Technische Universität München

Fakultät für Mathematik

Lehrstuhl für Optimalsteuerung

Efficient Higher Order Discontinuous Galerkin Time Discretizations for Parabolic Optimal Control Problems

Andreas Springer

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. Oliver Junge
Prüfer der Dissertation:	
	1. Univ.–Prof. Dr. Boris Vexler
	2. Univ.–Prof. Dr. Thomas Apel Universität der Bundeswehr München
	3. Jun.–Prof. Dr. Thomas Richter Ruprecht-Karls-Universität Heidelberg

Die Dissertation wurde am 8. Januar 2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 4. Mai 2015 angenommen.

Contents

1. Introduction

This work demonstrates the viability of constructing efficient discontinuous Galerkin (dG) time discretizations of higher order for optimal control problems governed by systems of parabolic partial differential equations (PDEs). In particular, we address the issue of efficient solution of the resulting large implicit time stepping equations and explore two ways of achieving rapid convergence of higher order dG methods in spite of the typically low regularity of solutions when additional inequality constraints on the control or state variables are present.

Using Galerkin-type methods for optimal control problems is desirable since in this setting discretization and optimization commute (see, e. g., [12]), i. e., discretizing the optimality system for the continuous problem yields the same result as deriving an optimality system for the already discretized problem. Compared to, e. g., the continuous Petrov-Galerkin (cG) methods, dG time stepping has the additional advantage that the adjoint time discretization is of the same form as the primal one since test and trial space are identical. This allows for unified treatment of state and adjoint equations in the analysis and also in the implementation.

Due to the inherent stiffness of parabolic PDEs, the strong A-stability of the dG method is an advantage compared to cG discretizations which are only A-stable. Since the trial space is discontinuous, dynamic meshes, i. e., spatial discretizations that vary over time, can be incorporated in a natural way into the variational formulation. This allows to resolve local phenomena that travel over time, for example travelling reaction fronts as seen in the problem given in Section 2.3.3.

When employing higher order versions of implicit single step methods like dG, a major practical issue, which of course also arises in the context of optimal control problems, is the efficient solution of the resulting large equation systems for each time step. We address this problem by an iterative process involving an approximation of the time stepping equation that can be decoupled. Our method is related to the *"Single Newton Process"* proposed by Perez-Rodriguez and co-workers (see, e. g., [87]) and represents an improvement over our previous approach presented in [91].

To achieve high order of convergence, higher order discretizations for PDEs typically require sufficient smoothness of the solution and in turn of the problem data. Whereas for solving PDEs without optimal control we can check the smoothness requirements on the data a priori, for optimal control problems, the regularity of the control, which enters the state equation as a datum, is determined by the problem itself. In particular

when inequality constraints on the control are present, its regularity can be limited at the boundaries between active and inactive sets.

We pursue two approaches to address this issue in the context of higher order dG methods. On the one hand, we construct an almost third order convergent method based on piecewise linear dG time discretization for a model problem exploiting additional regularity of the adjoint state and superconvergence properties of the dG solution. On the other hand, we propose to circumvent the issue of local lack of regularity by using an *hp*-adaptive discretization that resolves the parts of the time domain where the solution is less smooth by low order approximations with small step size while retaining fast convergence with few higher order time steps where the solution is smooth.

Subsequently, we give a brief overview of the organization of this thesis: In Chapter 2 we give a precise formulation for the class of optimal control problems considered in this work. Additionally we introduce a set of benchmark problems that we use for testing numerical algorithms in the later chapters.

Chapter 3 reviews optimality conditions and representation formulas for first and second derivatives. Subsequently we present the semismooth Newton algorithm we use for the solution of the optimal control problem. To improve convergence properties of the algorithm, a heuristic strategy similar to Steihaug cg is discussed. All considerations in this chapter apply to the undiscretized optimal control problem, ensuring mesh-independence of the resulting numerical methods.

The discretization of the optimal control problem is introduced in Chapter 4. We first discretize the state equation in time using discontinuous Galerkin schemes. In preparation for Chapter 7, where *hp* adaptivity with respect to time is investigated, we allow the order of the discretization to vary over time. Subsequently, the resulting semidiscrete state equation is discretized with respect to the spatial domain using standard finite elements. Finally we discuss two possibilities for treating the control variable, which is still left undiscretized at this point, and remark for each of them on the implications on the realization of the optimization algorithm.

In Chapter 5, we discuss the solution of the discrete time stepping equations for the state and auxiliary equations. Parts of this chapter have been previously published in Richter, Springer, and Vexler [91]. We start by analyzing the structure of Newton's method when applied to the time stepping equation. The resulting Newton update equation consists of a block structured linear system, where the number of blocks grows with the order of the discontinuous Galerkin scheme. We establish a connection to the Runge-Kutta Radau schemes and show that a decoupling of the blocks of the update equation introduces complex coefficients. To circumvent this issue, we propose an approximation of the Newton matrix that allows for decoupling of the blocks over the reals. Subsequently, an analysis of the convergence properties is carried out, first for linear and then for nonlinear problems with particular emphasis on semilinear equations. Afterwards we discuss the practical realization of the resulting time stepping schemes. The chapter is concluded with numerical tests assessing the performance of the decoupling approach.

Chapter 6 presents an a priori analysis of a time discretization scheme for a linear quadratic model problem with control constraints which—in spite of low temporal regularity of the control—converges with almost third order with respect to the fineness of the time discretization. To achieve this high order of convergence based on a first order discontinuous Galerkin discretization, we use a combination of several techniques. On the one hand, the control is treated with the variational approach due to Hinze [55], on the other hand we prove superconvergence properties of the adjoint solution. By a post-processing step using this higher order reconstruction of the adjoint solution, we obtain an improved control solution which we show to converge with almost third order. To complete the discussion, we analyze the spatial discretization error and back up the theoretical results with numerical evidence. The results presented in this chapter are already published in Springer and Vexler [104].

A completely different approach to resolving non-smooth features of optimal control problems while using higher order dG time discretizations, is taken in Chapter 7. We present an adaptive algorithm that contains provisions for assessing the local temporal smoothness of the solution and changing the order of the time discretization accordingly, thereby realizing *hp* adaptivity with respect to the time discretization. In the first section of the chapter, a posteriori error estimators based on the dual weighted residual approach as described in Becker and Rannacher [13] are derived, which also account for errors due to control constraints and numerical quadrature. Subsequently, we discuss a smoothness indicator, followed by a description of the complete adaptive algorithm. Numerical studies on three test problems with different non-smooth features complete the presentation.

Chapter 8 summarizes the presented results and discusses some ideas for possible extensions.

2. Problem formulation

In this chapter, after fixing basic notation, we give a precise definition of the problem class we consider. Subsequently we remark briefly on existence of solutions and discuss some simple examples that we will employ in the later chapters for numerical tests.

2.1. Basic notations

As usual we will denote L^p spaces and Sobolev spaces over some domain $\Omega \subseteq \mathbb{R}^d$, $d = 1, 2, 3$ by $L^p(\Omega)$ and $W^{k,p}(\Omega)$ with $1 \leq p \leq \infty$ and $k \in \mathbb{R}$. The corresponding Bochner spaces with values in a Banach space H are written as $L^p(\Omega, H)$ and $W^{k,p}(\Omega, H)$. For the case $p = 2$, we use the abbreviation $H^k(\Omega, H) = W^{k,2}(\Omega, H)$. The space of bounded linear operators mapping a Banach space U into another Banach space V is denoted by $B(U, V)$ and by $B(U)$ if $U = V$. For elements of some finite dimensional space $\mathbb{R}^n$, the p norms are written as $\|\cdot\|_p$.

To introduce the functional analytic setting for the state equation, we consider two Hilbert spaces V and H such that the embedding $V \hookrightarrow H$ is a continuous dense injection. If we identify H with its dual, then the spaces $V \hookrightarrow H \hookrightarrow V^*$ form a Gelfand triple, i. e., the second embedding is a dense injection as well.

With the Gelfand triple and a given finite time interval $I = (0, T)$ we define the usual space $X := W(I)$ by

$$X = \left\{ v \in L^2(I, V) \;\middle|\; \partial_t v \in L^2(I, V^*) \right\}. \tag{2.1}$$

This construction is commonly employed for analyzing parabolic PDEs, for details see, e. g., [27] or [117]. We note that the space X embeds continuously into $C(\bar{I}, H)$.

For inner products on a Hilbert space V we employ the notation $(\cdot, \cdot)_V$, for the space H we omit the subscript and write $(\cdot, \cdot)$. The corresponding norms are denoted by $\|\cdot\|_V$ and $\|\cdot\|$ respectively. Inner product and norm on the space $L^2(\hat{I}, H)$ for some interval $\hat{I}$ read $(\cdot, \cdot)_{\hat{I}}$ and $\|\cdot\|_{\hat{I}}$ respectively. For the duality pairing on a space V we write $\langle \cdot, \cdot \rangle_{V^* \times V}$.

2.2. Problem statement

Subsequently we state an abstract framework that all problems considered here can be cast into. In general, an optimal control problem consists of a cost functional J

depending on a state variable u and a control variable q that is to be minimized, equality constraints that couple q and u, typically in the form of a differential equation, and possibly further constraints. In our case we consider a parabolic PDE that couples the state to the control and optionally additional restrictions on the control variable q.

Here, the Hilbert space Q for the control is assumed to be of the form $Q = L^2(\Omega_Q)$ for some suitable underlying set Ω_Q. Note that the case of finite dimensional control is covered by this definition since we can choose a finite underlying set Ω_Q. The considered cost functional $J\colon Q \times X \to \mathbb{R}$ takes the form

$$J(q,u) = J_1(u) + J_2(u(T)) + \frac{\alpha}{2}\|q\|_Q^2,$$

with the continous, two times Fréchet-differentiable functionals $J_1\colon L^2(I,V) \to \mathbb{R}$ and $J_2\colon H \to \mathbb{R}$ and a quadratic control cost term weighted by a parameter $\alpha \geq 0$. When considering inverse problems, this term results from Tikhonov regularization.

As in the previous section, let V and H together with V^* form a Gelfand triple and let $A\colon I \times Q \times V \to V^*$ be a (possibly nonlinear) elliptic differential operator which is uniformly elliptic in the third argument. We pose the parabolic state equation in an abstract form as: Find $u \in X$ such that

$$\begin{aligned} \partial_t u(t) + A(t,q,u(t)) &= 0 \quad \text{for almost all } t \in I, \\ u(0) &= u_0(q). \end{aligned}$$

The initial datum has the form $u_0\colon Q \to H$. Note that since the differential operator is allowed to be non-linear, we can incorporate a non-homogeneous right hand side into the definition of A. For stating the weak form of this equation, we introduce the semilinear form $a\colon I \times Q \times V \times V \to \mathbb{R}$ given by

$$a(t,q,u)(\varphi) = \langle A(t,q,u), \varphi\rangle_{V^*\times V}.$$

For semilinear forms we adopt the convention that the form may be non-linear with respect to all arguments in the first parenthesis, whereas it is linear with respect to the arguments given in the second parenthesis. For a weak formulation in space and time, the obvious test space making all occuring terms well-defined is $L^2(I,V)$, resulting in

$$\begin{aligned} \int_I \langle \partial_t u, \varphi\rangle_{V^*\times V} + a(q,u)(\varphi)\,\mathrm{d}t = 0 \quad &\text{for any } \varphi \in L^2(I,V), \\ u(0) = u_0(q). \end{aligned} \tag{2.2}$$

Note that for simplicity of notation we do not state dependencies on t explicitly. Since the space X is dense in the space of test functions $L^2(I,V)$, we can restrict ourselves to consider only test functions from X. This has the benefit that, since $u \in X$ and X is embedded into $C(\bar{I},H)$, we can couple the initial condition to the weak formulation resulting in

$$\int_0^T \langle \partial_t u, \varphi\rangle_{V^*\times V} + a(q,u)(\varphi)\,\mathrm{d}t + (u(0),\varphi(0)) = (u_0(q),\varphi(0)) \quad \text{for any } \varphi \in X. \tag{2.3}$$

As it will turn out, having the initial values coupled to the weak formulation is beneficial when deriving optimality conditions.

Subsequently, we consider only $H = L^2(\Omega)$ where $\Omega \subseteq \mathbb{R}^d$, $d \in \{1, 2, 3\}$ is assumed to be a bounded Lipschitz domain. The space V is chosen to match the differential operator and to take possible Dirichlet boundary conditions into account.

As an additional constraint on the control we require it to be contained in the set of admissible controls Q_{ad} given by

$$Q_{\text{ad}} = \left\{ q \in Q \;\middle|\; q^a \leq q \leq q^b \text{ almost everywhere on } \Omega_Q \right\}, \tag{2.4}$$

where $q^a, q^b \colon \Omega_Q \to \mathbb{R} \cup \{\pm\infty\}$ are measurable functions with $q^a \leq q^b$ almost everywhere, i. e., we allow for pointwise box-constraints on the control.

Putting it all together, our optimization problem reads

$$\text{Minimize } J(q, u) \text{ subject to } \begin{cases} (q, u) \in Q \times X \text{ satisfying (2.3)}, \\ q \in Q_{\text{ad}}. \end{cases} \tag{2.5}$$

Whether this problem admits an optimal solution $(\bar{q}, \bar{u})$ and whether this solution is unique depends on the structure of the state cost term J_1, the spatial differential operator A, and possibly also on whether the control cost term is present (i. e., $\alpha > 0$) and on the properties of Q_{ad}. Existence proofs for solutions for a number of subsets of the considered problem class can be found for example in the textbooks by Tröltzsch [107] and Lions [69].

2.3. Example problems

In this section we introduce a few example problems that we use later on to test our numerical algorithms. The first example is the standard linear quadratic optimal control problem for the heat equation with time-dependent parameter control. Next we discuss a semilinear test problem with terminal observation. Our last problem models control of a combustion process by cooling and heating the boundary of the domain.

2.3.1. Linear heat equation with time parameter control

We consider the linear quadratic problem given by: Minimize the cost functional

$$J(q, u) = \frac{1}{2} \int_0^T \|u(t) - u_d(t)\|^2_{L^2(\Omega)} \, \mathrm{d}t + \frac{\alpha}{2} \int_0^T \sum_{i=0}^{d_Q} q_i(t)^2 \, \mathrm{d}t \tag{2.6a}$$

For the set of time-dependent control parameters we assume $q \in L^2(I, \mathbb{R}^{d_Q})$ for some $d_Q \in \mathbb{N}$, and the state variable u is subject to the linear heat equation

$$\begin{aligned} \partial_t u - \Delta u &= f + G^q q && \text{in } I \times \Omega \\ u &= 0 && \text{on } I \times \partial\Omega \\ u(0) &= u_0 && \text{in } \Omega. \end{aligned} \tag{2.6b}$$

The space V is chosen as $V = H_0^1(\Omega)$. For the control we have pointwise inequality constraints

$$q \in Q_{\text{ad}} = \left\{ q \;\middle|\; q^a \le q(t) \le q^b \text{ for almost all } t \in I \right\} \tag{2.6c}$$

The scalar value α is assumed to be positive, the bounds $q^a, q^b \in (\mathbb{R} \cup \{\pm\infty\})^{d_Q}$ are given fixed vectors with $q^a < q^b$ component-wise, and the linear operator $G^q \colon \mathbb{R}^{d_Q} \to H$ is given by $G^q q = \sum_{j=0}^{d_Q} q_j g_j$ with given functions $g_j \in V$. We extend G^q to time-dependent functions by setting $(G^q q)(t) := G^q(q(t))$. For the data we assume $u_0 \in V$ and $u_d, f \in L^2(I, H)$.

The existence of a unique solution for this problem is shown, e. g., by Tröltzsch [107]. To embed this problem into our abstract setting, we set

$$\begin{aligned} J_1(u) &= \frac{1}{2} \int_0^T \|u(t) - u_d(t)\|_{L^2(\Omega)}^2 \, \mathrm{d}t, \\ J_2(u(T)) &= 0, \\ \text{and} \quad a(t, q, u)(\varphi) &= (\nabla u, \nabla \varphi)_I - (f + G^q q, \varphi)_I, \end{aligned}$$

and note that $Q = L^2(I, \mathbb{R}^{d_Q})$ is isomorphic to $L^2(I^{d_Q})$.

