in \mathcal{U}$ uniquely defines a solution to (6.25) it is necessary that $B_{1k} \colon \mathcal{U} \to (\mathcal{V}_k)'$ is injective. This is not a restriction in theory, as B_{1k} is bounded, since by Proposition 4.6 V_k is the image of a bounded projection, and $\mathcal{U}$ is a Hilbert space. Thus, one can factor out the kernel of B_{1k}, cf. Remark 6.3.

Remark 6.40. The results of Section 6.1 establish conditions for existence of solutions for the abstract formal adjoint equation defined in Problem 6.4 and to the discrete formal adjoint equation defined in Problem 6.20. However, in their formulation, they are not applicable to prove convergence of solutions of the formal discrete optimality system (6.28) to a solution of the abstract optimality system (6.8). Here, a major obstacle is that the discrete states (v_k, p_k) only converge weakly towards the states (v, p) of the abstract equations (3.1)(Sym), cf. Theorem 4.39. For application of Theorem 6.34 that establishes convergence in the Galerkin approximations of the adjoint states defined in Problem 6.4, one needs strong convergence of the sequences $\mathcal{K}_{;v}(\cdot, v_k(\cdot))$ and $\mathcal{M}_{;v}(v_k(T))$ in $L^2(0, T; H')$ and H', respectively.

7. Iterative Solution of the Nonlinear Optimality System

In this section we investigate Newton schemes to solve the abstract optimality system (6.8). A direct application of a Newton iteration to the optimality system leads to a sequence of linear problems. We give sufficient conditions, that extend standard assumptions to the differential-algebraic setup, to prove existence of the Newton iterates and convergence of the sequence. As a side product, a direct proof for solvability of corresponding linear-quadratic optimal control problems is obtained. Under the same conditions, one may consider the reduced cost functional and a Newton scheme for the gradient of the reduced cost functional. In this case, having assumed solvability of the formal adjoint equations, convergence is proved as for equality constrained optimal control problems in Banach spaces.

All results are given for the abstract setting, but, in particular, apply for semidiscrete formulations.

We consider the optimal control problem defined in Problem 6.1 and the formal necessary optimality conditions (6.8) that are given in Corollary 6.8 and which we will write in short form as

$$\begin{aligned} \dot{v} - A(v) - J_1'p - B_1u &= f_v, \quad v(0) = v^0, \\ -J_2v &= f_p, \end{aligned} \tag{7.1a}$$

$$\begin{aligned} -\dot{\lambda} - A'_{;v}(v)\lambda - J_2'\mu + \mathcal{K}_{;v}(v) &= f_\lambda, \quad \lambda(T) + j\mathcal{M}_{;v}(v(T)) = 0, \\ -J_1\lambda &= 0, \end{aligned} \tag{7.1b}$$

$$-B_1'\lambda + \mathcal{K}_{;u}(u) = 0. \tag{7.1c}$$

We will write the state equations with the formal operator $\mathcal{G}\colon \mathcal{W}(0,T;V;V') \times L^2(0,T;Q_H) \times \mathcal{U} \to \mathcal{V}' \times {\mathcal{Q}_{\mathcal{H}}}' \times H$ via $\mathcal{G}(v,p,u) = 0$, if (v,p,u) solves (7.1a). Recall that the cost functional is assumed to be separated in v and u, i.e. $\mathcal{K}(v,u) = K_1(v) + K_2(u)$. We make the following assumptions:

Assumption 7.1. *Given $(v^*,u^*) \in \mathcal{W}(0,T;V;V') \times \mathcal{U}$ that, with a suitable $p \in L^2(0,T;Q_H)$, solves* (7.1a). *Then, for $(v_\varepsilon,u_\varepsilon)$ in a neighborhood of (v^*,u^*), we have*

- *(a) that $A(v_\varepsilon)$, $\mathcal{K}(v_\varepsilon,u_\varepsilon)$, and $\mathcal{M}(v_\varepsilon(T))$ are twice continuously Fréchet-differentiable in a neighborhood of (v^*,u^*) with Lipschitz-continuous second derivatives and $\mathcal{M}_{;vv}$ independent of $v_\varepsilon(T)$,*
- *(b) that the partial Fréchet derivative $\mathcal{G}_{;(v,p)}(v_\varepsilon,p_\varepsilon)$ of the state equations* (7.1a) *with respect to (v,p) is invertible and the kernel of $\mathcal{G}_{;(v,p,u)}$ splits the space $\mathcal{W} \times \mathcal{Q}_H \times U$,*
- *(c) that there exists a solution (λ,μ) to* (7.1b) *at v^*, such that $\mathcal{J}_{;(v,u)}(v^*,u^*) + \lambda\mathcal{G}_{;(v,u)}(v^*,u^*) = 0$ on $\mathcal{V} \times \mathcal{U}$, and*
- *(d) that there exists a constant $c > 0$ such that*

$$-\langle h_v(h_u), \lambda A'_{;vv}(v^*)h_v(h_u)\rangle_{\mathcal{V},\mathcal{V}'} + \mathcal{J}_{;(v,u)^2}(v^*,u^*)[h_v(h_u),h_u]^2 \geq c\|(h_v(h_u),h_u)\|^2_{\mathcal{V}\times\mathcal{U}}$$

for $(h_v(h_u),h_p,h_u)$ in the kernel of $\mathcal{G}_{;(v,p,u)}(v^*,u^*)$, where $h_p = h_p(h_v)$ is the associated algebraic variable.

Note that Assumption 7.1 makes (v^*,u^*) part of a local solution of (7.1), cf. Theorem 5.9, and that the parts (a) and (d) are standard for the analysis of Newton-like schemes for optimal control problems that do not contain the endpoint penalization $\mathcal{M}$, cf. [149, p. 209] and [95, 151]. Also, they imply local uniqueness of the solution, cf. Remark 5.10. Assuming $\mathcal{M}_{;vv}$ independent of $v(T)$, in the analysis of Newton schemes, we circumvent the difficulties that would come with state dependent boundary conditions.

Assumption 7.1(c) is stronger than the standard assumption that (v^*, u^*) are locally optimal, as it demands the existence of a solution of the formal adjoint equation, cf. Remark 6.9.

Assumption 7.1(b) is also more restricting than just assuming the submersion property of $\mathcal{G}_{;(v,p,u)}$ at (v^*, u^*) used to formulate necessary optimality conditions, cf. Theorem 5.2 and the discussion thereafter. However, invertibility of $\mathcal{G}_{;(v,p)}$ is often used to define well-posedness of nonlinear equations, cf. eg. [49, pp. 297] for the Navier-Stokes Equation, and can be established for concrete applications, cf., e.g., [1, 72, 149].

The reference point is the Newton scheme formulated in Banach spaces, cf. [81] and the textbooks [82] and [34]. The results on Newton schemes listed in [125] are formulated for finite dimensions with the perspective of being generalized to Banach spaces.

The steps of the Newton scheme applied to optimality systems as (7.1) can be interpreted as a sequence of linear-quadratic optimal control problems [149], often referred to as SQP. This interpretation and its implications for PDE constraints are investigated, e.g., in [4, 5, 69, 72, 148].

In optimal control a Newton step requires the Hessian $\hat{\mathcal{J}}_{;uu}$, i.e. in the setting of (7.1) the operator $[A'_{;v}(v)\lambda]_{;v}$. There are definitions to $\hat{\mathcal{J}}_{;uu}$ that lessen the computational effort while preserving the convergence properties, see [72, 95] for formulations of *Broyden's* or the *BFGS* update schemes (see [126] for an overview) for the approximated Hessian in infinite-dimensional spaces.

If the cost functional is quadratic as in (6.12), then one can obtain a linear scheme by linearizing the state equations and setting up the optimality system for the current linear-quadratic problem. This approach is the same as neglecting $[A'_{;v}(v)\lambda]_{;v}$ in a Newton scheme for (7.1). Then, if one solves the state equations (7.1a) and the adjoint equations (7.1b) consequently in an iterative fashion, one obtains a *Newton Gauss-Seidel* iteration, cf. [125], for the blocks (7.1a) and (7.1b). This decoupling, however, may destroy the convergence properties that may hold for the coupled equations [84, 118, 157].

A decoupling of the Newton scheme for (7.1), that is shown to be still locally quadratically convergent, and linearly convergent inexact variants are given with a class of so called *tangential block Newton* methods [29, 76, 118].

In implementations, rounding and discretization errors occur. For this reason, one considers *inexact Newton* methods, see, e.g., [34, 75, 137, 151] for formulations in function spaces.

7.1. **Linearizations and Newton Schemes for the Functional Equations.** For illustration, consider the task of finding x in a Banach space X, such that

$$P(x) = 0 \in Y, \tag{7.2}$$

where Y is a Banach space and $P\colon X \to Y$.

Given a starting value $x_0 \in X$, the Newton scheme defines a sequence $\{x^k\}_{k\in\mathbb{N}}$ via

$$P_{;v}(x^k)[x^{k+1} - x^k] = -P(x^k), \quad k = 1, 2, \cdots . \tag{7.3}$$

Common sufficient conditions for the convergence of Newton schemes establish the following generic situation.

Assumption 7.2. *For a starting value x^0, the sequence $\{x^k\}_{k\in\mathbb{N}}$ defined in* (7.3) *converges to a $x^* \in X$ solving* (7.2)*, if*

(a) x^0 is close to an x^*,
(b) $P_{;v}(x^0)$ is invertible, and
(c) $P_{;v}(x)$ is smooth in a neighborhood of x^0.

In the pioneering work [81] on Newton schemes in Banach spaces, the conditions

$$\|P_{;v}(x_0)P(x_0)\| \le \eta_0, \tag{7.4a}$$

$$\|P_{;v}(x_0)\| \le B_0, \tag{7.4b}$$

$$\|P_{;vv}(x)\| \le K \quad \text{on } \Omega_0, \tag{7.4c}$$

with $h_0 := \eta_0 B_0 K \le \frac{1}{2}$, $\Omega_0 = \overline{\mathscr{U}}_r(x_0)$, and $r = \frac{1-\sqrt{1-2h_0}}{h_0}\eta_0$, were used to prove quadratic convergence in Ω_0.

If (7.4b) can be replaced by a uniform bound

$$\|P_{;v}(x)\| \le B \quad \text{on } \Omega_0, \tag{7.4b*}$$

then the domain of quadratic convergence $\Omega_0 = \overline{\mathscr{U}}_r(x_0)$ is specified from the requirements $h = \eta_0 BK < 2$ and $r > \hat{r} = \sum_{k=0}^{\infty}\left(\frac{h}{2}\right)^{2k-1}$, cf. [82, Thm. 5, Ch. XVIII.2].

Throughout this section, for x_0 in a Banach space X and a $r > 0$, we will use the notation

$$\mathscr{U}_r(x_0) := \{x \in X : \|x - x_0\| < r\}$$

and its closure $\overline{\mathscr{U}}_r(x_0)$ in X to denote a suitable neighborhood of x_0.

Remark 7.3. Condition (7.4c) can be replaced by Lipschitz continuity of $P_{;v}(x)$, cf., e.g., [125, 12.6.2] for a general or [4, Sec. 3] for an optimal control setting, or by affine covariant Lipschitz conditions [34, Thms. 2.2, 2.3]. Given Lipschitz continuity of $P_{;v}$ and an estimate of $\|P_{;v}(x^*)^{-1}\|$ one can use the results in [132] to establish convergence.

7.2. **Newton for the Optimality Conditions.** We will establish conditions under which a Newton iteration for (7.1) can be analyzed via an uniform bound as in (7.4b*). Having reformulated (7.1) as an operator equation $P(x) = 0$, with $x := (v, p, \lambda, \mu, u)$, the Fréchet derivative of the optimality system taken at $x^k = (v^k, \lambda^k, u^k)$ reads

$$P_{;x}(x^k) = \begin{bmatrix} 0 & F_1(v^k) & B \\ F_1^{ad}(v^k) & W(v^k, \lambda^k) & 0 \\ B' & 0 & R(u^k) \end{bmatrix} : \begin{bmatrix} \mathcal{W} \\ \mathcal{Q} \end{bmatrix} \times_T \begin{bmatrix} \mathcal{W} \\ \mathcal{Q} \end{bmatrix} \times \mathcal{U} \to \begin{bmatrix} \mathcal{V}' \\ \mathcal{Q}' \end{bmatrix} \times \begin{bmatrix} \mathcal{V}' \\ \mathcal{Q}' \end{bmatrix} \times \mathcal{U}. \tag{7.5}$$

Here, and in what follows, we use the abbreviations $\mathcal{W} := \mathcal{W}(0, T; V; V')$ and $\mathcal{V} = L^2(0, T; V)$, $\mathcal{Q} = L^2(0, T; Q)$ and $\mathcal{U} = L^2(0, T; U)$, for the function spaces that were introduced in Section 3. The symbol $\times_T$ is used to denote that because of initial and end conditions, the domain of definition is restricted to

$$\mathcal{W} \times_T \mathcal{W} := \{(v, \lambda) \in \mathcal{W}^2 : v(0) = 0,\ \lambda(T) + \mathcal{M}_{;vv}(v^k(T))v(T) = 0\} \subset \mathcal{W}^2.$$

The operator components in (7.5) are given via

$$F_1(v^k)[v, p] + Bu = \begin{bmatrix} \dot{v} - A_{;v}(v^k)v - J_1' p \\ -J_2 v \end{bmatrix} + \begin{bmatrix} -B_1 u \\ 0 \end{bmatrix}, \tag{7.6a}$$

$$W(v^k, \lambda_v^k)[v, p] = \begin{bmatrix} -\lambda_v^k A_{;vv}(v^k)v + \mathcal{K}_{;vv}(v^k)v \\ 0 \end{bmatrix}, \tag{7.6b}$$

$$F_1^{ad}(v^k)[\lambda, \mu] = \begin{bmatrix} -\dot{\lambda} - A_{;v}'(v^k)\lambda - J_2' \mu \\ -J_1 \lambda \end{bmatrix}, \tag{7.6c}$$

and

$$B'[\lambda, \mu] + R(u^k)u = -B_1' \lambda + \mathcal{K}_{;uu}(u^k)u. \tag{7.6d}$$

Then, application of the Newton scheme as in (7.3), gives:

Algorithm 7.4 (Newton Scheme for the Optimality Conditions). *Given $x^0 = (v^0, \lambda^0, u^0) \in \mathcal{W} \times_T \mathcal{W} \times \mathcal{U}$, define $x^k = (v^k, p^k, \lambda^k, \mu^k, u^k)$, for $k = 1, 2, \cdots$ via the iterative scheme:*

(1) Solve $P_{;x}(x^k)\Delta_x^k = -P(x^k)$, i.e.

$$\begin{bmatrix} 0 & F_1(v^k) & B \\ F_1^{ad}(v^k) & W(v^k, \lambda^k) & 0 \\ B' & 0 & R(u^k) \end{bmatrix} \begin{bmatrix} (\Delta_\lambda^k, \Delta_\mu^k) \\ (\Delta_v^k, \Delta_p^k) \\ \Delta_u^k \end{bmatrix} = - \begin{bmatrix} \dot{v}^k - A(v^k) - J_1' p^k - B_1 u^k - f_v \\ -J_2 v^k - f_p \\ -\dot{\lambda} - A_{;v}'(v^k)\lambda^k - J_2'\mu^k + \mathcal{K}_{;v}(v^k) - f_\lambda \\ -J_1\lambda^k \\ -B_1'\lambda^k + \mathcal{K}_{;u}(u^k) \end{bmatrix}, \tag{7.7}$$

for Δx^k.

(2) $x^{k+1} = x^k + \Delta_x^k$.

Remark 7.5. Because the equations are linear in (p, μ), there exists no need for a starting value for these variables.

Remark 7.6. Because of the linearity of the algebraic equations imposed by F_1 and F_1^{ad}, one has that all iterates x^k, $k \geq 1$, are consistent, i.e. $J_2 v^k = f_p$ and $J_1 \lambda^k = 0$.

To prove existence and convergence of the sequence defined by Algorithm 7.4 to a candidate optimal solution x^*, we first establish invertibility of $P_{;x}(x)$ at x^* and then use a result from operator perturbation theory [83] to show that invertibility is also given for x^0 close to x^*.

To account for the differential-algebraic structure of the equations, we will give a direct proof. We extend the arguments from [97] to fit the setup of end point penalization in (6.1) and the standard second order sufficiency conditions (6.9).

In what follows, we will omit the dependencies of the components of (7.5) on the linearization point, i.e. we will write, e.g. F_1 instead of $F_1(v^*)$. Also, we will use the abbreviations

$$\mathcal{X} := \begin{bmatrix} \mathcal{W} \\ \mathcal{Q} \end{bmatrix} \times_T \begin{bmatrix} \mathcal{W} \\ \mathcal{Q} \end{bmatrix} \times \mathcal{U} \quad \text{and} \quad \mathcal{Y}' = \begin{bmatrix} \mathcal{V}' \\ \mathcal{Q}' \end{bmatrix} \times \begin{bmatrix} \mathcal{V}' \\ \mathcal{Q}' \end{bmatrix} \times \mathcal{U}.$$

Lemma 7.7. *Consider $P_{;x}$ as defined in* (7.5) *and consider $(v^*, u^*) \in \mathcal{W} \times \mathcal{U}$ for which Assumption 7.1 holds. Assume there exists an $\varepsilon > 0$ such that for all $v \in \mathscr{U}_\varepsilon(v^*)$, there exists a solution (λ, μ) to $F_1^{ad}[\lambda, \mu] = 0$ with the terminal condition $\lambda(T) + j\mathcal{M}_{;vv}v(T) = 0$. Then there is no sequence $x^k = (v^k, p^k, \lambda^k, \mu^k, u^k) \in \mathcal{X}$, $k = 1, 2, \ldots$, with $\|x^k\|_{\mathcal{X}} = 1$, such that $P_{;x}x^k \to 0$ as $k \to \infty$.*

Proof of Lemma 7.7. By Assumption 7.1(c), there exists a $\lambda^* \in \mathcal{W}$ associated with (v^*, u^*) via (λ^*, μ^*) solve (7.1b) at v^*, so that we can consider $P_{;x} := P_{;x}(v^*, \lambda^*, u^*)$. By Assumption 7.1(b), the operator $F_1 \colon \mathcal{W} \times \mathcal{Q}_H \to \mathcal{V}' \times \mathcal{Q}_H'$ is invertible. We first show, that with the given assumption on $j\mathcal{M}_{;vv}w$, the operator F_1^{ad} is invertible too, i.e. $F_1^{ad}[\lambda, \mu] = g$, with $\lambda(T) + j\mathcal{M}_{;vv}v(T) = 0$, has a unique solution for any $g \in \mathcal{V}' \times \mathcal{Q}_H'$ and for $v \in \mathscr{U}_\varepsilon(v^*)$. Note, that for zero terminal condition $\lambda(T) = 0$ one has $F_1^{ad} = F_1'$ and F_1' is invertible, as F_1 is invertible, cf. [82, Thm. 4(2.XII)]. Thus, for any $g \in \mathcal{V}' \times \mathcal{Q}_H'$ we find that $[\lambda, \mu] = [\lambda^h, \mu^h] + [\lambda^p, \mu^p]$ is a solution to $F_1^{ad}[\lambda, \mu] = g$, where $[\lambda^h, \mu^h]$ solves $F_1'[\lambda^h, \mu^h] = g$, $\lambda^h(T) = 0$ and $[\lambda^p, \mu^p]$ is a solution of $F_1^{ad}[\lambda^p, \mu^p] = 0$, $\lambda^p(T) + j\mathcal{M}_{;vv}w(T) = 0$, which exists by assumption. There cannot be a different solution to $F_1^{ad}[\lambda, \mu] = g$ as the difference z of two solutions must solve the homogeneous equation $F_1^{ad}z = F_1'z = 0$, which means that $z = 0$, since F_1' is invertible.

Now, assume the converse, i.e., there exists a sequence $\{x^k\}_{k=1,2,\dots} \subset \mathcal{X}$ with $\|x^k\|_{\mathcal{X}} = 1$, for all $k \in \mathbb{N}$, and with $P_{;x}x^k \to 0$ as $k \to \infty$.

By Assumption 7.1(b), the input to state map $x\colon u \mapsto (v(u), p(u))$, defined via $\mathcal{G}(v, p, u) = 0$, is Fréchet differentiable and, thus, locally Lipschitz-continuous [160, Thm. 4.B]. Thus, by Remark 5.8, we have that the tangent space $T_{(v_0,p_0,u_0)} = \{(x_{;u}(h_u) : h_u \in \mathcal{U}\}$ coincides with the kernel of the operator $\begin{bmatrix} F^1 & B \end{bmatrix}$ as defined in (7.6a). Since, by Assumption 7.1(b), $T_{(v_0,p_0,u_0)}$ splits the domain of $\begin{bmatrix} F^1 & B \end{bmatrix}$, we can assume that $(v^k, p^k, u^k) \in T_{(v_0,p_0,u_0)}$, as a part $(v_r^k, p_r^k, 0)$ in the complement of $T_{(v_0,p_0,u_0)}$ goes to zero with $F^1[v_r^k, p_r^k] \to 0$, as $k \to \infty$.

From $P_{;x}x^k \to 0$, computing the dual product of the three components of $P_{;x}x^k$ with $-(\lambda^k, \mu^k)$, (v^k, p^k), and u^k, we get that

$$\begin{aligned} &-\big\langle(\lambda^k, \mu^k), F_1[v^k, p^k]\big\rangle - \big\langle\lambda^k, B_1 u^k\big\rangle + \big\langle F_1^{ad}[\lambda^k, \mu^k], (v^k, p^k)\big\rangle + \\ &+\big\langle[\mathcal{K}_{;vv} - \lambda^0 A_{;vv}]v^k, v^k\big\rangle + \big\langle B_1'\lambda^k, u^k\big\rangle + \big\langle\mathcal{K}_{;uu}u^k, u^k\big\rangle \\ &\quad = -\big\langle\lambda^k, \dot{v}^k\big\rangle - \big\langle\dot{\lambda}^k, v^k\big\rangle + \big\langle[\mathcal{K}_{;vv} - \lambda^0 A_{;vv}]v^k, v^k\big\rangle + \big\langle\mathcal{K}_{;uu}u^k, u^k\big\rangle \\ &\quad = \big(j\mathcal{M}_{;vv}v^k(T), v^k(T)\big) + \big\langle[\mathcal{K}_{;vv} - \lambda^0 A_{;vv}]v^k, v^k\big\rangle + \big\langle\mathcal{K}_{;uu}u^k, u^k\big\rangle \\ &\quad \geq c\|(u^k, v^k)\|^2_{\mathcal{V}\times\mathcal{U}} \end{aligned} \tag{7.8}$$

goes to zero. This implies that $u^k \to 0$ as $k \to \infty$. Here we have used the properties of the dual operators, the symmetry of the dual product, and the optimality condition from Assumption 7.1(a). From the invertibility of F_1 and F_1^{ad}, we obtain that (v^k, p^k) and (λ^k, μ^k) go to zero, and, thus, a contradiction to the initial assumption that $\|x^k\|_{\mathcal{X}} = 1$ for all $k \in \mathbb{N}$. □

Lemma 7.8. *Consider $P_{;x}$ as defined in* (7.5) *and let the assumptions of Lemma 7.7 hold. Then,* $\ker P_{;x}' = \{0\}$.

Proof. Computing and comparing the corresponding dual products, we find that the dual operator of $\begin{bmatrix} 0 & F_1 \\ F_1^{ad} & W \end{bmatrix}$ is given via

$$\begin{bmatrix} 0 & F_1 \\ F_1^{ad} & W' \end{bmatrix} : \begin{bmatrix} \mathcal{W} \\ \mathcal{Q} \end{bmatrix} \times_{T'} \begin{bmatrix} \mathcal{W} \\ \mathcal{Q} \end{bmatrix} \to \begin{bmatrix} \mathcal{V}' \\ \mathcal{Q}' \end{bmatrix} \times \begin{bmatrix} \mathcal{V}' \\ \mathcal{Q}' \end{bmatrix},$$

with $\times_{T'}$ referring to the dual side conditions $j\mathcal{M}_{;vv}'w_\lambda(T) + w_v(T) = 0$ and $w_\lambda(0) = 0$. We assume the contrary, i.e. that there exists $0 \neq (w_v, w_p, w_\lambda, w_\mu, w_u) =: w \in D(P_{;x}')$, with

$$\begin{bmatrix} 0 & F_1 & B \\ F_1^{ad} & W' & 0 \\ B' & 0 & R' \end{bmatrix} \begin{bmatrix} (w_v, w_p) \\ (w_\lambda, w_\mu) \\ w_u \end{bmatrix} = 0. \tag{7.9}$$

The second block line in (7.9) implies $(w_\lambda, w_\mu, w_u) \in \ker \begin{bmatrix} F^1 & B \end{bmatrix} = T_{(v_0,p_0,u_0)}$. Computing the dual product of the components in (7.9) with $(-w_v, -w_p, w_\lambda, w_\mu, w_u)$, using the definition of the dual operators, and Assumption 7.1(d), we find that

$$\begin{aligned} 0 &= -\big\langle w_\lambda, \dot{w}_v\big\rangle + \big\langle w_\lambda, W'w_\lambda\big\rangle - \big\langle\dot{w}_\lambda, w_v\big\rangle + \big\langle w_u, R'w_u\big\rangle \\ &= \big(w_\lambda(T), j\mathcal{M}_{;vv}'w_\lambda(T)\big) + \big\langle w_\lambda, W'w_\lambda\big\rangle + \big\langle w_u, R'w_u\big\rangle \geq c\|(w_\lambda, w_u)\|^2_{\mathcal{V}\times\mathcal{U}}, \end{aligned}$$

and we conclude that $w_u = 0$ and subsequently that $w = 0$, which is a contradiction to the initial assumption. □

Theorem 7.9 (Cf. [97], Thm. 1)**.** *Let the assumptions of Lemma 7.7 hold. Then, the operator $P_{;x}(x^*)$ defined in* (7.5) *is invertible. In particular, there exists a constant $c^* > 0$ such that $\|P_{;x}(x^*)^{-1}\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} < \frac{1}{c^*}$.*

Proof. The assertion of Lemma 7.7 is equivalent to the statement: there exists a $c^* > 0$ such that

$$\|P_{;x}(x^*)x\|_{\mathcal{Y}'} > c^*\|x\|_{\mathcal{X}}, \tag{7.10}$$

for all $x \in \mathcal{X}$. This means, that $P_{;x}(x^*)$ is injective and, since $P_{;x}(x^*)$ is bounded and defined on a whole Banach space, that $\operatorname{im} P_{;x}(x^*)$ is closed in $\mathcal{Y}'$, cf. [3, Def. 2.1 and Thm. 2.5]. By invertibility of F_1 and F_1^{ad} we have that $\operatorname{im} P_{;x}(x^*) = (\mathcal{V} \times \mathcal{Q_H}') \times \mathcal{U}_P$, where $\mathcal{U}_P$ is the part of the range of $P_{;x}$ in $\mathcal{U}$, cf. (7.5). Since $\mathcal{U}$ is a Hilbert space, and $\mathcal{U}_P$ is closed, it has an orthogonal complement $\mathcal{U}_{P\perp}$, and we can write $\mathcal{Y}' = \operatorname{im} P_{;x}(x^*) \oplus Z$, where $Z = \{(0,0)\}^2 \times \mathcal{U}_{P\perp}$. Then, the dual is the direct sum of the annihilators $\mathcal{Y} = (\operatorname{im} P_{;x}(x^*))^0 \oplus Z^0$, cf. [83, Thm. IV.4.8]. By [83, Thm. IV.5.13] and Lemma 7.8 we have $(\operatorname{im} P_{;x}(x^*))^0 = \ker P'_{;x}(x^*) = \{0\}$ implying that $Z^0 = \mathcal{Y}$, i.e. $Z = \{0\}$, and, thus, $\operatorname{im} P_{;x}(x^*) = \mathcal{Y}'$.

Thus, $P_{;x}(x^*)$ is invertible and we can estimate

$$\|P_{;x}^{-1}y\|_{\mathcal{X}} \leq \frac{1}{c^*}\|P_{;x}P_{;x}^{-1}y\|_{\mathcal{Y}'} = \frac{1}{c^*}\|y\|_{\mathcal{Y}'}. \tag{7.11}$$

□

Remark 7.10. For the linear-quadratic case, where the cost functional is as in (6.12), Lemmata 7.7 and 7.8 and Theorem 7.9 establish a direct proof for the unique solvability of the corresponding optimality system (6.13).

So far, we have established invertibility of $P_{;x}(x^*)$ at an optimal solution $x^* = (v^*, p, \lambda^*, \mu, u^*)$. To apply a Newton iteration we need invertibility of $P_{;x}$ in a neighborhood of x^* and, in particular, at a starting point x^0. If the components of $P_{;x}(x)$ are smooth in x, then there exists a neighborhood $\mathscr{U}_\epsilon(x^*)$ on which $P_{;x}$ is invertible. To formulate this, we write $x = x^* + \Delta_x^*$ and $P_{;x}(x) = P_{;x}(x^*) + \Delta P_{;x}(x^*)$, where

$$\Delta P_{;x}(x^*) := P_{;x}(x^* + \Delta_x^*) - P_{;x}(x^*) =: \begin{bmatrix} \Delta F_{;v}(v^*) & 0 & 0 \\ \Delta W(v^*, \lambda^*) & \Delta F_{;v}^{ad}(v^*) & 0 \\ 0 & 0 & \Delta R(u^*) \end{bmatrix}, \tag{7.12}$$

with

$$\Delta F_1(v^*)[v,p] : = \begin{bmatrix} -[A_{;v}(v^* + \Delta_v^*) - A_{;v}(v^*)]v \\ 0 \end{bmatrix}$$

$$\Delta W(v^*, \lambda_v^*)[v,p] : =$$

$$\begin{bmatrix} -\big[[\lambda_v^* + \Delta_\lambda^*]A_{;vv}(v^* + \Delta_v^*) - \lambda^* A'_{;vv}(v^*)\big]v + [\mathcal{K}_{;vv}(v^* + \Delta_v^*) - \mathcal{K}_{;vv}(v^*)]v \\ 0 \end{bmatrix},$$

$$\Delta F_1^{ad}(v^*)[\lambda, \mu] : = \begin{bmatrix} [A'_{;v}(v^* + \Delta_v^*) - A'_{;v}(v^*)]\lambda \\ 0 \end{bmatrix}, \quad \text{and}$$

$$\Delta R(u^*)u : = [\mathcal{K}_{;uu}(u^* + \Delta_u^*) - \mathcal{K}_{;uu}(u^*)]u.$$

Lemma 7.11. *If the assumptions of Lemma 7.7 hold, i.e., in particular, there exists a locally optimal solution $(v^*, u^*) \in \mathcal{W} \times \mathcal{U}$ to Problem 6.1 for which Assumption 7.1 holds and the adjoint equation is solvable for v in a neighborhood of v^*. Consider $x^* = (v^*, p^*, \lambda^*, \mu^*, u^*) \in \mathcal{X}$, where (p^*, λ^*, μ^*) are the variables that complete (v^*, u^*) to a solution of* (7.1)*. Then there exists a $r_1 < 0$, such that $\|\Delta P_{;x}(x^*)\|_{\mathcal{X}} \leq c_1 < c_0$, for all $\|\Delta_x^*\| < r_1$, where ΔP is defined in* (7.12) *and c_0 is the constant that bounds $P_{;x}(x^*)$ from below, and such that $P_{;x}(x)$ is invertible for all $x \in \mathscr{U}_{r_1}(x^*)$ with*

$$\|P_{;x}(x)^{-1}\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \leq \frac{1}{c_0(1 - \frac{c_1}{c_0})}.$$

Proof. By Assumption and by Theorem 5.6, the operators setting up $\Delta P_{;x}(x^*)$ are Lipschitz-continuous so that $\|\Delta P_{;x}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}'} \to 0$ as Δ_x^* goes to zero. Thus, there exists a radius r_1 and $c_1 < c_0$, such that $\|\Delta_x^*\|_{\mathcal{X}} < r_1$ implies $\|\Delta P_{;x}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}'} \leq c_1$. Invertibility of $P_{;x}(x) = P_{;x}(x^*) + \Delta P_{;x}(x^*)$ and the uniform bound for its inverse on $\mathscr{U}_{r_1}(x^*)$ then follows by [83, Thm. IV.1.16]. □

Remark 7.12. If $A(v)$ is quadratic implying $A_{;v}(v)$ is linear and $A_{;vv}(v)$ is constant, the contributions of A to $\Delta P_{;x}$ reduce to $A_{;v}[\Delta_v^*]$, $A'_{;v}[\Delta_v^*]$, and $\Delta_\lambda^* A_{;vv}$. If the cost functional is quadratic, it does not contribute to $\Delta P_{;x}$. Then, in particular, smallness of $\Delta P_{;x}$ does not depend on u.

Having established uniform invertibility of $P_{;x}(x)$ in a neighborhood of x^* satisfying Assumption 7.1, we can formulate conditions sufficient for the convergence of Newton's method applied to the optimality system (7.1).

Corollary 7.13. *[Cf. [34, Thm. 2.2]] Let the assumptions of Lemma 7.11 hold. Consider* (v^*, u^*) *for which Assumption 7.1 holds and let* $x^* := (v^*, p^*, \lambda^*, \mu^*, u^*)$ *make up the corresponding solution of* (7.1). *If* x^0 *is such that for constants* r_1, B_0, $\eta_0 > 0$,

(a) $\|x^* - x^0\|_{\mathcal{X}} < r_1$, *where* r_1 *is as specified in Lemma 7.11 so that there exists a constant* B *with* $\|P_{;x}^{-1}(x)\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \leq B$ *on* $\mathscr{U}_{r_1}(x^*)$,

(b) $\|P(x^0)\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} \leq \eta_0$, *and*

(c) $LB\eta_0 < 2$, *where* L *is the Lipschitz constant of* $P_{;x}$ *on* $\mathscr{U}_{r_1}(x^*)$,

then the iterates of Algorithm 7.4 converge quadratically towards a solution $\bar{x}^*$ *of* (7.1) *and the following estimates are valid:*

$$\begin{aligned} \|x^{k+1} - x^k\| \leq& \frac{1}{2} LB \|x^k - x^{k-1}\|^2, \\ \|x^k - \bar{x}^*\| \leq& \frac{\|x^k - x^{k+1}\|}{1 - \frac{1}{2} LB \|x^k - x^{k-1}\|}. \end{aligned}$$

Proof. The proof for [34, Thm. 2.2] holds also for the current assumptions. In particular the uniform bound on the inverse and Lipschitz continuity of $P_{;x}$ imply the affine covariant Lipschitz condition used in [34]. For the extension to infinite dimension one can resort to the arguments used in the proof of [34, Thm. 2.1]. □

Corollary 7.13 bases on the conditions (7.4a), (7.4b*), and (7.4c) replaced by Lipschitz continuity of $P'(x)$. Convergence can be also assured by the variants using condition (7.4b) or an estimate of $\|P'(x^*)^{-1}\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})}$, cf. Remark 7.3.

7.3. **Newton for the Reduced Cost Functional.** We consider the optimal control problem Problem 6.1 and assume that Assumption 7.1 holds. By Assumption 7.1(b), the input to state map $u \mapsto (v(u), p(u))$ is well defined [160, Thm. 4.B], and we can define the reduced cost functional for (6.1) via $\hat{\mathcal{J}}(u) := \mathcal{J}(v(u), u)$. By Assumption 7.1(d), one has local convexity of the problem and thus local uniqueness of an optimal control u^* and that the necessary optimality condition $\hat{\mathcal{J}}_{;u}(u^*) = 0$ is also sufficient, cf. Theorem 5.9.

We formulate the Newton scheme (7.3) directly for the optimality condition. Given a starting value u^0, we compute the sequence u^k, $k = 1, 2, \cdots$ via

$$\hat{\mathcal{J}}_{;uu}(u^k)\Delta_u^k = -\hat{\mathcal{J}}_{;u}(u^k), \quad u^{k+1} = u^k + \Delta_u^k. \tag{7.13}$$

By Theorem 5.3 and, in particular, by Lemma 6.5 we have that

$$\hat{\mathcal{J}}_{;u}(u^k) = \mathcal{J}_{;u}(v, u^k) - \lambda B_1,$$

where $v = v(u^k)$ is the solution of the state equations (3.1) at u^k, and where $\lambda = \lambda(v(u^k))$ solves the adjoint system (6.3) at $v(u^k)$). With the assumption $\mathcal{J}_{;uv} = 0$ one has

$$\hat{\mathcal{J}}_{;uu}(u^k)\Delta_u^k = \mathcal{J}_{;uu}(v, u^k)\Delta_u^k - \delta\lambda_v B_1, \tag{7.14}$$

where $\delta\lambda_v := \lambda_{v;u}\Delta_u^k$ solves the linearized adjoint equation:

$$-\begin{bmatrix} \dot{\delta\lambda_v} \\ 0 \end{bmatrix} - \begin{bmatrix} A_{;v}(v)' & J_2' \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} \delta\lambda_v \\ \mu \end{bmatrix} + \begin{bmatrix} \mathcal{K}_{;vv}(v)\delta v - [A_{;vv}(v)\delta v]'\lambda \\ 0 \end{bmatrix} = 0,$$
$$\delta\lambda_v(T) + j\mathcal{M}_{;vv}(v(T))\delta v(T) = 0, \tag{7.15}$$

where $\delta v := v_{;u}[\Delta_u^k]$ solves the linearized about u^k state equations with input Δ_u^k:

$$\begin{bmatrix} \dot{\delta v} \\ 0 \end{bmatrix} - \begin{bmatrix} A_{;v}(v) & J_1' \\ J_2 & 0 \end{bmatrix} \begin{bmatrix} \delta v \\ p \end{bmatrix} + \begin{bmatrix} B_1\Delta_u^k \\ 0 \end{bmatrix} = 0,$$
$$\delta v(0) = 0. \tag{7.16}$$

For the relation of the derivative of the input to state maps and the linearized state equations see Remark 5.11.

Thus, one arrives at Algorithm 7.14 for one Newton step $u^k \to u^{k+1}$:

Algorithm 7.14 (Newton Scheme for the Reduced Cost Functional)**.**

(1) Compute $v = v(u^k)$ and then $\lambda = \lambda(v)$,

(2) Solve

$$\begin{bmatrix} F_1(v) & 0 & B \\ W(v,\lambda) & F_1^{ad}(v) & 0 \\ 0 & B' & R(u^k) \end{bmatrix} \begin{bmatrix} (\delta v, p) \\ (\delta\lambda, \mu) \\ \Delta_u^k \end{bmatrix} = - \begin{bmatrix} 0 \\ 0 \\ \mathcal{J}_{;u}(v, u^k) - \lambda B_1 \end{bmatrix} \tag{7.17}$$

for Δ_u^k,

(3) $u^{k+1} = u^k + \Delta_u^k$.

The operators used in (7.17) are as defined for (7.5).

By Assumption 7.1, we have regularity of the cost functional and the constraints, we have a local solution u^* such that $\hat{\mathcal{J}}(u^*) = 0$, and we have the submersion property of the state equations at the optimal solution. If, as for the convergence result Corollary 7.13, we additionally assume solvability of the formal adjoint equation (7.1b) in a neighborhood of $v(u^*)$, we obtain local quadratic convergence of the sequence generated by Algorithm 7.14, see [73, Thm 2.15].

Remark 7.15. Each iteration in Algorithm 7.14 requires one solve of the nonlinear state equation and the adjoint equation. If one resorts to iterative methods to solve (7.17), the action of $\hat{\mathcal{J}}_{;uu}$ onto an iterate can be determined by subsequently computing δv and $\delta\lambda$ in order to make use of relation (7.14).

8. Optimal Control of Finite-dimensional Index-2 DAEs

As in the abstract setting, one can can formally set up necessary optimality conditions by means of the formal adjoint equation, cf. Section 6.

Again, unlike the ODE case, for DAE constraints the existence and uniqueness of solutions to the involved adjoint equations and thus to the optimality system is in general not guaranteed, cf. [10, 12, 32, 98] and in particular [88] for the linear quadratic case. To provide necessary and sufficient conditions for the existence of optimal controls one can for example exploit the special structure of semi-explicit equations, cf. [32, 47, 48], or consider linear DAEs with properly stated leading term, cf. [9, 10, 11, 13, 98, 113]. The special case of Riccati-feedback solutions was investigated in [99]. The general way is to regularize the DAEs and formulate the conditions for the resulting strangeness-free system [27, 91]. In [89, 90] conditions and procedures for the construction of state feedbacks are presented such that the system is *strangeness-free* in the *behavior formulation.*

In our approach we make use of the structure of an optimality system that is stated in the original variables and the original equations. We will apply a decoupling only for theoretical considerations. The obtained results regarding optimality of the solutions to the Euler-Lagrange equations are already covered by [9, 10]. The innovations we propose base on the specific structure of the semi-explicit index-2 formulation, as it arises in linearized Navier-Stokes equations. We use the structure to prove the existence of an optimal solution directly. Thereto we introduce a differential-algebraic matrix Riccati equation that seems suitable for numerical computations, as it is stated in the original system matrices. Also, we mention how the necessity gap between the considered formal and the *true* [92] optimality conditions can be closed in applications.

We start with investigating classical solutions to a class of finite-dimensional DAEs and related optimal control problems that include the spatially discretized Problems discussed in Section 3 and 6. In fact, consider Problem 4.7 and let $\{\psi^i\}_i^{n_v}$ and $\{\phi^j\}_j^{n_p}$ be bases of V_k and Q_k. Then, a solution v_k and p_k to (4.6a-b) always has the representation $v_k(t) = \sum_{i=1}^{n_v} v^i(t)\phi^i$ and $p_k(t) = \sum_{j=1}^{n_p} p^j(t)\psi^j$ where the coordinate vectors $v(t) = \begin{bmatrix} v^1(t) & \dots & v^{n_v}(t)\end{bmatrix} \in \mathbb{R}^{n_v}$ and $p(t) = \begin{bmatrix} p^1(t) & \dots & p^{n_p}(t)\end{bmatrix} \in \mathbb{R}^{n_p}$ are solutions to

$$\begin{aligned} M\dot{v}(t) - A(t, v(t)) - J_1^{\mathsf{T}} p(t) &= f(t), \\ -J_2 v(t) &= g(t), \end{aligned}$$

on $(0, T)$, with coefficient matrices

$$\begin{aligned} M &= [m_{nm}] \in \mathbb{R}^{n_v, n_v}, \quad m_{nm} := \big(\phi^n, \phi^m\big)_H, \\ J_1^{\mathsf{T}} &= [j_{1nl}] \in \mathbb{R}^{n_v, n_p}, \quad j_{1nl} := \big(\phi^n, J_1'\psi^l\big)_H, \quad \text{and} \\ J_2 &= [j_{2lm}] \in \mathbb{R}^{n_p, n_v}, \quad j_{2ml} := \big(\psi^l, J_2\phi^m\big)_Q \end{aligned} \tag{8.1}$$

and with the nonlinearity

$$A(t, v(t)) = [a_n] \in \mathbb{R}^{n_v}, \quad a_n := \big(\phi^n, A(t, \sum_{m=1}^{n_v} v^i(t)\phi^m)\big)_H$$

and right hand sides

$$f(t) = [\big(\phi^n, f_k(t)\big)_H] \in \mathbb{R}^{n_v} \quad \text{and} \quad g(t) = [\big(g_k(t), \psi^l\big)_Q] \in \mathbb{R}^{n_p}.$$

In particular, we will consider the case where the nonlinearity can be written as $A(t, v(t))[v(t)]$ as it comes out of the semi discretization of the Navier Stokes Equation (3.6). We will allow for time dependent coefficients M, J_1^{T}, J_2, as they may arise in the case of time dependent basis functions.

Having at hand the explicit matrix formulations of all operators, we can give explicit representations of the decoupling operators that were considered in the infinite-dimensional and in general semi-discrete setting in Section 3 and 4.

We will also define an associated optimal control problem and set up sufficient optimality conditions via a formal adjoint equation, cf. Section 6 and 7. For the linear-quadratic case as it occurs for linear state equations and quadratic costs or in the course of Newton iterations for the general nonlinear optimality systems, see Section 7, we state the existence of a Riccati decoupling of the optimality system. Using this decoupling of states and adjoint states, we derive necessary and sufficient conditions for the existence of optimal solutions.

In the main part, we will consider cost functionals that only depend on the variable v. It will turn out that the direct inclusion of p in the cost functional will require additional regularity of p. We will discuss how to formally include p without needing more regularity and properties of the optimality system with the direct inclusion of p in the end of this section.

8.1. **The Finite-dimensional State Equations.** We start with stating assumptions that ensure that the DAEs under consideration are of tractability index 2. Then, we introduce a decoupling of the equations that identifies the differential and algebraic parts to read off necessary conditions for consistency and regularity of the data.

We will consider a semi-explicit semi-linear DAE with distributed control of the form:

Problem 8.1. *Let $T > 0$ and let n_v, n_p, $n_u \in \mathbb{N}$. Let $M \in \mathcal{C}(0,T;\mathbb{R}^{n_v,n_v})$ be pointwise invertible, let $A\colon (0,T) \times \mathbb{R}^{n_v} \to \mathbb{R}^{n_v,n_v}$ be a smooth function, let J_1, $J_2 \in \mathcal{C}(0,T;\mathbb{R}^{n_p,n_v})$, and let $B_1 \in \mathcal{C}(0,T;\mathbb{R}^{n_v,n_u})$ and $B_2 \in \mathcal{C}(0,T;\mathbb{R}^{n_p,n_u})$. For given $\alpha \in \mathbb{R}^{n_v}$, $u \in \mathcal{C}(0,T;\mathbb{R}^{n_u})$ and right hand sides f and g in $\mathcal{C}(0,T;\mathbb{R}^{n_v})$ and $\mathcal{C}(0,T;\mathbb{R}^{n_p})$, respectively, find $(v,p) \in \mathcal{C}^1(0,T;\mathbb{R}^{n_v}) \times \mathcal{C}(0,T;\mathbb{R}^{n_p})$ that fulfills*

$$M(t)\dot{v}(t) - A(t,v(t))v(t) - J_1(t)^{\mathsf{T}}p(t) - B_1(t)u(t) = f(t), \tag{8.2a}$$

$$-J_2(t)v(t) - B_2(t)u(t) = g(t), \tag{8.2b}$$

on $(0,T)$, and

$$v(0) = \alpha. \tag{8.2c}$$

In order to guarantee existence of solutions $(v,p) \in \mathcal{C}^1(0,T;\mathbb{R}^{n_v}) \times \mathcal{C}(0,T;\mathbb{R}^{n_p})$ of (8.2), we make the following assumption:

Assumption 8.2. *Consider Problem 8.1. We assume*

(a) that $S := J_2 M^{-1} J_1^{\mathsf{T}}$ is pointwise invertible,
(b) sufficient regularity of the data and the input, namely g, $B_2 u$, $M^{-1}J_1^{\mathsf{T}}S^{-1}$, and J_2 are differentiable, and
(c) consistency of the data and the input, i.e. $J_2(0)v(0) = g(0) - B_2(0)u(0)$.

As a direct application of Proposition 2.6 we can state:

Proposition 8.3. *Consider Problem 8.1. If Assumption 8.2(a) holds, then,for any input function $u \in \mathcal{C}(0,T;\mathbb{R}^{n_u})$, the DAE (8.2a,b) has tractability index $i_\mu = 2$.*

Remark 8.4. If M is the mass matrix of a spatial discretization, then it is, typically, symmetric and strictly positive definite as is (8.1) if $\{\phi^i\}_{i=1}^{n_v}$ is a basis. Also, if one considers stable discretization schemes, i.e. schemes such that J_1 and J_2 and V_k and Q_k fulfill Assumptions 4.8 and 4.13, then also Assumption 8.2(a) is fulfilled. This is in particular the case for *discrete LBB stable* finite element schemes for the Navier Stokes equations, cf. Remark 4.15.

Remark 8.5. If also Assumption 8.2(a) holds, then one can define the *differentiation index* which will be $i_\nu = 2$, cf. Remark 2.9 and [155, Exa. 2].

The following Theorem 8.6 gives a solution representation by means of the inherent ODE and algebraic equations, cf. Lemma 3.30 for the abstract setting, that will be used to ensure existence and uniqueness in the linear case of System (8.2).

Theorem 8.6. *Consider Problem 8.1. Each solution* (v, p) *of* (8.2) *can be represented as* $(v_\mathcal{P} + \mathcal{Q}v, p)$*, where*

$$\mathcal{Q}v = -M^{-1}J_1^\mathsf{T}S^{-1}[B_2u + g], \tag{8.3a}$$

$$p = -\mathcal{Q}^-[M^{-1}[A(\mathcal{Q}v + v_\mathcal{P})[\mathcal{Q}v + v_\mathcal{P}] + B_1u + f] + \dot{\mathcal{Q}v}], \tag{8.3b}$$

and $v_\mathcal{P} := \mathcal{P}v$ *solves the ODE*

$$\begin{aligned}\dot{v_\mathcal{P}} - \big[\tfrac{d}{dt}\mathcal{P} + \mathcal{P}M^{-1}A(\mathcal{Q}v + v_\mathcal{P})\big][\mathcal{Q}v + v_\mathcal{P}] - \mathcal{P}M^{-1}[B_1u + f] = 0,\\ v_\mathcal{P}(0) = \mathcal{P}v^0.\end{aligned} \tag{8.3c}$$

with $\mathcal{P} := I - \mathcal{Q}$*,* $\mathcal{Q} := M^{-1}J_1^\mathsf{T}S^{-1}J_2$ *and* $\mathcal{Q}^- := S^{-1}J_2$*.*

Proof. We rewrite (8.2) as

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} A(v) & J_1^\mathsf{T} \\ J_2 & 0 \end{bmatrix}\begin{bmatrix} v \\ p \end{bmatrix} = \begin{bmatrix} B_1u + f \\ B_2u + g \end{bmatrix} \tag{8.4}$$

and we compute the sequence of operators from Definition 2.3 as given in the proof of Proposition 2.6. In particular, with the projectors $\mathcal{Q} = M^{-1}J_1^\mathsf{T}S^{-1}J_2$, which satisfies

$$\mathcal{Q}^2 = \mathcal{Q}, \quad J_2\mathcal{Q} = J_2, \quad \mathcal{Q}M^{-1}J_1^\mathsf{T} = M^{-1}J_1^\mathsf{T}, \quad \text{and} \quad \mathcal{Q}^-\mathcal{Q} = \mathcal{Q}^-,$$

and $\mathcal{P} = I - \mathcal{Q}$, we define

$$\mathcal{E}_2^{-1} = \begin{bmatrix} \mathcal{P}M^{-1} & [I - \mathcal{P}M^{-1}A]M^{-1}J_1^\mathsf{T}S^{-1} \\ \mathcal{Q}^-M^{-1} & -[I + \mathcal{Q}^-M^{-1}AM^{-1}J_1^\mathsf{T}]S^{-1} \end{bmatrix} \tag{8.5}$$

for any $A = A(v)$. Scaling the state equations (8.4) by $\mathcal{E}_2^{-1}$ we get

$$\begin{bmatrix} \mathcal{P} & 0 \\ \mathcal{Q}^- & 0 \end{bmatrix}\begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \Big[\begin{bmatrix} \mathcal{P}M^{-1}A\mathcal{P} & 0 \\ \mathcal{Q}^-M^{-1}A\mathcal{P} & 0 \end{bmatrix} + \begin{bmatrix} \mathcal{Q} & 0 \\ -\mathcal{Q}^- & I \end{bmatrix}\Big]\begin{bmatrix} v \\ p \end{bmatrix} = \mathcal{E}_2^{-1}\begin{bmatrix} M^{-1}[B_1u + f] \\ B_2u + g \end{bmatrix}. \tag{8.6}$$

Having applied the projectors $\mathscr{Q}_1$, $\mathscr{Q}_0\mathscr{P}_1$ and $\mathscr{P}_0\mathscr{P}_1$, cf. Definition 2.3, to (8.6) we obtain the three subsystems

$$\begin{aligned} -\begin{bmatrix} \mathcal{Q} & 0 \\ -\mathcal{Q}^- & 0 \end{bmatrix}\begin{bmatrix} v \\ p \end{bmatrix} &= \mathscr{Q}_1\mathcal{E}_2^{-1}\begin{bmatrix} B_1u + f \\ B_2u + g \end{bmatrix} \\ &= \begin{bmatrix} M^{-1}J_1^\mathsf{T}S^{-1}[B_2u + g] \\ -S^{-1}[B_2u + g] \end{bmatrix}, \end{aligned} \tag{8.7a}$$

$$\begin{aligned} \begin{bmatrix} 0 & 0 \\ \mathcal{Q}^- & 0 \end{bmatrix}\begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \mathcal{Q}^-M^{-1}A\mathcal{P} & I \end{bmatrix}\begin{bmatrix} v \\ p \end{bmatrix} &= \mathscr{Q}_0\mathscr{P}_1\mathcal{E}_2^{-1}\begin{bmatrix} B_1u + f \\ B_2u + g \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ \mathcal{Q}^-M^{-1}[B_1u + f - AM^{-1}J_1^\mathsf{T}S^{-1}[B_2u + g]] \end{bmatrix} \end{aligned} \tag{8.7b}$$

and

$$\begin{aligned} \begin{bmatrix} \mathcal{P} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} \mathcal{P}M^{-1}A\mathcal{P} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} v \\ p \end{bmatrix} &= \mathscr{P}_0\mathscr{P}_1\mathcal{E}_2^{-1}\begin{bmatrix} B_1u + f \\ B_2u + g \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{P}M^{-1}[B_1u + f - AM^{-1}J_1^\mathsf{T}S^{-1}[B_2u + g]] \\ 0 \end{bmatrix}, \end{aligned} \tag{8.7c}$$

respectively. Since $\mathscr{Q}_1+\mathscr{P}_0\mathscr{P}_1+\mathscr{Q}_0\mathscr{P}_1 = I$, Equations (8.7) contain all information of (8.6) and vice versa. We decompose $v = v_{\mathcal{P}} + \mathcal{Q}v$, where $v_{\mathcal{P}} := \mathcal{P}v$ so that from (8.7a) we can deduce that

$$\mathcal{Q}v = -M^{-1}J_1^{\mathsf{T}}S^{-1}[B_2u + g] \tag{8.8}$$

and that $\mathcal{Q}v$ is differentiable by assumption. With $\dot{v} = \dot{\mathcal{Q}v} + \dot{v_{\mathcal{P}}}$ and $\mathcal{Q}^{-}\dot{v_{\mathcal{P}}} = 0$, Equation (8.7b) gives

$$p = -\mathcal{Q}^{-}M^{-1}[A(\mathcal{Q}v + v_{\mathcal{P}})[\mathcal{Q}v + v_{\mathcal{P}}] + B_1u + f] + \mathcal{Q}^{-}\dot{\mathcal{Q}v}, \tag{8.9}$$

while (8.7c) defines the inherent ODE for $v_{\mathcal{P}} := \mathcal{P}v$ via

$$\dot{v_{\mathcal{P}}} - \big[\tfrac{d}{dt}\mathcal{P} + \mathcal{P}M^{-1}A(\mathcal{Q}v + v_{\mathcal{P}})\big][\mathcal{Q}v + v_{\mathcal{P}}] = \mathcal{P}M^{-1}[B_1u + f], \quad v_{\mathcal{P}}(0) = \mathcal{P}v^0. \tag{8.10}$$

□

Remark 8.7. Note the necessity of the consistency condition in Assumption 8.2(c), since by (8.8) the condition

$$J_2v(0) = J_2[\mathcal{Q}v(0) + \mathcal{P}v(0)] = J_2\mathcal{Q}v(0) = -B_2u(0) - g(0),$$

must hold and note, that an initial condition for p would have to fulfill (8.9) at $t = 0$.