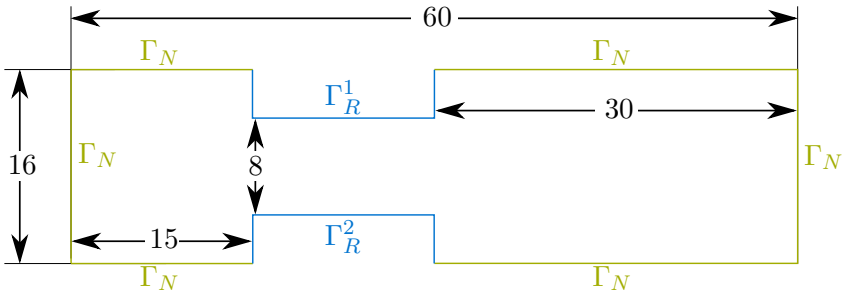
2.3.2. Semilinear problem with terminal observation

Also for this problem, the control consists of time-dependent parameters entering a source term. Correspondingly, the control space is set as $Q = L^2(I, \mathbb{R}^{d_Q})$ for some $d_Q \in \mathbb{N}$. The cost functional is given by

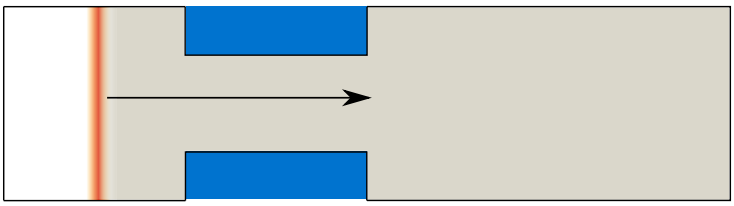
$$J(q, u) = \frac{1}{2} \|u(T) - u_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \int_0^T \sum_{i=0}^{d_Q} q_i(t)^2 \, \mathrm{d}t \tag{2.7a}$$

with some desired terminal state $u_d \in L^2(I, H)$ subject to the semilinear equation

$$\begin{aligned} \partial_t u - \Delta u + u^3 &= f + G^q q && \text{in } I \times \Omega, \\ u &= 0 && \text{on } I \times \partial\Omega, \\ u(0) &= u_0 && \text{in } \Omega, \end{aligned} \tag{2.7b}$$



(a) Domain Ω with boundary designations



(b) Initial configuration. Grey indicates fluid; the cooled rods are displayed in blue

Figure 2.1.: Domain Ω and initial configuration for the combustion example

and the control constraints

$$q \in Q_{\text{ad}} = \left\{ q \;\middle|\; q^a \leq q(t) \leq q^b \text{ for almost all } t \in I \right\} \tag{2.7c}$$

with $q^a, q^b \in (\mathbb{R})^{d_Q}$ The control-to-right-hand-side operator $G^q \colon \mathbb{R}^{d_Q} \to H$ is assumed to take the form $G^q q = \sum_{j=0}^{d_Q} q_j g_j$ with given functions $g_j \in L^\infty(\Omega)$. For the right hand side and the initial datum we require $f \in L^\infty(I \times \Omega)$ and $u_0 \in L^\infty(\Omega) \cap V$.

Due to the fact that the nonlinear term is not well defined for $W(0,T) \hookrightarrow L^{\frac{4}{3}}(I \times \Omega)$, this problem cannot be treated within the functional analytic setting proposed in Section 2.2. Instead, the space $X \cap L^\infty(\Omega)$ can be used for the state. For details on the existence theory for this type of semilinear problem, we refer for example to Neitzel and Vexler [85] or Chapter 5 in Tröltzsch [107].

2.3.3. Temperature control in combustion

As a practical example of an optimization problem governed by a system of parabolic partial differential equations we are going to control a combustion process by cooling at the boundary. The mathematical model for the combustion process is taken from Lang [65]. We note that our problem setup—involving a time dependent control—is

different from the parameter estimation problem considered by Meidner and Vexler [78], which was built around the same model.

The problem we consider uses a simplified model of a gaseous combustion process which was derived in Lang [65]. Under the low Mach number hypothesis, the density of the fluid becomes independent from its pressure. After further approximations, the motion of the fluid is independent of concentration and temperature and it enters the equation system for the other two quantities only via a convection term. Here we consider a stationary fluid, i. e., constant velocity zero. Assuming constant diffusion coefficients, we introduce the dimensionless temperature variable $\theta = \frac{T - T_{\text{unburnt}}}{T_{\text{burnt}} - T_{\text{unburnt}}}$ and the fluid concentration Y. Then the combustion process is modelled by the two equations

$$\partial_t \theta - \Delta\theta = \omega(Y, \theta) \qquad \text{in } I \times \Omega, \tag{2.8a}$$

$$\partial_t Y - \frac{1}{\text{Le}} \Delta Y = -\omega(Y, \theta) \qquad \text{in } I \times \Omega, \tag{2.8b}$$

where Le is the Lewis number which indicates the ratio of the diffusivities of mass and temperature. The reaction rate ω on the right hand side is modelled by an Arrhenius law for a simple one-species reaction process with an approximation for large activation energy. It is given by

$$\omega(Y, \theta) = Y \begin{cases} \frac{\beta^2}{2\text{Le}} \mathrm{e}^{\frac{\beta(\theta-1)}{1+\alpha_c(\theta-1)}}, & \theta > \frac{\alpha_c - 1}{\alpha_c} \\ 0, & \theta \leq \frac{\alpha_c - 1}{\alpha_c}. \end{cases}$$

Compared to the reaction term used in [65], we removed the singularity at $\theta = \frac{\alpha_c - 1}{\alpha_c}$ by continuing with 0 for smaller θ. This avoids problems if temperature drops below T_{unburnt} and is justified from the modelling point of view since for low temperatures the reaction should come to a stop. We note that the modified reaction term is continuously differentiable.

The configuration we consider for the control problem is a freely propagating laminar flame in two space dimensions passing through an obstacle formed by a set of two parallel rods that can be cooled or heated. We assume that the temperature of each rod can be controlled individually over time, so the control variable consists of two time-dependent temperature parameters q_1 and q_2. The considered spatial domain Ω, along with a visualization of the configuration at initial time can be seen from Figure 2.1. For simplicity we assume that the heat exchange at the rods can be modelled by Newton's law of cooling which leads to a boundary condition of Robin type. We omitted the Dirichlet boundary conditions on the left boundary of the domain proposed in [65] since they are not required here. With the boundary designations as indicated in Figure 2.1(a) we have the boundary conditions

$$\begin{aligned} &\partial_n \theta = 0 && \text{on } \Gamma_N \times (0,T), && \partial_n Y = 0 \text{ on } \Gamma_N \times (0,T), \\ &\partial_n \theta = k_\theta(q_1 - \theta) && \text{on } \Gamma_R^1 \times (0,T), && \partial_n Y = 0 \text{ on } \Gamma_R^1 \times (0,T), \\ &\partial_n \theta = k_\theta(q_2 - \theta) && \text{on } \Gamma_R^2 \times (0,T), && \partial_n Y = 0 \text{ on } \Gamma_R^2 \times (0,T) \end{aligned} \tag{2.9}$$

Figure 2.2.: Desired mass distribution Y_{ref} at final time $T = 40$

with the two components of the control entering the Robin boundary conditions on Γ_R^1 and Γ_R^2. As initial condition we set the analytic solution for a one-dimensional right-travelling flame in the limit $\beta \to \infty$ located left of the obstacle:

$$\begin{aligned} \theta(0,x) &= \begin{cases} 1 & \text{for } x_1 \leq \tilde{x}_1 \\ \mathrm{e}^{\tilde{x}_1 - x_1} & \text{for } x_1 > \tilde{x}_1 \end{cases} && \text{on } \Omega, \\ Y(0,x) &= \begin{cases} 0 & \text{for } x_1 \leq \tilde{x}_1 \\ 1 - \mathrm{e}^{\tilde{x}_1 - x_1} & \text{for } x_1 > \tilde{x}_1 \end{cases} && \text{on } \Omega. \end{aligned} \tag{2.10}$$

For our computations, we set $\tilde{x}_1 = 9$. Following Lang [65], the remaining parameters are chosen as $\mathrm{Le} = 1$, $\alpha_c = 0.8$, $\beta = 10$ and $k_\theta = 0.1$.

To embed the state equations into the abstract setting proposed in Section 2.2, we specify the state vector as $u := (\theta, Y)$ from the state space X as defined in (2.1) with the spaces V and H

$$V = H^1(\Omega)^2, \quad H = L^2(\Omega)^2.$$

We note that this is the same construction as employed in Meidner [76].

As control space we use $Q_{\text{ad}} = \left\{ q \in L^2(I, \mathbb{R}^2) \;\middle|\; q^a \leq q \leq q^b \right\}$. Since we want to control the progress of the combustion at the end of the simulation time, the cost functional consist of a tracking term with terminal observation at time $T = 40$ of the mass distribution and a L^2 cost term for the control.

$$J(q,u) = \frac{1}{2} \left\| Y(T) - Y_{\text{ref}} \right\|^2_{L^2(\Omega))} + \frac{\alpha}{2} \left\| q \right\|^2_{L^2(I,\mathbb{R}^2)}.$$

with the control cost parameter $\alpha = 1$. The desired mass distribution Y_{ref} at final time is given by

$$Y_{\text{ref}}(x) = \begin{cases} 0, & x_1 - x_2 \leq \bar{x}, \\ 1 - e^{\bar{x} + x_2 - x_1}, & x_1 - x_2 > \bar{x}, \end{cases}$$

where $\bar{x} = 37$. It is visualized in Figure 2.2. The control constraints are chosen as $q^a = -0.1$ and $q^b = 0.5$.

3. Optimality conditions and optimization algorithms

In the first section of this chapter we discuss optimality conditions for the optimization problem (2.5) and give cheaply computable representation formulas for first and second derivatives which can be used in second order optimization algorithms.

The second section presents a semismooth Newton algorithm which is based on a reformulation of the first order optimality condition in terms of the normal map. This approach was briefly mentioned by Ulbrich [108] and described in some more detail by Kunisch, Pieper, and Rund [62] and Kunisch and Rund [64]. We formulate the algorithm in the continuous setting on the Banach space. How to apply this algorithm in the discretized setting will be discussed in Section 4.2.

In Section 3.2.2 we describe a practical realization of the algorithm, along with an extension similar to the Steihaug conjugate gradient method that increases the radius of convergence.

3.1. Optimality conditions and representation of derivatives

Subsequently we assume that the state equation (2.3) is well posed for every control $q \in Q_{\mathrm{ad}}$ in the sense that there exists a unique solution and, furthermore, the state depends continuously on the control. Therefore we can define the continuous *solution operator* $S\colon Q_{\mathrm{ad}} \to X, q \mapsto u$, which is frequently also called the *control-to-state mapping* in the literature. Later on we will also use the notation $u(q) := S(q)$ to denote the state resulting from the control q. We will use analogous notations for further quantities depending on q via auxiliary equations. Inserting the solution operator into the cost functional, we can formulate the reduced optimization problem

$$\min j(q) \text{ s. t. } q \in Q_{\mathrm{ad}}, \tag{3.1}$$

where the *reduced cost functional* is given by $j(q) = J(q, S(q))$. Evidently, this formulation is equivalent to (2.5). To state optimality conditions and formulate derivative-based optimization algorithms, we have to make certain differentiability assumptions on the reduced cost functional. While for linear-quadratic problems, Fréchet differentiability can be analyzed in a straight-forward fashion, it can be a delicate issue in the non-linear setting. For our purposes we will subsequently assume that all stated directional

derivatives are well defined. We denote the directional derivative of a semilinear form with respect to a variable by the corresponding subscript and a prime and add the direction at the beginning of the list of linear arguments. For example, the derivative of $a(q,u)(\varphi)$ with respect to u in direction δu reads

$$a'_u(q,u)(\delta u,\varphi).$$

For higher order derivatives, the directions are given in the order of differentiation from left to right.

To state optimality conditions and the optimization algorithm, we will use some linear auxiliary equations. Well-posedness of those equations for given $(q,u) \in Q_{\mathrm{ad}} \times X$ is ensured by the following assumption.

Assumption 3.1. For any pair $(q,u) \in Q_{\mathrm{ad}} \times X$, any $f \in L^2(I,V^*)$ and any $v_0, w_T \in H$, the problems

$$\int_I \langle \partial_t v, \varphi\rangle_{V^*\times V} + a'_u(q,u)(v,\varphi)\,\mathrm{d}t + (v(0),\varphi(0)) = \int_I \langle f,\varphi\rangle_{V^*\times V}\,\mathrm{d}t + (v_0,\varphi(0))$$

for all $\varphi \in X$ and

$$\int_I -\langle \partial_t w, \varphi\rangle_{V^*\times V} + a'_u(q,u)(\varphi,w)\,\mathrm{d}t + (\varphi(T),w(T)) = \int_I \langle f,\varphi\rangle_{V^*\times V}\,\mathrm{d}t + (w_T,\varphi(T))$$

for all $\varphi \in X$ admit unique solutions v and w in X.

Remark 3.2. This assumption holds for example if we require that for all admissible pairs (q,u) with $q \in Q_{\mathrm{ad}}$ and $u = S(q)$, the semilinear form $\tilde{a}\colon I \times V \times V \to \mathbb{R}$ given by $(t,\varphi,\psi) \mapsto a_u(t,q,u(t))(\psi,\varphi)$ is measurable with respect to t for all $\psi,\varphi \in V$ and there are positive constants C_1, C_2, and C_3 such that

$$\begin{aligned} |\tilde{a}(t)(\psi,\varphi)| &\le C_1 \|\psi\|_V \|\varphi\|_V && \text{for all } \psi,\varphi \in V \text{ and almost all } t \in I, \text{ and} \\ \tilde{a}(t)(\psi,\psi) + C_2\|\psi\|^2 &\ge C_3 \|\psi\|_V^2 && \text{for all } \psi \in V \text{ and almost all } t \in I. \end{aligned}$$

The corresponding standard existence theorem for linear parabolic equations can be found for example as Theorem 11.7 in [23].

3.1.1. Optimality conditions for the reduced problem

For completeness, we quote the standard necessary first order optimality condition for the reduced cost functional j as given for example in [107].

Lemma 3.3 (Necessary first order optimality condition)**.** *Let Q be a Banach space, $Q_{\mathrm{ad}} \subseteq Q$ and $j\colon Q \to \mathbb{R}$ be Fréchet-differentiable in an open set containing Q_{ad}. If j attains its minimum in Q_{ad} at $\bar{q}$, then the first order necessary optimality condition*

$$j'(\bar{q})(q-\bar{q}) \ge 0 \quad \textit{for any } q \in Q_{\mathrm{ad}} \tag{3.2}$$

holds true. If j is convex, then conversely condition (3.2) *implies optimality.*

This condition can be reformulated into an equality which will be exploited for the semismooth Newton method discussed in the next section. Assuming Gâteaux differentiability of the cost functional, we can introduce the gradient $\nabla j(q)$ of the reduced cost functional defined via the Riesz representation theorem by

$$(\nabla j(q), \delta q)_Q = j'(q)(\delta q) \quad \text{for all } \delta q \in Q.$$

Furthermore, we define the L^2 projection onto the admissible set $P_{Q_{\mathrm{ad}}} \colon Q \to Q_{\mathrm{ad}}$ given by the condition

$$\|q - P_{Q_{\mathrm{ad}}}(q)\|_Q \leq \|q - p\|_Q \quad \text{for any } p \in Q_{\mathrm{ad}}.$$

Since Q_{ad} is convex, $P_{Q_{\mathrm{ad}}}$ is well defined.

Both parts of the following simple lemma are shown for example in the textbook [57] (Lemma 1.11 and 1.12).

Lemma 3.4. *1. For any $\gamma_G > 0$, the first order optimality condition* (3.2) *is equivalent to the condition*

$$\bar{q} = P_{Q_{\mathrm{ad}}}\left(\bar{q} - \gamma_G \nabla j(\bar{q})\right). \tag{3.3}$$

2. For Q_{ad} given as in (2.4)*, the projection $P_{Q_{\mathrm{ad}}}$ has the explicit representation*

$$P_{Q_{\mathrm{ad}}}(q) = \max(q^a, \min(q^b, q)) \tag{3.4}$$

where the superposition operators min *and* max *denote the pointwise minimum and maximum of the arguments respectively.*

For stating the semismooth Newton method we will use another reformulation of the optimality condition.

Proposition 3.5. *Let $\gamma_G > 0$. A control $\bar{q} \in Q_{\mathrm{ad}}$ satisfies the first order optimality condition* (3.2) *if and only if there exists $\bar{p} \in Q$ such that $\bar{q} = P_{Q_{\mathrm{ad}}}(\bar{p})$ and the condition*

$$\mathcal{N}(\bar{p}) = 0 \tag{3.5}$$

holds true, where the map $\mathcal{N} \colon Q \to Q$ is given by

$$\mathcal{N}(p) = p - P_{Q_{\mathrm{ad}}}(p) + \gamma_G \nabla j(P_{Q_{\mathrm{ad}}}(p)). \tag{3.6}$$

Following the naming introduced by Robinson, see, e. g., [92], we call $\mathcal{N}$ the normal map.

Proof. Let us assume we have $\bar{p} \in Q$ satisfying $\mathcal{N}(\bar{p}) = 0$. Setting $\bar{q} = P_{Q_{\mathrm{ad}}}(\bar{p})$ we obtain from (3.6)

$$\bar{p} = \bar{q} - \gamma_G \nabla j(\bar{q}).$$

Hence the identity (3.3) holds true for $\bar{q}$.

Conversely, for a stationary point $\bar{q} \in Q_{\mathrm{ad}}$ satisfying (3.3) we set

$$\bar{p} = \bar{q} - \gamma_G \nabla j(\bar{q})$$

and obtain on the one hand $\bar{q} = P_{Q_{\mathrm{ad}}}(\bar{p})$ and on the other hand $\mathcal{N}(\bar{p}) = 0$. □

Remark 3.6. According to Proposition 3.5, instead of looking for a point $\bar{q}$ satisfying the first order optimality condition, we can equivalently search for some $\bar{p}$ that fulfills (3.5). The semismooth Newton algorithm described in Section 3.2 employs that idea.

3.1.2. Representation of derivatives and optimality system

The approach taken here for deriving a computationally efficient representation for the first derivative and the first order optimality system is standard and similar derivations can be found for example in the textbooks [57, 59, 107]. For the representation formula for the second derivative we follow Becker et al. [12].