Remark 8.8. In the setting of the Navier-Stokes Equation, the projector $\mathcal{Q}$ realizes the discrete Helmholtz-decomposition that splits a vector field into a divergence free part and a part that can be expressed as the gradient of a scalar potential, cf. [49, Cor. 3.4]. If J_2 is the discrete divergence operator, then the decomposition $v = \mathcal{Q}v + \mathcal{P}v =: \mathcal{Q}v + v_{\mathcal{P}}$ delivers that $J_2v_{\mathcal{P}} = 0$ and $\mathcal{Q}v$ is in the range of $M^{-1}J_1^{\mathsf{T}}$, which is the discrete gradient operator in many discretization schemes. The matrix $\mathcal{Q}^{-}$ is a generalized left inverse of $M^{-1}J_1^{\mathsf{T}}$ and can be seen as the operator that maps the potential field $\mathcal{Q}v = M^{-1}J_1^{\mathsf{T}}\rho$ onto its potential ρ. Accordingly, (8.3b) is the discrete Pressure Poisson Equation, cf. [50].

Corollary 8.9. *If $B_2 = 0$, then the solutions of* (8.2) *do not depend on the time derivative of the input. The condition $B_2 = 0$ is also necessary for the existence of solutions for all continuous inputs.*

Proof. The first assertion of Corollary 8.9 follows from the representation of the solution as given in Theorem 8.6. For the converse direction, one concludes that the solution component $\frac{d}{dt}(B_2u)$ can only exist for all continuous u if $B_2 = 0$. □

For the results of the next sections we will always require $B_2 = 0$ which by Corollary 8.9 is necessary and sufficient for the admissibility of inputs that are only continuous. This is what in [9, 10] and [134] is also assumed and referred to as *causality*.

8.2. **Optimal Control of Semi-explicit DAEs.** We define the semi-discrete optimal control problem:

Problem 8.10. *Consider the setup of Problem 8.1. Let $\mathcal{U} := \mathcal{C}(0, T; \mathbb{R}^{n_u})$ be the space of input functions and let $\mathcal{M} \colon \mathbb{R}^{n_v} \to \mathbb{R}$ and $\mathcal{K} \colon (0, T) \times \mathbb{R}^{n_v} \times \mathbb{R}^{n_p} \times \mathbb{R}^{n_u} \to \mathbb{R}$ be differentiable functionals. Consider the problem of finding an input $u \in \mathcal{U}$ such that the cost functional*

$$\mathcal{J}(v, p, u) = \mathcal{M}(v(T)) + \int_0^{\mathsf{T}} \mathcal{K}(t, v(t), p(t), u(t))\ \mathrm{d}t. \tag{8.11}$$

is minimal, where (v, p) and u are constrained via the DAE (8.2) *for given right hand sides f and g and a given initial value α.*

For Problem 8.10, one can formally derive the associated *Euler-Lagrange Equations*, cf. [91] and the formulation in infinite-dimensions given in Corollary 6.8:

Problem 8.11. *Consider Problem 8.10 and assume that A is differentiable. Find $(v,p) \in \mathcal{C}^1(0,T;\mathbb{R}^{n_v}) \times \mathcal{C}(0,T;\mathbb{R}^{n_p})$, $(\lambda,\mu) \in \mathcal{C}^1(0,T;\mathbb{R}^{n_v}) \times \mathcal{C}(0,T;\mathbb{R}^{n_p})$ and $u \in \mathcal{U}$ that fulfill the following equations:*

$$
\begin{aligned}
M\dot v - A(v)v - J_1^{\mathsf{T}} p - B_1 u &= f, && \text{(8.12a)}\\
-J_2 v - B_2 u &= g, && \text{(8.12b)}\\
v(0) &= \alpha, && \text{(8.12c)}\\
-\tfrac{d}{dt}(M^{\mathsf{T}}\lambda) - A_{;v}^{\mathsf{T}}(v)\lambda - J_2^{\mathsf{T}}\mu + \mathcal{K}_{;v}^{\mathsf{T}}(v,p,u) &= 0 && \text{(8.12d)}\\
M^{\mathsf{T}}\lambda(T) &= -\mathcal{M}_{;v}^{\mathsf{T}}(v(T)), && \text{(8.12e)}\\
-J_1\lambda + \mathcal{K}_{;p}^{\mathsf{T}}(v,p,u) &= 0, && \text{(8.12f)}\\
\mathcal{M}_{;p}(v(T)) &= 0, && \text{(8.12g)}\\
\mathcal{K}_{;u}(v,p,u) - B_1^{\mathsf{T}}\lambda - B_2^{\mathsf{T}}\mu &= 0, && \text{(8.12h)}
\end{aligned}
$$

where the time dependencies are dropped.

Remark 8.12. Equation system (8.12) is commonly referred to as *Euler-Lagrange equations.* If it possesses a solution, then it provides necessary optimality conditions, cf. Theorem 5.5 for a general formulation and [9, 92, 113] for the finite-dimensional linear case. Note, in particular, that the optimal control problem can be solvable also if (8.12) is not well posed [92].

We will not investigate existence of solutions here, but for the special case of linear-quadratic control.

8.3. **Linear-quadratic Optimal Control.**

In this section, we formulate an optimality system and determine necessary and sufficient conditions for optimal solutions in terms of the original equations rather than for a *strangeness-free* reformulation, cf. [27, 91].

We investigate a linearized version of (8.2), i.e. $A(t,v) = A(t)$, and a quadratic cost functional:

Problem 8.13. *Consider Problem 8.10 with the cost functional $\mathcal{J}$ be given as*

$$
\mathcal{J}(v,p,u) = \frac{1}{2}\begin{bmatrix} v\\ p\end{bmatrix}^{\mathsf{T}}\begin{bmatrix} V_1 & V_{12}\\ V_{21} & V_2\end{bmatrix}\begin{bmatrix} v\\ p\end{bmatrix}\Bigg|_{t=T} + \frac{1}{2}\int_0^T \begin{bmatrix} v\\ p\\ u\end{bmatrix}^{\mathsf{T}}\begin{bmatrix} W_1 & W_{12} & S_{vu}\\ W_{21} & W_2 & S_{pu}\\ S_{uv} & S_{up} & R\end{bmatrix}\begin{bmatrix} v\\ p\\ u\end{bmatrix}\,\mathrm{d}t, \tag{8.13}
$$

with R invertible and symmetric positive semi-definite weighting matrices $\begin{bmatrix} V_1 & V_{12}\\ V_{21} & V_2\end{bmatrix}$ and $\begin{bmatrix} W_1 & W_{12} & S_{vu}\\ W_{21} & W_2 & S_{pu}\\ S_{uv} & S_{up} & R\end{bmatrix}$ and the state equations given as

$$
\begin{aligned}
M(t)\dot v(t) - A(t)v(t) - J_1(t)^{\mathsf{T}}p(t) - B_1(t)u(t) &= f(t),\\
-J_2(t)v(t) - B_2(t)u(t) &= g(t),
\end{aligned}
$$

on $(0,T)$, and

$$
v(0) = \alpha.
$$

In the setting of Problem 8.13 the formal Euler-Lagrange equations, cf. [91], are given by

$$
\begin{aligned}
&M\dot{v} - Av - J_1^\mathsf{T} p - B_1 u = g, \quad v(0) = v^0 && \text{(8.14a)}\\
&-J_2 v - B_2 u = g && \text{(8.14b)}\\
&-\tfrac{d}{dt}(M^\mathsf{T}\lambda_1) - A^\mathsf{T}\lambda_1 - J_2^\mathsf{T}\lambda_2 + W_1 v + W_{12} p + S_{vu} u = 0, &&\\
&\qquad M^\mathsf{T}\lambda_1(T) = -V_1 v\Big|_{t=T} - V_{12} p\Big|_{t=T}, && \text{(8.14c)}\\
&-J_1\lambda_1 + W_{21} v + W_2 p + S_{pu} u = 0, \quad 0 = V_{21} v\Big|_{t=T} + V_2 p\Big|_{t=T}, && \text{(8.14d)}\\
&-B_1^\mathsf{T}\lambda_1 - B_2^\mathsf{T}\lambda_2 + S_{uv} v + S_{up} p + Ru = 0. && \text{(8.14e)}
\end{aligned}
$$

If system (8.14) possesses a solution, then it provides necessary and sufficient conditions for an optimal input u, cf. Remark 8.12. In what follows, we will establish conditions for existence of solutions of (8.14).

Remark 8.14. Since we consider state solutions $(v, p) \in \mathcal{C}^1 \times \mathcal{C}$ and inputs $u \in \mathcal{C}$ candidate solutions of (8.14) must not contain $\dot{u}$ or $\dot{p}$. Thus, by Corollary 8.9 it is necessary that

$$B_2 = 0, \quad W_2 = 0 \quad \text{and} \quad S_{pu} = S_{up}^\mathsf{T} = 0. \tag{8.15}$$

Another necessary condition for the existence of solutions, is the consistency of the initial values. The true optimality conditions, cf. [92], necessitate that $\operatorname{span}\begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \subset \operatorname{span}\begin{bmatrix} M^\mathsf{T} & 0 \\ 0 & 0 \end{bmatrix}$, i.e. V_{22} and $V_{12} = V_{21}^\mathsf{T}$ must be zero. By combining (8.14d) and the terminal condition for λ_1, we find

$$-J_1\lambda_1(T) = -W_{21} v(T) = J_1 M^{-\mathsf{T}} V_1 v(T).$$

We will ensure this condition by requiring

$$J_1 M^{-\mathsf{T}} V_1 = 0, \quad \text{and} \quad W_{21} = W_{12}^\mathsf{T} = 0. \tag{8.16}$$

The latter condition means that V_1 acts only on the dynamical part of v as it is given by (8.3c). Note that these conditions are equivalent to the assumptions that were made in [9].

Remark 8.15. In theory, setting W_2, $W_{12} = W_{21}^\mathsf{T}$ to zero, does not cause a loss of generality, as p is an affine linear function of v and u, cf. Theorem 8.6. Thus, in the cost functional, all terms in p can be replaced by terms in v and u. Furthermore, the cross terms of v and u can be formally eliminated by an input shift, cf. Section 8.5. However, for applications, the exclusion of p from the cost functional is a restriction.

We use the invertibility of R to express u via

$$u = R^{-1}[B_1^\mathsf{T}\lambda_1 + B_2^\mathsf{T}\lambda_2 - S_{uv} v - S_{up} p]$$

and write (8.14) in matrix vector form

$$\begin{bmatrix} f \\ g \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} M\dot{v} \\ 0 \\ -\frac{d}{dt}(M^{\mathsf{T}}\lambda_1) \\ 0 \end{bmatrix} -$$
$$\begin{bmatrix} B_1R^{-1}B_1^{\mathsf{T}} & B_1R^{-1}B_2^{\mathsf{T}} & A - B_1R^{-1}S_{uv} & J_1^{\mathsf{T}} - B_1R^{-1}S_{up} \\ B_2R^{-1}B_1^{\mathsf{T}} & B_2R^{-1}B_2^{\mathsf{T}} & J_2 - B_2R^{-1}S_{uv} & -B_2R^{-1}S_{up} \\ A^{\mathsf{T}} - S_{vu}R^{-1}B_1^{\mathsf{T}} & J_2^{\mathsf{T}} - S_{vu}R^{-1}B_2^{\mathsf{T}} & S_{vu}R^{-1}S_{uv} - W_1 & S_{uv}R^{-1}S_{up} - W_{12} \\ J_1 - S_{pu}R^{-1}B_1^{\mathsf{T}} & -S_{pu}R^{-1}B_2^{\mathsf{T}} & S_{pu}R^{-1}S_{uv} - W_{21} & S_{up}R^{-1}S_{up} - W_2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ v \\ p \end{bmatrix} \tag{8.17}$$

If (8.15) holds, the corresponding terms in (8.17) vanish:

$$\begin{bmatrix} M\dot{v} \\ 0 \\ -\frac{d}{dt}(M^{\mathsf{T}}\lambda_1) \\ 0 \end{bmatrix} - \begin{bmatrix} B_1R^{-1}B_1^{\mathsf{T}} & 0 & A - B_1R^{-1}S_{uv} & J_1^{\mathsf{T}} \\ 0 & 0 & J_2 & 0 \\ A^{\mathsf{T}} - S_{vu}R^{-1}B_1^{\mathsf{T}} & J_2^{\mathsf{T}} & -W_1 + S_{vu}R^{-1}S_{uv} & -W_{12} \\ J_1 & 0 & -W_{21} & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ v \\ p \end{bmatrix}$$
$$= \begin{bmatrix} f^{\mathsf{T}} & g^{\mathsf{T}} & 0 & 0 \end{bmatrix}^{\mathsf{T}}. \tag{8.18}$$

Remark 8.16. By inverting the mass matrices and permuting the rows and the columns, System (8.18) can be brought into the form of (8.2). Then, with Assumption 8.2(a) we have that the (8.18) is of tractability index 2. This follows from Proposition 2.6 and from

$$\begin{bmatrix} 0 & J_2 \\ J_1 & -W_{21} \end{bmatrix} \begin{bmatrix} 0 & M \\ -M^{\mathsf{T}} & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 & J_1^{\mathsf{T}} \\ J_2^{\mathsf{T}} & -W_{12} \end{bmatrix} =$$
$$\begin{bmatrix} 0 & J_2M^{-1}J_1^{\mathsf{T}} \\ -J_1M^{-\mathsf{T}}J_2^{\mathsf{T}} & -W_{21}M^{-1}J_1^{\mathsf{T}} - J_1M^{-\mathsf{T}}W_{12} \end{bmatrix}$$

being invertible by Assumption 8.2(a).

Assuming further that $W_{21} = 0$, cf. (8.16), we can write the system as $u = R^{-1}B_1^{\mathsf{T}}\lambda_1$,

$$\begin{bmatrix} M\dot{v} \\ 0 \\ -\frac{d}{dt}(M^{\mathsf{T}}\lambda_1) \\ 0 \end{bmatrix} - \begin{bmatrix} G & 0 & F & J_1^{\mathsf{T}} \\ 0 & 0 & J_2 & 0 \\ F^{\mathsf{T}} & J_2^{\mathsf{T}} & H & 0 \\ J_1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ v \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \\ 0 \\ 0 \end{bmatrix}, \tag{8.19a}$$

$$v(0) = v^0 \quad \text{and} \quad M^{\mathsf{T}}\lambda_1(T) = -V_1v(T), \tag{8.19b}$$

with $F := A - B_1R^{-1}S_{uv}$, symmetric matrices $G := B_1R^{-1}B_1^{\mathsf{T}}$ and $H := -W_1 + S_{vu}R^{-1}S_{uv}$.

8.4. **Existence and Representations of Optimal Solutions.** In this section, we introduce a Riccati-decoupling for the optimality system. Using the projectors from Section 8.1, we determine differential and algebraic parts of the obtained differential-algebraic matrix Riccati equation and prove well-posedness. As a side-product we establish the unique solvability of the corresponding optimality system.

We have that if (8.15) holds, then the considered Euler-Lagrange equations are in the form (8.2), cf. Remark 8.16. Therefore, one may apply Theorem 8.6 to identify the inherent ODE (8.10), one may use the theory for ODEs to state the existence of solutions to the obtained linear boundary value problem, cf. [8, Thm. 3.26]. However, the reformulation as used in Theorem 8.6 will not preserve the symmetry of (8.19) and thus makes it more difficult to investigate whether the boundary values admit the existence of a solution. We will use a reformulation that preserves the *Hamiltonian* structure such that the existence of a solution can be obtained via a differential Riccati equation, cf. [35, 36, 37, 130].

Lemma 8.17. *Consider the semi-explicit linear DAE of index 2*

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} A & J_1^\mathsf{T} \\ J_2 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} - \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u = \begin{bmatrix} f \\ g \end{bmatrix}, \quad v(0) = v^0 \tag{8.20}$$

and a cost functional

$$\mathcal{J}(v,u) = \frac{1}{2}[v - v^*]^\mathsf{T} V_1 [v - v^*]\Big|_{t=T} + \frac{1}{2}\int_0^T \begin{bmatrix} v - v^* \\ u \end{bmatrix}^\mathsf{T} \begin{bmatrix} W_1 & S_{vu} \\ S_{uv} & R \end{bmatrix} \begin{bmatrix} v - v^* \\ u \end{bmatrix} \mathrm{d}t, \tag{8.21}$$

which does not act onto the algebraic variable p and with symmetric positive semi-definite weighting matrices and R symmetric positive definite. Define the matrix functions $F := A - B_1 R^{-1} S_{uv}$, $G := B_1 R^{-1} B_1^\mathsf{T}$ and $H := -W_1 + S_{vu} R^{-1} S_{uv}$.

1.) Each solution $(v, p, \lambda_1, \lambda_2)$ of the associated Euler-Lagrange equations as given by (8.19) *has a representation $(v, p) = (v_\mathcal{P} + \mathcal{Q}v, p)$ and $(M^\mathsf{T}\lambda_1, \lambda_2) = (\lambda_\mathcal{P} + \mathcal{Q}^\mathsf{T} M^\mathsf{T} \lambda_1, \lambda_2)$ given by the decoupled system*

$$\mathcal{Q}v = -M^{-1} J_1^\mathsf{T} S^{-1} g, \tag{8.22a}$$

$$\mathcal{Q}^\mathsf{T} M^\mathsf{T} \lambda_1 = 0, \tag{8.22b}$$

$$\lambda_2 = -S^{-\mathsf{T}} J_1 M^{-\mathsf{T}} \big[H[\mathcal{Q}v + v_\mathcal{P}] + F^\mathsf{T} M^{-\mathsf{T}} \lambda_\mathcal{P}\big], \tag{8.22c}$$

$$\begin{aligned} p = {} & -\mathcal{Q}^-[M^{-1}[F[\mathcal{Q}v + v_\mathcal{P}] + f] - \tfrac{d}{dt}(\mathcal{Q}v)] - \\ & - \mathcal{Q}^- M^{-1} G M^{-\mathsf{T}} [\lambda_\mathcal{P} + \mathcal{Q}^\mathsf{T} F M^{-\mathsf{T}} \lambda_\mathcal{P} + J_2^\mathsf{T} \lambda_2], \end{aligned} \tag{8.22d}$$

and

$$\begin{gathered} \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} \dot{\lambda}_\mathcal{P} \\ \dot{v}_\mathcal{P} \end{bmatrix} - \begin{bmatrix} G_0 & F_0 \\ F_0^\mathsf{T} & H_0 \end{bmatrix} \begin{bmatrix} \lambda_\mathcal{P} \\ v_\mathcal{P} \end{bmatrix} = \begin{bmatrix} \mathcal{P} M^{-1}[f - F M^{-1} J_1^\mathsf{T} S^{-1} g] \\ \mathcal{P}^\mathsf{T} H M^{-1} J_1^\mathsf{T} S^{-1} g \end{bmatrix}, \\ v_\mathcal{P}(0) = \mathcal{P} v^0 \quad \text{and} \quad \lambda_\mathcal{P}(T) = -\mathcal{P}^\mathsf{T} V_1 v(T), \end{gathered} \tag{8.22e}$$

where $F_0 := \frac{d}{dt}\mathcal{P} + \mathcal{P} M^{-1} F \mathcal{P}$, $G_0 = G_0^\mathsf{T} := \mathcal{P} M^{-1} G M^{-\mathsf{T}} \mathcal{P}^\mathsf{T}$, $H_0 = H_0^\mathsf{T} := \mathcal{P}^\mathsf{T} H \mathcal{P}$ and $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{Q}^-$ and S as defined in Theorem 8.6.

2.) If in addition

$$J_2 v^0 = g(0) \quad \text{and} \quad J_1 M^{-\mathsf{T}} V_1 = 0, \tag{8.23}$$

then the Euler-Lagrange equations (8.19) *possess a unique solution.*

3.) If in addition f and g are zero, then (8.19) *can be decoupled via*

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2^\mathsf{T} \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix}, \tag{8.24}$$

where $X_1 = X_1^\mathsf{T}$ and X_2 fulfill the differential-algebraic Riccati equation

$$\begin{aligned} & \tfrac{d}{dt} M^\mathsf{T} X_1 M + M^\mathsf{T} X_1 F + F^\mathsf{T} X_1 M + M^\mathsf{T} X_1 G X_1 M + H + \\ & \qquad + M^\mathsf{T} X_2^\mathsf{T} J_2 + J_2^\mathsf{T} X_2 M = 0, \end{aligned} \tag{8.25a}$$

with the terminal condition

$$M^\mathsf{T} X_1(T) M = -V_1, \tag{8.25b}$$

and the algebraic constraints

$$M^\mathsf{T} J_1 X_1 = 0 \quad \text{and} \quad J_1 X_1 M = 0. \tag{8.25c}$$

Equations (8.25a-c) uniquely define a symmetric negative semi-definite X_1.

4.) If, however, f, g and v^* are not identically zero, then the solution of (8.19) *decouples via*

$$\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}X_1 & X_2^\mathsf{T}\\ X_2 & 0\end{bmatrix}\begin{bmatrix}M & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} + \begin{bmatrix}w_1\\ w_2\end{bmatrix}, \tag{8.26}$$

where X_1 and X_2 are given by a solution of (8.25) *and (w_1, w_2) is the unique solution of*

$$-\tfrac{d}{dt}(\begin{bmatrix}M^\mathsf{T} w_1\\ 0\end{bmatrix}) - \begin{bmatrix}M^\mathsf{T} X_1 G + F^\mathsf{T} & J_2^\mathsf{T}\\ J_1 & 0\end{bmatrix}\begin{bmatrix}w_1\\ w_2\end{bmatrix} = \begin{bmatrix}f_{\lambda_1} + M^\mathsf{T}[X_1\tilde{f}_v + X_2 g]\\ 0\end{bmatrix}, \tag{8.27}$$
$$M^\mathsf{T} w_1(T) = V_1 v^*(T),$$

with $\tilde{f}_v := f + B_1 R^{-1} S_{uv} v^*$ and $f_{\lambda_1} := [W_1 - S_{vu} R^{-1} S_{uv}]v^*$.

Proof. ad *1.)* We write the Euler-Lagrange system, cf. (8.19), as

$$\begin{bmatrix}0 & 0 & I & 0\\ 0 & 0 & 0 & 0\\ -I & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\frac{d}{dt}\begin{bmatrix}M^\mathsf{T}\lambda_1\\ \lambda_2\\ v\\ p\end{bmatrix} - \begin{bmatrix}M^{-1}GM^{-\mathsf{T}} & 0 & M^{-1}F & M^{-1}J_1^\mathsf{T}\\ 0 & 0 & J_2 & 0\\ F^\mathsf{T} M^{-\mathsf{T}} & J_2^\mathsf{T} & H & 0\\ J_1 M^{-\mathsf{T}} & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}M^\mathsf{T}\lambda_1\\ \lambda_2\\ v\\ p\end{bmatrix}$$
$$= \begin{bmatrix}f^\mathsf{T} M^{-\mathsf{T}} & g^\mathsf{T} & 0 & 0\end{bmatrix}^\mathsf{T},$$
$$v(0) = v^0 \quad \text{and} \quad M^\mathsf{T}\lambda_1(T) = -V_1 v(T).$$

In order to preserve the self-adjoint structure, cf. [94], only congruence transformations should be applied, i.e. a scaling of the equations must be accompanied by the transpose inverse scaling of the variables. In accordance to (8.6) we congruently transform the system by

$$S_2 := \begin{bmatrix}\mathcal{E}_2^{-1} & & \\ & I & \\ & & I\end{bmatrix} = \begin{bmatrix}\mathcal{P} & [I - \mathcal{P}M^{-1}F]M^{-1}J_1^\mathsf{T} S^{-1} & 0 & 0\\ \mathcal{Q}^- & -[I + \mathcal{Q}^- M^{-1}FM^{-1}J_1^\mathsf{T}]S^{-1} & 0 & 0\\ 0 & 0 & I & 0\\ 0 & 0 & 0 & I\end{bmatrix},$$

where $\mathcal{E}_2 = \begin{bmatrix}I + M^{-1}F\mathcal{Q} & M^{-1}J_1^\mathsf{T}\\ J_2 & 0\end{bmatrix}$ as defined in Definition 2.3 with the inverse given in (8.5), up to a scaling by M^{-1}. The summand that comes from the time-dependency in the variable transformation S_2^T is given by

$$S_2\begin{bmatrix}0 & 0 & I & 0\\ 0 & 0 & 0 & 0\\ -I & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\dot{S}_2^\mathsf{T} = \begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ -\frac{d}{dt}\mathcal{P}^\mathsf{T} & -\dot{\mathcal{Q}}^{\mathsf{T}-} & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}.$$

With this we get the scaled and transformed system

$$\tilde{f} = \begin{bmatrix}0 & 0 & \mathcal{P} & 0\\ 0 & 0 & \mathcal{Q}^- & 0\\ -\mathcal{P}^\mathsf{T} & -\mathcal{Q}^{\mathsf{T}-} & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\dot{\tilde{\lambda}}_1\\ \dot{\tilde{\lambda}}_2\\ \dot{v}\\ \dot{p}\end{bmatrix} + \begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ -\frac{d}{dt}\mathcal{P}^\mathsf{T} & -\dot{\mathcal{Q}}^{\mathsf{T}-} & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\tilde{\lambda}_1\\ \tilde{\lambda}_2\\ v\\ p\end{bmatrix} -$$
$$\begin{bmatrix}\mathcal{P}M^{-1}GM^{-\mathsf{T}}\mathcal{P}^\mathsf{T} & M^{-1}GM^{-\mathsf{T}}\mathcal{Q}^{\mathsf{T}-} & \mathcal{P}M^{-1}F\mathcal{P} + \mathcal{Q} & 0\\ \mathcal{Q}^- M^{-1}GM^{-\mathsf{T}} & 0 & \mathcal{Q}^- M^{-1}F\mathcal{P} - \mathcal{Q}^- & I\\ \mathcal{P}^\mathsf{T} F^\mathsf{T} M^{-\mathsf{T}}\mathcal{P}^\mathsf{T} + \mathcal{Q}^\mathsf{T} & \mathcal{P}^\mathsf{T} F^\mathsf{T} M^{-\mathsf{T}}\mathcal{Q}^{\mathsf{T}-} - \mathcal{Q}^{\mathsf{T}-} & H & 0\\ 0 & I & 0 & 0\end{bmatrix}\begin{bmatrix}\tilde{\lambda}_1\\ \tilde{\lambda}_2\\ v\\ p\end{bmatrix} \tag{8.29}$$

with the transformed state and scaled right hand side

$$\begin{bmatrix}\tilde\lambda_1\\ \tilde\lambda_2\\ v\\ p\end{bmatrix} := S_2^{-\mathsf{T}}\begin{bmatrix}M^\mathsf{T}\lambda_1\\ \lambda_2\\ v\\ p\end{bmatrix} = \begin{bmatrix}[I+\mathcal{Q}^\mathsf{T}F^\mathsf{T}M^{-\mathsf{T}}]M^T\lambda_1 + J_2^\mathsf{T}\lambda_2\\ J_1\lambda_1\\ v\\ p\end{bmatrix}$$

and $\tilde f := S_2\begin{bmatrix}f^\mathsf{T}M^{-\mathsf{T}} & g^\mathsf{T} & 0 & 0\end{bmatrix}^\mathsf{T}$, respectively. From the last line in (8.29) we find that $\tilde\lambda_2 = 0$ so that we can rewrite the equations for $(\tilde\lambda_1, v, p)$ as

$$\begin{bmatrix}\mathcal{P} & 0\\ \mathcal{Q}^- & 0\end{bmatrix}\begin{bmatrix}\dot v\\ \dot p\end{bmatrix} - \begin{bmatrix}\mathcal{P}M^{-1}GM^{-\mathsf{T}}\mathcal{P}^\mathsf{T} & 0\\ \mathcal{Q}^-M^{-1}GM^{-\mathsf{T}} & 0\end{bmatrix}\begin{bmatrix}\tilde\lambda_1\\ \tilde\lambda_2\end{bmatrix} - \begin{bmatrix}\mathcal{P}M^{-1}F\mathcal{P}+\mathcal{Q} & 0\\ \mathcal{Q}^-M^{-1}F\mathcal{P}-\mathcal{Q}^- & I\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} = \mathcal{E}_2^{-1}\begin{bmatrix}M^{-1}f\\ g\end{bmatrix} \tag{8.30a}$$

and

$$-\tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1) - [\mathcal{P}^\mathsf{T}F^\mathsf{T}M^{-\mathsf{T}}\mathcal{P}^\mathsf{T} + \mathcal{Q}^\mathsf{T}]\tilde\lambda_1 - Hv = 0. \tag{8.30b}$$

Analogously to (8.7) we apply the projectors

$$\mathscr{Q}_1 = \begin{bmatrix}\mathcal{Q} & 0\\ -\mathcal{Q}^- & 0\end{bmatrix},\quad \mathscr{Q}_0\mathscr{P}_1 = \begin{bmatrix}0 & 0\\ \mathcal{Q}^- & I\end{bmatrix}\quad\text{and}\quad \mathscr{P}_0\mathscr{P}_1 = \begin{bmatrix}\mathcal{P} & 0\\ 0 & 0\end{bmatrix}$$

to (8.30a) to obtain the three subsystems

$$-\begin{bmatrix}\mathcal{Q} & 0\\ -\mathcal{Q}^- & 0\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} = \begin{bmatrix}M^{-1}J_1^\mathsf{T}S^{-1}g\\ -S^{-1}g\end{bmatrix}, \tag{8.31a}$$

$$\begin{bmatrix}0 & 0\\ \mathcal{Q}^- & 0\end{bmatrix}\begin{bmatrix}\dot v\\ \dot p\end{bmatrix} - \begin{bmatrix}0 & 0\\ \mathcal{Q}^-M^{-1}GM^{-\mathsf{T}} & 0\end{bmatrix}\begin{bmatrix}\tilde\lambda_1\\ \tilde\lambda_2\end{bmatrix} - \begin{bmatrix}0 & 0\\ \mathcal{Q}^-M^{-1}F\mathcal{P} & I\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} = \begin{bmatrix}0\\ \mathcal{Q}^-M^{-1}[f - FM^{-1}J_1^\mathsf{T}S^{-1}g]\end{bmatrix} \tag{8.31b}$$

and

$$\begin{bmatrix}\mathcal{P} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\dot v\\ \dot p\end{bmatrix} - \begin{bmatrix}\mathcal{P}M^{-1}GM^{-\mathsf{T}}\mathcal{P}^\mathsf{T} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\tilde\lambda_1\\ \tilde\lambda_2\end{bmatrix} - \begin{bmatrix}\mathcal{P}M^{-1}F\mathcal{P} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} = \begin{bmatrix}\mathcal{P}M^{-1}[f - FM^{-1}J_1^\mathsf{T}S^{-1}g]\\ 0\end{bmatrix}, \tag{8.31c}$$

respectively. Using the projector property $\mathcal{P}^\mathsf{T} = \mathcal{P}^\mathsf{T}\mathcal{P}^\mathsf{T}$ to obtain the relation

$$\tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1) = \dot{\mathcal{P}}^\mathsf{T}\mathcal{P}^\mathsf{T}\tilde\lambda_1 + \mathcal{P}^\mathsf{T}\tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1) = \tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1) - \mathcal{Q}^\mathsf{T}\tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1) + \dot{\mathcal{P}}^\mathsf{T}\mathcal{P}^\mathsf{T}\tilde\lambda_1$$

we split (8.30b) into the two subsystems

$$\mathcal{Q}^\mathsf{T}\tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1)\ - \mathcal{Q}^\mathsf{T}\tilde\lambda_1 - \mathcal{Q}^\mathsf{T}Hv = 0 \tag{8.32a}$$

and

$$-\tfrac{d}{dt}(\mathcal{P}^\mathsf{T}\tilde\lambda_1) - \tfrac{d}{dt}(\mathcal{P}^\mathsf{T})\mathcal{P}^\mathsf{T}\tilde\lambda_1 - \mathcal{P}^\mathsf{T}F^\mathsf{T}M^{-\mathsf{T}}\mathcal{P}^\mathsf{T}\tilde\lambda_1 - \mathcal{P}^\mathsf{T}Hv = 0. \tag{8.32b}$$

If we then define $v_\mathcal{P} := \mathcal{P}v$ and $\tilde\lambda_\mathcal{P} := \mathcal{P}^\mathsf{T}\tilde\lambda_1$ and decompose $\tilde\lambda_1 = \tilde\lambda_\mathcal{P} + \mathcal{Q}^\mathsf{T}\tilde\lambda_1$ and $v = v_\mathcal{P} + \mathcal{Q}v$ we find that (8.31a-b) and (8.32a) define algebraic relations for

$$\mathcal{Q}v = -M^{-1}J_1^\mathsf{T}S^{-1}g, \tag{8.33a}$$

$$\mathcal{Q}^\mathsf{T}\tilde\lambda_1 = -\mathcal{Q}^\mathsf{T}[H\mathcal{Q}v + Hv_\mathcal{P}] + \mathcal{Q}^\mathsf{T}\dot{\tilde\lambda}_\mathcal{P} \tag{8.33b}$$

and, with $\mathcal{Q}^-\dot{v_\mathcal{P}}$

$$p = -\mathcal{Q}^-[M^{-1}F[\mathcal{Q}v + v_\mathcal{P}] + M^{-1}f + M^{-1}GM^{-\mathsf{T}}\tilde\lambda_1 - \tfrac{d}{dt}(\mathcal{Q}v)], \tag{8.33c}$$

while $\tilde{\lambda}_{\mathcal{P}}$ and $v_{\mathcal{P}}$ are defined by the coupled ODEs given by (8.32b) and (8.31c):

$$-\dot{\tilde{\lambda}}_{\mathcal{P}} - \big[\tfrac{d}{dt}\mathcal{P}^{\mathsf{T}} + \mathcal{P}^{\mathsf{T}}F^{\mathsf{T}}M^{-\mathsf{T}}\mathcal{P}^{\mathsf{T}}\big]\tilde{\lambda}_{\mathcal{P}} - \mathcal{P}^{\mathsf{T}}H\mathcal{P}v_{\mathcal{P}} = \mathcal{P}^{\mathsf{T}}HM^{-1}J_1^{\mathsf{T}}S^{-1}g \tag{8.34a}$$

and

$$\begin{aligned}\dot{v}_{\mathcal{P}} - \mathcal{P}M^{-1}GM^{-\mathsf{T}}\mathcal{P}^{\mathsf{T}}\tilde{\lambda}_{\mathcal{P}} - \big[\tfrac{d}{dt}\mathcal{P} + \mathcal{P}M^{-1}F\mathcal{P}\big]v_{\mathcal{P}} = \\ \mathcal{P}M^{-1}[f - FM^{-1}J_1^{\mathsf{T}}S^{-1}g].\end{aligned} \tag{8.34b}$$

Note that we have used the projector property $\mathcal{P} = \mathcal{P}^2$ to keep the symmetry in (8.34) obvious.

In view of expressing the obtained relations in terms of the original variables (λ_1, λ_2) we observe that

$$\tilde{\lambda}_{\mathcal{P}} = \mathcal{P}^{\mathsf{T}}\tilde{\lambda}_1 = \mathcal{P}^{\mathsf{T}}[M^{\mathsf{T}}\lambda_1 + \mathcal{Q}^{\mathsf{T}}F^{\mathsf{T}}\lambda_1 + J_2^{\mathsf{T}}\lambda_2] = \mathcal{P}^{\mathsf{T}}M^{\mathsf{T}}\lambda_1 =: \lambda_{\mathcal{P}}.$$

From $\tilde{\lambda}_2 = J_1\lambda_1 = 0$ we confer

$$\mathcal{Q}^{\mathsf{T}}M^{\mathsf{T}}\lambda_1 = J_2^{\mathsf{T}}S^{-\mathsf{T}}J_1\lambda_1 = 0.$$

For λ_2 we use $\mathcal{Q}^{\mathsf{T}}\tilde{\lambda}_1 = \mathcal{Q}^{\mathsf{T}}[I + \mathcal{Q}^{\mathsf{T}}FM^{-\mathsf{T}}]M^{\mathsf{T}}\lambda_1 + \mathcal{Q}^{\mathsf{T}}J_2^{\mathsf{T}}\lambda_2 = \mathcal{Q}^{\mathsf{T}}FM^{-\mathsf{T}}\lambda_{\mathcal{P}} + J_2^{\mathsf{T}}\lambda_2$, relation (8.33b), and the invertibility of $S^{\mathsf{T}} = J_1M^{-1}J_2^{\mathsf{T}}$ to get

$$\lambda_2 = -S^{-\mathsf{T}}J_1M^{-\mathsf{T}}[H[\mathcal{Q}v + v_{\mathcal{P}}] + FM^{-\mathsf{T}}\lambda_{\mathcal{P}}].$$

Note that $J_1^{\mathsf{T}}M^{-\mathsf{T}}\mathcal{P}^{\mathsf{T}} = 0$, so that $\dot{\lambda}_{\mathcal{P}} \in \operatorname{span}\mathcal{P}^{\mathsf{T}}$ does not appear.

Similarly, one can express the equation for p in terms of (λ_1, λ_2) which completes the derivation of Equations (8.22).

ad *2.)*

First, we show that for any v^0 and $\mathcal{P}^{\mathsf{T}}V_1$ symmetric positive semi-definite the decoupled system (8.22) has a unique solution $(v_{\mathcal{P}}, \mathcal{Q}v, p, \lambda_{\mathcal{P}}, \mathcal{Q}^{\mathsf{T}}M^{\mathsf{T}}\lambda_1, \lambda_2)$. Second, we confer that under the consistency conditions (8.23) the solution of (8.22) provides a solution of the Euler-Lagrange equations (8.19). Finally, by *1.)* every solution of (8.19) has a representation in (8.22), such that in summary the Euler-Lagrange equations must possess a unique solution.

We first consider in (8.22e-f) the case with a zero right hand side. With the Riccati ansatz $\lambda_{\mathcal{P}} = X_0(t)v_{\mathcal{P}}(t)$ these equations can be rewritten as the differential Riccati equation

$$\dot{X}_0 = -X_0G_0X_0 - X_0F_0 - F_0^{\mathsf{T}}X_0 - H_0, \quad X_0(T) = -\mathcal{P}^{\mathsf{T}}V_1, \tag{8.35}$$

which has a unique solution, cf. [2, Thm. 4.1.6], since $\mathcal{P}^{\mathsf{T}}V_1$, G_0, and $-H_0$ are symmetric positive semi-definite. With this X_0 we get $v_{\mathcal{P}}$ and $\lambda_{\mathcal{P}}$ as the solution of $\dot{v}_{\mathcal{P}} - [G_0X_0 + F_0]v_{\mathcal{P}} = 0, v_{\mathcal{P}}(0) = \mathcal{P}v^0$ and $\lambda_{\mathcal{P}} = X_0v_{\mathcal{P}}$, respectively.

One can show that if there exists a solution to (8.22e-f) with a zero right hand side, then it is unique. This is equivalent to the fact that the linear part of the affine boundary conditions are stated such, that (8.22e-f) with $\mathcal{P}^{\mathsf{T}}V_1$ symmetric positive semi-definite, has a unique solution, cf. [8, Thm. 3.26], for any continuous right hand side.

By construction, a solution of (8.19) uniquely defines a solution to (8.22). The converse is true if, and only if, the algebraic variables fulfill the initial and terminal conditions, i.e.,

$$\mathcal{Q}v(0) = \mathcal{Q}v_{\mathcal{P}} = M^{-1}J_1^{\mathsf{T}}S^{-1}J_2v_{\mathcal{P}} \quad \text{and} \tag{8.36a}$$

$$\mathcal{Q}^{\mathsf{T}}M^{\mathsf{T}}\lambda_1(T) = -\mathcal{Q}^{\mathsf{T}}V_1v(T) = -J_2^{\mathsf{T}}S^{-\mathsf{T}}J_1M^{-\mathsf{T}}V_1v(T). \tag{8.36b}$$

By (8.22a) we have that $\mathcal{Q}v(0) = M^{-1}J_1^{\mathsf{T}}S^{-1}g(0)$ such that $J_2v(0) = g(0)$ is necessary and sufficient for (8.36a). By (8.22b) we have that $\mathcal{Q}^{\mathsf{T}}M^{\mathsf{T}}\lambda_1 = 0$ such that $J_1M^{-\mathsf{T}}V_1 = 0$ is sufficient but in general not necessary for (8.36b). Note, however, that in this case we can infer that $J_1^{\mathsf{T}}M^{-\mathsf{T}}V_1 = 0$, so that $V_1M^{-1}J_1 = 0$

or $V_1\mathcal{Q} = 0$, which means that $V_1 v = V_1\mathcal{P}v$, such that in (8.22f), $\mathcal{P}^\mathsf{T} V_1$ can be replaced by $\mathcal{P}^\mathsf{T} V_1 \mathcal{P}$. Thus, condition (8.23) implies the symmetry in the terminal condition that was sufficient for the existence of X_0 in (8.35).

ad *3.)*

With the ansatz

$$\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix} = \begin{bmatrix}X_1 & X_2^\mathsf{T}\\ X_2 & 0\end{bmatrix}\begin{bmatrix}M & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} \tag{8.37}$$

we obtain that

$$\tfrac{d}{dt}(\begin{bmatrix}M^\mathsf{T} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}) = \begin{bmatrix}\frac{d}{dt}M^\mathsf{T}X_1M & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}v\\ p\end{bmatrix} + \begin{bmatrix}M^\mathsf{T}X_1 & M^\mathsf{T}X_2^\mathsf{T}\\ 0 & 0\end{bmatrix}\begin{bmatrix}M & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\dot v\\ \dot p\end{bmatrix}. \tag{8.38}$$

In (8.38) we replace $\frac{d}{dt}(\begin{bmatrix}M^\mathsf{T} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix})$ and $\begin{bmatrix}M & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\dot v\\ \dot p\end{bmatrix}$ via the relations given in (8.19) and every occurrence of $\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}$ by the ansatz (8.37) to obtain $\mathscr{X}\begin{bmatrix}v\\ p\end{bmatrix} = 0$, where

$$\mathscr{X} := \begin{bmatrix}\begin{array}{c}\frac{d}{dt}(M^\mathsf{T}X_1M) + F^\mathsf{T}X_1M + M^\mathsf{T}X_1F +\\ M^\mathsf{T}X_1GX_1M + H + J_2^\mathsf{T}X_2M + M^\mathsf{T}X_2^\mathsf{T}J_2\end{array} & M^\mathsf{T}X_1J_1^\mathsf{T}\\ \\ J_1X_1M & 0\end{bmatrix}. \tag{8.39}$$

Since $\mathscr{X}\begin{bmatrix}v\\ p\end{bmatrix} = 0$ must hold for every state trajectory, one requires $\mathscr{X} = 0$ which gives the equations for X_1 and X_2:

$$\begin{aligned}\tfrac{d}{dt}M^\mathsf{T}X_1M + M^\mathsf{T}X_1F + F^\mathsf{T}X_1M + M^\mathsf{T}X_1GX_1M + H+&\\ +M^\mathsf{T}X_2^\mathsf{T}J_2 + J_2^\mathsf{T}X_2M = 0&,\\ M^\mathsf{T}X_1(T)M = -V_1&, \qquad (8.40\text{a})\\ M^\mathsf{T}J_1X_1 = 0 \quad\text{and}\quad J_1X_1M = 0&. \qquad (8.40\text{b})\end{aligned}$$

The terminal condition in (8.40a) is defined via (8.19b) and (8.24):

$$M^\mathsf{T}\lambda_1(T) = M^\mathsf{T}X_1(T)Mv(T) = -V_1v(T) \quad\Rightarrow\quad M^\mathsf{T}X_1(T)M = -V_1.$$

To show that (8.40) has a solution, we consider Equation (8.40a) in the transformed variables $X := -M^\mathsf{T}X_1M$ and $Y := X_2M$:

$$\begin{aligned}-\dot X - F^\mathsf{T}M^{-\mathsf{T}}X - XM^{-1}F + XM^{-1}G^{-\mathsf{T}}M^{-\mathsf{T}}X + H + J_2^\mathsf{T}Y + Y^\mathsf{T}J_2 = 0&,\\ X(T) = V_1&. \qquad (8.41)\end{aligned}$$

By means of the projector $\mathcal{Q} := M^{-1}J_1^\mathsf{T}[J_2M^{-1}J_1^\mathsf{T}]^{-1}J_2$ we write $X = [\mathcal{Q}^\mathsf{T} + \mathcal{P}^\mathsf{T}]X[\mathcal{P} + \mathcal{Q}]$. From (8.40b) we obtain that $\mathcal{Q}^\mathsf{T}X = X\mathcal{Q} = 0$ and thus X is completely defined via $X_0 := \mathcal{P}^\mathsf{T}X\mathcal{P}$. Applying $\mathcal{P}^\mathsf{T}$ and $\mathcal{P}$ to (8.41) from the left and the right, respectively, we get a standard differential Riccati equation

$$\begin{aligned}-\dot X_0 - F_0^\mathsf{T}X_0 - X_0F_0 + X_0M^{-1}GM^{-\mathsf{T}}X_0 + \mathcal{P}^\mathsf{T}H\mathcal{P} = 0&,\\ X_0(T) = \mathcal{P}^\mathsf{T}V_1\mathcal{P}&, \qquad (8.42)\end{aligned}$$

which has a unique and symmetric positive semi-definite solution, cf. [2, Thm. 4.1.6], since V_1, G and $-H$ are symmetric positive semi-definite. Again, the consistency condition (8.23) ensures that $X_0(T)$ also satisfies the initial condition and the algebraic constraints in (8.40). Since $\mathcal{Q}^\mathsf{T}X = 0$ and $X\mathcal{Q} = 0$, we have $X_1 = -M^{-\mathsf{T}}XM^{-1}$ is unique and symmetric negative semi-definite.

Application of $\mathcal{P}^\mathsf{T}$ from the left and $\mathcal{Q}$ from the right to (8.41) gives

$$-X_0\dot{\mathcal{Q}} - X_0M^{-1}F\mathcal{Q} + \mathcal{P}^\mathsf{T}H\mathcal{Q} = -\mathcal{P}^\mathsf{T}Y^\mathsf{T}J_2\mathcal{Q} = -\mathcal{P}^\mathsf{T}Y^\mathsf{T}J_2,$$

which is uniquely solvable for $\mathcal{P}^\mathsf{T} Y^\mathsf{T}$. The projected equation obtained via $\mathcal{Q}^\mathsf{T}$ and $\mathcal{P}$ is the transpose of the above equation and bears no additional information.

Finally, one can determine $\mathcal{Q}^\mathsf{T} Y^\mathsf{T}$ from the projection of (8.41) onto the range of $\mathcal{Q}^\mathsf{T}$ and $\mathcal{Q}$ which reads

$$\mathcal{Q}^\mathsf{T} H \mathcal{Q} + \mathcal{Q}^\mathsf{T} Y^\mathsf{T} J_2 \mathcal{Q} + \mathcal{Q}^\mathsf{T} J_2^\mathsf{T} Y \mathcal{Q} = 0. \tag{8.43}$$

With $J_2 \mathcal{Q} = J_2$, we find that (8.43) is of the form $[Y\mathcal{Q}]^\mathsf{T} J_2 + J_2^\mathsf{T} [Y\mathcal{Q}] = -\mathcal{Q}^\mathsf{T} H \mathcal{Q}$ that was investigated in [24]. With $\mathcal{Q}^- := M^{-1} J_1^\mathsf{T} [J_2 M^{-1} J_1^\mathsf{T}]^{-1}$ being a generalized inverse to J_2, we obtain the projectors $P_1 := \mathcal{Q}^- J_2 = \mathcal{Q}$ and $P_2 := J_2 \mathcal{Q}^- = I$ and the existence of solutions to (8.43) follows by [24, Thm. 1], since $\mathcal{Q}^\mathsf{T} H \mathcal{Q}$ is symmetric and $[I - P_1]^\mathsf{T} \mathcal{Q}^\mathsf{T} H \mathcal{Q} [I - P_1] = 0$.

The general solution to (8.43) is given by

$$Y\mathcal{Q} = \frac{1}{2} [J_1 M^{-\mathsf{T}} J_2^\mathsf{T}]^{-1} J_1 M^{-\mathsf{T}} H \mathcal{Q} + Z J_2,$$

where Z is arbitrary with $Z^\mathsf{T} = -Z$. Thus existence of $M^\mathsf{T} X_1 M$ and $M^\mathsf{T} X_2^\mathsf{T} = Y^\mathsf{T} = \mathcal{P}^\mathsf{T} Y^\mathsf{T} + \mathcal{Q}^\mathsf{T} Y^\mathsf{T}$ and therefore X_1 and X_2 is proved.

By construction, with X_1 and X_2 as determined above, the solution of

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \left(\begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 & X_2^\mathsf{T} \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} F & J_1^\mathsf{T} \\ J_2 & 0 \end{bmatrix} \right) \begin{bmatrix} v \\ p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad v(0) = v^0,$$

and

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2^\mathsf{T} \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix}$$

gives the solution of (8.19) with a zero right-hand side.

ad *4.)* The result for the affine-linear case is obtained via the affine-linear ansatz (8.26) using similar arguments as in the proof of *3.)*. Proceeding analogously to the first steps for part *3.)*, but with the affine linear ansatz (8.26) instead of the linear (8.24), we come to the expression

$$\mathscr{X} \begin{bmatrix} v \\ p \end{bmatrix} + \tfrac{d}{dt} \left(\begin{bmatrix} M^\mathsf{T} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \right) - \begin{bmatrix} M^\mathsf{T} X_1 G + F^\mathsf{T} & J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} f_{\lambda_1} + M^\mathsf{T} X_1 \tilde{f}_v + M^\mathsf{T} X_2 g \\ f_{\lambda_2} \end{bmatrix}, \qquad M^\mathsf{T} w_1(T) = V_1 v^*(T), \tag{8.45}$$

where $\mathscr{X}$ is as in (8.39). Again, the requirement $\mathscr{X} = 0$ uniquely defines X_1 and $X_2 =: \tilde{X}_2 + Z J_2 M^{-1}$ up to an arbitrary skew-symmetric matrix Z. We write $M^\mathsf{T} X_2^\mathsf{T} g = M^\mathsf{T} \tilde{X}_2^\mathsf{T} g - J_2^\mathsf{T} Z f_p$ and define $\tilde{w}_2 := w_2 - Zg$. With this Equation (8.45) gives a system for $(w_1, \tilde{w}_2)$:

$$\begin{bmatrix} -\frac{d}{dt}(M^\mathsf{T} w_1) \\ 0 \end{bmatrix} - \begin{bmatrix} M^\mathsf{T} X_1 G + F^\mathsf{T} & J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ \tilde{w}_2 \end{bmatrix} = \begin{bmatrix} f_{\lambda_1} + M^\mathsf{T} X_1 \tilde{f}_v + M^\mathsf{T} \tilde{X}_2 g \\ f_{\lambda_2} \end{bmatrix}, \qquad M^\mathsf{T} w_1(T) = -V_1 v^*(T), \tag{8.46}$$

which is of type (8.2). Since by (8.23) the terminal condition is consistent, system (8.46) has a unique solution $(w_1, \tilde{w}_2)$. In particular, w_1 is independent of Zg, cf. (8.8) and (8.10). For the solution w_2 of (8.45) we have $w_2 = \tilde{w}_2 + Z f_p$. Thus the existence of the functions used for the ansatz (8.26) is shown.

By construction, we have that the ansatz (8.26) leads to the solution of the Euler-Lagrange equations (8.19) via the decoupled system

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \left(\begin{bmatrix} G & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 & X_2^\mathsf{T} \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} F & J_1^\mathsf{T} \\ J_2 & 0 \end{bmatrix} \right) \begin{bmatrix} v \\ p \end{bmatrix} = \begin{bmatrix} f + B_1 R^{-1} S_{uv} v^* + G w_1 \\ g \end{bmatrix},$$

$$v(0) = v^0,$$

and

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2^\mathsf{T} \\ X_2 & 0 \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

□

Remark 8.18. The solution of (8.25) is unique up to an additive term ZJ_2M^{-1} in X_2, with an arbitrary matrix Z, that fulfills $Z^\mathsf{T} = -Z$. However, this does not contradict the unique solvability of the Euler-Lagrange equations, since λ_1 and λ_2 as defined via (8.24) are independent of any choice of Z.