To make the first order optimality condition accessible for computational evaluation, we reformulate it in terms of the *Lagrangian* $\mathcal{L}\colon Q \times X \times X \to \mathbb{R}$, which for our problem is given by

$$\mathcal{L}(q,u,z) = J(q,u) + (u_0(q), z(0)) - (u(0), z(0)) - \int_I \left[\langle \partial_t u, z\rangle_{V^*\times V} + a(q,u)(z)\right] \mathrm{d}t. \tag{3.7}$$

Obviously, for $q \in Q_{\mathrm{ad}}$, $u = S(q)$, and arbitrary $z \in X$, the identity $j(q) = \mathcal{L}(q,u,z)$ holds true. Since u solves the state equation, the partial derivative with respect to z in direction φ

$$\mathcal{L}'_z(q,u,z)(\varphi) = (u_0(q), \varphi(0)) - (u(0), \varphi(0)) - \int_I \left[\langle \partial_t u, \varphi\rangle_{V^*\times V} + a(q,u)(\varphi)\right] \mathrm{d}t.$$

vanishes for any $\varphi \in X$. Hence, assuming Fréchet differentiability of J and a with respect to q and u and Gâteaux-differentiability of the solution operator S, we can differentiate the Lagrangian with the chain rule. Hence, for any $z \in X$ and δq with $q + \delta q \in Q_{\mathrm{ad}}$, the identity

$$j'(q)(\delta q) = \mathcal{L}'_q(q,u,z)(\delta q) + \mathcal{L}'_u(q,u,z)(\delta u)$$

holds true where $\delta u = S'(q)(\delta q)$. Since $z \in X$ was arbitrary, we can choose it in such a way that the second term vanishes, i. e., that the equation $\mathcal{L}'_u(q,u,z)(\varphi) = 0$, which can be stated equivalently as

$$\int_I -\langle \varphi, \partial_t z\rangle_{V\times V^*} + a'_u(q,u)(\varphi, z)\,\mathrm{d}t + (\varphi(T), z(T)) = J'_1(u)(\varphi) + J'_2(u(T))(\varphi(T)), \tag{3.8}$$

holds true for *any* $\varphi \in X$. The representation (3.8) is obtained by integrating the term involving the temporal derivative by parts. That this is admissible on the space X

is shown for example in Wloka [117]. Assumption 3.1 ensures the well-posedness of this *adjoint equation*. To see that the right hand side and terminal value satisfy the requirements stated there we note that $J_1'(u)$ is in $(L^2(I,V))^* \cong L^2(I,V^*)$ and rewrite the functional $J_2'(u)$ on H by its Riesz representation.

Summarizing, we get the following statement:

Lemma 3.7 (Representation of the first derivative)**.** *Let Assumption 3.1 hold for $q \in Q_{\mathrm{ad}}$ and $u = S(q)$. We assume further that J_1 and J_2 are Fréchet-differentiable and that the semilinear form a is continuously Fréchet-differentiable with respect to q and u. Then, the derivative of the reduced cost functional in a direction δq is given by*

$$j'(q)(\delta q) = \mathcal{L}_q'(q,u,z)(\delta q) = \alpha(q,\delta q)_Q + (u_0'(q)(\delta q), z(0)) - \int_I a_q'(q,u)(\delta q, z)\,dt \tag{3.9}$$

where the adjoint state z is a solution of (3.8)*.*

Proof. To apply the above derivation, we need to ensure that the solution operator is Gâteaux differentiable. With continuous Fréchet-differentiability of the state equation and Assumption 3.1 this follows from the implicit function theorem (see, e. g., Dieudonne [29, Theorem 10.2.1]). □

Remark 3.8. 1. For many nonlinear problems, the assumption of continuous Fréchet differentiability of the semilinear form is too strong. In some of those cases Gâteaux differentiability of the solution operator can be shown directly without relying on the implicit function theorem, see, e. g., Neitzel and Vexler [85] or Wachsmuth [112].
2. We point out that the adjoint equation inherits the parabolic structure from the state equation and constitutes a linear parabolic equation with time running backwards. Later we will see that, thanks to the properties of the discontinuous Galerkin discretization, also the discrete adjoint equation has the same algebraic structure as the linearization of the discrete state equation. Hence we can employ the same techniques to solve it numerically.
3. Once for given q the corresponding state u and adjoint state z are known, the representation (3.9) is explicit with respect to the direction δq. For numerical realization after discretization this means that no further PDEs have to be solved to evaluate the derivative for various directions.

For later use, we introduce the gradient G_{impl} of the implicitly defined part $J_1(u(q)) + J_2(u(q)(T))$ of the reduced cost functional j, which is given by

$$(G_{\mathrm{impl}}(q,u,z),\delta q)_Q = (u_0'(q)(\delta q), z(0)) - \int_I a_q'(q,u)(\delta q, z)\,\mathrm{d}t \quad \text{for all } \delta q \in Q. \tag{3.10}$$

In terms of G_{impl}, the representation formula for the first derivative reads

$$j'(q)(\delta q) = (\alpha q + G_{\mathrm{impl}}(q, u(q), z(q)), \delta q)_Q.$$

Later on, we will use the short hand notation $G_{\text{impl}}(q)$ to refer to the implicit part of the gradient $G_{\text{impl}}(q, u(q), z(q))$ for given control q and corresponding state and adjoint solution $u(q)$ and $z(q)$.

Denoting by $\bar{u}$ and $\bar{z}$ the state and adjoint variables corresponding to the optimal control $\bar{q}$, the first order optimality system in terms of the Lagrangian reads

$$\mathcal{L}'_z(\bar{q}, \bar{u}, \bar{z})(\varphi) = 0 \qquad \text{for all } \varphi \in X \text{ (state equation)} \tag{3.11a}$$

$$\mathcal{L}'_u(\bar{q}, \bar{u}, \bar{z})(\varphi) = 0 \qquad \text{for all } \varphi \in X \text{ (adjoint equation)} \tag{3.11b}$$

$$\mathcal{L}'_q(\bar{q}, \bar{u}, \bar{z})(q - \bar{q}) \geq 0 \qquad \text{for all } q \in Q_{\text{ad}} \text{ (gradient condition).} \tag{3.11c}$$

Analogous to the optimality condition for the reduced problem, the last condition can be equivalently stated as

$$\bar{q} = P_{Q_{\text{ad}}}\left(\bar{q} - \gamma_G \alpha \bar{q} - \gamma_G G_{\text{impl}}(\bar{q}, \bar{u}, \bar{z})\right). \tag{3.11c*}$$

We note that for $\alpha > 0$ and $\gamma_G = \frac{1}{\alpha}$ the first two terms in the argument of the projection in (3.11c*) cancel out.

For an optimality system in terms of the unprojected variable p we rewrite the normal map in terms of G_{impl} yielding

$$\mathcal{N}(p, u, z) = p + (\gamma_G \alpha - 1)\, P_{Q_{\text{ad}}}(p) + \gamma_G G_{\text{impl}}(P_{Q_{\text{ad}}}(p), u, z).$$

Then, at a local optimum, $(\bar{p}, \bar{u}, \bar{z})$ satisfy

$$\mathcal{L}'_z\left(P_{Q_{\text{ad}}}(\bar{p}), \bar{u}, \bar{z}\right)(\varphi) = 0 \quad \text{for all } \varphi \in X, \tag{3.12a}$$

$$\mathcal{L}'_u\left(P_{Q_{\text{ad}}}(\bar{p}), \bar{u}, \bar{z}\right)(\varphi) = 0 \quad \text{for all } \varphi \in X, \tag{3.12b}$$

$$\mathcal{N}(\bar{p}, \bar{u}, \bar{z}) = 0. \tag{3.12c}$$

For solving the Newton update equations arising in the semismooth Newton method, we need to solve for the Hessian of the reduced cost functional. We use an iterative solver for this purpose since assembling the full Hessian is only cost-effective when the (discretized) control possesses no more than a handful of degrees of freedom. What is needed is an efficient way to compute an appropriate representation of functionals of the form

$$\delta q \mapsto j''(q)(\delta q, \tau q)$$

for a given iterate $q \in Q_{\text{ad}}$ and a direction $\tau q \in Q$. We introduce two more linearized auxiliary problems which are well-posed under Assumption 3.1.

Definition 3.9. For given $q, \tau q \in Q$ and $u \in X$, the *tangent equation* is given as: find $\tau u \in X$ satisfying

$$\int_I \langle \partial_t \tau u, \varphi \rangle_{V^* \times V} + a'_u(q, u)(\tau u, \varphi)\, \mathrm{d}t + (\tau u(0), \varphi(0)) = -\int_I a'_q(q, u)(\tau q, \varphi)\, \mathrm{d}t + (u'_0(q)(\tau q), \varphi(0)) \tag{3.13}$$

for all $\varphi \in X$. Assuming a and J to be twice Fréchet-differentiable, the *additional adjoint equation* for given $q, \tau q \in Q$ and $u, z, \tau u \in X$ reads: find $\tau z \in X$ such that

$$\begin{aligned}(3.14)\quad \int_I -\langle \varphi, \partial_t \tau z\rangle_{V\times V^*} + a'_u(q,u)(\varphi,\tau z)\,\mathrm{d}t + (\varphi(T),\tau z(T)) \\ = -\int_I \left[a''_{uq}(q,u)(\varphi,\tau q,z) + a''_{uu}(q,u)(\varphi,\tau u,z)\right]\,\mathrm{d}t \\ + J_1''(u)(\tau u,\varphi) + J_2''(u(T))(\varphi(T),\tau u(T))\end{aligned}$$

for any $\varphi \in X$.

Lemma 3.10. *We assume J and a to be twice continuously Fréchet-differentiable. For given $q \in Q_{\mathrm{ad}}$ and corresponding $u = u(q) \in X$ let Assumption 3.1 hold. Furthermore let $z \in X$ be the solution of the adjoint equation* (3.8) *and for a direction $\tau q \in Q$ let τu and τz solve the tangent and additional adjoint equations* (3.13) *and* (3.14) *respectively. Then, for $\delta q \in Q$ the second derivative of the reduced cost functional admits the representation*

$$\begin{aligned}(3.15)\quad j''(q)(\delta q,\tau q) = \alpha(\delta q,\tau q)_Q + (u_0''(q)(\delta q,\tau q),z(0)) \\ - \int_I \left[a''_{qq}(q,u)(\delta q,\tau q,z) + a''_{qu}(q,u)(\delta q,\tau u,z) + a'_q(q,u)(\delta q,\tau z)\right]\,dt.\end{aligned}$$

Proof. Applying the implicit function theorem to the state equation shows that the derivative of the solution map $S\colon q \mapsto u$ in q in direction τq is given as the solution τu of the tangent equation (3.13). In the same way, the derivative of the adjoint map $q \mapsto z$ in q in direction τq is found from the adjoint equation (3.8) through the implicit function theorem. It is given as the solution τz of the additional adjoint equation (3.14) with τu as above. Hence, the representation formula (3.15) for the second derivative is obtained by taking the total derivative of the representation formula (3.9) in direction τq. □

Corollary 3.11. *Under the assumptions of Lemma 3.10, the Hessian $H_{\mathrm{impl}}\colon Q_{\mathrm{ad}} \to B(Q)$ of the implicitly defined part of the reduced cost functional is given by the identity*

$$\begin{aligned}(3.16)\quad (H_{\mathrm{impl}}(q)\delta q,\tau q)_Q = (u_0''(q)(\delta q,\tau q),z(0)) \\ - \int_I \left[a''_{qq}(q,u)(\delta q,\tau q,z) + a''_{qu}(q,u)(\delta q,\tau u,z) + a'_q(q,u)(\delta q,\tau z)\right]\,dt\end{aligned}$$

for all directions $\delta q, \tau q \in Q$, where τu and τz are the solutions of the tangent and additional adjoint equations corresponding to τq.

Proof. With the same reasoning as above, the identity (3.16) results from total differentiation of equation (3.10). □

3.2. Semismooth Newton method

In this section we outline the semismooth Newton method that we use to solve the optimization problem (2.5). Using a semismooth Newton approach for optimal control problems with inequality constraints on the control can be considered well-established practice. This fact is also resembled by the wealth of publications available on this topic including the works by Ito and Kunisch [59], Hinze et al. [57], and [108] by Ulbrich.

Algorithms for optimal control problems can be classified by whether they try to solve a first order optimality system like (3.11) treating q, u, and z as optimization variables or whether they work on the reduced problem (3.1) with the sole optimization variable q instead while u and z are dependent quantities. The first type of algorithms is commonly referred to as *all-at-once* methods while the latter are called *black box* methods, for an early use of those terms, see Frank and Shubin [42] where the two approaches are compared for an airfoil design problem.

Here we opt for the black-box approach since it allows to use standard PDE solution algorithms. Furthermore a memory-efficient implementation, which is of particular importance for time-dependent problems, is more straight-forward than with all-at-once methods. Another important consideration is that sequential quadratic programming (SQP) methods, which employ the all-at-once idea, usually only offer a significant gain in efficiency if the cost of solving linearized partial differential equations is considerably lower than for solving the full non-linear equation, see Hinze and Kunisch [56]. However, for the time-stepping method we consider in Chapter 5, solutions of the linearized equations and of the full state equation have roughly the same cost.

The fundamental idea behind semismooth Newton methods is to apply Newton's method for solving equations involving maps that are not globally differentiable in the classical sense. If the considered maps are *semismooth* with respect to a suitable generalized derivative, i. e., they satisfy a certain weakened differentiability requirement, then a Newton type method employing this generalized derivative exhibits local superlinear convergence. In the standard approach to a semismooth Newton method for our reduced optimal control problem, Newton's method is applied to the optimality condition (3.3) to solve for the control q. Here, we base on the reformulated optimality condition (3.5) instead, solving for the unprojected variable p. An obvious benefit of this formulation is that the active sets of the control constraints can be deduced immediately from p. An algorithm using the reduced formulation in terms of p for PDE-constrained optimization problems was published in [62] and [64], however a related idea was presented earlier by Schiela in Section 6 of [97]. There, a model problem is considered for which p coincides with the adjoint state z. This fact is used to eliminate the control from the optimality system (3.11) yielding a semismooth optimality system in u and $p = z$.

For the reduced formulation in terms of p, we will see that, interpreted in the right way, the operator that has to be inverted for computing the semismooth Newton update is self-adjoint and close to the optimum also positive definite. Therefore, in practice, the

conjugate gradient method can be employed for its inversion instead of generic solvers like GMRES which have higher memory consumption.

To ensure robustness of the method, in particular since we intend to solve problems with non-linear state equations, a globalization strategy is desirable. For the standard approach to semismooth Newton solving for the control q, such a strategy built around a trust region framework is discussed in detail by Ulbrich [108]. Hinze and Vierling [58] use the approach by Gräser and Kornhuber [48] to give a globalization for linear quadratic problems with control constraints. This approach is based on a dual optimization problem and has the benefit that the control q does not have to be computed and stored explicit which makes it feasible in practice for the so called *variational* treatment of the control (see Section 4.2.1). However, it is limited to problems with linear state equation.

For our computations we use a simpler heuristic strategy by Pieper (see [62,89]) to increase the radius of convergence. It is similar to the Steihaug cg method, see Steihaug [105]. This strategy is motivated by the approach taken to solve the Newton update equation.

3.2.1. Theoretical considerations

In this section, we derive the semismooth Newton method and prove local superlinear convergence, given that some assumptions on the problem are satisfied. Under stronger assumptions it is also possible to quantify the rate of convergence. For clarity of presentation, we do not discuss the corresponding extensions of the results here. However they are straight-forward with the techniques shown in, e. g., [108] or [97].

We use the following definition for semismoothness of operators on Banach spaces.

Definition 3.12. Let Q, P denote Banach spaces, and $G\colon Q \to P$ and $\partial G\colon Q \to B(Q, P)$ given mappings. The operator G is called ∂G-semismooth at a point $q \in Q$ if G is continuous in a neighbourhood of q and

$$\|G(q + \delta q) - G(q) - \partial G(q + \delta q)(\delta q)\|_P = o(\|\delta q\|_Q)$$

for $\|\delta q\|_Q \to 0$.

Remark 3.13. 1. This definition is identical to the one given in [108] apart from the fact that there, the mapping ∂G is set-valued and the defining property is enforced for all representatives of the set. Since for the purpose of stating a semismooth Newton algorithm, a single fixed generalized derivative ∂G is all we need, our simplification poses no restriction in this context.
2. Evidently, continuously Fréchet differentiable operators are semismooth with respect to their Fréchet derivative.

To solve (3.5) with the semismooth Newton method we need to ensure that the normal map $\mathcal{N}$ is semismooth with respect to a suitable generalized derivative. As a preparation, we need a chain rule for semismooth functions, which we quote from [108], and

semismoothness of the pointwise projection operator $P_{Q_{\mathrm{ad}}}$. For the sake of completeness, we give a short proof for the latter result employing the techniques proposed in [97].

Lemma 3.14 ([108, Proposition 3.7]). *Let P, Q, R be Banach spaces, $U \subseteq P$, $V \subseteq Q$ be open subsets, $F\colon U \to Q$ Lipschitz continuous around $p \in U$, ∂F-semismooth in p and $F(U) \subseteq V$. Let furthermore $G\colon V \to R$ be ∂G-semismooth in $q = F(p)$ and ∂G be bounded around q. Then the composition $H = G \circ F$ is ∂H-semismooth in p with*

$$\partial H\colon P \to B(P,R),\ p \mapsto \partial G(F(p))\partial F(p).$$

Lemma 3.15 (Semismoothness of $P_{Q_{\mathrm{ad}}}$). *Let $1 \leq \hat{r} < r < \infty$ and let the operator $\chi_{\mathcal{I}}\colon L^r(\Omega_Q) \to B(L^r(\Omega_Q), L^{\hat{r}}(\Omega_Q))$ be given as the characteristic function of the inactive sets defined by*

$$\chi_{\mathcal{I}}(p)(\delta q)(x) = \begin{cases} \delta q(x), & \text{if } q^a(x) < p(x) < q^b(x), \\ 0, & \text{otherwise} \end{cases}$$

for almost all $x \in \Omega_Q$. Then the operator $P_{Q_{\mathrm{ad}}}\colon L^r(\Omega_Q) \to L^{\hat{r}}(\Omega_Q)$ defined by (3.4) is $\chi_{\mathcal{I}}$-semismooth.