In view of optimal control, the above results can be summarized as follows. To obtain an optimal input u for (8.20) with respect to a cost functional of type (8.21) it is sufficient to have a solution of the associated Euler-Lagrange equations (8.19), cf. [113]. By Lemma 8.17 it follows that for the considered state equations and cost functionals this solution exists, that it is unique, that it can be obtained via the separation ansatz (8.24), and that an optimal u is obtained via expression (8.19c). For the inhomogeneous and for the trajectory tracking case, one can use an affine linear Riccati-ansatz, cf. [93]. Thus, we can state the following theorem:

Theorem 8.19. *Let $T > 0$ and consider the time interval $(0, T]$, let n_u, n_v, $n_p \in \mathbb{N}$, $n_v > n_p$, $M \in \mathcal{C}(0, T; \mathbb{R}^{n_v,n_v})$ pointwise invertible, $A \in \mathcal{C}(0, T; \mathbb{R}^{n_v,n_v})$, and let J_1, $J_2 \in \mathcal{C}(0, T; \mathbb{R}^{n_p,n_v})$, such that $J_2M^{-1}J_1^\mathsf{T}$ is invertible and such that $M^{-1}J_1^\mathsf{T}S^{-1}J_2$ is differentiable. Let W_1,$V_1 \in \mathbb{R}^{n_v,n_v}$ be symmetric positive semi-definite, $S_{uv} = S_{vu}^\mathsf{T} \in \mathbb{R}^{n_u,n_v}$ an arbitrary matrix, and let $R \in \mathbb{R}^{n_u,n_u}$ symmetric positive definite.*

For a given $v^* \in \mathcal{C}^1(0, T; \mathbb{R}^{n_v})$ consider the linear-quadratic optimal control problem of finding $u \in \mathcal{C}(0, T; \mathbb{R}^{n_u})$ such that

$$\frac{1}{2} \left[v - v^*\right]^\mathsf{T} V_1 \left[v - v^*\right] \Big|_{t=T} + \frac{1}{2} \int_0^T \begin{bmatrix} v - v^* \\ u \end{bmatrix}^\mathsf{T} \begin{bmatrix} W_1 & S_{vu} \\ S_{uv} & R \end{bmatrix} \begin{bmatrix} v - v^* \\ u \end{bmatrix} \,\mathrm{d}t,$$

is minimal, where v on $(0, T]$ satisfies the state equations

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} A & J_1^\mathsf{T} \\ J_2 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} - \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u = \begin{bmatrix} f \\ g \end{bmatrix}, \quad v(0) = v^0.$$

If $f \in \mathcal{C}(0, T; \mathbb{R}^{n_v})$, $g \in \mathcal{C}^1(0, T; \mathbb{R}^{n_p})$ and if

$$J_2 v^0 = g(0) \quad \text{and} \quad J_1 M^{-\mathsf{T}} V_1 = 0,$$

then the optimal control problem is solvable and an optimal input u is given via the feedback-law

$$u = R^{-1}[B_1^\mathsf{T}[X_1 M v + w_1] - S_{uv}(v - v^*)],$$

where $X_1 = X_1^\mathsf{T}$, negative semi-definite, and w_1 are the unique solutions of

$$\begin{aligned}\tfrac{d}{dt}M^\mathsf{T}X_1M + F^\mathsf{T}X_1M + M^\mathsf{T}X_1F + M^\mathsf{T}X_1GX_1M + H+&\\ +J_2^\mathsf{T}X_2M + M^\mathsf{T}X_2^\mathsf{T}J_2 &= 0,\\ M^\mathsf{T}X_1(T)M &= -V_1,\\ J_1X_1M &= 0,\end{aligned}$$

and

$$\begin{aligned}-\tfrac{d}{dt}(M^\mathsf{T}w_1) - [M^\mathsf{T}X_1G + F^\mathsf{T}]w_1 - J_2^\mathsf{T}w_2 &= f_{\lambda_1} + M^\mathsf{T}X_1\tilde{f}_v + M^\mathsf{T}X_2g,\\ M^\mathsf{T}w_1(T) &= V_1v^*(T),\\ J_1w_1 &= 0,\end{aligned}$$

respectively, where $F := A - B_1R^{-1}S_{uv}$, $G := B_1R^{-1}B_1^\mathsf{T}$ and $H := -W_1 + S_{vu}R^{-1}S_{uv}$ and with $\tilde{f}_v := f + B_1R^{-1}S_{uv}v^*$ and $f_{\lambda_1} := [W_1 - S_{vu}R^{-1}S_{uv}]v^*$.

Theorem 8.19 gives – in particular – sufficient optimality conditions in terms of the original variables and coefficients. Necessity of these conditions is not guaranteed in general, as an inconsistent V_1 renders them ill-posed, although for well-posed state equations a solution of the optimal control problem always exists, cf. the true optimality system defined in [93].

For practical applications, the following modification that closes the gap between sufficiency and necessity of the optimality conditions in Theorem 8.19 may be considered.

Remark 8.20. By Theorem 8.6 one has that if v solves (8.20), then it can be expressed as $v = \mathcal{P}v - c$, with $c := M^{-1}J_1^\mathsf{T}S^{-1}g$ independent of u and v and that the terminal point penalization in the cost functional (8.21) can be replaced like

$$\frac{1}{2}v^\mathsf{T}(T)V_1v(t) \quad \leftarrow \quad \frac{1}{2}[\mathcal{P}v(t) - c(t)]^\mathsf{T}V_1[\mathcal{P}v(T) - c(T)].$$

With this equivalent formulation, the terminal condition on $M^\mathsf{T}\lambda_1$ in (8.19b) coming from the variation of the cost functional with respect to v reads $M^\mathsf{T}\lambda_1(T) = -\mathcal{P}^\mathsf{T}V_1[\mathcal{P}v(T) - c(T)]$. Then the end condition for the gain matrix X_1 is given via $\mathcal{P}^\mathsf{T}V_1\mathcal{P}$ and for the affine part w_1 via $M^\mathsf{T}w_1(T) = \mathcal{P}^\mathsf{T}V_1[v^*(T) + M^{-1}J_1^\mathsf{T}S^{-1}g(T)$. Both conditions are consistent as $J_1M^{-\mathsf{T}}\mathcal{P}^\mathsf{T} = 0$. With this modification, in Theorem 8.19, the restriction $J_1M^{-\mathsf{T}}V_1 = 0$ is obsolete and the given optimality conditions are equivalent to the true optimality conditions.

8.5. **Crossterms and the Algebraic Variable in the Cost Functional.** As mentioned in Remark 8.15, in the linear case, one can theoretically reformulate any cost functional of type (8.13) as an equivalent cost weighting without the algebraic variable p and without cross terms in the integral part.

To illustrate this, we assume that the right hand sides f and g in the state equations (8.2) are zero. If f and g are not zero, the same approach will lead to a linear in v and u term in the cost functional, which then appears in the right hand side of the adjoint equation as does W_1v^* in the affine linear formulation in Theorem 8.19. If they are zero, then, by Theorem 8.6, we have

$$p = -\mathcal{Q}^- M^{-1}Av - \mathcal{Q}^- M^{-1}B_1u$$

and the trajectory weighting $\begin{bmatrix} v \\ p \\ u \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} W_1 & W_{12} & S_{vu} \\ W_{21} & W_2 & S_{pu} \\ S_{uv} & S_{up} & R \end{bmatrix} \begin{bmatrix} v \\ p \\ u \end{bmatrix}$ in Problem 8.13 can be reformulated as $\begin{bmatrix} v \\ u \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \tilde{W}_1 & \tilde{S}_{vu} \\ \tilde{S}_{uv} & \tilde{R} \end{bmatrix} \begin{bmatrix} v \\ u \end{bmatrix}$ with

$$\begin{bmatrix} \tilde{W}_1 & \tilde{S}_{vu} \\ \tilde{S}_{uv} & \tilde{R} \end{bmatrix} :=$$

$$\begin{bmatrix} I & 0 \\ -Q^{-}M^{-1}A & -Q^{-}M^{-1}B_1 \\ 0 & I \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} W_1 & W_{12} & S_{vu} \\ W_{21} & W_2 & S_{pu} \\ S_{uv} & S_{up} & R \end{bmatrix} \begin{bmatrix} I & 0 \\ -Q^{-}M^{-1}A & -Q^{-}M^{-1}B_1 \\ 0 & I \end{bmatrix}.$$

If then $\tilde{R} = \begin{bmatrix} -Q^{-}M^{-1}B_1 \\ I \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} W_2 & S_{pu} \\ S_{up} & R \end{bmatrix} \begin{bmatrix} -Q^{-}M^{-1}B_1 \\ I \end{bmatrix}$ was invertible, this would give the situation of that was considered in Section 8.3.

Then, due to the linearity of the problem, one could also formally eliminate the crossterms given by $\tilde{S}_{uv}$ in the cost functional by shifting the input and considering $\tilde{u} = u + \tilde{R}^{-1}\tilde{S}_{uv}^{\mathsf{T}}v$. This gives an equivalent formulation of the optimal control problem Problem 8.13 without endpoint penalization: Find $u \in \mathcal{U}$, such that

$$\mathcal{J}(v,p,\tilde{u}) = \frac{1}{2}\int_0^{\mathsf{T}} v^{\mathsf{T}}[\tilde{W} - \tilde{S}_{uv}\tilde{R}^{-1}\tilde{S}_{uv}^{\mathsf{T}}]v + \tilde{u}^{\mathsf{T}}\tilde{R}\tilde{u}\ \mathrm{d}t \to \min \tag{8.48a}$$

subject to

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} A - B_1\tilde{R}^{-1}\tilde{S}_{uv}^{\mathsf{T}} & J_1^{\mathsf{T}} \\ J_2 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} - \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \tilde{u} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad v(0) = v^0, \tag{8.48b}$$

where $\tilde{u} = u + \tilde{R}^{-1}\tilde{S}_{uv}^{\mathsf{T}}v$.

However, the presence of a cross term S_{up}, may make $\tilde{R}$ indefinite, such that the theory of Section 8.3 is not applicable.

Remark 8.21. As illustrated above, if R is invertible, in the setting of linear-quadratic systems as Problem 8.13, one can always eliminate cross terms S_{uv} by shifting the input as $\tilde{u} \leftarrow u + R^{-1}S_{uv}^{\mathsf{T}}v$.

The elimination of the linking of p and u by a shift of the input is not simply possible in general. In the shifted system with $\tilde{u} = u + R^{-1}S_{up}^{\mathsf{T}}p$, the state equations read

$$\begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{p} \end{bmatrix} - \begin{bmatrix} A & J_1^{\mathsf{T}} - B_1R^{-1}S_{up}^{\mathsf{T}} \\ J_2 & 0 \end{bmatrix} \begin{bmatrix} v \\ p \end{bmatrix} - \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \tilde{u} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad v(0) = v^0.$$

This shows, that if one shifts u via p, a different approach is necessary, since the index of the state equations may change, cf. Proposition 8.3 and Assumption 8.2(a). In particular, if $J_1^{\mathsf{T}} - B_1R^{-1}S_{up}^{\mathsf{T}}$ is not of full rank, the state equations (8.48b) are not uniquely solvable.

8.6. **Optimal Control Including the Algebraic Variables.** In this section, we consider linear-quadratic optimal control problems that explicitly includes the variable p in the cost functional. This will need additional regularity of p, cf. Remark 8.14. We will not discuss solvability but investigate the differential algebraich structure of the associated formal Euler-Lagrange equations.

Consider Problem 8.13 with a cost functional of type

$$\mathcal{J}(v,p,u) = \frac{1}{2}\begin{bmatrix} v-v^* \\ p-p^* \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} V_1 & V_{12} \\ V_{21} & V_2 \end{bmatrix} \begin{bmatrix} v-v^* \\ p-p^* \end{bmatrix}\Bigg|_{t=T} + \\ \frac{1}{2}\int_0^T \begin{bmatrix} v-v^* \\ p-p^* \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} W_1 & W_{12} \\ W_{21} & W_2 \end{bmatrix} \begin{bmatrix} v-v^* \\ p-p^* \end{bmatrix} + u^{\mathsf{T}} Ru \;\mathrm{d}t,$$

subject to (8.2), with weighting matrices $\begin{bmatrix} V_1 & V_{12} \\ V_{21} & V_2 \end{bmatrix}$, $\begin{bmatrix} W_1 & W_{12} \\ W_{21} & W_2 \end{bmatrix}$ being positive semi definite and R invertible. We do not include cross terms, because the coupling of u to v only means a shift in the input, while the coupling of u and p causes changes in the system structure, cf. Remark 8.21. The formal Euler-Lagrange equations, cf. [91], read

$$M\dot{v} - Av - J_1^{\mathsf{T}} p - B_1 u = f, \quad v|_{t=0} = v^0 \tag{8.49a}$$

$$-J_2 v = g \tag{8.49b}$$

$$-\tfrac{d}{dt}(M^{\mathsf{T}}\lambda_1) - A^{\mathsf{T}}\lambda_1 - J_2^{\mathsf{T}}\lambda_2 + W_1[v-v^*] + W_{12}[p-p^*] = 0, \tag{8.49c}$$

$$M^{\mathsf{T}}\lambda_1(T) = -V_1[v-v^*]\Big|_{t=T} - V_{12}[p-p^*]\Big|_{t=T}, \tag{8.49d}$$

$$-J_1\lambda_1 + W_{21}[v-v^*] + W_2[p-p^*] = 0, \tag{8.49e}$$

$$V_{21}[v-v^*]\Big|_{t=T} + V_2[p-p^*]\Big|_{t=T} = 0, \tag{8.49f}$$

$$-B_1^{\mathsf{T}}\lambda_1 + Ru = 0. \tag{8.49g}$$

Unlike in Section 8.3 we now consider arbitrary smooth data. Nevertheless, for the existence of solutions, the initial and terminal conditions have to be consistent with the algebraic constraints. We assume that the target state is consistent at the endpoint, i.e. $-J_2 v^*(T) = g(T)$ to conclude that

$$v(T) - v^*(T) \in \mathcal{P}\mathbb{R}^{n_v}, \quad \mathcal{Q}[v(T)-v^*(T)] = 0 \quad \text{and} \quad p(T)-p^*(T) \in \mathbb{R}^{n_p}, \tag{8.50}$$

using the projectors $\mathcal{P}$ and $\mathcal{Q}$ that were defined in Theorem 8.6. We will not consider the possibility of cancelling a term that contains $v - v^*$ by a term that contains $p - p^*$.

Thus, combining the terminal condition (8.49d) and the algebraic constraints (8.49e) on λ_1 we obtain consistency conditions for the p related coefficients via $\big[J_1 M^{-1} V_{12} + W_2\big][p(T) - p^*(T)] = 0$ or, eqivalently, $V_{12} = -J_2^{\mathsf{T}} S^{-T} W_2 + \mathcal{P}^{\mathsf{T}} Z_1$, with an arbitrary $Z_1 \in \mathbb{R}^{n_v, n_p}$. By symmetry and (8.49f) we further need

$$V_{21}[v(T) - v^*(T)] = V_{12}^{\mathsf{T}}[v(T) - v^*(T)] = Z_1^{\mathsf{T}}\mathcal{P}[v(T) - v^*(T)] = 0$$

which can only hold for $Z_1^{\mathsf{T}}\mathcal{P} = 0$. Thus the one obtains the necessary condition

$$V_{21}^{\mathsf{T}} = V_{21} = -J_2^{\mathsf{T}} S^{-T} W_2. \tag{8.51}$$

From (8.49f) we also conclude that necessarily

$$V_2 = 0. \tag{8.52}$$

Regarding v, the conditions (8.49d) and (8.49e) give $\big[J_1 M^{-1} V_1 + W_{21}\big][v(T) - v^*(T)] = 0$ or, equivalently, $V_1\mathcal{P} = -J_2^{\mathsf{T}} S^{-T} W_{21}\mathcal{P} + \mathcal{P}^{\mathsf{T}} Z_2 \mathcal{P}$, as, by assumption, $v(T) - v^*(T)$ takes on values only in the subspace $\mathcal{P}\mathbb{R}^{n_v}$. This means that for given W_{21}, the only condition on V_1 is

$$Q^{\mathsf{T}} V_1 \mathcal{P} = -J_2^{\mathsf{T}} S^{-T} W_{21}\mathcal{P}, \tag{8.53}$$

while $\mathcal{P}^{\mathsf{T}} V_1 \mathcal{P}$ and $V_1\mathcal{Q}$ can be chosen arbitrarily.

Remark 8.22. If $W_2 \neq 0$ and $g(T) \neq 0$, then the regulator problem, i.e. $v^* = 0$, may not be solvable in this setting. This follows from $V_{21}v(T) = -W_2 S^{-1} J_2 v(T) = W_2 S^{-1} g(T)$ which is not necessarily 0 and thus possibly violates (8.49f).

Using relation (8.49g) to express u in terms of λ_1 and assuming that (8.51), (8.52), and (8.53) hold, we find that the matrices

$$\mathcal{E} = \begin{bmatrix} 0 & -M^\mathsf{T} & 0 & 0 \\ M & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{A} = \begin{bmatrix} -W_1 & F^\mathsf{T} + \dot{M}^\mathsf{T} & -W_{12} & J_2^\mathsf{T} \\ F & G & J_1^\mathsf{T} & 0 \\ -W_{21} & J_1 & -W_2 & 0 \\ J_2 & 0 & 0 & 0 \end{bmatrix}, \tag{8.54}$$

with $F := A$ and $G := B_1 R^{-1} B_1^\mathsf{T}$, represent system (8.49) without the boundary conditions (8.49d) and (8.49f) via

$$\mathcal{E}\dot{x} - \mathcal{A}x = f, \tag{8.55}$$

for the state vector and right hand side

$$x = \begin{bmatrix} v \\ \lambda_1 \\ p \\ \lambda_2 \end{bmatrix} \quad \text{and} \quad f = \begin{bmatrix} f_{\lambda_1} \\ f \\ f_{\lambda_2} \\ g \end{bmatrix} := \begin{bmatrix} W_1 v^* + W_{12} p^* \\ f \\ W_{21} v^* + W_2 p^* \\ g \end{bmatrix},$$

respectively. Since $\mathcal{E}^\mathsf{T} = -\mathcal{E}$ and $\mathcal{A}^\mathsf{T} = \mathcal{A} + \dot{\mathcal{E}}$ hold, by [94, Def. 3.9] the pair $(\mathcal{E}, \mathcal{A})$ of matrix functions is self-adjoint. Thus (8.55) can be reduced to the canonical form developed in [94] to state the determine the differentiation index of (8.49). In view determining the index, we state the following technical lemma.

Lemma 8.23. *Let $M \in \mathbb{R}^{n_v, n_v}$ be invertible and A, $B \in \mathbb{R}^{n_p, n_v}$ such that $AM^{-1}B^\mathsf{T}$ is invertible. Then for all Q_A, $Q_B \in \mathbb{R}^{n_v, n_v - n_p}$, where the columns of Q_A, Q_B form a basis of* $\ker A$, $\ker B$, *respectively, the matrix $Q_B^\mathsf{T} M Q_A$ is invertible.*

Proof of Lemma 8.23. We have, $AM^{-1}B^\mathsf{T}$ is invertible or $\ker A \cap \operatorname{im} M^{-1}B^\mathsf{T} = \{0\}$ implying that $\left[Q_A, M^{-1}B^\mathsf{T}\right]$ is invertible. Also, $BM^{-T}A^\mathsf{T}$ is invertible, what gives $\ker BM^{-T} \cap A^\mathsf{T} = \{0\}$ and, thus, $\left[M^\mathsf{T} Q_B, A^\mathsf{T}\right]$ is invertible. Consequently

$$\left[M^\mathsf{T} Q_B, A^\mathsf{T}\right]^\mathsf{T} \left[Q_A, M^{-1}B^\mathsf{T}\right] = \begin{bmatrix} Q_B^\mathsf{T} M Q_A & 0 \\ 0 & AM^{-1}B^\mathsf{T} \end{bmatrix}$$

and in particular $Q_B^\mathsf{T} M Q_A$ is invertible. □

Now we can determine the *differentiation index* of (8.49), which, by Remark 2.9, in this case, coincides with the *tractability index*.

Lemma 8.24. *Consider Equation* (8.49). *Let M be pointwise symmetric positive definite and let $J_2 M^{-1} J_1^\mathsf{T}$ be invertible. If $W_{21} = W_{12}^\mathsf{T}$ and W_1, W_2, and G are symmetric positive semi-definite and if conditions* (8.51), (8.52), *and* (8.53) *hold and if the coefficients are sufficiently smooth, then* (8.49) *is of differentiation index* $\nu_d = 3$.

Proof. We write (8.49) as $\mathcal{E}\dot{x} - \mathcal{A}x = f$, cf. (8.55). Since J_1 and J_2 are of full rank, there are orthogonal matrices $Q_i = \begin{bmatrix} Q_i^1 \\ Q_i^2 \end{bmatrix}$ such that $J_i \begin{bmatrix} Q_i^{1\mathsf{T}} & Q_i^{2\mathsf{T}} \end{bmatrix} = \begin{bmatrix} R_{J_i} & 0 \end{bmatrix}$, with R_{J_i} invertible, $i = 1, 2$. With this, a congruence transformation with $Q := \mathtt{diag}(Q_2, Q_1, I, I)$ of system (8.55) reads

$$\hat{\mathcal{E}}\dot{\hat{x}} - \hat{\mathcal{A}}\hat{x} = \hat{f},$$

where

$$\hat{\mathcal{E}} := \begin{bmatrix} 0 & 0 & -Q_2^1 M^T Q_1^{1\mathsf{T}} & -Q_2^1 M^\mathsf{T} Q_1^{2\mathsf{T}} & 0 & 0 \\ 0 & 0 & -Q_2^2 M^\mathsf{T} Q_1^{1\mathsf{T}} & -Q_2^2 M^\mathsf{T} Q_1^{2\mathsf{T}} & 0 & 0 \\ Q_1^1 M Q_2^{1\mathsf{T}} & Q_1^1 M Q_2^{2\mathsf{T}} & 0 & 0 & 0 & 0 \\ Q_1^2 M Q_2^{1\mathsf{T}} & Q_1^2 M Q_2^{2\mathsf{T}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \hat{x} := \begin{bmatrix} Q_2^1 v \\ Q_2^2 v \\ Q_1^1 \lambda_1 \\ Q_1^2 \lambda_1 \\ p \\ \lambda_2 \end{bmatrix},$$

$$\hat{\mathcal{A}} := \begin{bmatrix} -Q_2^1 W_1 Q_2^{1\mathsf{T}} & -Q_2^1 W_1 Q_2^{2\mathsf{T}} & Q_2^1 F_M^\mathsf{T} Q_1^{1\mathsf{T}} & Q_2^1 F_M^\mathsf{T} Q_1^{2\mathsf{T}} & -Q_2^1 W_{12} & R_{J_1}^\mathsf{T} \\ -Q_2^2 W_1 Q_2^{1\mathsf{T}} & -Q_2^2 W_1 Q_2^{2\mathsf{T}} & Q_2^2 F_M^\mathsf{T} Q_1^{1\mathsf{T}} & Q_2^2 F_M^\mathsf{T} Q_1^{2\mathsf{T}} & -Q_2^2 W_{12} & 0 \\ Q_1^1 F Q_2^{1\mathsf{T}} & Q_1^1 F Q_2^{2\mathsf{T}} & Q_1^1 G Q_1^{1\mathsf{T}} & Q_1^1 G Q_1^{2\mathsf{T}} & R_{J_1}^\mathsf{T} & 0 \\ Q_1^2 F Q_2^{1\mathsf{T}} & Q_1^2 F Q_2^{2\mathsf{T}} & Q_1^2 G Q_1^{1\mathsf{T}} & Q_1^2 G Q_1^{2\mathsf{T}} & 0 & 0 \\ -W_{21} Q_2^{1\mathsf{T}} & -W_{21} Q_2^{2\mathsf{T}} & R_{J_1} & 0 & -W_2 & 0 \\ R_{J_2} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} + Q\mathcal{E}\dot{Q}, \tag{8.56}$$

and where $F_M := F + \dot{M}$ and $\hat{f} := Qf$, cf. [94]. Note that congruence transformations preserve the structure, i.e. $(\hat{\mathcal{E}}, \hat{\mathcal{A}})$ is again self-adjoint. Since R_{J_2} is invertible, $Q_1^2 v$ is readily determined. Also λ_2 is decoupled from the solution of the remaining variables so we can proceed with investigating the subsystem with the coefficients

$$\hat{\mathcal{E}}_{22} := \begin{bmatrix} 0 & -Q_2^2 M^\mathsf{T} Q_1^{1\mathsf{T}} & -Q_2^2 M^\mathsf{T} Q_1^{2\mathsf{T}} & 0 \\ Q_1^1 M Q_2^{2\mathsf{T}} & 0 & 0 & 0 \\ Q_1^2 M Q_2^{2\mathsf{T}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{and}$$

$$\hat{\mathcal{A}}_{22} := \begin{bmatrix} -Q_2^2 W_1 Q_2^{2\mathsf{T}} & Q_2^2 (F^\mathsf{T} + \dot{M}^\mathsf{T}) Q_1^{1\mathsf{T}} & Q_2^2 (F^\mathsf{T} + \dot{M}^\mathsf{T}) Q_1^{2\mathsf{T}} & -Q_2^2 W_{12} \\ Q_1^1 F Q_2^{2\mathsf{T}} & Q_1^1 G Q_1^{1\mathsf{T}} & Q_1^1 G Q_1^{2\mathsf{T}} & R_{J_1}^\mathsf{T} \\ Q_1^2 F Q_2^{2\mathsf{T}} & Q_1^2 G Q_1^{1\mathsf{T}} & Q_1^2 G Q_1^{2\mathsf{T}} & 0 \\ -W_{21} Q_2^{2\mathsf{T}} & R_{J_1} & 0 & -W_2 \end{bmatrix} + [Q\mathcal{E}\dot{Q}^\mathsf{T}]_{22}, \tag{8.57}$$

where $[Q\mathcal{E}\dot{Q}]_{22}$ is the corresponding submatrix of $Q\mathcal{E}\dot{Q}$. Since by Lemma 8.23 the matrix $Q_1^2 M Q_2^{2\mathsf{T}}$ and its transpose are invertible, a congruence transformation via $U = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & -Q_1^1 M Q_2^{2\mathsf{T}} (Q_1^2 M Q_2^{2\mathsf{T}})^{-1} & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$ can be applied to (8.57) to obtain the equivalent subsystem with coefficients as follows:

$$\hat{\hat{\mathcal{E}}}_{22} := \begin{bmatrix} 0 & 0 & -Q_2^2 M^\mathsf{T} Q_1^{2\mathsf{T}} & 0 \\ 0 & 0 & 0 & 0 \\ Q_1^2 M Q_2^{2\mathsf{T}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \text{and}$$

$$\hat{\hat{\mathcal{A}}}_{22} := \begin{bmatrix} * & * & * & * \\ * & \hat{G} & * & R_{J_1}^\mathsf{T} \\ * & * & * & * \\ * & R_{J_1} & * & -W_2 \end{bmatrix} + U[Q\mathcal{E}\dot{Q}^\mathsf{T}]_{22} \dot{U}^\mathsf{T}. \tag{8.58}$$

with some possibly nonzero entries $*$ and

$$\hat{G} := \begin{bmatrix} Q_1^1 & -Q_1^1 M Q_2^{2\mathsf{T}} (Q_1^2 M Q_2^{2\mathsf{T}})^{-1} Q_1^2 \end{bmatrix} G \begin{bmatrix} Q_1^{1\mathsf{T}} \\ -Q_1^{2\mathsf{T}} [Q_1^1 M Q_2^{2\mathsf{T}} (Q_1^2 M Q_2^{2\mathsf{T}})^{-1}]^\mathsf{T} \end{bmatrix}.$$

Since $U[Q\mathcal{E}\dot{Q}^\mathsf{T}]_{22}\dot{U}^\mathsf{T}$ has the same pattern of zero entries as $\mathcal{E}_{22}$, the part of $\hat{\hat{\mathcal{A}}}_{22}$, that lies in the left nullspace of $\hat{\hat{\mathcal{E}}}_{22}$, is given by $\begin{bmatrix} \hat{G} & R_{J_1}^\mathsf{T} \\ R_{J_1} & -W_2 \end{bmatrix}$. Considering that $\hat{G}$ and R_{J_1} are invertible, by symmetry and positive definiteness of $\hat{G}$ and W_2 we find that $\hat{\hat{\mathcal{E}}}_{22}$ is invertible. Thus, subsystem (8.57) is strangeness-free. Since the transformation by U does not affect the overall structure of system (8.56), one can read off the differentiation index of (8.56) as follows, cf. [94, Cor. 4.4]. By the last line $Q_2^1 v$ is defined in terms of the right hand side f. Due to the nonzero entries in the first column of $\hat{\mathcal{E}}$, the first derivative $\dot{f}$ appears in solution of the index-1 subsystem $(\hat{\mathcal{E}}_{22}, \hat{\mathcal{A}}_{22})$. Finally, because of nonzero entries in the first row of $\hat{\mathcal{E}}$ the second derivative $\ddot{f}$ enters the definition of the λ_2. Thus, (8.56) has differentiation index $\nu_d = 3$, since a general solution will depend on the second time derivative of the right-hand side Qf. As the differentiation index is invariant under equivalence transformations, cf. [90, Thm. 3.38], also (8.55) has $i_\nu = 3$. □

9. Numerical Algorithms and Application Example

In this section we provide algorithms for solving the finite-dimensional linear-quadratic optimal control system, cf. Section 8.3, using the results stated with Lemma 8.17. We illustrate the algorithms using the implicit Euler scheme for the time integration. In practice, one rather calls on schemes of higher order to integrate the state equations [50]. For the time integration of large-scale differential Riccati equations, *Rosenbrock* schemes are the method of choice [117].

9.1. **Linear-Quadratic Setup and Algorithms.** For illustration we consider an equation system as it comes from a semi-discretization and linearization of the Navier-Stokes Equation, cf. (8.1).

$$M\dot{v} + [A + N(t)]v - J^{\mathsf{T}}p = f + B_1u, \quad v(0) = v^0 \in \mathbb{R}^{n_v}, \tag{9.1a}$$

$$Jv = g, \tag{9.1b}$$

$$y = Cv. \tag{9.1c}$$

Here, A is the discrete Laplacian, N is the convection matrix and time dependent. The matrices B, and C represent the input and the output operators. We assume, that we have used a stable space discretization, so that equations (9.1a-b) fulfill Assumption 8.2. In particular the boundary conditions, are resolved in the right hand side.

In what follows, we will make use of the projector

$$\mathcal{P} := I - M^{-1}J_1^{\mathsf{T}}S^{-1}J_2$$

as it was defined in Theorem 8.6.

9.1.1. *Cost Functional.* Given a target output y^* with $y^* \in Y$. For all $t \in (0, T)$, we assume that $y^*(t) \in C\mathcal{P}\mathbb{R}^{n_v}$ and for formal considerations – in the implementation we will never use it – we define $v^*(t) = \mathcal{P}v^*(t) := C^-y^*(t)$. We consider cost functionals of type

$$\mathcal{J}(v, u) = \frac{1}{2}v_\Delta^{\mathsf{T}}(T)Vv_\Delta(T) + \frac{1}{2}\int_0^T v_\Delta^{\mathsf{T}}Wv_\Delta + \gamma u^{\mathsf{T}}M_Uu \,\mathrm{d}t, \tag{9.2}$$

with a scalar parameter γ and

$$V := \mathcal{P}^{\mathsf{T}}C^{\mathsf{T}}M_YC\mathcal{P}, \quad W := C^{\mathsf{T}}M_YC. \tag{9.3}$$

9.1.2. *The linear part – the Riccati DAE.* For the derivation we assume $\gamma = 1$. Then, cf. Theorem 8.19, an optimal input fulfills

$$u = M_U^{-1}B_1^{\mathsf{T}}[X_1Mv + w_1], \tag{9.4}$$

where X_1 solves

$$\begin{aligned} M^{\mathsf{T}}\dot{X}_1M - M^{\mathsf{T}}X_1[A + N(t)] - [A + N(t)]^{\mathsf{T}}X_1M + M^{\mathsf{T}}X_1\tilde{B}\tilde{B}^{\mathsf{T}}X_1M - \tilde{C}^{\mathsf{T}}\tilde{C}& \\ -M^{\mathsf{T}}X_2^{\mathsf{T}}J - J_2^{\mathsf{T}}X_2M = 0,& \\ M^{\mathsf{T}}X_1(T)M = -P^{\mathsf{T}}C^{\mathsf{T}}M_YCP,& \end{aligned} \tag{9.5a}$$

$$M^{\mathsf{T}}X_1J^{\mathsf{T}} = 0 \quad \text{and} \quad JX_1M = 0, \tag{9.5b}$$

with

$$\tilde{B} := B_1M_U^{-1/2} \quad \text{and} \quad \tilde{C} := M_Y^{1/2}C. \tag{9.6}$$

9.1.3. *The affine part.* Having computed the feedback matrices X_1 and X_2, we can compute the affine correction w_1 or the so called feedforward term as the solution of

$$-\tfrac{d}{dt}\begin{bmatrix} M^{\mathsf{T}} w_1 \\ 0 \end{bmatrix} - \begin{bmatrix} M^{\mathsf{T}} X_1 \tilde{B} \tilde{B}^{\mathsf{T}} - [A+N(t)]^{\mathsf{T}} & -J^{\mathsf{T}} \\ J & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \tilde{C}^{\mathsf{T}} \tilde{C} v^* + M^{\mathsf{T}}[X_1 f + X_2 g] \\ 0 \end{bmatrix}, \tag{9.7}$$

$$M^{\mathsf{T}} w_1(T) = P^{\mathsf{T}} \tilde{C}^{\mathsf{T}} \tilde{C} v^*(T),$$

cf. (8.27). Note that here $\tilde{f}_v = f$ – since there are no cross terms in the cost functional – and that by assumption and by definition we have $y^*(t) = Cv^*(t) = CPv^*(t)$.

9.1.4. *Time Integration and Newton Iteration for the Differential-algebraic Riccati Equation.* Given linear state equations and a quadratic cost functional, we need to integrate system (9.5). For the time discretization, we will apply an implicit time-stepping algorithm. This gives an algebraic Riccati equation at every time instance. Application of the standard Newton scheme, then leads to a Lyapunov equation to be solved in every Newton iteration. For the solution of the inner Lyapunov equation, we employ a *low-rank ADI iteration* [19, 127, 153]. The main feature of low-rank ADI is that a solution X is approximated by means of a factor $Z \in \mathbb{R}^{n_v, n_c}$, with $n_c \ll n_v$, such that $X \approx ZZ^{\mathsf{T}}$. This makes the computation of solution to the Riccati equations accessible for large-scale problems, since typically X is dense.

Since, the discretization and the linearization preserve the structure, also the Lyapunov equations appear as generalized or constrained Lyapunov equations. We will give a thorough derivation of the ADI algorithm in Section 9.2.

We illustrate the algorithm using implicit Euler. For other schemes see [20, 117]. Let τ be the time step length. By X^p, X^c, X^n we denote the numerically computed value of X_1 at the previous, current, next time instance $t-\tau$, t, $t+\tau$.

We start with $X^c := -M^{-T}\mathcal{P}^{\mathsf{T}}\tilde{C}^{\mathsf{T}}\tilde{C}\mathcal{P}M^{-1}$.

We approximate the current time derivative term by $\frac{1}{\tau}M^{\mathsf{T}}[X^c - X^p]M$, so that the new value X^p is obtained as the solution of

$$\begin{aligned} -\frac{1}{\tau} M^{\mathsf{T}} X^p M - M^{\mathsf{T}} X^p [A+N^p] - [A+N^p]^{\mathsf{T}} X^p M + \\ M^{\mathsf{T}} X^p \tilde{B}\tilde{B}^{\mathsf{T}} X^p M - M^{\mathsf{T}} {Y^c}^{\mathsf{T}} J - J^{\mathsf{T}} Y^c M = -\frac{1}{\tau} M^{\mathsf{T}} X^c M + \tilde{C}^{\mathsf{T}} \tilde{C} \end{aligned} \tag{9.8}$$

and

$$M^{\mathsf{T}} X^p J^{\mathsf{T}} = 0 \quad \text{and} \quad J X^p M = 0.$$

We rewrite (9.8) as

$$-M^{\mathsf{T}} X^p F^p - [F^p]^{\mathsf{T}} X^p M + \tau M^{\mathsf{T}} X^p \tilde{B}\tilde{B}^{\mathsf{T}} X^p M - L(Y^c) = -M^{\mathsf{T}} X^c M + \tau \tilde{C}^{\mathsf{T}} \tilde{C}, \tag{9.9}$$

with

$$F^p := \tfrac{1}{2} M + \tau[A + N^p] \quad \text{and} \quad L(Y) := \tau[M^{\mathsf{T}} Y^{\mathsf{T}} J + J^{\mathsf{T}} Y M].$$

To be in line with standard literature e.g. [20, Eqn. (8)] we make another change of variables and coefficients:

$$\boxed{X \leftarrow -X^p, \quad F \leftarrow -F^p, \quad B \leftarrow \sqrt{\tau}\tilde{B}, \quad W \leftarrow \left[M^{\mathsf{T}}\sqrt{-X^c}, \sqrt{\tau}\tilde{C}^{\mathsf{T}}\right]} \tag{9.10}$$

With this, having multiplied the equations by -1, Equation (9.9) reads

$$\begin{aligned} M^\mathsf{T} XF + F^\mathsf{T} XM - M^\mathsf{T} XBB^\mathsf{T} XM + L(Y) &= -WW^\mathsf{T} \\ JXM = 0, \quad M^\mathsf{T} XJ^\mathsf{T} &= 0. \end{aligned} \tag{9.11}$$

Applying a Newton scheme to (9.11), for the current Newton iterate X_N, we find the updated X_{N+1} by solving the following equations:

$$\begin{aligned} M^\mathsf{T} X_{N+1} F_N + [F_N]^\mathsf{T} X_{N+1} M + L(Y^c_{N+1}) &= -WW^\mathsf{T} - M^\mathsf{T} X_N BB^\mathsf{T} X_N M, \\ M^\mathsf{T} X_{N+1} J^\mathsf{T} = 0, \quad \text{and} \quad JX_{N+1}M &= 0, \end{aligned} \tag{9.12a}$$

where the coefficient F_N is defined as

$$F_N := F - [M^\mathsf{T} X_N BB^\mathsf{T}]^\mathsf{T} = F - BB^\mathsf{T} X_N M. \tag{9.13}$$

Thus, the solution of the implicit time update (9.8) by means of a Newton iteration, requires the solution of a sequence of constrained Lyapunov equations as given in (9.12).

The solution of these Lyapunov equations by means of an iteration for factors Z, such that $X \approx ZZ^H$, is described in Section 9.2. In particular, given $X_N = Z_N[Z_N]^H$, we can directly apply Algorithm 9.1 with

$$F \leftarrow F_N \quad \text{and} \quad W \leftarrow \begin{bmatrix} W & M^\mathsf{T} X_N B \end{bmatrix}. \tag{9.14}$$

to compute Z_{N+1}, approximating the solution $X_{N+1} \approx Z_{N+1}[Z_{N+1}]^H$ of (9.12). The relation to the actual quantities are given via the substitution rule (9.10).

9.1.5. *Implicit Euler for the Feedforward.* In the case of that the right hand sides and the target output is different from zero, the optimal control takes the form of a state feedback plus a feedforward term, cf. 9.4.

The feedforward variable w_1 is defined via Equation 9.7. We approximate $-\frac{d}{dt}(M^\mathsf{T} w_1)$ by $-M^\mathsf{T}\frac{w_1^c - w_1^p}{\tau}$. With this, given X^p, starting from

$$w_1^c = M^{-T}\mathcal{P}^\mathsf{T}\tilde{C}^\mathsf{T}\tilde{C}\mathcal{P}v^*(T) = M^{-T}\mathcal{P}^\mathsf{T} C^\mathsf{T} M_Y y^*(T),$$

we can advance w_1^c to w_1^p by solving

$$\begin{bmatrix} M^\mathsf{T} + \tau[A + N^p]^\mathsf{T} + \tau M^\mathsf{T} X^p \tilde{B}\tilde{B}^\mathsf{T} & -\tau J^\mathsf{T} \\ J & 0 \end{bmatrix} \begin{bmatrix} w_1^p \\ w_2^p \end{bmatrix} = \\ \begin{bmatrix} M^\mathsf{T} w_1^c + \tau\big[C^\mathsf{T} M_Y y^{*p} - M^\mathsf{T}[X^p \tilde{f}_v^p + Y^p g^p]\big] \\ 0 \end{bmatrix}. \tag{9.15}$$

Note that $X^p \approx -X_1^p$, cf. (9.10).

9.1.6. *The Closed Loop System.* Implementing the control law (9.4) in the controlled state equations (9.1) one obtains the closed loop system

$$\begin{aligned} M\dot{v} + [A + N(t) - B_1 M_U^{-1} B_1^\mathsf{T} X_1(t) M]v - J^\mathsf{T} p = f + B_1 M^{-1} B_1^\mathsf{T} w_1, \\ v(0) = v^0 \in \mathbb{R}^{n_v}, \end{aligned} \tag{9.16a}$$

$$Jv = g, \tag{9.16b}$$

$$y = Cv. \tag{9.16c}$$

Starting with $v^c = v^0$, we advance v^c and $y^c = Cv^c$ via the solution of

$$\begin{bmatrix} M + \tau[A + N^n] + \tau\tilde{B}\tilde{B}^\mathsf{T} X^n M & -\tau J^\mathsf{T} \\ J & 0 \end{bmatrix} \begin{bmatrix} v^n \\ p^n \end{bmatrix} = \begin{bmatrix} Mv^c + \tau f^n + \tau\tilde{B}\tilde{B}^\mathsf{T} w^n \\ g^n \end{bmatrix}$$

and

$$y^n = Cv^n.$$

Note that $X^n \approx -X_1^n$, cf. (9.10).

9.1.7. *Particular Implementation Issues.* We mention some distinct features of the implementation. The general point is that in the low-rank ADI iteration, one cannot access the full matrix $X = ZZ^H$ directly.

Boundary Conditions. In the numerical approximation, we resolve the Dirichlet boundaries as follows:

Let `A` be the Laplacian and that contains all degrees of freedom of `V`. Let `Ii`, `Ib` be the inner and the boundary nodes. Let `bvals` be the values at the boundary. We resolve the boundary values, by excluding the boundary nodes equations from the system and moving the known boundary variables to the right hand side. In the implementation this is realized like:

```
auxv = numpy.zeros((NV,1))
auxv[Ib,:] = bcvals
fbc = A[Ii,:][:,:]*auxv
Ac = A[Ii,:][:,Ii]
```

And the right hand side to the condensed system becomes:

```
fvbc = fv[Ii,:] - fbc
```

For the convection matrix N that changes with time, we conduct the same procedure at every time step.

9.1.8. *Solution of the Feedback System via the Sherman-Morrison Formula.* In general, one cannot use the feedback matrix $X = ZZ^{\mathsf{T}}$ it self, as this will be a dense matrix. In iterative solves, one can simply use the factorization. For direct solves, one can resort to the *Sherman-Morrison Formula.*

The above algorithms require the repeated solves of systems with the same or similar coefficient matrices that write like

$$\begin{bmatrix} F^p - \tau X_N^p \tilde{B}\tilde{B}^{\mathsf{T}} & \tau J^{\mathsf{T}} \\ J & 0 \end{bmatrix} = \begin{bmatrix} F^p & \tau J^{\mathsf{T}} \\ J & 0 \end{bmatrix} - \begin{bmatrix} \tau X_N^p \tilde{B} \\ 0 \end{bmatrix} \begin{bmatrix} \tilde{B}^{\mathsf{T}} & 0 \end{bmatrix} := \mathcal{F} - UV$$

cf. (9.12). Its inverse is given via

$$\begin{aligned} (\mathcal{F} - UV)^{-1} &= \mathcal{F}^{-1} + \mathcal{F}^{-1}U \left(I - V\mathcal{F}^{-1}U\right)^{-1} V\mathcal{F}^{-1} \\ &= \mathcal{F}^{-1}[I + U(I - V\mathcal{F}^{-1}U)^{-1}V\mathcal{F}^{-1}]. \end{aligned}$$

We assume that the inverse of $\mathcal{F}$ is available. To apply the inverse of $\mathcal{F} + UV$ to a vector Z, we compute the correction $\tilde{Z} = I + U(I - V\mathcal{F}^{-1}U)^{-1}\mathcal{F}^{-1}Z$ and then $\mathcal{F}^{-1}\tilde{Z}$.

9.1.9. *Compression of the columns of Z.* If at the current time instance t^c the number of columns of Z^c is ${n_z}^c$, then at the next time instance ${n_z}^n = N_{adi}(n_z + n_u + n_y)$, cf. the relations (9.10) and (9.14), where N_{adi} is the number of ADI iteration that were used to approximate the solution of the inner Lyapunov equations. Thus, e.g. for a fixed number of inner ADI iterations, the number of columns of Z grows exponentially with the time steps.

Calling on the assumption, that X^c can be approximated by low rank factors, we use a truncated singular value decomposition of `Z_itc` to compress it to a matrix `Z_itc_red` with `k` columns.

```
nny = Z_itc.shape[1]
U, s, V = numpy.linalg.svd(Z, full_matrices=False)

S = scipy.sparse.dia_matrix((s,0), (nny,nny)).tocsr()
Sred = S[:k, :][:, :k]

Z_itc_red = np.dot(U[:, :k], Sred)
```

Note that the right singular vectors can be left out of the definition of the reduced `Z_itc_red` since they do not affect the product of `Z_itc_red` with its transpose.

9.1.10. *Norm of the Newton updates.* To monitor the convergence of the Newton scheme for (9.11) we check the norm of the Newton updates via $\|Z_{N+1}Z_{N+1}^H - Z_N Z_N^H\|_F$.

For large scale problems, one cannot setup ZZ^H, but one can compute the Frobenius norm of the difference in the factored matrices via

$$\begin{aligned}\|Z_1Z_1^H - Z_2Z_2^H\|_F^2 &= \operatorname{tr}([Z_1Z_1^H - Z_2Z_2^H]^H[Z_1Z_1^H - Z_2Z_2^H]) \qquad (9.17)\\ &= \operatorname{tr}([Z_1^HZ_1]^2) - 2\operatorname{tr}(Z_1^HZ_2Z_2^HZ_1) + \operatorname{tr}([Z_2^HZ_2]^2)\end{aligned}$$

which can be evaluated with setting up matrices only of the size of the number of columns of Z_1 or Z_2. The above formula can be derived using the linearity of the trace and that $\operatorname{tr}(Z_1^HZ_2) = \operatorname{tr}(Z_2Z_1^H)$.

Nevertheless, the computation of the norm of the update is a bottleneck in terms of CPU time and memory requirements. For this reason, in the first Newton steps we approximate the norm of the Newton update, cf. (9.17), by difference of the application to a random vector v:

$$\|Z_1Z_1^Hv - Z_2Z_2^Hv\|_2^2$$

.

9.2. **Solution of Constrained Lyapunov Equations.** In this section, we investigate the iterative solution of constrained Lyapunov equations of type

$$F^\mathsf{T}XM + M^\mathsf{T}XF - J_2^\mathsf{T}YM - M^\mathsf{T}Y^\mathsf{T}J_2 = -WW^H, \qquad (9.18a)$$

$$J_1XM = 0, \quad \text{and} \quad M^\mathsf{T}XJ_1^\mathsf{T} = 0, \qquad (9.18b)$$

that are the outcome of the application of an implicit time-stepping and a Newton scheme in the course of the numerical integration the differential-algebraic Riccati equation.

The variant of Smith's method, see e.g. [111], proposed in this section for solving generalized Lyapunov equations of type (9.18) is designed to achieve consistency with the algebraic constraints for every iterate on the way to the actual solution of the differential-algebraic Riccati equation.

We want to point out that the algorithm derived here is identical with the single-shift ADI iteration developed in [64]. In view of model reduction via balanced truncation the authors of [64] consider a class of projected Lyapunov equations that define the observability and controllability Gramians for linear time-invariant semi-explicit index-2 systems with observation and control. For an efficient numerical solution in [64] the projected equations are interpreted as saddle-point systems. Here, the same saddle-point systems arise within the iterative solution of the Cayley- or Stein-transformed generalized Lyapunov equations. We will use the interpretation as projected Lyapunov equations to state the convergence of the numerical scheme.

The algorithm for the solution of the projected Lyapunov equations, has also been applied for the solution of projected algebraic Riccati equations for the stabilization of incompressible flows [14, 15].

The presented derivation assumes that all coefficients and iterants are real and only the shift parameter μ is complex. However, if one uses complex shifts, the iterants will become complex as well, see [17] for implementation variants with complex arithmetics. All derivations of the following section remain valid for complex valued coefficients with the transpose, e.g. A^{T}, replaced by $\overline{A}^{\mathsf{T}}$, where the overline denotes the complex-conjugate.

9.2.1. *Smith's Method and Low-Rank Factor Iteration.* With $\mathbf{M} := \begin{bmatrix} M & 0 \\ 0 & I \end{bmatrix}$ and for a complex parameter μ we rewrite equations (9.18) as

$$\begin{aligned}\mathbf{M}^{-\mathsf{T}} \begin{bmatrix} F^{\mathsf{T}} + \mu M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} X & 0 \\ Y & 0 \end{bmatrix} =& \mathbf{M}^{-\mathsf{T}} \begin{bmatrix} -WW^H & 0 \\ 0 & 0 \end{bmatrix} \mathbf{M}^{-1} \\ &- \begin{bmatrix} X & Y^{\mathsf{T}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} F - \mu M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix} \mathbf{M}^{-1}. \end{aligned} \tag{9.19}$$

Assuming that $F^{\mathsf{T}} + \mu M^{\mathsf{T}}$ and $J_1[F^{\mathsf{T}} + \mu M^{\mathsf{T}}]^{-1} J_2^{\mathsf{T}}$ are invertible and that X is symmetric we find by inverting and transposing that (9.19) is equivalent to

$$\begin{aligned}&\begin{bmatrix} X & Y^{\mathsf{T}} \\ 0 & 0 \end{bmatrix} = \\ &\mathbf{M}^{-\mathsf{T}} \Big(\begin{bmatrix} -WW^H & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} F^{\mathsf{T}} - \mu M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} X & 0 \\ Y & 0 \end{bmatrix} \mathbf{M} \Big) \begin{bmatrix} F + \mu M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix}^{-1}. \end{aligned} \tag{9.20}$$

Expression (9.20) plugged into the backward shifted equations

$$\begin{aligned}\mathbf{M}^{-\mathsf{T}} \begin{bmatrix} F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} X & 0 \\ Y & 0 \end{bmatrix} =& \mathbf{M}^{-\mathsf{T}} \begin{bmatrix} -WW^H & 0 \\ 0 & 0 \end{bmatrix} \mathbf{M}^{-1} \\ &- \begin{bmatrix} X & Y^{\mathsf{T}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} F - \bar{\mu} M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix} \mathbf{M}^{-1} \end{aligned} \tag{9.21}$$

gives after a premultiplication by $\begin{bmatrix} F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix}^{-1} \mathbf{M}^T$, of which we assume that it exists,

$$\begin{aligned}\begin{bmatrix} X & 0 \\ Y & 0 \end{bmatrix} = \begin{bmatrix} F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix}^{-1} \Bigg(& \begin{bmatrix} -WW^H & 0 \\ 0 & 0 \end{bmatrix} \mathbf{M}^{-1} - \\ - \Big(& \begin{bmatrix} -WW^H & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} F^{\mathsf{T}} - \mu M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} X & 0 \\ Y & 0 \end{bmatrix} \mathbf{M} \Big) \times \\ \times & \begin{bmatrix} F + \mu M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix}^{-1} \begin{bmatrix} F - \bar{\mu} M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix} \mathbf{M}^{-1} \Bigg). \end{aligned} \tag{9.22}$$

With the Schur complement $S_{\bar{\mu}} := J_1[F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}}]^{-1} J_2^{\mathsf{T}}$ and S_μ defined analogously we have the formulas for the inverse matrices

$$\begin{bmatrix} F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix}^{-1} = \tag{9.23a}$$

$$\begin{bmatrix} \big[I - [F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}}]^{-1} J_2^{\mathsf{T}} S_{\bar{\mu}}^{-1} J_1\big][F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}}]^{-1} & [F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}}]^{-1} J_2^{\mathsf{T}} S_{\bar{\mu}}^{-1} \\ -S_{\bar{\mu}}^{-1} J_1 [F^{\mathsf{T}} + \bar{\mu} M^{\mathsf{T}}]^{-1} & S_{\bar{\mu}}^{-1} \end{bmatrix}$$

and

$$\begin{bmatrix} F+\mu M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix}^{-1} = \tag{9.23b}$$
$$\begin{bmatrix} \big[I-[F+\mu M]^{-1}J_1^{\mathsf{T}}S_\mu^{-\mathsf{T}}J_2\big][F+\mu M]^{-1} & -[F+\mu M]^{-1}J_1^{\mathsf{T}}S_\mu^{-\mathsf{T}} \\ S_\mu^{-\mathsf{T}}J_2[F+\mu M]^{-1} & S_\mu^{-\mathsf{T}} \end{bmatrix}$$

Since

$$\mathcal{I}_\mu^l := \begin{bmatrix} F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} F^{\mathsf{T}}-\mu M^{\mathsf{T}} & -J_2^{\mathsf{T}} \\ J_1 & 0 \end{bmatrix} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix} \tag{9.24a}$$

and

$$\mathcal{I}_\mu^r := \begin{bmatrix} F+\mu M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix}^{-1} \begin{bmatrix} F-\bar{\mu}M & J_1^{\mathsf{T}} \\ -J_2 & 0 \end{bmatrix} = \begin{bmatrix} * & 0 \\ * & * \end{bmatrix} \tag{9.24b}$$

have a zero block in the upper-right position the matrix equation (9.22) yields one expression for X independent of Y, one equation for Y and two tautologies $0=0$.

We consider the expression for X and examine the term containing WW^H and the term containing X separately.