Proof. We define the function $\pi_{Q_{\mathrm{ad}}}\colon \mathbb{R} \times \Omega_Q \to \mathbb{R}$ by

$$\pi_{Q_{\mathrm{ad}}}(s,x) = \min(q^b(x), \max(q^a(x), s))$$

and the function $\hat{\chi}_{\mathcal{I}}\colon \mathbb{R} \times \Omega_Q \to \mathbb{R}$ by

$$\hat{\chi}_{\mathcal{I}}(s,x) = \begin{cases} 1, & \text{if } q^a(x) < s < q^b(x), \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, $P_{Q_{\mathrm{ad}}}(p)(x) = \pi_{Q_{\mathrm{ad}}}(p(x), x)$, i. e., $P_{Q_{\mathrm{ad}}}$ is the Nemytskii operator corresponding to $\pi_{Q_{\mathrm{ad}}}$, and $\chi_{\mathcal{I}}(p)(\delta q)(x) = \hat{\chi}_{\mathcal{I}}(p(x), x)\delta q(x)$. We note that $\pi_{Q_{\mathrm{ad}}}$ is a *Carathéodory function*, i. e., it is continuous with respect to the first argument for almost all $x \in \Omega_Q$ and measurable with respect to the second argument for any $s \in \mathbb{R}$. The function $\hat{\chi}_{\mathcal{I}}$ belongs to the larger class of *Baire-Carathéodory functions* consisting of functions that can be expressed as pointwise limits of Carathéodory functions almost everywhere (see [6, Section 1.4]).

To show semismoothness of $P_{Q_{\mathrm{ad}}}$ at a fixed point $p^* \in L^r(\Omega_Q)$, we introduce the auxiliary function $\psi_*\colon \mathbb{R} \times \Omega_Q \to \mathbb{R}$ given by

$$\psi_*(s,x) = \begin{cases} \frac{\pi_{Q_{\mathrm{ad}}}(s,x) - \pi_{Q_{\mathrm{ad}}}(p^*(x),x) - \hat{\chi}_{\mathcal{I}}(s,x)(s - p^*(x))}{|s - p^*(x)|}, & \text{if } s \neq p^*(x), \\ 0, & \text{if } s = p^*(x). \end{cases}$$

The idea is now to show that the corresponding Nemytskii operator Ψ_* is well defined and continuous at p^* and to conclude semismoothness in p^* from this fact.

First, we show that for any $x \in \Omega_Q$, ψ_* is continuous with respect to the first argument at $s = p^*(x)$. For this purpose we consider the characteristic functions $\hat{\chi}_{\mathcal{A}}^-, \hat{\chi}_{\mathcal{A}}^+ \colon \mathbb{R} \times \Omega_Q \to \mathbb{R}$ of the active sets with respect to the lower and upper bounds defined by

$$\hat{\chi}_{\mathcal{A}}^-(s,x) = \begin{cases} 1, & \text{if } s < q^a(x), \\ 0, & \text{otherwise,} \end{cases}$$
$$\hat{\chi}_{\mathcal{A}}^+(s,x) = \begin{cases} 1, & \text{if } s > q^b(x), \\ 0, & \text{otherwise,} \end{cases}$$

and note that

$$\hat{\chi}_{\mathcal{I}}(s,x) = 1 - \hat{\chi}_{\mathcal{A}}^-(s,x) - \hat{\chi}_{\mathcal{A}}^+(s,x),$$
$$\text{and } \pi_{Q_{\mathrm{ad}}}(s,x) = \hat{\chi}_{\mathcal{I}}(s,x)s + \hat{\chi}_{\mathcal{A}}^-(s,x)q^a(x) + \hat{\chi}_{\mathcal{A}}^+(s,x)q^b(x).$$

Plugging these identities into the definition of ψ_* yields for $s \neq p^*(x)$

$$\begin{aligned}\psi_*(s,x)|s - p^*(x)| = \left(q^b(x) - p^*(x)\right)\left(\hat{\chi}_{\mathcal{A}}^+(p^*(x),x) - \hat{\chi}_{\mathcal{A}}^+(s,x)\right) \\ + \left(q^a(x) - p^*(x)\right)\left(\hat{\chi}_{\mathcal{A}}^-(p^*(x),x) - \hat{\chi}_{\mathcal{A}}^-(s,x)\right).\end{aligned}$$

The first summand vanishes if $q^b(x) = p^*(x)$ and otherwise vanishes for s in a sufficiently small neighbourhood of $p^*(x)$. In the same way, the second summand vanishes for s sufficiently close to $p^*(x)$, which shows that ψ_* is continuous with respect to s at $s = p^*(x)$. Hence it is easy to see that ψ_* is a Baire-Carathéodory function. Furthermore, we can estimate

$$\begin{aligned}\left|\left(q^b(x) - p^*(x)\right)\left(\hat{\chi}_{\mathcal{A}}^+(p^*(x),x) - \hat{\chi}_{\mathcal{A}}^+(s,x)\right)\right| \\ = \begin{cases} |q^b(x) - p^*(x)|, & \text{if } s < q^b(x) < p^*(x) \text{ or } p^*(x) < q^b(x) < s, \\ 0, & \text{otherwise} \end{cases} \\ \leq |s - p^*(x)|.\end{aligned}$$

If we estimate the second summand in the same way, we can conclude $\psi_*(s,x) \leq 2$ for all $p^* \in L^r(\Omega_Q)$, $s \in \mathbb{R}$ and almost all $x \in \Omega_Q$. Therefore, the Nemytskii operator $\Psi_* \colon L^r(\Omega_Q) \to L^\infty(\Omega_Q)$ given by $\Psi_*(p)(x) = \psi_*(p(x),x)$ is well-defined.

According to Lemma 3.1 in [97], since ψ_* is a Baire-Carathéodory function and the image of Ψ_* is in $L^\infty(\Omega_Q)$, continuity of ψ_* with respect to the first argument at $s = p^*(x)$ for almost all $x \in \Omega_Q$ implies continuity of $\Psi_* \colon L^r(\Omega_Q) \to L^{r'}(\Omega_Q)$ at $p = p^*$ for any $r' < \infty$.

Noting that

$$P_{Q_{\mathrm{ad}}}(p) - P_{Q_{\mathrm{ad}}}(p^*) - \chi_{\mathcal{I}}(p)(p - p^*) = \Psi_*(p)|p - p^*|,$$

where the absolute value $|\cdot|$ is to be read as the corresponding Nemytskii operator, we choose r' such that $\frac{1}{r'} + \frac{1}{r} = \frac{1}{\hat{r}}$. Since $\hat{r} < r$, we have $1 < r' < \infty$ and Hölder's inequality yields

$$\begin{aligned}\|P_{Q_{\mathrm{ad}}}(p) - P_{Q_{\mathrm{ad}}}(p^*) - \chi_{\mathcal{I}}(p)(p-p^*)\|_{L^{\hat{r}}(\Omega_Q)} \\ \leq \|\Psi_*(p)\|_{L^{r'}(\Omega_Q)} \|p-p^*\|_{L^r(\Omega_Q)} = o(\|p-p^*\|_{L^r(\Omega_Q)})\end{aligned}$$

as $\|p-p^*\|_{L^r(\Omega_Q)}$ tends to zero. Hence we have shown that $P_{Q_{\mathrm{ad}}}$ is $\chi_{\mathcal{I}}$-semismooth at p^* where $p^* \in L^r(\Omega_Q)$ was chosen arbitrary. □

With these preparations we can show semismoothness of the normal map given that the implicit part of the reduced gradient satisfies a suitable regularity property. Furthermore we have to assume $\alpha > 0$ and $\gamma_G = \frac{1}{\alpha}$ such that the optimality condition (3.5) takes the form

$$p + \frac{1}{\alpha} G_{\mathrm{impl}}(P_{Q_{\mathrm{ad}}}(p)) = 0.$$

Lemma 3.16. *We assume that the image of the implicit part of the reduced gradient $G_{\mathrm{impl}}(q) = G_{\mathrm{impl}}(q, u(q), z(q))$ is contained in $L^r(\Omega_Q)$ for some $r > 2$ and additionally that G_{impl} is continuously Fréchet differentiable when considered as a map from $L^2(\Omega_Q)$ to $L^r(\Omega_Q)$. Furthermore, let $\alpha > 0$ and $\gamma_G = \frac{1}{\alpha}$. Then the following three statements hold true.*

1. Any unprojected control $p \in Q$ that satisfies the first order optimality condition (3.5) is contained in $L^r(\Omega_Q)$.

2. The image of the restriction of the normal map $\mathcal{N}$ to $L^r(\Omega_Q)$ is contained in $L^r(\Omega_Q)$.

3. The restriction of the normal map to the space $L^r(\Omega_Q)$, $\mathcal{N}\colon L^r(\Omega_Q) \to L^r(\Omega_Q)$, is $\partial\mathcal{N}$-semismooth with $\partial\mathcal{N}\colon L^r(\Omega_Q) \to B(Q)$ given by

$$\partial\mathcal{N}(p) = \mathrm{Id} + \frac{1}{\alpha} H_{\mathrm{impl}}(P_{Q_{\mathrm{ad}}}(p))\chi_{\mathcal{I}}(p). \tag{3.17}$$

where $\chi_{\mathcal{I}}\colon Q \to B(Q)$ is defined as in Lemma 3.15.

Remark 3.17. 1. For many practically relevant optimization problems of the form (2.5) the gradient of the implicitly defined part of the reduced cost functional can be shown to have the required smoothing property. As seen from the identity (3.10) it is expressed in terms of the state and adjoint solutions, which usually possess higher regularity than the data.
2. For the case $\alpha = 0$ or other choices for γ_G, the normal map is not semismooth in general because the so called "norm gap" between $L^2(\Omega_Q)$ and $L^r(\Omega_Q)$ is essential for semismoothness of the projection operator $P_{Q_{\mathrm{ad}}}$.

Proof. Under the stated assumptions, the first order optimality condition (3.5) simplifies to

$$p + \frac{1}{\alpha} G_{\mathrm{impl}}(P_{Q_{\mathrm{ad}}}(p)) = 0. \tag{3.18}$$

With the regularity assumptions on G_{impl} we obtain immediately

$$p = -\frac{1}{\alpha} G_{\text{impl}}(P_{Q_{\text{ad}}}(p)) \in L^q(\Omega_Q)$$

and hence statements 1 and 2 are shown.

The first term on the left-hand side of (3.18) is obviously semismooth with the identity operator Id as generalized derivative, for the second term we verify the assumptions of the chain rule (Lemma 3.14). According to Lemma 3.15, the inner function $P_{Q_{\text{ad}}} \colon L^r(\Omega_Q) \to L^2(\Omega_Q)$ is $\chi_{\mathcal{I}}$-semismooth. It is easy to see that it is Lipschitz continuous with constant 1. The outer function G_{impl} is by assumption H_{impl}-semismooth and since the derivative was assumed to be continuous, it is bounded locally. Hence, part 3 of the claim follows by invoking the chain rule. □

Given semismoothness of the normal map, we can state the following local convergence result. A proof is given for example in [108] or [97], however, since it is short, we restate it here for completeness.

Theorem 3.18. *Let $\bar{p} \in L^r(\Omega_Q)$ satisfy the first order optimality condition (3.5) and assume that there is a neighbourhood of $\bar{p}$ where $\partial\mathcal{N}$ has a bounded inverse in $B(L^r(\Omega_Q))$ and the assumptions of Lemma 3.16 hold. Then, there is an open ball $B_\varepsilon(\bar{p})$ around $\bar{p}$ in $L^r(\Omega_Q)$ such that for any $p^0 \in B_\varepsilon(\bar{p})$ the semismooth Newton method defined by*

$$p^{k+1} = p^k - \partial\mathcal{N}(p^k)^{-1}\mathcal{N}(p^k) \tag{3.19}$$

converges q-superlinearly towards $\bar{p}$.

Proof. We choose ε such that for fixed $\kappa < 1$ and any $p \in B_\varepsilon(\bar{p})$, the estimate

$$\left\|\partial\mathcal{N}(p)^{-1}\right\|_{B(L^r(\Omega_Q))} \left\|\mathcal{N}(p) - \mathcal{N}(\bar{p}) - \partial\mathcal{N}(p)(p-\bar{p})\right\|_{L^r(\Omega_Q)} \leq \kappa \left\|p - \bar{p}\right\|_{L^r(\Omega_Q)}$$

holds true. This is possible since the first factor on the left-hand side is bounded by assumption and the second one satisfies the defining equation of semismoothness.

Hence, assuming the iterate p^k is contained in $B_\varepsilon(\bar{p})$ and using the fact that $\mathcal{N}(\bar{p}) = 0$, we get for the next iterate

$$\begin{aligned}
\left\|p^{k+1} - \bar{p}\right\|_{L^p(\Omega_Q)} &= \left\|p^k - \partial\mathcal{N}(p^k)^{-1}\mathcal{N}(p^k) - \bar{p}\right\|_{L^r(\Omega_Q)} \\
&\leq \left\|\partial\mathcal{N}(p^k)^{-1}\right\|_{B(L^r(\Omega_Q))} \left\|\partial\mathcal{N}(p^k)(p^k - \bar{p}) - \mathcal{N}(p^k) + \mathcal{N}(\bar{p})\right\|_{L^r(\Omega_Q)} \\
&\leq \kappa \left\|p^k - \bar{p}\right\|_{L^r(\Omega_Q)}
\end{aligned}$$

and hence $p^{k+1} \in B_\varepsilon(\bar{p})$. Since $\kappa < 1$, we have shown linear convergence towards $\bar{p}$. By definition of semismoothness, the contraction factor approaches zero as p^k approaches $\bar{p}$, i. e., the convergence is superlinear. □

To ensure that Theorem 3.18 is applicable, we should discuss the assumption of local boundedness of the inverse of $\partial\mathcal{N}$ around $\bar{p}$ in some more detail. We show that the following second order sufficient optimality condition ensures that this inverse is bounded in a neighbourhood of $\bar{p}$.

Assumption 3.19. There is a constant $\mu > 0$ such that

$$j''(P_{Q_{\mathrm{ad}}}(\bar{p}))(\tau p, \tau p) \geq \mu \|\tau p\|_Q^2 \tag{3.20}$$

holds true for any direction $\tau p \in Q$.

Lemma 3.20. *Let the assumptions of Lemma 3.16 be satisfied and let $\bar{p} \in L^r(\Omega_Q)$ be a point fulfilling the first order optimality condition* (3.5) *and the second order sufficient condition stated in Assumption 3.19. Then there is a $L^r(\Omega_Q)$ neighbourhood of $\bar{p}$ where the inverse of $\partial\mathcal{N}(p)$ is well-defined and bounded in $B(L^r(\Omega_Q))$ for any p in this neighbourhood.*

Remark 3.21. Obviously, Assumption 3.19 is always satisfied for linear-quadratic problems with $\alpha > 0$ by choosing $\mu = \alpha$. Furthermore, we note that the above result can be generalized to weaker second order sufficient optimality conditions. We refer to Tröltzsch [107, Section 4.10] and the references therein for a detailed discussion of such conditions.

Lemma 3.20 is essentially a special case of Lemma 5.9 in [62]. For convenience of the reader we will give a complete proof nevertheless. As a preparation, we need some insights into the structure of the operator $\partial\mathcal{N}$.

Let $p, b \in L^r(\Omega_Q)$ be fixed and consider the equation

$$\partial\mathcal{N}(p)\tau p = b. \tag{3.21}$$

In order to keep the notation simple, we drop the dependencies on p where appropriate. The operator $\partial\mathcal{N} = \partial\mathcal{N}(p)$ has the explicit form

$$\partial\mathcal{N} = \mathrm{Id} + \frac{1}{\alpha} H_{\mathrm{impl}} \chi_{\mathcal{I}}$$

with $H_{\mathrm{impl}} = H_{\mathrm{impl}}(P_{Q_{\mathrm{ad}}}(p))$ and $\chi_{\mathcal{I}} = \chi_{\mathcal{I}}(p)$. If $\chi_{\mathcal{I}} = \mathrm{Id}$, i. e., no constraints are active, then $\partial\mathcal{N} = \frac{1}{\alpha}\nabla^2 j(p)$ is self-adjoint. Together with positive definiteness as required in Assumption 3.19 this means that there is an inverse in $B(Q)$. Clearly, if constraints are active on part of the domain, $\partial\mathcal{N}$ is not self-adjoint. The idea is now to restrict the operator to the inactive set $\mathcal{I} = \left\{x \in \Omega_Q \;\middle|\; q^a(x) < p(x) < q^b(x)\right\}$ to obtain a self-adjoint operator again. Formally, this can be accomplished by moving to a quotient space. As we shall discuss later, the point of view taken here will be useful for the algorithmic realization of the semismooth Newton update.