Using the formulas for the inverses given in (9.23a-b) we can write the contribution of WW^H to the equation for X in (9.22) as

$$-\big[I-[F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}J_2^{\mathsf{T}}S_{\bar{\mu}}^{-1}J_1\big][F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}WW^H\big[M^{-1}-[\mathcal{I}_\mu^r]_{11}M^{-1}\big], \tag{9.25}$$

where

$$[\mathcal{I}_\mu^r]_{11} := \big[I-[F+\mu M]^{-1}J_1^{\mathsf{T}}S_\mu^{-\mathsf{T}}J_2\big][F+\mu M]^{-1}[F-\bar{\mu}M]+[F+\mu M]^{-1}J_1^{\mathsf{T}}S_\mu^{-\mathsf{T}}J_2 \tag{9.26}$$

is the left-upper block of $\mathcal{I}_\mu^r$ as defined in (9.24a). With

$$[F+\mu M]^{-1}[F-\bar{\mu}M] = [F+\mu M]^{-1}[F+\mu M-\mu M-\bar{\mu}M] = I-2\,\mathrm{Re}(\mu)[F+\mu I]^{-1}M \tag{9.27}$$

we find that

$$I-[\mathcal{I}_\mu^r]_{11} = 2\,\mathrm{Re}(\mu)\big[I-[F+\mu M]^{-1}J_1^{\mathsf{T}}S_\mu^{-\mathsf{T}}J_2\big][F+\mu M]^{-1}M \tag{9.28}$$

and that (9.25) is the same as

$$-2\,\mathrm{Re}(\mu)[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}WW^H[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}^H, \tag{9.29}$$

where $[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}$ is the upper-left block of the matrix given in (9.23a) and the superscript $*$ denotes the transpose and complex conjugate.

With this, we can extract an implicit relation for X from (9.22):

$$X = [\mathcal{I}_\mu^l]_{11}X[\mathcal{I}_\mu^r]_{11}-2\,\mathrm{Re}(\mu)[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}WW^H[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}^H. \tag{9.30}$$

Through (9.23a) and (9.24a) we obtain an explicit expression for $[\mathcal{I}_\mu^l]_{11}$ and using (9.27) we can directly confirm that

$$\begin{aligned}
[\mathcal{I}_\mu^l]_{11} &= \big[I-[F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}J_2^{\mathsf{T}}S_{\bar{\mu}}^{-1}J_1\big][F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}[F^{\mathsf{T}}-\mu M^{\mathsf{T}}]^{-1}+ \\
&\qquad\qquad + [F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}J_2^{\mathsf{T}}S_{\bar{\mu}}^{-1}J_1 \\
&= \big[I-[F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}J_2^{\mathsf{T}}S_{\bar{\mu}}^{-1}J_1\big]\big[I-2\,\mathrm{Re}(\mu)[F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}M^{\mathsf{T}}\big]+ \\
&\qquad\qquad + [F^{\mathsf{T}}+\bar{\mu}M^{\mathsf{T}}]^{-1}J_2^{\mathsf{T}}S_{\bar{\mu}}^{-1}J_1 \\
&= I-2\,\mathrm{Re}(\mu)\big[I-[F^{\mathsf{T}}+\bar{\mu}M]^{-1}J_2^{\mathsf{T}}S_{\bar{\mu}}^{-1}J_1\big][F^{\mathsf{T}}+\bar{\mu}M]^{-1}M^{\mathsf{T}} \\
&= \Big[I-2\,\mathrm{Re}(\mu)M[F+\mu M]^{-1}\big[I-J_1^{\mathsf{T}}S_\mu^{-\mathsf{T}}J_2[F+\mu M]^{-1}\big]\Big]^H \\
&= \big[M[\mathcal{I}_\mu^r]_{11}M^{-1}\big]^H,
\end{aligned}$$

where the last equality follows from Equation (9.28). We write $[I_\mu]_{11} := M[I^r_\mu]_{11}M^{-1}$ so that (9.30) becomes

$$X = [\mathcal{I}_\mu]_{11} X [\mathcal{I}_\mu]_{11}^H - 2\,\mathrm{Re}(\mu)[\mathcal{F}_{\bar{\mu}}^{-1}]_{11} W W^H [\mathcal{F}_{\bar{\mu}}^{-1}]_{11}^H, \tag{9.31}$$

which is a Stein or discrete-time Lyapunov equation for X.

A formal solution to (9.31) is given via

$$X = -2\,\mathrm{Re}(\mu) \sum_{k=0}^{\infty} [\mathcal{I}_\mu]_{11}^k [\mathcal{F}_{\bar{\mu}}^{-1}]_{11} W W^H [\mathcal{F}_{\bar{\mu}}^{-1}]_{11}^H [\mathcal{I}_\mu]_{11}^{*k} \tag{9.32}$$

The series in (9.32), truncated after the n-th summand, can be written as

$$X_n = U_n U_n^H, \quad \text{where } U_n := \sqrt{-2\,\mathrm{Re}(\mu)}(\tilde{W}, [\mathcal{I}_\mu]_{11}\tilde{W}, \dots, [\mathcal{I}_\mu]_{11}^{n-1}\tilde{W}) \tag{9.33}$$

and $\tilde{W} := [\mathcal{F}_{\bar{\mu}}^{-1}]_{11} W$.

If the series in (9.32) converges, then $[\mathcal{I}_\mu]_{11}^n \to 0$ as $n \to \infty$ and X_n as defined in (9.33) can serve as an approximate solution to (9.18), cf. [127].

Instead of computing the matrix inverses for the definition of $[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}$ and $[\mathcal{I}_\mu]_{11}$ one rather computes the products $\tilde{W} = [\mathcal{F}_{\bar{\mu}}^{-1}]_{11}$ and $[\mathcal{I}_\mu]_{11}\tilde{W}$ directly as solutions of linear equation systems.

The definition of $[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}$ as the upper-left block of the matrix inverse defined in (9.23a) implicates the following lemma:

Lemma 9.1. *Consider* $[\mathcal{F}_{\bar{\mu}}^{-1}]_{11}$ *as in Equation* (9.29)*. Then, for a given matrix* W *of suitable size,* $Z := [\mathcal{F}_{\bar{\mu}}^{-1}]_{11} W$ *is the solution of*

$$\begin{bmatrix} F^\mathsf{T} + \bar{\mu} M^\mathsf{T} & -J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} Z \\ * \end{bmatrix} = \begin{bmatrix} W \\ 0 \end{bmatrix}$$

and in particular $J_1 \tilde{W} = J_1 Z = 0$.

For the products $[\mathcal{I}_\mu]_{11}^k \tilde{W}$ one can conclude from the definition of $\mathcal{I}_\mu$ in (9.24a):

Lemma 9.2. *If for* $\tilde{Z}$ *it holds* $J_1 \tilde{Z} = 0$*, then* $Z := [\mathcal{I}_\mu]_{11}\tilde{Z}$ *is given as the solution of*

$$\begin{bmatrix} F^\mathsf{T} + \bar{\mu} M^\mathsf{T} & -J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} Z \\ * \end{bmatrix} = \begin{bmatrix} [F^\mathsf{T} - \mu M^\mathsf{T}]\tilde{Z} \\ 0 \end{bmatrix} \tag{9.34}$$

and in particular $J_1 Z = 0$.

Proof. By definition of $[\mathcal{I}_\mu]_{11}$ one has

$$\begin{bmatrix} Z \\ * \end{bmatrix} = \mathcal{I}_\mu \begin{bmatrix} \tilde{Z} \\ 0 \end{bmatrix} = \begin{bmatrix} F^\mathsf{T} + \bar{\mu} M^\mathsf{T} & -J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} F^\mathsf{T} - \mu M^\mathsf{T} & -J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{Z} \\ 0 \end{bmatrix},$$

which is equivalent to

$$\begin{bmatrix} F^\mathsf{T} + \bar{\mu} M^\mathsf{T} & -J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} Z \\ * \end{bmatrix} = \begin{bmatrix} F^\mathsf{T} - \mu M^\mathsf{T} & -J_2^\mathsf{T} \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{Z} \\ 0 \end{bmatrix},$$

which coincides with (9.34) if $J_1 \tilde{Z} = 0$. □

Thus, we can formulate an algorithm without explicitly calling on the inverse matrices (9.23):

Remark 9.3. Algorithm 9.1 can be extended to varying shift parameters, see, e.g. [15]. The use of multivariate shifts can speed up convergence and, simultaneously, reduce memory consumption, in particular, for (sub)optimally chosen shift parameters.

Algorithm 9.1 : For a shift parameter μ, $\mathrm{Re}(\mu) < 0$, coefficient matrices F, J_1, J_2 and a right-hand side WW^H, compute an approximate solution to (9.18) via relation (9.33).

```
Step 1:
    $U \leftarrow Z, \quad \text{where } Z \text{ solves } \begin{bmatrix} F^\intercal + \bar{\mu} M^\intercal & -J_2^\intercal \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} Z \\ * \end{bmatrix} = \begin{bmatrix} W \\ 0 \end{bmatrix}.$
Step 2: (repeat until convergence)
    a.)
        $\tilde{Z} \leftarrow [F^\intercal - \mu M^\intercal] Z$
    b.)
        $U \leftarrow [U, Z], \quad \text{where } Z \text{ solves } \begin{bmatrix} F^\intercal + \bar{\mu} M^\intercal & -J_2^\intercal \\ J_1 & 0 \end{bmatrix} \begin{bmatrix} Z \\ * \end{bmatrix} = \begin{bmatrix} \tilde{Z} \\ 0 \end{bmatrix}$
Step 3:
    $U \leftarrow \sqrt{-2\,\mathrm{Re}(\mu)} U$
```

Remark 9.4. If $J_1 = J_2$, then Algorithm 9.1 coincides with Algorithm 5.6 in [64] and one can carry over the convergence results derived in [64]. For general $J_1 \neq J_2$ one cannot expect convergence of the approximations X_n as defined in (9.33). For possibly differing J_1 and J_2, as they occur in discretizations of flow equations, convergence has to be investigated separately.

9.3. **Distributed Control of a Driven Cavity.** As the example problem, we consider a two dimensional *driven cavity* with Reynolds number 100 on the unit square $\Omega = [0,1]^2$.

We consider the evolution of the flow for t within $t_0 = 0$ and $T = 1$, taking the steady state solution of the Stokes problem as the intial value.

Then we linearize the dynamics about the solution, so that the state equations are of the form (9.1), i.e. linear with time-varying coefficients.

The control setup is similar to [62] but with different input and output spaces.

For $n_u \in \mathbb{N}$, we set the input space $\mathcal{U} := \mathcal{C}(0,T;U \times U)$, where U is spanned by n_u linear hat functions equally distributed on the unit interval $[0,1]$.

We define the domain of control to be $\Omega_c = [0.4, 0.6] \times [0.2, 0.3]$, cf. Figure 3, and the input operator $B\colon \mathcal{U} \to \mathcal{C}(0,T;\mathcal{C}(\Omega;\mathbb{R}^2))$ via

$$B_1 u(t;x_1,x_2) = \begin{cases} \begin{bmatrix} u_1(t;\theta(x_1)) \\ u_2(t,\theta(x_1)) \end{bmatrix}, & \text{if } (x_1,x_2) \in \Omega_c, \\ 0, & \text{elsewhere,} \end{cases} \tag{9.35}$$

with the affine linear function θ_c mapping $[0.4, 0.6]$ onto $[0,1]$.

For a parameter n_y we define the output space $\mathcal{Y}$ similar to $\mathcal{U}$. As the domain of observation, we use $\Omega_o = [0.45, 0.55] \times [0.5, 0.7]$, cf. Figure 3, and for a $v \in \mathcal{C}(0,T;\mathcal{C}(\Omega;\mathbb{R}^2))$, we define the observation operator $C\colon v \to y \in \mathcal{Y}$ via

$$Cv(t)(\eta) = \begin{bmatrix} y_1(t;\eta) \\ y_1(t;\eta) \end{bmatrix} = \int_{0.45}^{0.55} \begin{bmatrix} v_1(t;x_1,\theta_o(\eta)) \\ v_2(t;x_1,\theta_o(\eta)) \end{bmatrix} \mathrm{d}x_1, \tag{9.36}$$

where θ_o is an affine linear mapping adjusting $[0,1]$ to $[0.5, 0.7]$.

In the presented example we have chosen $n_u = n_y = 4$, meaning that the x and y components of both input and output signal are described by 4 nodal values each.

The weighting matrices in the cost functional where defined via the mass matrices of $\mathcal{Y}$ and $\mathcal{U}$, see (9.3). The scalar parameter γ controlling the control effort was set to $\gamma = 1e - 7$.

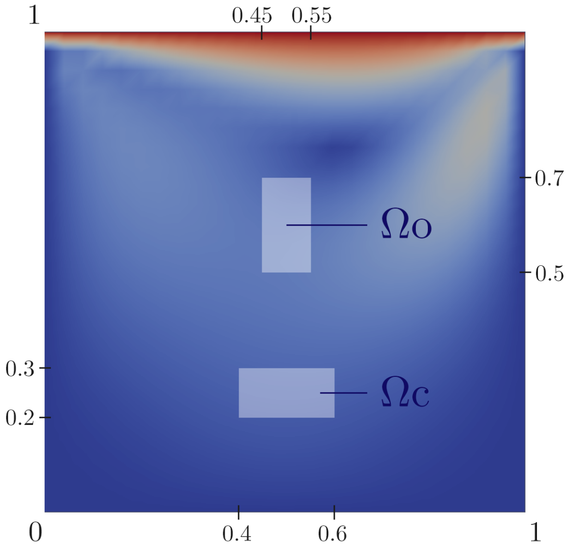

Figure 3. The computational domain of the driven cavity with the domains of control Ω_c and observation Ω_o.

For the spatial discretization we used a regular triangulation with 2500 cells of the unit square and *Taylor-Hood* finite elements [141] with 5202 velocity and 676 pressure nodes.

For the time discretization we used the *implicit Euler* scheme. The time interval $[0, 1]$ was discretized with 41 points. To account for the fast changes of the states at the beginning and at the end of the interval, the time points were clustered towards the margins.

The approximate solution to the differential-algebraic Riccati equation was obtained as described in Section 9.1. The inner Lyapunov equations were solved via a multishift ADI-iteration, cf. Remark 9.3, using real shifts $\mu \in \{-10, -5, -3, -1\}$ and a direct sparse solver. The inner ADI-iteration was stopped when the relative norm of the update of the factor dropped below 10^{-5}. The inner Newton iteration was continued until the relative norm of the update dropped below $4 \cdot 10^{-8}$. To compress the factors Z^c of every time iteration, we computed a truncated singular value decomposition cutting off all singular values smaller than $5 \cdot 10^{-5}$.

Once the feedback matrix X_1^c is available at every time instance, for a given target signal $y^* \in \mathcal{C}(0, 1; Y)$, one can compute the feedthrough w_1^c for all time instances and, subsequently, the corresponding optimal states.

We illustrate the applicability of the algorithm in Figure 4 for target states $y^* = \begin{bmatrix} y_x \\ y_y \end{bmatrix}$, with the spatial components y_x, $y_y \in \{0, 0.1, -0.1\}$ constant in space and time.

The plots in Figure 4 show that the derived algorithms are well applicable in distributed control of flows. When the control is active, the states immediately approach the targets. The plots also show a strong effect of the endpoint penalization in the final phase. In particular for large deviations of the uncontrolled state, the terminal forcing to the target leads to fast changes.

The implementation [61] was done in Python. For the spatial discretization we used the Python interface *dolfin* to *FeNiCS* [110]. The major computational work,

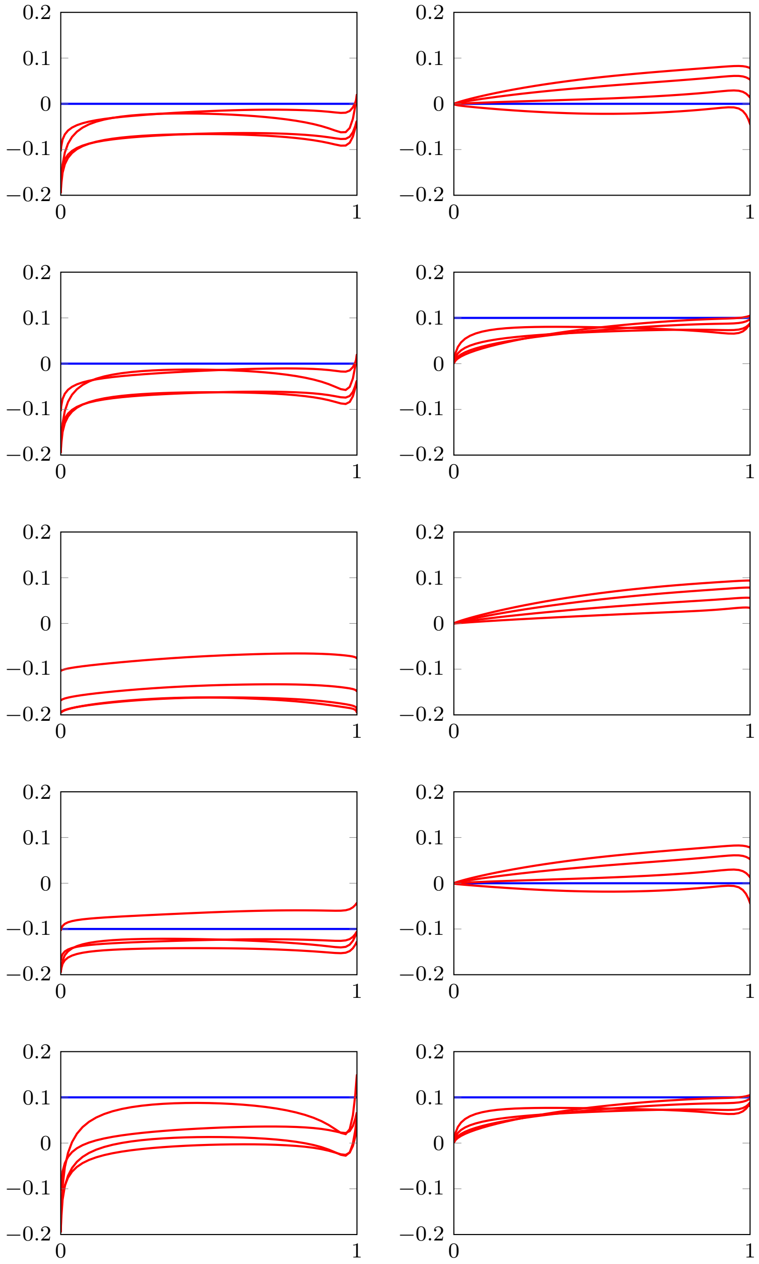

Figure 4. Time evolution of the measured signals $y \in C(0,T;Y)$ for different target signals y^*. The red lines show the x-component (left) and y-components (right) each given as 4 nodal values in Y. The blue lines denote the values of the different target signals. The middle row shows the signal for the uncontrolled system.

the precomputation of the factors of the feedback matrices for every time instance, took about 16000 seconds or 4.5 hours on a laptop with 8 GB RAM and an *Intel Core i3-3110M* processor with 2.4 GHz.

10. Discussion and Outlook

In the chapters on the infinite-dimensional differential-algebraic equations, we have derived a decoupling that separates differential and algebraic equations and proved that the taken assumptions hold for reasonable cases of a weak formulation of the Navier-Stokes Equations. The provided decoupling is compliant with the same approach for finite-dimensional approximations. With the decoupling at hand, we have read off consistency and smoothness conditions, i.e. necessary conditions for the existence of solutions. We have shown, that under common stability assumptions on the semi-discretization and under uniform monotonicity assumption on the nonlinearity in the equation, the finite-dimensional approximations converge weakly to a solution of the continuous problem. The convergence holds both for the native and for the decoupled formulations. This, in particular, gave an answer to the question on what is the Pressure Poisson Equation in infinite dimensions.

Since the derived decoupling happens in space, the extension of the theory to second order systems, that may be quantified as index-3 systems, is straight forward. As an example consider the dynamic elasticity problem

$$\ddot{v}(t) - A(\dot{v}(t), v(t)) - J'\lambda(t) = f(t) \quad \text{in } (W^{1,2}(\Omega))', \text{ a.e. in } (0,T), \tag{10.1}$$

$$Jv(t) = g(t) \quad \text{in } W^{1/2,2}(\partial\Omega), \text{ a.e. in } (0,T), \tag{10.2}$$

as it was considered in [6]. The solution v is considered in $W^{1,2}(\Omega)$ and the operator $J\colon W^{1,2}(\Omega) \to W^{1/2,2}(\partial\Omega)$ is associated with the *trace* operator. It is well known, cf. [6], that J fulfills an *LBB-condition* and, thus, admits a splitting of the solution space. However, the extension to $\bar{J}$ defined on $L^2(\Omega)$, i.e. the definition of the trace for functions in $L^2(\Omega)$, has been investigated only for domains with a C^∞ boundary [107, Ch. 9]. A topic of future work is to derive the splitting for the trace operator. This will also enable us to apply the setting to boundary control setups, if the control is included via a multiplier.

In the chapters on optimal control in the infinite-dimensional setup, we have shown that the decoupling for the state equations is also applicable for the adjoint equations. Carrying over the convergence results of the state equations to the adjoint, we have shown existence of solutions to the formal adjoint equation. This is of particular importance, since the formal adjoint can not be used in the statement of necessary optimality conditions, if it does not have a solution. For particular choices of the linearization point in the definition of the adjoint equation, we have laid out that the formal optimality system for the semi-discrete equations coincide with a semi-discretization of the formal optimality system. However, for the convergence of the adjoint state of the semi-discretization to the adjoint state of a continuous solutions, the provided results require strong convergence of the approximations of the state variables. So, the question on the mutual convergence of the semi discrete states and it adjoints is not answered. Also, the convergence of the input that is optimal for the semi-discrete equations to an input that is optimal for the continuous problem, as it was proven in the linear evolution case in [106], has not been established yet.

The chapter on linearization schemes for the solution of the nonlinear optimality system has provided two variants of a Newton scheme. As a corollary of the results on convergence of the Newton iterates, we have stated the existence of a unique solution to a class of linear-quadratic optimal control problems with abstract differential-algebraic constraints and with endpoint penalization. The taken way to state convergence of the Newton scheme via uniform positivity in the region of interest comes with a certain robustness against perturbations. Nevertheless, an immediate extension towards applicability of the results for numerical approximations will base on the considerations of inexact Newton schemes [75]. In view of incorporating also constraints on the control, one will have to examine nonsmooth Newton methods, cf. [31] for the Navier-Stokes setup.

Eventually, every numerical approximation will carry out a discretization and a linearization. For this reason, we have particularly focussed on the finite-dimensional linear-quadratic optimal control problem. The results from the abstract settings immediately apply here. Moreover, the explicit availability of the operators allow for explicit representations of the decoupling projectors. Using the particular structure, we have proven existence of solutions to formal optimality conditions by means of a differential algebraic Riccati ansatz.

We have used the linear-quadratic case, to also state necessary smoothness conditions on the data and inputs that come with the differential-algebraic structure and that are influenced by the choice of the cost functional.

As a general result for semi-linear, semi-explicit state equations, we have found that the action of the control in the algebraic constraint requires differentiability of the input and also constraints the value of the input at time $t = 0$, see Remark 8.7 and Corollary 8.9.

Adjusting the cost functional, may well influence the index of the equations, and, thus, the necessary smoothness constraints. In Remark 8.16, we have given sufficient conditions for *tractability index* $i_\mu = 2$ that, basically, exclude the algebraic variable p from the cost functional.

Also we have discussed, how one can formally reformulate the cost functional, so that p is included also in the index-2 case. This reformulation, however, may be infeasible, as it needs the explicit computation of projections. In particular we have shown, that a direct inclusion of p in the cost functional leads to an optimality system of index 3. For problems of higher index, the generally applicable reduction to strangeness-free formulations [27], are probably the way to go. However, it may be worth investigating particularly structured optimal control problems, as it was done here for an index-2 case, in view of efficient numerical algorithms.

Another crucial point is the incorporation of constraints for the control as well as lower regularity. This has been investigated in [134] for a large class of optimal control problems of type (8.13) subject to semi-explicit DAEs of type

$$M\dot{v} - f(t, v, p, u) = 0, \quad v(0) = v^0, \tag{10.3a}$$

$$g(t, v, p, u) = 0, \tag{10.3b}$$

and in particular the finite-dimensional cases that are considered here. The formulation of a maximum principle in [134] bases on an equivalent index 1 representation of the state equations that can be formally obtained for the semi-explicit case. In the index 2 case it is given via

$$M\dot{v} - f(t, v, p, u) = 0, \quad v(0) = v^0, \tag{10.4a}$$

$$G(t, v, p, u) = 0, \tag{10.4b}$$

where the function $G(t, v, p, u) := \dot{g} + gM^{-1}f$ allows for an implicit function representation of $p = F(v, u)$. The maximum principle in [134] states that a solution (v, p, u) of the associated optimal control problem with a cost-functional as in (8.13) and pointwise constraints on u is a maximizer for a specifically chosen Hamilton function, i.e. (v, p, u) solves

$$\max_{(p,u)\in\mathcal{D}(t,v)} \mathcal{H}(t, p, u; v, \lambda) = \max_{(p,u)\in\mathcal{D}(t,v)} \lambda_1^T f(t, v, p, u) - \mathcal{K}(t, v, p, u), \tag{10.5}$$

where

$$\mathcal{D}(t, v) := \big\{(p, u) : \ u \text{ is admissible and } G(t, v, p, u) = 0\big\},$$

and where λ_1 solves the adjoint equation of the reduced system (10.4). In order to compare it to our result, we assume for the moment that u is unconstrained and that the first variation, cf. [149], of f and $\mathcal{K}$ with respect to u exists. With the implicit function theorem we write $p = F(v,u)$ and the constrained maximization problem in (10.5) becomes unconstrained:

$$\max_u \mathcal{H}(t, F(v,u), u; v, \lambda) = \max_u \lambda_1^T \tilde{f}(t,v,u) - \tilde{\mathcal{K}}(t,v,u),$$

with the reduced functions $\tilde{f}$ and $\tilde{\mathcal{K}}$. Now a candidate solution may be obtained by setting the first variation to zero, i.e. to find an u such that

$$\lambda_1^T \tilde{f}_u - \tilde{\mathcal{K}}_u^T = 0. \tag{10.6}$$

Thus, Equation (10.6) replaces the specification of the maximum in (10.5) and together with the state equations (10.4) and its adjoint equations it gives the Euler-Lagrange equations, cf. (8.12). This means that under the assumptions that were necessary for the derivation of our results, namely that the cost functional does not depend on p, the maximum principle may also provide a system for the computation of an optimal input, however, on the base of an index-1 formulation.

The numerical example of Section 9 has shown applicability of the derived Riccati decoupling in optimal control of flows. Beside the statement in the original variables and coefficients, its main advantage is the availability of low-rank ADI iterations. This avoids the high requirement of memory of other common approaches to boundary value problems of type (8.14) as there are finite differences, collocation, and shooting methods cf. [8]. In terms of speed and memory consumption, there is space for improvement of the presented implementation in terms of computing optimized shifts and residual evaluations, adapting the recent results of [18] to the generalized setup. However, especially for large time horizons and unsteady targets, the memory requirement for storing the factors at every time instance will inevitably grow. A remedy might be the application of checkpointing schemes, see e.g. [154], that store only snapshots from which needed values are computed on demand.

In view of real-world applications, there are two remaining major issues to resolve. Firstly, the uniqueness of the input to state response is tied to the existence of unique solutions to the Navier-Stokes equations which, in the interesting three dimensional case, is guaranteed only for very particular setups. Secondly, the assumption of bounded input operators in Problem 6.1 and also in the finite dimensional equations is only met *distributed control* which has but a few applications [71]. In practice, however, flows are acted upon typically via the boundary [85, 86]. To make the presented theory applicable to practical applications, one needs to adopt recent theoretical and numerical results [15, 129] that show the successful modelling of boundary feedback control via terms of distributed control.

Acknowledgment

This work was supervised by Prof. Volker Mehrmann and supported by a scholarship of “Studienstiftung des deutschen Volkes”.

Index

References

[1] F. Abergel and R. Temam. On some control problems in fluid mechanics. *Theor. Comput. Fluid Dyn.*, 1(6):303–325, 1990.

[2] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank. *Matrix Riccati equations in control and systems theory.* Birkhäuser, Basel, Switzerland, 2003.

[3] Y. A. Abramovich and C. D. Aliprantis. *An Invitation to Operator Theory.* Graduate Studies in Mathematics, V. 50. American Math. Soc., 2002.

[4] W. Alt. The Lagrange-Newton method for infinite-dimensional optimization problems. *Numer. Funct. Anal. Optim.*, 11(3-4):201–224, 1990.

[5] W. Alt and K. Malanowski. The Lagrange-Newton method for nonlinear optimal control problems. *Comput. Optim. Appl.*, 2(1):77–100, 1993.

[6] R. Altmann. Index reduction for operator differential-algebraic equations in elastodynamics. *Z. Angew. Math. Mech.*, 93(9):648–664, 2013.

[7] R. Altmann and J. Heiland. Finite element decomposition and minimal extension for flow equations. Preprint 2013–11, Technische Universität Berlin, Germany, 2013.

[8] U. M. Ascher, R. M. Mattheij, and R. D. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* SIAM, Philadelphia, 1995.

[9] A. Backes. Optimale Steuerung der linearen DAE im Fall Index 2. Preprint 2003-04, Institut für Mathematik, Humboldt-Universität zu Berlin, 2003.

[10] A. Backes. *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen.* PhD thesis, Institut für Mathematik, Humboldt-Universität zu Berlin, Germany, 2006.

[11] K. Balla, G. A. Kurina, and R. März. Index criteria for differential algebraic equations arising from linear-quadratic optimal control problems. *J. Dyn. Control Syst.*, 12(3):289–311, 2006.

[12] K. Balla and V. H. Linh. Adjoint pairs of differential-algebraic equations and Hamiltonian systems. *Appl. Numer. Math.*, 53(2-4):131–148, 2005.

[13] K. Balla and R. März. A unified approach to linear differential algebraic equations and their adjoints. *Z. Anal. Anwendungen*, 21(3):783–802, 2002.

[14] E. Bänsch and P. Benner. Stabilization of incompressible flow problems by Riccati-based feedback. Technical report, Universität Magdeburg, Germany, 2010.

[15] E. Bänsch, P. Benner, J. Saak, and H. K. Weichelt. Riccati-based boundary feedback stabilization of incompressible Navier-Stokes flow. Preprint SPP1253-154, DFG-SPP1253, 2013.

[16] V. Barbu. *Analysis and control of nonlinear infinite dimensional systems.* Academic Press, Boston, 1993.

[17] P. Benner, P. Kürschner, and J. Saak. Efficient Handling of Complex Shift Parameters in the Low-Rank Cholesky Factor ADI Method. *Numer. Algorithms*, 62(2):225–251, 2013.

[18] P. Benner, P. Kürschner, and J. Saak. Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations. Preprint MPIMD/13-18, Max Planck Institute Magdeburg, 2013.

[19] P. Benner, J.-R. Li, and T. Penzl. Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems. *Numer. Linear Algebra Appl.*, 15(9):755–777, 2008.

[20] P. Benner and H. Mena. BDF methods for large-scale differential Riccati equations. In B. Moor, B. Motmans, J. Willems, P. Dooren, and V. Blondel, editors, *Proc. of Mathematical Theory of Network and Systems, MTNS 2004*, 2004.

[21] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter. *Representation and control of infinite-dimensional systems. Vol. I.* Birkhäuser, 1992.

[22] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter. *Representation and control of infinite-dimensional systems. Vol. II.* Birkhäuser, 1993.

[23] H. Bounit and A. Idrissi. Time-varying regular bilinear systems. *SIAM J. Control Optim.*, 47(3):1097–1126, 2008.

[24] H. Braden. The equations $A^TX \pm X^TA = B$. *SIAM J. Matrix Anal. Appl.*, 20(2):295–302, 1998.

[25] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations.* Springer, New York, NY, 2011.

[26] S. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18(4):1101–1115, 1987.

[27] S. Campbell, P. Kunkel, and V. Mehrmann. *Regularization of Linear and Nonlinear Descriptor Systems*, chapter 2, pages 17–36. SIAM, Philadelphia, PA, 2012.

[28] L. Cesari. *Optimization - Theory and Applications. Problems with Ordinary Differential Equations.* Springer, New York, NY, 1983.

[29] T. F. Chan. An approximate Newton method for coupled nonlinear systems. *SIAM J. Numer. Anal.*, 22:904–913, 1985.

[30] P. Ciarlet jun. Augmented formulations for solving Maxwell equations. *Comput. Methods Appl. Mech. Eng.*, 194(2-5):559–586, 2005.

[31] J. C. De Los Reyes and K. Kunisch. A semi-smooth Newton method for control constrained boundary optimal control of the Navier-Stokes equations. *Nonlinear Anal.*, 62(7):1289–1316, 2005.

[32] M. do R. de Pinho and R. Vinter. Necessary conditions for optimal control problems involving nonlinear differential algebraic equations. *J. Math. Anal. Appl.*, 212(2):493–516, 1997.

[33] K. Deckelnick and M. Hinze. Semidiscretization and error estimates for distributed control of the instationary Navier-Stokes equations. *Numer. Math.*, 97(2):297–320, 2004.

[34] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms.* Springer Series in Computational Mathematics 35. Berlin: Springer, 2004.

[35] L. Dieci. Numerical integration of the differential Riccati equation and some related issues. *SIAM J. Numer. Anal.*, 29(3):781–815, 1992.

[36] L. Dieci. On the decoupling of dichotomic linear Hamiltonians. Considerations on integrating symmetric differential Riccati equations. *J. Comp. Appl. Math.*, 45(1–2):47 – 63, 1993.

[37] L. Dieci, M. R. Osborne, and R. D. Russell. A Riccati transformation method for solving linear BVPs. I: Theoretical aspects. *SIAM J. Numer. Anal.*, 25(5):pp. 1055–1073, 1988.

[38] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics.* Oxford University Press, Oxford, UK, 2005.

[39] E. Emmrich. *Analysis von Zeitdiskretisierungen des inkompressiblen Navier-Stokes-Problems.* Cuvillier Verlag, Göttingen, Germany, 2001.

[40] E. Emmrich. *Gewöhnliche und Operator-Differentialgleichungen.* Vieweg, Wiesbaden, Germany, 2004.

[41] E. Emmrich and V. Mehrmann. Operator differential-algebraic equations arising in fluid dynamics. *Comp. Methods Appl. Math.*, 13(4):443–470, 2013.

[42] D. Estévez Schwarz. *Consistent Initialization for Index-2 Differential-algebraic Equations and its Application to Circuit Simulation.* PhD thesis,

Institut für Mathematik, Humboldt-Universität zu Berlin, Germany, 2000.

[43] H. O. Fattorini. *Infinite dimensional linear control systems. The time optimal and norm optimal problems.* Elsevier, Amsterdam, Netherlands, 2005.

[44] O. Forster. *Analysis 2. Differentialrechnung im $\mathbb{R}^n$, gewöhnliche Differentialgleichungen.* Vieweg, Wiesbaden, Germany, 1999.

[45] H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, volume 38 of *Mathematische Lehrbücher und Monographien.* Akademie-Verlag, Berlin, Germany, 1974. (Nonlinear Operator Equations and Operator Differential Equations).

[46] A. Galántai. *Projectors and Projection Methods.* Kluwer Academic Publishers, Boston, MA, 2004.

[47] M. Gerdts. Optimal control and real-time optimization of mechanical multibody systems. *Z. Angew. Math. Mech.*, 83(10):705–719, 2003.

[48] M. Gerdts. Local minimum principle for optimal control problems subject to differential-algebraic equations of index two. *J. Optim. Theory Appl.*, 130(3):443–462, 2006.

[49] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms.* Springer, Berlin, Germany, 1986.

[50] P. M. Gresho and R. L. Sani. *Incompressible Flow and the Finite Element Method. Vol. 2: Isothermal Laminar Flow.* Wiley, Chichester, UK, 2000.

[51] E. Griepentrog and R. März. *Differential-algebraic Equations and Their Numerical Treatment.* Teubner, Leipzig, Germany, 1986.

[52] A. Griewank. The local convergence of broyden-like methods on lipschitzian problems in hilbert spaces. *SIAM J. Numer. Anal.*, 24(3):pp. 684–705, 1987.

[53] C. Großmann and H.-G. Roos. *Numerische Behandlung partieller Differentialgleichungen.* Teubner, Wiesbaden, Germany, 2005.

[54] R. Guberovic, C. Schwab, and R. Stevenson. Space-time variational saddle point formulations of Stokes and Navier-Stokes equations. *ESAIM: Math. Model. Numer. Anal.*, eFirst, 12 2013.

[55] M. Gunzburger and S. Manservisi. The velocity tracking problem for Navier-Stokes flows with bounded distributed controls. *SIAM J. Control Optim.*, 37(6):1913–1945, 1999.

[56] M. D. Gunzburger. *Perspectives in Flow Control and Optimization*, volume 5. SIAM, Philadelphia, PA, 2003.

[57] W. Hackbusch. *Elliptic Differential Equations: Theory and Numerical Treatment.* Springer, Berlin, Germany, 1992.

[58] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-algebraic Systems by Runge-Kutta Methods.* Springer, Berlin, Germany, 1989.

[59] J. Heiland. Differential-algebraic Riccati decoupling for linear-quadratic optimal control problems for semi-explicit index-2 DAEs. Preprint 2012–06, TU Berlin, Institut für Mathematik, 2012.

[60] J. Heiland. Example for an L2-FEM-projection of a solenoidal functions. `http://nbviewer.ipython.org/url/highlando.github.io/misc-scripts/projOfDivFree.ipynb`, 2013.

[61] J. Heiland. optconpy - a Python module for the solution of DAE-Riccati equations via Newton-ADI and application example, v0.0. `https://github.com/highlando/optconpy`, 2013.

[62] J. Heiland and V. Mehrmann. Distributed control of linearized Navier-Stokes equations via discretized input/output maps. *Z. Angew. Math. Mech.*, 92(4):257–274, 2012.

[63] J. Heiland, V. Mehrmann, and M. Schmidt. A new discretization framework for input/output maps and its application to flow control. In R. King, editor,

Active Flow Control. Papers contributed to the Conference "Active Flow Control II 2010", Berlin, Germany, May 26 to 28, 2010, pages 375–372. Springer, Berlin, 2010.

[64] M. Heinkenschloss, D. C. Sorensen, and K. Sun. Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations. *SIAM J. Sci. Comput.*, 30(2):1038–1063, 2008.

[65] M. R. Hestenes. *Calculus of Variations and Optimal Control Theory.* Huntington, New York, NY, 1980.

[66] J. S. Hesthaven and T. Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*, volume 54. Springer, New York, NY, 2008.

[67] H. Heuser. *Funktionalanalysis. Theorie und Anwendung.* Teubner, Wiesbaden, Germany, 2006.

[68] J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier-Stokes problem. I. regularity of solutions and second-order error estimates for spatial discretization. *SIAM J. Numer. Anal.*, 19(2):275–311, 1982.

[69] M. Hintermüller and M. Hinze. A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier–Stokes system subject to control constraints. *SIAM J. Optim.*, 16(4):1177–1200, 2006.

[70] M. Hinze. Optimal and instantaneous control of the instationary navier-stokes equations. Habilitationsschrift, Institut für Mathematik, Technische Universität Berlin, 2000.

[71] M. Hinze. Control of weakly conductive fluids by near wall Lorentz forces. SFB609- Preprint 19-2004, Technische Universität Dresden, 2004.

[72] M. Hinze and K. Kunisch. Second order methods for optimal control of time-dependent fluid flow. *SIAM J. Control Optim.*, 40(3):925–946, 2001.

[73] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints.* Springer, Dordrecht, Netherlands, 2009.

[74] J. Hoffman. Adaptive finite element methods for the unsteady Maxwell's equations. preprint 2000-001, Chalmers University of Technology, Göteborg Sweden, 2000.

[75] A. Hohmann. *Inexact Gauss Newton methods for parameter dependent nonlinear problems.* PhD thesis, Freie Universität Berlin, Germany, 1994.

[76] W. Hoyer and J. W. Schmidt. Newton-type decomposition methods for equations arising in network analysis. *Z. Angew. Math. Mech.*, 64:397–405, 1984.

[77] J. A. Infante and E. Zuazua. Boundary observability for the space semi-discretizations of the 1-D wave equation. *M2AN Math. Model. Numer. Anal.*, 33(2):407–438, 1999.

[78] K. Ito and K. Kunisch. Augmented Lagrangian-SQP-methods in Hilbert spaces and application to control in the coefficients problems. *SIAM J. Optim.*, 6(1):96–125, 1996.

[79] K. Ito and K. Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications.* SIAM, Philadelphia, PA, 2008.

[80] S. Kakutani. Some characterizations of Euclidean space. *Jap. J. Math.*, 16:93–97, 1939.

[81] L. W. Kantorowitsch. О методе Ньютона для функционалных уравнений. *Dokl. Akad. Nauk. SSSR*, 59(7):1237–1240, 1948.

[82] L. W. Kantorowitsch and G. P. Akilow. *Funktionalanalysis in normierten Räumen.* Akademie-Verlag, Berlin, Germany, 1978.

[83] T. Kato. *Perturbation theory for linear operators.* Springer, New York, NY, 1966.

[84] J. Y. Kim, N. R. Aluru, and D. A. Tortorelli. Improved multi-level newton solvers for fully coupled multi-physics problems. *Internat. J. Numer. Methods Engrg.*, 58(3):463–480, 2003.

[85] R. King (editor). *Active flow control. Papers contributed to the conference 'Active flow control 2006', Berlin, Germany, September 27–29, 2006.* Springer, Berlin, Germany, 2007.

[86] R. King (editor). *Active Flow Control II: Papers Contributed to the Conference 'Active Flow Control II 2010', Berlin, Germany, May 26–28, 2010.* Notes on Numerical Fluid Mechanics and Multidisciplinary Design. Springer, Berlin, Germany, 2010.

[87] M. A. Krasnosel'skii. Топологические методы в теории нелинейных интегральных уравнений. National Publisher of technical-theoretical literature, Moscow, SSSR, 1956.

[88] P. Kunkel and V. Mehrmann. The linear quadratic optimal control problem for linear descriptor systems with variable coefficients. *Math. Control Signals Syst.*, 10(3):247–264, 1997.

[89] P. Kunkel and V. Mehrmann. Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems. *Math. Control Signals Syst.*, 14(3):233–256, 2001.

[90] P. Kunkel and V. Mehrmann. *Differential-algebraic Equations. Analysis and Numerical solution.* European Mathematical Society Publishing House, Zürich, Switzerland, 2006.

[91] P. Kunkel and V. Mehrmann. Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index. *Math. Control Signals Syst.*, 20(3):227–269, 2008.

[92] P. Kunkel and V. Mehrmann. Formal adjoints of linear DAE operators and their role in optimal control. *Electron. J. Linear Algebra*, 22:672–693, 2011.

[93] P. Kunkel and V. Mehrmann. Optimal control for linear descriptor systems with variable coefficients. In P. Van Dooren, S. P. Bhattacharyya, R. H. Chan, V. Olshevsky, and A. Routray, editors, *Numerical Linear Algebra in Signals, Systems and Control*, pages 313–339. Springer, 2011.

[94] P. Kunkel, V. Mehrmann, and L. Scholz. Self-adjoint differential-algebraic equations. *Math. Control Signals Systems*, pages 1–30, 2013.

[95] F.-S. Kupfer. An infinite-dimensional convergence theory for reduced SQP methods in Hilbert space. *SIAM J. Optim.*, 6(1):126–163, 1996.

[96] A. J. Kurdila and M. Zabarankin. *Convex Functional Analysis.* Birkhäuser, Boston, MA, 2005.

[97] G. A. Kurina. Invertibility of an operator appearing in the control theory for linear systems. *Math. Notes*, 70:206–212, 2001.

[98] G. A. Kurina and R. März. On linear-quadratic optimal control problems for time-varying descriptor systems. *SIAM J. Control Optim.*, 42(6):2062–2077, 2004.

[99] G. A. Kurina and R. März. Feedback solutions of optimal control problems with DAE constraints. *SIAM J. Control Optim.*, 46(4):1277–1298, 2007.

[100] O. A. Ladyzhenskaya. *The Mathematical Theory of Viscous Incompressible Flow.* Gordon and Breach Science Publishers, New York, NY, 1969.

[101] R. Lamour, R. März, and C. Tischendorf. PDAEs and further mixed systems as abstract differential algebraic systems. Preprint 01-11, Humboldt-Universität zu Berlin, Berlin, Germany, 2001.

[102] R. Lamour, R. März, and C. Tischendorf. *Differential-algebraic equations: a projector based analysis.* Differential-Algebraic Equations Forum. Springer, Heidelberg, Germany, 2013.

[103] W. Layton. *Introduction to the Numerical Analysis of Incompressible Viscous Flows.* SIAM, Philadelphia, PA, 2008.

[104] R. L. Lee and N. K. Madsen. A mixed finite element formulation for Maxwell's equations in the time domain. *J. Comput. Phys.*, 88(2):284–304, 1990.

[105] L. León and E. Zuazua. Boundary controllability of the finite-difference space semi-discretizations of the beam equation. *ESAIM Control Optim. Calc. Var.*, 8:827–862, 2002.

[106] J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations.* Springer, New York, NY, 1971.

[107] J.-L. Lions and E. Magenes. *Non-Homogeneous Boundary Value Problems and Applications. Vol. I.* Springer, Berlin, Germany, 1972.

[108] J.-L. Lions and E. Zuazua. Exact boundary controllability of Galerkin's approximations of Navier-Stokes equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 26(4):605–621, 1998.

[109] L. A. Ljusternik. Об условных экстремумах функционалов. *Math. Sb.*, 41:390–401, 1934.

[110] A. Logg, K. B. Ølgaard, M. E. Rognes, and G. N. Wells. FFC: the FEniCS form compiler. In *Automated Solution of Differential Equations by the Finite Element Method*, pages 227–238. Springer-Verlag, Berlin, Germany, 2012.

[111] A. Lu and E. Wachspress. Solution of Lyapunov equations by alternating direction implicit iteration. *Comput. Math. Appl.*, 21(9):43 – 58, 1991.

[112] D. Luenberger. *Optimization by Vector Space Methods.* John Wiley and Sons, New York, NY, 1969.

[113] R. März. Adjoint equations of differential-algebraic systems and optimal control problems. *Proceedings of Belarussian National Academy of Sciences, Institute of Mathematics*, 7:88–97, 2001.

[114] R. März. Differential algebraic systems anew. *Appl. Numer. Math.*, 42(1-3):315–335, 2002.

[115] M. Matthes. *Numerical Analysis of Nonlinear Partial Differential-Algebraic Equations: A Coupled and an Abstract Systems Approach.* PhD thesis, Universität zu Köln, 2012.

[116] V. Mehrmann. Index concepts for differential-algebraic equations. Technical Report 3–2012, Institut für Mathematik, Technische Universität Berlin, 2012.

[117] H. Mena and P. Benner. Rosenbrock methods for solving differential Riccati equations. *IEEE Trans. Automat. Control,*, 2013, to appear.

[118] J. Menck. An approximate newton-like coupling of subsystems. *Z. Angew. Math. Mech.*, 82(2):101–114, 2002.

[119] P. Monk. A comparison of three mixed methods for the time-dependent Maxwell's equations. *SIAM J. Sci. Statist. Comput.*, 13(5):1097–1122, 1992.

[120] P. Monk. *Finite element methods for Maxwell's equations.* Oxford University Press, Oxford, UK, 2003.

[121] F. J. Murray. On complementary manifolds and projections in spaces L_p and l_p. *Trans. Amer. Math. Soc.*, 41(1):138–152, 1937.

[122] J. Nečas. *Direct Methods in the Theory of Elliptic Equations.* Springer, Berlin, Germany, 2012.

[123] M. Opmeer, T. Reis, and W. Wollner. Finite-rank ADI iteration for operator Lyapunov equations. *SIAM J. Control Optim.*, 51(5):4084–4117, 2013.

[124] S. J. Orfanidis. *Electromagnetic waves and antennas.* `www.ece.rutgers.edu/~orfanidi/ewa`, 2002.

[125] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* SIAM, Philadelphia, PA, 2000.

[126] J. M. Papakonstantinou. *Historical development of the BFGS secant method and its characterization properties.* PhD thesis, Rice University, Texas, 2009.

[127] T. Penzl. *Numerische Lösung großer Lyapunov-Gleichungen.* PhD thesis, Technische Universität Chemnitz, 1998.

[128] J. W. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory. A Behavioral Approach.* Springer, New York, NY, 1997.

[129] J.-P. Raymond. Local boundary feedback stabilization of the Navier-Stokes equations. In *Control Systems: Theory, Numerics and Applications, Rome, 30 March – 1 April 2005*, Proceedings of Science. SISSA, `http://pos.sissa.it`, 2005.

[130] W. T. Reid. *Riccati Differential Equations.* Academic Press, New York, NY, 1972.

[131] T. Reis. *Systems Theoretic Aspects of PDAEs and Applications to Electrical Circuits.* Shaker, Aachen, Germany, 2006.

[132] W. C. Rheinboldt. An adaptive continuation process for solving systems of nonlinear equations. Math. Models and numer. Methods, Banach Cent. Publ. 3, 129-142, 1978.

[133] T. Roubíček. *Nonlinear Partial Differential Equations with Applications.* Birkhäuser, Basel, Switzerland, 2005.

[134] T. Roubiček and M. Valášek. Optimal control of causal differential-algebraic systems. *J. Math. Anal. Appl.*, 269(2):616–641, 2002.

[135] V. Runde. A new and simple proof of Schauder's theorem. *arXiv:1010.1298*, 2010.

[136] M. Sal Moslehian. A Survey on the Complemented Subspace Problem. *arXiv:math/0501048*, 2005.

[137] A. Schiela and A. Günther. An interior point algorithm with inexact step computation in function space for state constrained optimal control. *Numer. Math.*, 119:373–407, 2011.

[138] M. Schmidt. *Systematic Discretization of Input/Output Maps and other Contributions to the Control of Distributed Parameter Systems.* PhD thesis, Technische Universität Berlin, 2007.

[139] E. D. Sontag. *Mathematical Control Theory. Deterministic Finite Dimensional Systems.* Springer, New York, NY, 1998.

[140] L. Tartar. *An Introduction to Navier-Stokes Equation and Oceanography.* Springer, New York, NY, 2006.

[141] 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.

[142] R. Temam. *Numerical Analysis.* Kluwer, Dordrecht, Netherlands, 1973.

[143] R. Temam. *Navier-Stokes Equations. Theory and Numerical Analysis.* North-Holland, Amsterdam, Netherlands, 1977.

[144] A. R. Terrel, L. R. Scott, M. G. Knepley, R. C. Kirby, and G. N. Wells. Finite elements for incompressible fluids. In A. Logg, K.-A. Mardal, and G. Wells, editors, *Automated Solution of Differential Equations by the Finite Element Method*, pages 385–397. Springer, Berlin, Germany, 2012.

[145] C. Tischendorf. Coupled systems of differential algebraic and partial differential equations in circuit and device simulation. modeling and numerical analysis. Habilitationsschrift, Institut für Mathematik, Humboldt-Universität zu Berlin, 2004.

[146] F. Tröltzsch. On convergence of semidiscrete Ritz-Galerkin schemes applied to the boundary control of parabolic equations with nonlinear boundary condition. *Z. Angew. Math. Mech.*, 72(7):291–301, 1992.

[147] F. Tröltzsch. Semidiscrete Ritz-Galerkin approximation of nonlinear parabolic boundary control problems—strong convergence of optimal controls. *Appl. Math. Optim.*, 29(3):309–329, 1994.

[148] F. Tröltzsch. On the Lagrange–Newton–SQP method for the optimal control of semilinear parabolic equations. *SIAM J. Control Optim.*, 38(1):294–312, 1999.

[149] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen.* Vieweg+Teubner, Wiesbaden, Germany, 2009.

[150] S. Turek. *Efficient Solvers for Incompressible Flow Problems. An Algorithmic and Computational Approach.* Springer, Berlin, Germany, 1999.

[151] S. Volkwein and M. Weiser. Affine invariant convergence analysis for inexact augmented Lagrangian-SQP methods. *SIAM J. Control Optim.*, 41(3):875–899, 2002.

[152] D. Wachsmuth and T. Roubíček. Optimal control of planar flow of incompressible non-Newtonian fluids. *Z. Anal. Anwend*, 29(3):351–376, 2010.

[153] E. L. Wachspress. Iterative solution of the Lyapunov matrix equation. *Appl. Math. Lett.*, 1(1):87 – 90, 1988.

[154] A. Walther and A. Griewank. Advantages of binomial checkpointing for memory-reduced adjoint calculations. In *Numerical mathematics and advanced applications*, pages 834–843. Springer, 2004.

[155] J. Weickert. *Applications of the theory of differential-algebraic equations to partial differential equations of fluid dynamics.* PhD thesis, Fakultät für Mathematik, Technische Universität Chemnitz, 1997.

[156] R. Whitley. An elementary proof of the Eberlein-Šmulian theorem. *Math. Ann.*, 172:116–118, 1967.

[157] A. Yeckel, L. Lun, and J. J. Derby. An approximate block Newton method for coupled iterations of nonlinear solvers: Theory and conjugate heat transfer applications. *J. Comput. Phys.*, 228(23):8566 – 8588, 2009.

[158] I. Yousept. Optimal control of quasilinear $\mathbf{H}(\mathbf{curl})$-elliptic partial differential equations in magnetostatic field problems. *SIAM J. Control Optim.*, 51(5):3624–3651, 2013.

[159] E. Zeidler. *Nonlinear functional analysis and its applications. III: Variational methods and optimization.* Springer, Berlin, Germany, 1985.

[160] E. Zeidler. *Nonlinear functional analysis and its applications. I: Fixed-point theorems.* Springer, Berlin, Germany, 1986.

[161] E. Zeidler. *Nonlinear functional analysis and its applications. II/A: Linear monotone operators.* Springer, Berlin, Germany, 1990.

[162] E. Zeidler. *Nonlinear functional analysis and its applications. II/B: Nonlinear monotone operators.* Springer, Berlin, Germany, 1990.

Institut für Mathematik MA4-5, Technische Universität Berlin, Strasse des 17. Juni 136, D–10623 Berlin, Germany

E-mail address: `heiland@math.tu-berlin.de`