We define equivalence classes on Q by grouping all functions that take identical values on the current inactive sets, i. e., for given $q \in Q$ we consider

$$[q] = \{\hat{q} \in Q \mid \chi_{\mathcal{I}} q = \chi_{\mathcal{I}} \hat{q} \text{ almost everywhere}\}.$$

The corresponding quotient space can be defined as

$$Q_{\mathcal{I}} = \{[q] \mid q \in Q\}\,.$$

It is easily verified that the bilinear form $(\cdot,\cdot)_{\mathcal{I}}\colon Q_{\mathcal{I}} \times Q_{\mathcal{I}} \to \mathbb{R}$ given by

$$([p_1],[p_2])_{\mathcal{I}} = (p_1, \chi_{\mathcal{I}} p_2)_Q \quad \text{for } p_1, p_2 \in Q$$

is an inner product on $Q_{\mathcal{I}}$. As usual, the corresponding norm is denoted by $\|\cdot\|_{\mathcal{I}}$. Obviously, $Q_{\mathcal{I}}$ with this inner product is isometric to the space $L^2(\mathcal{I})$. We define the operator $[\partial\mathcal{N}]\colon Q_{\mathcal{I}} \to Q_{\mathcal{I}}$ by $[\partial\mathcal{N}][\tau p] = [\partial\mathcal{N}\tau p]$. Noting that $\chi_{\mathcal{I}}$ is a self-adjoint projection operator and hence $(\cdot, \chi_{\mathcal{I}}\cdot)_Q = (\chi_{\mathcal{I}}\cdot, \cdot)_Q = (\chi_{\mathcal{I}}\cdot, \chi_{\mathcal{I}}\cdot)_Q$ and using the fact that H_{impl} is self-adjoint, we verify for $p_1, p_2 \in Q$

$$\begin{aligned}([\partial\mathcal{N}][p_1],[p_2])_{\mathcal{I}} &= (\chi_{\mathcal{I}} \left(\mathrm{Id} + \tfrac{1}{\alpha} H_{\mathrm{impl}}\chi_{\mathcal{I}}\right) p_1, p_2)_Q \\ &= (p_1, \left(\chi_{\mathcal{I}} + \tfrac{1}{\alpha}\chi_{\mathcal{I}} H_{\mathrm{impl}}\chi_{\mathcal{I}}\right) p_2)_Q = ([p_1],[\partial\mathcal{N}][p_2])_{\mathcal{I}},\end{aligned}$$

that is, $[\partial\mathcal{N}]$ is self-adjoint.

The idea is now to construct a solution to the original equation (3.21) from a solution of

$$[\mathcal{N}][\tau p] = [b] \tag{3.22}$$

on the quotient space $Q_{\mathcal{I}}$. Let us assume we are given such a solution $[\tau p]$ with an arbitrary representative $\tau p \in [\tau p]$. Then we set

$$\widetilde{\tau p} = b - \frac{1}{\alpha} H_{\mathrm{impl}}\chi_{\mathcal{I}}\tau p. \tag{3.23}$$

Since τp is evaluated only on the inactive set, $\widetilde{\tau p}$ is uniquely determined. We verify that

$$\chi_{\mathcal{I}}\widetilde{\tau p} = \chi_{\mathcal{I}} b - \chi_{\mathcal{I}}\frac{1}{\alpha} H_{\mathrm{impl}}\chi_{\mathcal{I}}\tau p = \chi_{\mathcal{I}}\tau p$$

and hence $\widetilde{\tau p} \in [\tau p]$. Hence, by construction, $\widetilde{\tau p}$ is the solution of (3.21).

Proof of Lemma 3.20. We first note that due to continuity of the second derivative of j and of the projection $P_{Q_{\mathrm{ad}}}$, there is a neighbourhood of $\bar{p}$ such that condition (3.20) holds for all p from that neighbourhood with some constant $\hat{\mu} < \mu$.

We fix a p from that neighbourhood and check that the corresponding operator $[\partial\mathcal{N}]$ is positive definite by verifying for arbitrary $[\tau p] \in Q_{\mathcal{I}}$ that

$$\begin{aligned}([\tau p],[\partial\mathcal{N}][\tau p])_{\mathcal{I}} &= (\chi_{\mathcal{I}}\tau p, \chi_{\mathcal{I}} \left(\mathrm{Id} + \tfrac{1}{\alpha} H_{\mathrm{impl}}\chi_{\mathcal{I}}\right) \tau p)_Q \\ &= (\chi_{\mathcal{I}}\tau p, \left(\mathrm{Id} + \tfrac{1}{\alpha} H_{\mathrm{impl}}\right) \chi_{\mathcal{I}}\tau p)_Q \\ &= \frac{1}{\alpha} j''(P_{Q_{\mathrm{ad}}}(p))(\chi_{\mathcal{I}}\tau p, \chi_{\mathcal{I}}\tau p) \geq \frac{\hat{\mu}}{\alpha}\,\|\chi_{\mathcal{I}}\tau p\|_Q^2 = \frac{\hat{\mu}}{\alpha}\,\|[\tau p]\|_{\mathcal{I}}^2\,.\end{aligned}$$

Therefore, for any $b \in L^r(\Omega_Q)$, the equation $[\partial\mathcal{N}][\tau p] = [b]$ admits a unique solution in $Q_\mathcal{I}$ satisfying

$$\|[\tau p]\|_\mathcal{I} \leq \frac{\alpha}{\hat{\mu}} \|[b]\|_\mathcal{I} \leq \frac{\alpha}{\hat{\mu}} \|b\|_{L^r(\Omega_Q)}.$$

We set $\widetilde{\tau p}$ as in (3.23). As noted above, $\widetilde{\tau p}$ solves $\partial\mathcal{N}(q)\widetilde{\tau p} = b$ and using the smoothing property of H_{impl}, we can estimate

$$\|\widetilde{\tau p}\|_{L^r(\Omega_Q)} = \left\| b - \frac{1}{\alpha} H_{\text{impl}} \chi_\mathcal{I} \tau p \right\|_{L^r(\Omega_Q)} \leq \|b\|_{L^r(\Omega_Q)} + \frac{1}{\hat{\mu}} \|H_{\text{impl}}\|_{B(Q,L^r(\Omega_Q))} \|b\|_{L^r(\Omega_Q)}.$$

Since the constant $\hat{\mu}$ is independent of the choice of p and H_{impl} is continuous, this shows the claim that the inverse of $\partial\mathcal{N}(p)$ is uniformly bounded in $B(L^r(\Omega_Q))$ for any p from the chosen neighbourhood. □

3.2.2. Algorithmic realization

In this section we discuss how to solve the semismooth Newton update equation and present a heuristics based on the Steihaug conjugate gradient method that enlarges the radius of convergence and makes the optimization procedure more robust. The described algorithm was proposed by Kunisch, Pieper, and Rund (see [64] and [62]).

The Newton update $\tau p^k = p^{k+1} - p^k$ according to (3.19) satisfies

$$\partial\mathcal{N}(p^k)\tau p^k = -\mathcal{N}(p^k). \tag{3.24}$$

As noted in the previous section, a solution of this equation can be found by first solving

$$[\partial\mathcal{N}(p^k)][\widehat{\tau p}] = -[\mathcal{N}(p^k)] \tag{3.25}$$

for $[\widehat{\tau p}] \in Q_\mathcal{I}$ and then picking the right representative by setting

$$\tau p^k = -\mathcal{N}(p^k) - \frac{1}{\alpha} H_{\text{impl}}(p^k)\chi_\mathcal{I}(p^k)\widehat{\tau p} = \widehat{\tau p} - \mathcal{N}(p^k) - \partial\mathcal{N}(p^k)\widehat{\tau p}. \tag{3.26}$$

From an algorithmic point of view, this formulation has the benefit that the operator in (3.25) is self-adjoint positive definite. Therefore we can employ the conjugate gradient (cg) method (see, e. g., Hestenes and Stiefel [53]) for its solution. Arithmetic operations on Ω_Q are realized by performing the corresponding operations on some representative in Q. The only difference to a cg method operating on Q is that we use the inner product $(\cdot,\cdot)_\mathcal{I}$. For the active step (3.26) we point out that the second formulation contains the residual of the Newton update equation for the computed representative $\widehat{\tau p}$, which is evaluated anyways during the cg algorithm.

A semismooth Newton method implemented in this way requires not only that the initial value p^0 is within the radius of convergence, but also that for all iterates, the operator $[\partial\mathcal{N}(p^k)]$ remains positive definite. In order to increase robustness of the optimization

procedure if any of the two conditions is not met, we embed it into a trust region-type algorithm similar to Steihaug conjugate gradient (see Steihaug [105]).

For this purpose, we model the change of the reduced cost functional around the current iterate given by $j(P_{Q_{\mathrm{ad}}}(p^k + \tau q)) - j(P_{Q_{\mathrm{ad}}}(p^k))$ by the quadratic model

$$M_{p^k}(\tau q) = \frac{1}{2}([\tau q], [\partial\mathcal{N}(p^k)][\tau q])_{\mathcal{I}} + ([\tau q], [\mathcal{N}(p^k])_{\mathcal{I}}.$$

Note that the modelling error here consists not only of higher order terms in the Taylor expansion of j but also of the error caused by the change of the active and inactive sets which is not accounted for in the model M_{p^k}.

In every step, we fix a trust region radius Δ_k, compute a descent direction τp^k satisfying $\left\|\tau p^k\right\|_Q \leq \Delta_k$ through a modification of the cg based linear solver outlined above, and compare the actual change $j(P_{Q_{\mathrm{ad}}}(p^k + \tau q^k)) - j(P_{Q_{\mathrm{ad}}}(p^k))$ to the decrease predicted by the model $M_{p^k}(\tau q^k)$. In the usual way, the decision about accepting the step is based on whether the quotient ρ_k of both is larger than a chosen constant. The radius for the next step is computed based on how close ρ_k is to one, i. e., how well the model predicts the behaviour of the cost functional. The details of the algorithm can be seen from the listing in Algorithm 3.1.

In case for an iterate, $\left\|[\mathcal{N}(p^k)]\right\|_{\mathcal{I}}$ becomes zero, we can not expect to achieve a descent for the model M_{p^k}. Hence Algorithm 3.1 fails in this case. This would happen if the current iterate is optimal when fixing the current inactive sets but the inactive sets are not yet correctly determined or if p^k is active everywhere. Obviously this condition arises when the initial iterate is completely active and when the optimal solution is in fact completely active. The first case can be resolved by choosing a proper initial iterate and the latter case can be detected by attempting to perform the active step (3.26). If, in the local optimum, the bounds are active everywhere, the residual vanishes after this step. For other cases for which $\left\|[\mathcal{N}(p^k)]\right\|_{\mathcal{I}}$ becomes zero it is less obvious how to continue the optimization in a meaningful way. During our computations, this issue was not encountered.

What is left to discuss is the computation of the trust region step τp^k. As long as the matrix $[\partial\mathcal{N}(p^k)]$ is positive definite and the iterates of the cg iteration do not leave the trust region, we want to solve the Newton update equation (3.24). Otherwise, the proposed algorithm proceeds similar to the inner solver of Steihaug cg: In case the cg method encounters a direction with non-positive curvature, its last iterate is scaled up to the trust region radius and returned as τp^k. If a step leaves the trust region radius, τp^k is taken as a linear combination of this step and the preceding one such that its norm equals the trust region radius Δ_k. In those cases we do not perform the active step (3.26) since it is only valid if Equation (3.25) holds.

If the modified cg iteration terminates with "cg converged", the active step (3.26) is performed in the end. Since this step produces the same result for any representative of the equivalence class $[\tau p^k]$, it does not matter what the algorithm does on the active set

Algorithm 3.1 Heuristic trust region algorithm

Input: initial value $p^0 \in Q$, initial radius Δ_0

fix $0 < \kappa_0 \leq \kappa_1 < 1 < \kappa_2$, $0 < \eta_1 \leq \eta_2 \leq 1$, $\rho_{\min} > 0$, and $\Delta_{\max} > 0$
for $k = 0, 1, 2, \ldots$ **do**
    compute $\mathcal{N}(p^k)$
    if $\left\|\mathcal{N}(p^k)\right\|_Q$ small enough **then**
        return p^k
    end if
    if $\left\|[\mathcal{N}(p^k)]\right\|_{\mathcal{I}} = 0$ **then**
        abort
    end if
    compute τq^k by means of Algorithm 3.2 with radius Δ_k
    compute $\rho_k = \frac{j(P_{Q_{\mathrm{ad}}}(p^k)) - j(P_{Q_{\mathrm{ad}}}(p^k + \tau q^k))}{-M_{p^k}(\tau q^k)}$
    if $\rho_k < \rho_{\min}$ **then**
        $p^{k+1} = p^k$
        $\Delta_{k+1} = \kappa_0 \min(\left\|\tau q^k\right\|_Q, \Delta_k)$
    else
        $p^{k+1} = p^k + \tau q^k$
        $\delta_k = |\rho_k - 1|$
        $$\Delta_{k+1} = \begin{cases} \min(\kappa_2 \min(\left\|\tau q^k\right\|_Q, \Delta_k), \Delta_{\max}), & \text{if } \delta_k < \eta_1, \\ \min(\left\|\tau q^k\right\|_Q, \Delta_k), & \text{if } \eta_1 \leq \delta_k \leq \eta_2, \\ \kappa_1 \min(\left\|\tau q^k\right\|_Q, \Delta_k), & \text{if } \eta_2 < \delta_k \end{cases}$$
    end if
end for

$\mathcal{A} = \Omega_Q \setminus \mathcal{I}$. Hence, it would be possible as well to set the active part of the iterates x^l to zero and restrict the algorithm to the inactive sets. However, this would introduce jumps at the transition between active and inactive sets into the iterates x^l while the update τq^k is frequently smooth. Compared to that, our algorithm, which performs all operations on the complete iterates, preserves this property also if the cg method terminates preliminary with "non-positive curvature" or "trust region left" and the active step is not performed.

Numerical tests indicate that for many common test problems, after convergence of the cg method proposed here, the difference $x - \tilde{x}$ between the final iterate and the corrected variable $\tilde{x}$ obtained from (3.26) is almost zero. However, a systematic explanation for this observation is still lacking. If it turned out to be substantiated, this would provide another justification for operating on full vectors instead of just the inactive parts during the cg iteration.

We remark that in practice, the tolerance for the cg iteration is not always sufficiently small to ensure the validity of the relation (3.26), hence to increase robustness, one can

Algorithm 3.2 Inner solver for Algorithm 3.1

Input: $b = -\mathcal{N}(p^k)$, $A = \partial\mathcal{N}(p^k)$, Δ

$r^0 = b$
$d^0 = b$
$x^0 = 0$
for $l = 0, 1, \ldots$ **do**
    evaluate Ad^l
    if $([d^l], [A][d^l])_{\mathcal{I}} \leq 0$ **then**
        $x^{l+1} = x^l + \theta d^l$ such that $\left\|x^{l+1}\right\|_Q = \Delta$ {Go to boundary}
        return $\tau q = x^{l+1}$, "non-positive curvature"
    end if
    $\beta_l = \frac{\left\|[r^l]\right\|_{\mathcal{I}}^2}{([d^l],[A][d^l])_{\mathcal{I}}}$
    if $\left\|x^l + \beta_l d^l\right\|_Q \geq \Delta$ **then**
        $x^{l+1} = x^l + \theta d^l$ such that $\left\|x^{l+1}\right\|_Q = \Delta$
        return $\tau q = x^{l+1}$, "trust region left"
    end if
    $x^{l+1} = x^l + \beta_l d^l$
    $r^{l+1} = r^l - \beta_l A d^l$
    if tolerance reached **then**
        return $\tau q = x^{l+1} + b - Ax^{l+1}$, "cg converged"
    end if
    $d^{l+1} = r^{l+1} + \frac{\left\|[r^{l+1}]\right\|_{\mathcal{I}}^2}{\left\|[r^l]\right\|_{\mathcal{I}}^2} d^l$
end for

perform a line search in the direction of the residual instead of adding it with factor one to x. The cost functional for the line search is the norm of the final residual $b - A\tilde{x}$.

4. Discretization

So far we considered our optimal control problem in the continuous setting which in general is infinite-dimensional. For the numerical solution, we replace it by a finite dimensional approximation. To ensure that the discretized optimality conditions for the continuous problem are the same as the optimality conditions arising for the discretized optimal control problem, i. e., *"discretize-then-optimize=optimize-then-discretize"* holds true, we use a Galerkin-type discretization for the state variable in both, space and time. Our approach for the treatment of the state variable follows mostly Meidner and Vexler [76, 78, 80] with a discontinuous Galerkin discretization in time and standard conforming finite elements in space. However, we allow the order of the time discretization to vary over time enabling *hp* adaptivity.

For the treatment of the control variable, we discuss two possible approaches, the *variational approach* first proposed by Hinze, see [55], which avoids a separate control discretization, and an explicit discretization matching the discretization of the state variable. The first approach will be used for the almost-third order scheme discussed in Chapter 6, while for our exploration of *hp* adaptivity in Chapter 7, we will discretize the control explicitly.

4.1. Discretization of the state variable

The state variable is discretized according to Rothe's method. This means that first, the time dimension is discretized resulting in a *semidiscrete* formulation which consists of a system of continuous elliptic equations in space. Subsequently those equations are discretized on the spatial domain resulting in a fully discrete problem. Compared to the *method of lines*, i. e., first discretizing with respect to the spatial domain and then discretizing the resulting ordinary differential equation (ODE) system in time, this has the obvious benefit that the spatial discretization can be varied over time. This feature is essential for an efficient adaptive discretization of, e. g., travelling fronts.

4.1.1. Semidiscretization in time with *hp* discontinuous Galerkin methods

To discretize the optimization problem (2.5) in time direction, we consider the discontinuous Galerkin method. Due to continuity of the exact solution space, it is obviously a non-conforming discretization.

The first published application of discontinuous Galerkin methods was the (spatial) discretization of the neutron transport equation, see Lesaint and Raviart [67]. The idea of using them for time discretization of parabolic PDEs can be traced back to a work by Jamet [60] in 1978. It is interesting to note that in this early work, the primary motivation for a discontinuous time discretization was the desire to vary the space discretization over time. A series of papers by Eriksson, Johnson, and coworkers [32–37] subsequently built a systematic theory of discontinuous Galerkin time stepping including a priori and a posteriori error estimation. For a detailed overview of the historical development of the method, we refer to the survey article [25] by Cockburn et al.

To define the time discretization, we partition the time interval $I = (0, T)$ by the temporal nodes

$$0 = t_0 < t_1 < t_2 < \cdots < t_{M-1} < t_M = T$$

into open intervals $I_m = (t_{m-1}, t_m)$ for $m = 1, \dots, M$. We denote by $k \in \mathbb{R}^M$ the vector of time step sizes $k_m = |I_m|$. By abuse of notation we refer to the discretization parameter, i. e., the maximum of all time steps by k as well. Since we want to allow for varying the order of discretization over time, we introduce the order vector $r \in \mathbb{N}_0^M$ that assigns a polynomial order $r_m \in \mathbb{N} \cup \{0\}$ to each time interval I_m. Then, the semidiscrete test and trial space is given by

$$X_k^r = \left\{ v \in L^2(I, V) \;\middle|\; v|_{I_m} \in \mathcal{P}_{r_m}(I_m, V) \right\},$$

where $\mathcal{P}_{r_m}(I_m, V)$ denotes the space of polynomials of maximal degree r_m over I_m with values in V. For functions $v \in X_k^r$ we introduce the abbreviations

$$v(t)^- = \lim_{\varepsilon \searrow 0} v(t - \varepsilon), \qquad v(t)^+ = \lim_{\varepsilon \searrow 0} v(t + \varepsilon),$$
$$v_m^{+/-} = v(t_m)^{+/-}, \qquad \text{and} \quad [v]_m = v_m^+ - v_m^-.$$

We adopt the convention that for some $j \in \mathbb{Z}$ and an order vector r, the notation $r + j$ indicates that each component of r is increased by j. Similarly, replacing r by a single natural number indicates a discretization with constant order.

The standard semidiscrete formulation for the state equation reads: given $q_k \in Q$, find $u_k \in X_k^r$, such that

$$\begin{aligned} (4.1) \quad \sum_{m=1}^{M} (\partial_t u_k, \varphi)_{I_m} + \sum_{m=1}^{M-1} ([u_k]_m, \varphi_m^+) + \int_0^T a(q_k, u_k)(\varphi)\, \mathrm{d}t \\ + (u_{k,0}^+, \varphi_0^+) = (u_0(q_k), \varphi_0^+) \quad \text{for any } \varphi \in X_k^r. \end{aligned}$$

We note that, compared to the state equation (2.3) on the continuous level, jump terms were added to account for the discontinuities of the solution u_k at the interval boundaries. For a linear state equation with coercive bilinear form, the existence of a unique solution to the semidiscrete equation was shown for example in [100, Proposition 1.7]. Existence of a solution of the semidiscrete state equation for a certain class of semilinear equations

was shown in [85] in the case $r_m = 0$. However, the proof can be generalized also to higher orders. Note that both quoted results do not impose any restrictions on the size of the time step.

To state the semidiscrete optimization problem, we replace the state space X by X_k^r and the state equation by its semidiscrete equivalent resulting in

$$\text{(4.2)} \qquad \text{Minimize } J(q_k, u_k) \text{ subject to } \begin{cases} (q_k, u_k) \in Q \times X_k^r \text{ satisfying (4.1)}, \\ q_k \in Q_{\mathrm{ad}}. \end{cases}$$

To derive the optimality system for this problem, we formulate the semidiscrete Lagrangian $\hat{\mathcal{L}}\colon Q \times X_k^r \times X_k^r \to \mathbb{R}$

$$\begin{aligned} \text{(4.3)} \quad \hat{\mathcal{L}}(q_k, u_k, z_k) = J(q_k, u_k) + (u_0(q_k), z_{k,0}^+) - (u_{k,0}^+, z_{k,0}^+) \\ - \sum_{m=1}^{M} (\partial_t u_k, z_k)_{I_m} - \int_0^T a(q_k, u_k)(z_k)\,\mathrm{d}t - \sum_{m=1}^{M-1} ([u_k]_m, z_{k,m}^+). \end{aligned}$$

With the same reasoning as for the derivation of the continuous optimality system (3.11), we obtain that for a local optimum, the triple $(\bar{q}_k, \bar{u}_k, \bar{z}_k) \in Q \times X_k^r \times X_k^r$ satisfies

$$\text{(4.4a)} \qquad \hat{\mathcal{L}}_z'(\bar{q}_k, \bar{u}_k, \bar{z}_k)(\varphi) = 0 \qquad \text{for all } \varphi \in X_k^r,$$

$$\text{(4.4b)} \qquad \hat{\mathcal{L}}_u'(\bar{q}_k, \bar{u}_k, \bar{z}_k)(\varphi) = 0 \qquad \text{for all } \varphi \in X_k^r,$$

$$\text{(4.4c)} \qquad \hat{\mathcal{L}}_q'(\bar{q}_k, \bar{u}_k, \bar{z}_k)(q_k - \bar{q}_k) \geq 0 \qquad \text{for all } q_k \in Q_{\mathrm{ad}}.$$

Condition (4.4a) is the semidiscrete state equation again. For the adjoint equation (4.4b) of the semidiscrete problem we obtain after interval-wise integration by parts: given $q_k \in Q$ and $u_k \in X_k^r$, find $z_k \in X_k^r$, such that for all $\varphi \in X_k^r$

$$\begin{aligned} \text{(4.5)} \quad - \sum_{m=1}^{M} (\partial_t z_k, \varphi)_{I_m} - \sum_{m=1}^{M-1} ([z_k]_m, \varphi_m^-) + \int_I a_u'(q_k, u_k)(z_k, \varphi)\,\mathrm{d}t + (z_{k,M}^-, \varphi_M^-) = \\ J_1'(u_k)(\varphi) + J_2'(u_{k,M}^-)(\varphi_M^-). \end{aligned}$$

As expected, this is precisely the discontinuous Galerkin discretization of the continuous adjoint equation (3.8). The gradient condition is the same as in the continuous case apart from the fact that all continuous quantities are replaced by their semidiscrete counterparts. Hence we see that the approaches "discretize-then-optimize" and "optimize-then-discretize" commute.

The reformulation of the optimality condition in terms of the unprojected variable p_k and the normal map yields the same condition as (3.12c) with all continuous quantities replaced by the corresponding semidiscrete quantities. In the same way, the representation formula (3.15) for the second derivative of the reduced cost functional can be translated. The necessary auxiliary equations are obtained by semidiscretization of their continuous

counterparts (3.13) and (3.14). Their explicit forms can be found for example in [12] or [76].

For the a priori analysis in Chapter 6 and also for the evaluation of the a posteriori error indicators presented in Chapter 7, we need certain temporal interpolation and reconstruction operators. On each discretization interval I_m, they are defined in terms of the nodes of the $r_m + 1$ point Radau quadrature rule (see, e. g., Abramowitz and Stegun [1, p. 888]), which we subsequently refer to as *Radau points*. We will make use of the operators to exploit superconvergence properties of the discontinuous Galerkin methods at these points. Since the adjoint equation is formulated backwards in time, also the interpolation nodes have to be reversed such that we get two different versions of each operator—one for the state solution and one for the adjoint solution.

To give a characterization of the Radau points, we define the $(r_m + 1)^{\text{th}}$ *right Radau polynomial* $\mathcal{R}_{r_m+1}$ on the unit interval $[0, 1]$ by

$$\mathcal{R}_{r_m+1}(\tau) = \mathcal{L}_{r_m+1}(\tau) - \mathcal{L}_{r_m}(\tau),$$

where $\mathcal{L}_j$ is the Legendre polynomial of degree j. We denote the $r_m + 1$ roots of this polynomial by $\tau_0^{r_m}, \dots, \tau_r^{r_m} = 1$. Transforming them onto the interval I_m gives the points $\theta_{m,j}^{\mathrm{S}} := t_{m-1} + k_m \tau_j^{r_m}$ for j in $0, \dots, r_m$.

With these preparations, we can define the interpolation operator $\pi_k^{\mathrm{S}} \colon C(\bar{I}, V) \to X_k^r$ given interval-wise by

$$\pi_k^{\mathrm{S}} v(\theta_{m,j}^{\mathrm{S}})^- = v(\theta_{m,j}^{\mathrm{S}}) \quad \forall\ j = 0, \dots, r_m,\ m = 1, \dots, M.$$

Besides, we introduce a reconstruction operator into a continuous space with higher polynomial order. Since the reconstruction depends on the initial value, which we did not include in the definition of the semidiscrete space, the operator takes it as an additional argument. So the reconstruction is given by $\hat{\pi}_k^{\mathrm{S}} \colon V \times \left(X_k^r \cup C(\bar{I}, V)\right) \to X_k^{r+1} \cap C(\bar{I}, V)$ satisfying the conditions

$$\begin{aligned} \hat{\pi}_k^{\mathrm{S}}(v_0, v)(\theta_{m,j}^{\mathrm{S}}) &= v(\theta_{m,j}^{\mathrm{S}})^- \quad \forall j = 0, \dots, r_m,\ m = 1, \dots, M, \quad \text{and} \\ \hat{\pi}_k^{\mathrm{S}}(v_0, v)(0) &= v_0. \end{aligned}$$

The operator $\hat{\pi}_k^{\mathrm{S}}$ has been used previously, for example by Akrivis et al. [3], Schötzau and Wihler [102], and Fidkowski [40] to construct a posteriori error estimates or by Adjerid [2] and Matthies and Schieweck [73] to recover improved solutions from the computed ones. In Figure 4.1, the effect of the operators π_k^{S} and $\hat{\pi}_k^{\mathrm{S}}$ for fixed order $r = 1$ is visualized.

For the adjoint operators, we use the time-reversed nodes $\theta_{m,j}^{\mathrm{D}} := t_m - k_m \tau_j^{r_m}$ with $j = 0, \dots, r_m$ instead. Correspondingly, the adjoint interpolation operator $\pi_k^{\mathrm{D}} \colon C(\bar{I}, V) \to X_k^r$ fulfills

$$\pi_k^{\mathrm{D}} v(\theta_{m,j}^{\mathrm{D}})^+ = v(\theta_{m,j}^{\mathrm{D}}) \quad \forall\ j = 0, \dots, r_m,\ m = 1, \dots, M,$$

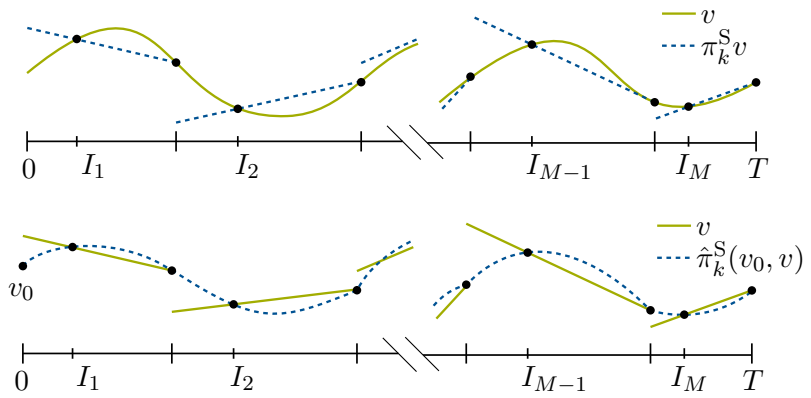


Figure 4.1.: Visualization of the operators π_k^S and $\hat{\pi}_k^S$ for $r = 1$.

and the adjoint reconstruction operator $\hat{\pi}_k^D \colon V \times \left(X_k^r \cup C(\bar{I}, V)\right) \to X_k^{r+1} \cap C(\bar{I}, V)$ is given by

$$\hat{\pi}_k^D(v_T, v)(\theta_{m,j}^D) = v(\theta_{m,j}^D)^+ \quad \forall j = 0, \dots, r_m,\ m = 1, \dots, M, \quad \text{and}$$
$$\hat{\pi}_k^D(v_T, v)(T) = v_T.$$

4.1.2. Discretization in space with continuous elements on dynamic meshes

With respect to the spatial domain, semidiscrete functions from the space X_k^r are still infinite-dimensional, so in order to get a fully discrete state variable, we approximate the space V by a discrete subspace. Since the space X_k^r allows for discontinuities at the temporal nodes, it adds no theoretical difficulty if we allow for different discrete subspaces $V_m \subseteq V$ for each time interval I_m. Here, we limit our considerations to spaces of piecewise polynomials with fixed order on grids consisting of intervals, quadrilaterals, or hexahedrons respectively.

We first discuss how to construct an appropriate triangulation of the domain Ω. For simplicity we consider only the case of a polyhedral domain $\Omega \subseteq \mathbb{R}^d$, where $d \in \{1, 2, 3\}$. This restriction ensures that that the domain can be covered by a polyhedral triangulation. Techniques for treating domains with curved boundaries are discussed, e. g. by Braess [19]. We partition the domain into open intervals, quadrilaterals, or hexahedrons (more precisely convex cuboids) respectively in one, two, and three dimensions. Subsequently we refer to these geometric primitives as *cells* denoted by $K \subseteq \Omega$. For the resulting triangulation, we write

$$\mathcal{T}_h = \{K\}$$

where the mesh parameter h is given as the maximal cell diameter.

We need to impose some assumptions on the triangulation. Since the adaptive refinement procedure considered in Chapter 7 introduces *hanging nodes*, we need to weaken the

canonical definition of a regular mesh (see, e. g., Braess [19] or Ciarlet [24]), which does not allow for hanging nodes. Additionally, to evaluate a posteriori error indicators, the mesh needs to possess a patch structure.

To state the modified assumptions on the mesh rigorously, we define for two cells $K_i, K_j \in \mathcal{T}_h$ with $\overline{K}_i \cap \overline{K}_j \neq \emptyset$ the affine space

$$M_{ij} = \left\{ \lambda x + (1-\lambda) y \;\middle|\; \lambda \in \mathbb{R},\ x, y \in \overline{K}_i \cap \overline{K}_j \right\}.$$

Furthermore we call a collection of 2^d disjoint cells K_i a refinement of a cell K if $\bigcup_{i=1}^{2^d} \overline{K}_i = \overline{K}$ and for each i, $\overline{K}_i$ contains exactly one corner of the cell K.

Assumption 4.1. For a triangulation $\mathcal{T}_h$ to be admissible, it has to satisfy the following conditions.

1. $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} \overline{K}$.
2. For $K_0, K_1 \in \mathcal{T}_h$ with $K_0 \neq K_1$, we have $K_0 \cap K_1 = \emptyset$.
3. If $\overline{K}_0 \cap \overline{K}_1 \neq \emptyset$, then at least one of the conditions $\overline{K}_0 \cap M_{01} \subseteq \overline{K}_1$ or $\overline{K}_1 \cap M_{01} \subseteq \overline{K}_0$ holds true.
4. Let $\overline{K}_0 \cap \overline{K}_1 \neq \emptyset$, $d_M := \dim M_{01} > 0$, and $\overline{K}_1 \cap M_{01} \subsetneq \overline{K}_0 \cap M_{01}$. Let furthermore F_0 denote the relative interior of $\overline{K}_0 \cap M_{01}$ in M_{01}. We note that F_0 can be regarded as a d_M-dimensional cell in M_{01}. Then, for $l = 2^{d_M}$, we assume there are $l-1$ cells $K_2, \ldots, K_l \in \mathcal{T}_h$ such that $F_1, \ldots, F_l$ defined in the same way as F_0 above from $K_1, \ldots, K_l$ form a refinement of F_0.
5. The mesh possesses a patch structure, that is, there is a coarser mesh $\mathcal{T}_{2h}$ satisfying Assumptions 1 to 4 such that $\mathcal{T}_h$ can be obtained as a global refinement of $\mathcal{T}_{2h}$.

Remark 4.2. If we modify the third assumption to require both, $\overline{K}_0 \cap M_{01} \subseteq \overline{K}_1$ and $\overline{K}_1 \cap M_{01} \subseteq \overline{K}_0$, this results in the requirement that whenever edges or faces of cells have common points, the cells share the whole edge or face respectively. Together with Assumptions 1 and 2, this is equivalent to the canonical definition of a regular mesh. In order to allow for adaptive mesh refinement, we relaxed this requirement. Condition 4 permits to refine cells one more time than their neighbours, resulting in hanging nodes, i. e., degrees of freedom on the boundary of a cell that do not possess a counterpart on the neighbouring cell. For standard first order elements, these degrees of freedom correspond to cell corners that lie in the relative interior of edges or faces of neighbouring cells.

A collection of cells of $\mathcal{T}_h$ that together form a refinement of a cell in the coarse mesh $\mathcal{T}_{2h}$ as defined in the last condition of the assumption is called a *patch*.

The local approximation spaces on the individual cells are constructed by transformation from a reference cell. For given $s \in \mathbb{N}$, we define the polynomial space $\hat{\mathcal{Q}}_s(\hat{K})$ on the reference cell $\hat{K} = (0,1)^d$ by

$$\hat{\mathcal{Q}}_s(\hat{K}) = \operatorname{span}\left\{ x \mapsto \prod_{j=1}^{d} x_j^{\alpha_j} \;\middle|\; \alpha_j \leq s,\ j = 1, \dots, d \right\}.$$

Let $T_K \colon \hat{K} \to K$ be a linear, bilinear, or trilinear transformation respectively, i. e., $T_K \in \hat{\mathcal{Q}}_1(\hat{K})^d$, that maps $\hat{K}$ to K. Then the local approximation space on K is given by

$$\mathcal{Q}_s(K) = \left\{ v \colon K \to \mathbb{R} \;\middle|\; v \circ T_K \in \hat{\mathcal{Q}}_s(\hat{K}) \right\}.$$

Due to the symmetry properties of the reference cell, this space is well defined although the choice of T_K is not unique. Note that the elements of $\mathcal{Q}_s(K)$ are in general not polynomials but rational functions. However their restrictions to the edges of K are polynomials of maximal order s.

For the standard spatial interpolation estimates that the a priori results for the spatial discretization quoted in Chapter 6 are based on to hold, a regularity condition for the mesh is needed. Here, we require that all cells can be obtained by affine transformations of the reference cell, i. e., we restrict ourselves to parallelogram- or parallelepiped-shaped cells respectively. With this restriction it is sufficient to assume that there is a constant $\rho > 0$ such that uniformly for all $K \in \mathcal{T}_h$, the condition

$$\frac{1}{\rho} \leq \frac{|\det T_K'|}{|K|} \leq \rho$$

holds true. Conditions for more general cell shapes along with the corresponding interpolation estimates can be found, e. g., in Apel [4, Section 2.4f.].

With these preparations, we can define the V-conforming and continuous finite element space V_h^s for some polynomial order $s \in \mathbb{N}$ by

$$V_h^s = \left\{ v \in H_0^1(\Omega) \cap C(\Omega) \;\middle|\; v|_K \in \mathcal{Q}_s(K),\ K \in \mathcal{T}_h \right\}.$$

The continuity requirement on the finite element functions implies that there are no degrees of freedom associated with the hanging nodes since the corresponding values are determined by the values of the finite element function on the coarser cell.

For each time interval I_m, we fix a triangulation $\mathcal{T}_h^m$ with associated finite element space $V_{h,m}^s$. Then the fully discrete space for the state variable reads

$$X_{k,h}^{r,s} = \left\{ v \in L^2(I, V) \;\middle|\; v|_{I_m} \in \mathcal{P}_{r_m}(I_m, V_{h,m}^s) \right\}.$$

Obviously $X_{k,h}^{r,s} \subseteq X_k^r$ and therefore the discrete state equation can be obtained from the semidiscrete formulation by restricting test and trial space. Hence we have: given $q_{kh} \in Q$, find $u_{kh} \in X_{k,h}^{r,s}$ such that

$$\sum_{m=1}^{M} (\partial_t u_{kh}, \varphi)_{I_m} + \sum_{m=1}^{M-1} ([u_{kh}]_m, \varphi_m^+) + \int_0^T a(q_{kh}, u_{kh})(\varphi)\,\mathrm{d}t + (u_{kh,0}^+, \varphi_0^+) = (u_0(q_{kh}), \varphi_0^+) \quad \text{for any } \varphi \in X_{k,h}^{r,s}. \tag{4.6}$$

Apart from the jump terms, all terms can be evaluated on each time step individually. So for practical realization of dynamic meshes, only the capability of evaluating H inner products of functions residing on different meshes is needed. For further details on the practical realization, we refer to Schmich [98]. To discuss the existence of solutions to the fully discrete equation, one can apply the results mentioned for the semidiscrete case. Janssen and Wihler [61] show existence of solutions for fully discrete problems under more general assumptions on the continuous problem given that the time steps are chosen sufficiently small. However we point out that for second order parabolic problems their approach requires a CFL condition.

The optimization problem with fully discrete state reads

$$\text{Minimize } J(q_{kh}, u_{kh}) \text{ subject to } \begin{cases} (q_{kh}, u_{kh}) \in Q \times X_{k,h}^{r,s} \text{ satisfying (4.6)}, \\ q_{kh} \in Q_{\mathrm{ad}}. \end{cases} \tag{4.7}$$

The Lagrangian for this problem is identical to the one for the semidiscrete problem, we only restrict the spaces for state and adjoint equation to $X_{k,h}^{r,s}$ resulting in a map $\hat{\mathcal{L}}\colon Q \times X_{k,h}^{r,s} \times X_{k,h}^{r,s} \to \mathbb{R}$. Therefore, all optimality conditions and derivative representations can be transferred from the semidiscrete setting by restricting all occurrences of X_k^r to $X_{k,h}^{r,s}$ and replacing all variables by their discrete counterparts. In particular, we refer to the adjoint state and the unprojected control for the discrete problem as $z_{kh} \in X_{k,h}^{r,s}$ and $p_{kh} \in Q$.

4.2. Control discretization

In the previous section we only discretized the state variable, the control remained an element of the original space Q. Assuming that this space was not finite dimensional to begin with, we still have to discuss how to treat the control. In the first part of this section, we discuss the so called variational approach due to Hinze. It is based on the observation that for certain problems, after moving to a discrete state variable, a control satisfying the corresponding gradient condition has simple enough structure to be treated computationally without separate discretization. A more conventional treatment of the control by separate discretization is presented in the second part. For both approaches we remark on the implications for the realization of the semismooth Newton algorithm.

4.2.1. Variational treatment of the control

The variational approach proposed by Hinze [55] is applicable to problems where the control enters the state equation linearly. For the parabolic setting considered here, that means the state equation (2.3) takes the form

(4.8)
$$\int_0^T \langle \partial_t u, \varphi\rangle_{V^*\times V} + \hat{a}(u)(\varphi)\,\mathrm{d}t + (u(0),\varphi(0)) = (u_0 + B_0 q, \varphi(0)) + (B_1 q, \varphi)_I \quad \text{for any } \varphi \in X.$$

with a semilinear form $\hat{a}\colon X \times X \to \mathbb{R}$, initial value $u_0 \in H$, and linear operators $B_0 \in B(Q,H)$ and $B_1 \in B(Q, L^2(I,H))$. Furthermore we assume $\alpha > 0$. Then the implicit part of the gradient for the problem with discretized state reads

$$G_{\mathrm{impl}}(q_{kh}, u_{kh}, z_{kh}) = B_0^* z_{kh,0}^+ + B_1^* z_{kh}.$$

Since $z_{kh} \in X_{k,h}^{r,s}$, we have $G_{\mathrm{impl}}(q_{kh}, u_{kh}, z_{kh}) \in Q_{\mathrm{var}}$ with the linear space $Q_{\mathrm{var}} = \operatorname{span}\left(B_0^*(V_{h,1}^s) \cup B_1^*(X_{k,h}^{r,s})\right)$. As image of discrete spaces, this space is discrete in itself. The discrete version of the optimality condition (3.12c) with $\gamma_G = \frac{1}{\alpha}$ gives

$$\bar{p}_{kh} = -\frac{1}{\alpha}\left[B_0^* \bar{z}_{kh,0}^+ + B_1^* \bar{z}_{kh}\right],$$

that is, $\bar{p}_{kh}$ is a discrete quantity from the space Q_{var}. Hence the necessary optimality condition for the problem with discrete state equation is equivalent to $(\bar{p}_{kh}, \bar{u}_{kh}, \bar{z}_{kh}) \in Q_{\mathrm{var}} \times X_{k,h}^{r,s} \times X_{k,h}^{r,s}$ satisfying the discrete version of the optimality system (3.12). This is a fully discrete problem, however for the practical realization of Algorithm 3.1, it has to be ensured that the projection operator $P_{Q_{\mathrm{ad}}}$ and the characteristic function $\chi_{\mathcal{I}}$ of the inactive sets can be evaluated on Q_{var}. Considering neither $P_{Q_{\mathrm{ad}}}(Q_{\mathrm{var}})$ nor $\chi_{\mathcal{I}}(Q_{\mathrm{var}})$ are contained in discrete spaces, an exact realization of those two operators is only tractable if the control constraints q^a and q^b admit a suitable discrete representation and if the space Q_{var} has simple enough structure. In the case of high order discretizations or complicated operators B_0 and B_1 the implementation effort for an exact projection can be prohibitive. On the other hand, as we will see in Chapter 6, the variational approach can achieve better convergence rates than a discretized control in certain settings.

To illustrate the practical realization of variational control, we discuss it for the linear example problem with time-dependent parameter control presented in Section 2.3.1 which will also be analyzed in Chapter 6.

In this case, we have $B_0 = 0$ and $B_1 = G^q$. Therefore it can be seen easily that the space Q_{var} takes the form

$$Q_{\mathrm{var}} = \left\{ p \in Q \;\middle|\; p|_{I_m} \in \mathcal{P}_{r_m}(I_m, \mathbb{R}^{d_Q}) \right\}.$$

We require $q^a, q^b \in Q_{\mathrm{var}}$. In order to evaluate the normal map, the temporal integral $(G^q P_{Q_{\mathrm{ad}}}(p_{kh}), \varphi)_I$ has to be evaluated when solving the state equation. If the control

constraints are not active anywhere (and hence $P_{Q_{\mathrm{ad}}} = \mathrm{Id}$), then this is a temporal integral over piecewise polynomials on each discretization interval. It can be evaluated exactly when using sufficiently high order quadrature formulas on each interval. If the constraints change from active to inactive within a discretization interval, then in general, $P_{Q_{\mathrm{ad}}}(p_{kh})$ will have kinks where the polynomial $p_{kh}|_{I_m}$ is cut off. Hence, also high order quadrature formulas only deliver a first order approximation.

As a resolution, we identify the locations of the kinks and partition the discretization interval for numerical quadrature purposes by those points. Since the d_Q components of $p_{kh} - q^a$ and $q^b - p_{kh}$ are scalar polynomials, this amounts to finding the roots of those polynomials. The resulting partition of the discretization intervals can also be used for computing all temporal integrals containing the characteristic function $\chi_{\mathcal{I}}(p_{kh})$ that occur within the linear solver of the semismooth Newton algorithm. Naturally, with increasing order of discretization, computation of the partition becomes more involved. However, due to the regularity of the state and adjoint variables, which is limited by the presence of control constraints, very high orders of discretization are of limited use anyway. In Chapter 2.3.1 we will show that a piecewise linear discretization in time is sufficient to achieve the optimal order of convergence implied by the regularity of the adjoint state. The implementation effort for the variational approach is modest in this setting.

We point out that the quadrature formula resulting from subdividing the discretization intervals at the boundaries between active and inactive sets may only be used for integrating discrete quantities and not problem data. Otherwise during the course of the optimization, we effectively use a different interpolant of the problem data whenever the active sets change. That means the discrete optimization problem changes. As a result, the optimization procedure may not converge at all or converge slower. For more complicated settings than the considered test problem, where integrals of data and control can not be separated, one can work with a projection of the data onto the appropriate discrete space instead.

Remark 4.3. For the usual approach to semismooth Newton working on the control, a realization requires storing an explicit representation for the projected control variable q_{kh} and for the corresponding active sets (see Hinze and Vierling [58] for details). By contrast, our algorithm needs to store only discrete quantities explicitly while active sets and projected control are only needed as data for discretized PDEs. This makes it a very natural approach for the variational concept.

4.2.2. Explicit discretization of the control

As seen in the previous section, the variational treatment of the control is restricted to problems where the control enters the state equation linearly. Furthermore, it greatly increases complexity of the practical realization, in particular when higher order discretizations are employed. These limitations do not arise when the control space is discretized explicitly. On the other hand, an additional discretization error is introduced

and, since kinks are not resolved, the order of convergence on the boundary between active and inactive sets is restricted. However, this shortcoming can be addressed by an adaptive algorithm that takes the error resulting from the control discretization into account. Therefore, in the context of *hp* adaptivity, where feasibility of the implementation of higher discretization orders is required, explicit discretization of the control is the method of choice.

The resulting fully discrete optimal control problem takes the form

$$\text{(4.9)}\quad \text{Minimize } J(q_{khd}, u_{khd}) \text{ subject to } \begin{cases} (q_{khd}, u_{khd}) \in Q_d \times X_{k,h}^{r,s} \text{ satisfying (4.6)}, \\ q_{khd} \in Q_{d,\mathrm{ad}} \end{cases}$$

with $Q_d \subseteq Q$ the discrete control space and $Q_{d,\mathrm{ad}}$ a suitable discrete admissible set. As we will see later, it is not always advantageous to set $Q_{d,\mathrm{ad}}$ as the intersection of the continuous admissible set Q_{ad} and the discrete space Q_d.

Obviously, the choice of the discrete control space Q_d depends on the structure of the continuous control space Q. Subsequently we discuss control discretization for the case of a control consisting of a set of time-dependent parameters, that is, for a control space of the form $Q = L^2(I, \mathbb{R}^{d_Q})$. Most of the considerations presented apply analogously to the case of a spatially distributed control, which we briefly discuss at the end of this section.

For discretization of the space Q of time-dependent parameters, we use the same approach as for semidiscretization of the state in time. Let $I_1, \dots, I_{M^Q}$ denote a partition of the time interval I with associated step size vector $k^Q \in \mathbb{R}^{M^Q}$ and order vector $r^Q \in \mathbb{N}_0^{M^Q}$. Then the discrete control space is defined as

$$Q_d = Q_{k^Q}^{r^Q} = \left\{ q \in Q \;\middle|\; q|_{I_m} \in \mathcal{P}_{r_m^Q}(I_m, \mathbb{R}^{d_Q}) \text{ for } m = 1, \dots, M^Q \right\}.$$

In principle, the discretization parameters k^Q and r^Q for the control discretization can be chosen independent from the respective parameters k and r for the state. However, for simplicity of implementation, we only consider an identical discretization, i. e., $k^Q = k$ and $r^Q = r$.

Next, we have to specify the discrete admissible space $Q_{d,\mathrm{ad}}$. The obvious choice would be $Q_d \cap Q_{\mathrm{ad}}$. However, this option leads to problems with our semismooth Newton algorithm. This is because the latter relies on the reformulation of the optimality condition (3.2) in terms of the L^2 projection onto the admissible set. While in the continuous setting, the projection $P_{Q_{\mathrm{ad}}}$ can be written as a superposition operator acting pointwise, this is true in the discrete setting only for $r = 0$.

To see that, let us consider a model configuration for the case $r = 1$. We consider the space of linear polynomials on the unit interval $\mathcal{P}_1((0,1))$ and the admissible set $\{q \in \mathcal{P}_1((0,1)) \mid q \geq 0\}$. Then the L^2 projection of the polynomial $p\colon x \mapsto 2x - 1$ to the admissible set is $x \mapsto \frac{1}{2}x$, as opposed to the nodal projection, which would be $x \mapsto x$.

While for $r = 1$, a simple algebraic representation of the discrete admissible set is still possible by using a nodal basis with the left and right boundaries of the intervals as nodes and imposing the constraints on the nodal values, the same is not true for higher orders. Rewriting $Q_{d,\mathrm{ad}} = Q_d \cap Q_{\mathrm{ad}}$ in terms of a finite number of constraints on a representation vector of an element of Q_d results in non-linear constraints. For orders of $r = 6$ and greater it is even impossible to derive a closed form for those constraints since that would amount to finding extreme values of polynomials of degree 6 and greater.

As a conclusion, an implementation of the semismooth Newton algorithm from Chapter 3 for explicit control discretization with $Q_{d,\mathrm{ad}} = Q_d \cap Q_{\mathrm{ad}}$ and high discretization orders is complicated. This runs contrary to the main motivation behind discretizing the control explicitly, which was to allow for a simple realization. Hence, we propose to weaken the admissibility condition on the discrete control. This will allow us to retain the simple structure of the continuous L^2 projection also in the discrete setting. A potential drawback is that the resulting discrete optimal control is in general not admissible for the continuous problem. However, in Chapter 7, we will show that the error resulting from that can be controlled within an adaptive algorithm. If admissibility of the computed solution is a concern, it can be enforced by a post-processing step, i. e., we can use the exact pointwise projection of the computed optimal unprojected control $\bar{p}_{khd}$ as the final result.

For $m = 1, \dots, M$, let the nodes of the $(r_m + 1)$-point Gauß-Legendre quadrature rule on the interval I_m be denoted by

$$t^L_{m,j} \quad \text{with } j = 0, \dots, r_m.$$

Then we define the discrete admissible set by enforcing the bounds only at these nodes, i e., we set

$$Q_{d,\mathrm{ad}} = \left\{ q \in Q_d \;\middle|\; q^a(t^L_{m,j}) \le q(t^L_{m,j}) \le q^b(t^L_{m,j}) \quad \text{for } m = 1, \dots, M \text{ and } j = 0, \dots, r_m \right\}.$$

We use the Lagrange polynomials $\{\psi_{m,j}\}$ corresponding to the Legendre nodes $t^L_{m,j}$ with values in $\mathbb{R}^{d_Q}$ as a basis to represent the components of the control. A full basis of Q_d is given by

$$\{\psi_{m,j} \cdot e_i \mid m = 1, \dots, M,\ j = 0, \dots, r_m, i = 1, \dots, d_Q\}$$

with e_i denoting the unit vectors on $\mathbb{R}^{d_Q}$. Due to orthogonality of the polynomials, the respective mass matrix $\mathbf{M}_Q$ is a diagonal matrix with entries $\int_{I_m} (\psi_{m,j}(t))^2 \, \mathrm{d}t$ on the diagonal. For a discrete control $q_{khd} \in Q_d$, let $\mathbf{q} \in \mathbb{R}^{\dim Q_d}$ denote the representation vector of q with respect to the above basis. Then obviously, the condition $q_{khd} \in Q_{d,\mathrm{ad}}$ translates to component-wise inequality constraints on $\mathbf{q}$.

Next, we look at the structure of the L^2 projection $P_{Q_{d,\mathrm{ad}}} \colon Q_d \to Q_{d,\mathrm{ad}}$. Let $p_{khd} \in Q_d$ and $q_{khd} = P_{Q_{d,\mathrm{ad}}}(p_{khd})$ and consider the corresponding representation vectors $\mathbf{p}$ and $\mathbf{q}$. The vector $\mathbf{q}$ solves the minimization problem

$$\text{Minimize } (\mathbf{p} - \hat{\mathbf{q}})^T \mathbf{M}_Q (\mathbf{p} - \hat{\mathbf{q}}) \text{ s. t. } \hat{\mathbf{q}} \in Q_{d,\mathrm{ad}}.$$

Since the mass matrix $\mathbf{M}_Q$ is diagonal, this problem decouples and it is easy to see that $P_{Q_{d,\mathrm{ad}}}$ amounts to a component-wise projection of the representation vector $\mathbf{p}$ onto the admissible set. Assuming the bounds q^a and q^b possess sufficient regularity to allow for pointwise evaluation, let $\mathbf{q^a}, \mathbf{q^b} \in \mathbb{R}^{\dim Q_d}$ denote the representation vectors of the nodal interpolants of q^a and q^b. The projection $P_{Q_{d,\mathrm{ad}}}$ admits the representation

$$P_{Q_{d,\mathrm{ad}}}(\mathbf{p}) = \max(\mathbf{q^a}, \min(\mathbf{q^b}, \mathbf{p}))$$

where min and max operate component-wise. Therefore, the generalized derivative $\chi_{\mathcal{I},d}\colon \mathbb{R}^{\dim Q_d} \to B(\mathbb{R}^{\dim Q_d})$ of $P_{Q_{d,\mathrm{ad}}}$ satisfies

$$\left[\,\chi_{\mathcal{I},d}(\mathbf{p})(\boldsymbol{\delta q})\,\right]_i = \begin{cases} \boldsymbol{\delta q}_i, & \text{if } [\mathbf{q^a}]_i < \mathbf{p}_i < \left[\mathbf{q^b}\right]_i, \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \dots, \dim Q_d$. Since $P_{Q_{d,\mathrm{ad}}}$ and $\chi_{\mathcal{I},d}$ are simple nodal operations, realization of Algorithm 3.1 for the fully discrete problem (4.9) is a straightforward task. We remark that some care has to be taken in the discrete setting to distinguish between nodal representation vectors and load vectors. Where necessary, load vectors have to be converted to nodal vectors by multiplying with the inverse mass matrix $\mathbf{M}_Q^{-1}$.

Remark 4.4. If the state equation is of the form (4.8) with $B_0 = 0$, and $Q_d = B_1^*(X_{k,h}^{r,s})$, then the fully discrete problem (4.9) can be interpreted as the result of approximating all integrals involving the control in the state-discrete problem (4.7) by interval-wise Gauß quadrature with $r_m + 1$ points. So in this special case, the proposed control discretization is equivalent to treating the control variationally, but instead of resolving kinks in the control exactly, it is integrated with Gauß quadrature rules, which due to lack of differentiability of the integrand typically will not achieve their maximal order of accuracy on intervals with kinks.

As seen above, whenever there is an orthogonal nodal basis of Q_d corresponding to the nodes where $Q_{d,\mathrm{ad}}$ enforces the constraints, the corresponding mass matrix is diagonal. Consequently, the discrete projection $P_{Q_{d,\mathrm{ad}}}$ and its generalized derivative $\chi_{\mathcal{I},d}$ act component-wise on representation vectors with respect to that basis. This general principle also applies to a control distributed on the spatial domain, a boundary, or the space-time cylinder. A common approach to spatial control discretization is to use the same finite element space V_h^s as for the state variable. However, in general it is not feasible to construct a nodal basis consisting of L^2 orthogonal functions for such a space. Therefore, a discontinuous space is more practical in this context.

Here, for problems involving spatially distributed control, we opt for a piecewise constant discontinuous discretization in space. While only offering first-order accuracy, the mass matrix is diagonal for the standard basis in this setting. To specify the space Q_d for a distributed control in time and space, i. e., $\Omega_Q = I \times \Omega$, let $\mathcal{T}_h^m$ for $m = 1, \dots, M$ denote the triangulation used for the state discretization on each time interval of the semidiscretization. By Q_h^m we denote the spaces

$$Q_h^m = \left\{ v \in L^2(\Omega) \;\middle|\; v|_K \in \mathcal{P}_0(K),\ K \in \mathcal{T}_h^m \right\}$$

of piecewise constant discontinuous functions. Then we use as discrete control space

$$Q_d = \{q \in Q \mid q|_{I_m} \in \mathcal{P}_{r_m}(I_m, Q_h^m),\ m = 1, \ldots, M\}\,.$$

While the combination of low order discretization of the spatial domain and *hp* discretization in time is unsatisfying with respect to balancing the accuracy of the two discretizations, it is beyond the scope of this work to develop viable high order approximations for a spatially distributed control. We assume the most promising approach to this issue within the given framework would be to work with piecewise discontinuous higher order polynomials. A similar route was taken by Wachsmuth and Wurst [110]. Their optimization procedure is based on an interior point algorithm where the penalty function is only evaluated at the nodes of the chosen quadrature formula. The nodal values at those points serve as the discrete representation of the control.

5. Solution of the time stepping equations for higher order dG methods

In this chapter we focus on the issue of solving the equation systems resulting from the discrete state and auxiliary equations efficiently. Formulating Newton's method for the time stepping equation leads to a system of $r_m + 1$ coupled linear elliptic problems. This system is closely related to the Newton update equation for the well-known Runge-Kutta Radau IIA scheme. Since typically, the dimension of the spatial discretization is large and in the context of *hp* adaptivity the number of blocks may change in between time steps, it is desirable to decouple the linearized system for numerical evaluation. However, for $r_m \geq 1$, this is not possible without introducing complex coefficients. The same issue arises for Radau IIA methods.

As we will see, the Jacobians for dG and Radau IIA time stepping equations are in fact so closely related that any idea developed for treating one can be applied directly to the other. Therefore we summarize some approaches that have been investigated for Runge-Kutta methods along with parallel developments for Galerkin time stepping.

The Runge-Kutta methods of Radau IIA type were proposed in 1969 independently by Axelsson [7] and Ehle [31]. The most wide-spread implementations, the code RADAU5 by Hairer and Wanner [52, Section IV.8] and the variable order variant RADAU described in [51], use a simplified Newton method and follow the approach proposed independently by Butcher [21] and Bickart [17] for decoupling the Newton update system. The resulting decoupled system has complex coefficients. A very similar approach was developed by Schötzau and coworkers in [100,101,115] for *hp* discontinuous Galerkin time discretization of linear parabolic equations.

In 1974, Axelsson [8,9] considered approximating the result of a block Gauß elimination for Radau IIA methods of order 3 and 5 for linear ODE systems resulting from spatial discretization of parabolic PDEs. The constructed approximation is used to solve for the last solution component by means of a preconditioned Richardson iteration. In Richter, Springer, and Vexler [91], essentially the same approximation was used to construct an inexact Newton method for dG discretizations with polynomial orders up to $r = 3$ for nonlinear parabolic PDEs. We note that in the case of linear problems without time dependent coefficients, both methods are equivalent. Basting and Weller [10] subsequently used the same approximation as a preconditioner for a conjugate gradient method to solve linear PDEs with piecewise linear dG time discretization.

Cooper and Butcher [26] in 1983 followed a different idea. Instead of working with Gaussian elimination they proposed to approximate the system in such a way that a block similarity transform leads to a block triangular system with identical entries on the diagonal. They developed schemes for Gauß integrators with two to four stages. Their approach was picked up later by González-Pinto and co-workers to develop approximation schemes for various implicit Runge-Kutta methods, starting with a scheme for the two-stage Gauß method in [45] followed by schemes for various three stage methods including Radau IIA in [46] and for four stage methods in [47]. The same authors proposed a solver based on the same idea for the two stage Radau IIA method in the context of time discretization for PDEs, see [87].

We discuss an approach that is based on this idea of finding a suitable approximation, which after a similarity transform results in a block triangular system. Its main advantage is that it allows a unified treatment of all orders r within a reasonable range. Schemes with useful properties for polynomial degrees up to $r = 7$ are obtained. Although the schemes for dG(2) and dG(3) resulting from transferring the decouplings for three- and four-stage Radau IIA integrators given in [46] and [47] are expected to yield faster convergence for these particular cases, they are obtained by numerical solution of a heuristic optimization problem modelling some properties of an efficient decoupling. This precludes a rigorous analysis of their properties. In comparison, our approach has simple enough structure to do large parts of the analysis without relying on numerical approximations.

For the convergence analysis of our scheme we will rely on the assumption that the Jacobian of the spatial differential operator has positive real spectrum. However, where appropriate, we will also give generalized results that only require the spatial part of the differential operator to be positive. For convergence in the non-linear case, we adopt a result by Calvo et al. [22] showing convergence of the corresponding iterative schemes for Runge-Kutta methods.

5.1. Structure of Newton's method for the discrete time stepping equation

Parts of this section were previously published in [91], namely the derivation of the structure of the time stepping equation in Subsection 5.1.1 and the considerations leading to the result about the spectrum of the coefficient matrix in Subsection 5.1.3. Starting from the discrete state equation (4.6), we derive a time stepping formulation and state Newton's method for the resulting equation systems. After an approximation of the Jacobian we obtain the block structured system that will be the starting point for the decoupling procedure. Throughout this section we will frequently refer to matrices and vectors of dimension $r_m + 1$ relating to the temporal discretization. To make the index range for those consistent with the associated polynomial degree, we will use zero-based indexing for them.

5.1.1. Time stepping formulation for the state equation and Newton's method

Due to the discontinuity of the test function space $X_{k,h}^{r,s}$ at the interval boundaries, it has a basis consisting only of functions that are supported on a single interval each. Therefore it is sufficient to test with such functions and with the notation $u_{kh,0}^- = \Pi_{h,1} u_0(q_{kh})$ where $\Pi_{h,1}$ denotes the L^2 projection onto the space $V_{h,1}$, the discrete state equation (4.6) is equivalent to $u_{kh} \in X_{k,h}^{r,s}$ satisfying

$$(5.1)\quad (\partial_t u_{kh}, \varphi)_{I_m} + (u_{kh,m-1}^+, \varphi_{m-1}^+) + \int_{I_m} a(q_{kh}, u_{kh})(\varphi)\,\mathrm{d}t = (u_{kh,m-1}^-, \varphi_{m-1}^+) \quad \text{for all } \varphi \in \mathcal{P}_{r_m}(I_m, V_{h,m}^s) \text{ and } m = 1, \dots, M.$$

We note that the equations for a single m are sufficient to determine the solution $u_{kh}|_{I_m} \in \mathcal{P}_{r_m}(I_m, V_{h,m}^s)$ on the corresponding interval, given that the terminal value $u_{kh,m-1}^-$ on the previous interval is known. The time stepping equations for the auxiliary problems have the same basic structure apart from the fact that the dual problems are backward in time. For the remaining chapter we will restrict ourselves to consider only the equations for a single time step, that is, we work with a single fixed m. In order to simplify notation, we suppress the dependency on the control q_{kh}, write r and V_h^s instead of r_m and $V_{h,m}^s$, and set $N = \dim V_h^s$.

Let $\{\psi_i \in \mathcal{P}_r(I_m, \mathbb{R}) \,|\, i = 0, \dots, r\}$ be a basis of the polynomial space $\mathcal{P}_r(I_m, \mathbb{R})$ and $\{\Phi_n \in V_h^s \,|\, n = 1, \dots, N\}$ be a basis of V_h^s. Then, $\{\psi_i \Phi_n \,|\, i = 0, \dots, r,\ n = 1, \dots, N\}$ forms a basis of $\mathcal{P}_r(I_m, V_h^s)$ and the time stepping equation (5.1) can be rewritten as a system of $(r+1)N$ scalar nonlinear equations for each discretization interval

$$(5.2)\quad (\partial_t u_{kh}, \psi_i \Phi_n))_{I_m} + \int_I a(t, u_{kh})(\Phi_n) \cdot \psi_i\,\mathrm{d}t + \psi_i(t_{n-1})([u_{kh}]_{m-1}, \Phi_n) = 0, \quad i = 0, \dots, r,\ n = 1, \dots, N.$$

To solve the nonlinear system (5.2) numerically, we apply Newton's method. Starting from an initial guess for the solution on the current interval, a sequence of approximations is computed by repeatedly solving the Newton update equation. Since we restrict our considerations to the single time interval I_m and in order to keep notations simple, we will denote the Newton iterates by $u_{kh}^l \in \mathcal{P}_r(I_m, V_h^s)$ with the iteration index $l \in \mathbb{N}_0$ without indicating the interval. We define the Newton residuals as

$$(5.3)\quad R_{i,n}^l := -(\partial_t u_{kh}^l, \psi_i \Phi_n)_{I_m} - \int_{I_m} a(t, u_{kh}^l)(\Phi_n) \cdot \psi_i\,\mathrm{d}t + \psi_i(t_{m-1})(u_{kh,m-1}^- - u_{kh}^l(t_{m-1}), \Phi_n), \quad i = 0, \dots, r,\ n = 1, \dots, N.$$

Then the Newton update $w_{kh}^l := u_{kh}^{l+1} - u_{kh}^l$ solves the linear system

$$(5.4)\quad (\partial_t w_{kh}^l, \psi_i \Phi_n)_{I_m} + \int_{I_m} a_u'(t, u_{kh}^l)(w_{kh}^l, \psi_i \Phi_n)\,\mathrm{d}t + \psi_i(t_{m-1})(w_{kh}^l(t_{m-1}), \Phi_n) = R_{i,n}^l, \quad i = 0, \dots, r,\ n = 1, \dots, N.$$

Next, we write the Newton update in the chosen temporal and spatial basis, that is $w_{kh}^l = \sum_{j=0}^r \sum_{n'=1}^N \mathbf{w}_{j,n'}^l \psi_j \Phi_{n'}$ with real coefficients $\mathbf{w}_{j,n'}^l$. Collecting the temporal derivative and jump terms simplifies the update equation to

$$(5.5)\quad \sum_{j=0}^r \sum_{n'=1}^N \left[\left(\int_{I_m} \partial_t \psi_j \psi_i \,\mathrm{d}t + \psi_j(t_{m-1})\psi_i(t_{m-1}) \right) (\Phi_{n'}, \Phi_n) + \int_{I_m} a_u'(t, u_{kh}^l)(\Phi_{n'}, \Phi_n) \cdot \psi_j \psi_i \,\mathrm{d}t \right] \mathbf{w}_{j,n'}^l = R_{i,n}^l, \quad i = 0, \dots, r,\ n = 1, \dots, N.$$

For given i and j, evaluating the expression $\int_{I_m} a_u'(t, u_{kh}^l)(\Phi_{n'}, \Phi_n) \cdot \psi_j \psi_i \,\mathrm{d}t$ for all $n, n' = 0, \dots, N$ amounts to a weighted temporal integral over the stiffness matrix of the linearized spatial differential operator, which is usually too expensive to be computed numerically since it has to be evaluated for each combination of weights $\psi_j \psi_i$. Thus, we approximate it by a suitable mean value

$$\int_{I_m} a_u'(t, u_{kh}^l)(\Phi_{n'}, \Phi_n) \cdot \psi_j \psi_i \,\mathrm{d}t \approx \overline{a_u'(u_{kh}^l)(\Phi_{n'}, \Phi_n)} \int_{I_m} \psi_j \psi_i \,\mathrm{d}t.$$

Since we perform a Newton iteration even for linear problems and the residual is computed without this approximation, the accuracy of the computed final solution is not affected. Rather, the convergence behaviour of the Newton iteration changes. Introducing the midpoint $\tilde{t}_m := \frac{t_{m-1}+t_m}{2}$ of the current time interval, the most obvious choice for the mean value is

$$\overline{a_u'(u_{kh}^l)(\Phi_{n'}, \Phi_n)} := a_u'(\tilde{t}_m, u_{kh}^l(\tilde{t}_m))(\Phi_{n'}, \Phi_n),$$

that is, we evaluate the derivative once at the midpoint of the time interval.

Remark 5.1. For higher order implicit Runge-Kutta methods, typically assembly of the Newton system involves evaluating the stiffness matrix at several time points, which is considered too expensive as well. The standard way of dealing with this issue is to approximate all occurrences of the stiffness matrix by the stiffness matrix evaluated at the beginning of the time interval with the terminal solution from the last interval, see for example [52, Section IV.8]. The impact on the convergence of the Newton scheme is very similar to the approximation proposed here.

We introduce the mass matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ and the averaged stiffness matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ with the entries

$$\mathbf{M}_{n,n'} = (\Phi_{n'}, \Phi_n) \text{ and } \mathbf{A}_{n,n'} = \overline{a_u'(u_{kh}^l)(\Phi_{n'}, \Phi_n)} \text{ respectively.}$$

With the notations

$$\alpha_{ij} := \int_{I_m} \partial_t \psi_j \psi_i \,\mathrm{d}t + \psi_j(t_{m-1})\psi_i(t_{m-1}) \text{ and } \beta_{ij} = \frac{1}{k_m}\int_{I_m} \psi_j \psi_i \,\mathrm{d}t,$$

the Newton update equation with the approximation for the linearized form discussed above reads

(5.6)
$$\begin{pmatrix} \alpha_{00}\mathbf{M} + k_m\,\beta_{00}\mathbf{A} & \alpha_{01}\mathbf{M} + k_m\,\beta_{01}\mathbf{A} & \cdots & \alpha_{0r}\mathbf{M} + k_m\,\beta_{0r}\mathbf{A} \\ \alpha_{10}\mathbf{M} + k_m\,\beta_{10}\mathbf{A} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \alpha_{r0}\mathbf{M} + k_m\,\beta_{r0}\mathbf{A} & \cdots & \cdots & \alpha_{rr}\mathbf{M} + k_m\,\beta_{rr}\mathbf{A} \end{pmatrix} \begin{pmatrix} \mathbf{w}_0^l \\ \mathbf{w}_1^l \\ \vdots \\ \mathbf{w}_r^l \end{pmatrix} = \begin{pmatrix} R_0^l \\ R_1^l \\ \vdots \\ R_r^l \end{pmatrix},$$

where the vectors $\mathbf{w}_j^l := \begin{pmatrix} \mathbf{w}_{j,1}^l & \cdots & \mathbf{w}_{j,N}^l \end{pmatrix}^T$ and $R_i^l := \begin{pmatrix} R_{i,1}^l & \cdots & R_{i,N}^l \end{pmatrix}^T$ collect the corrections and residual terms respectively for one temporal basis function.

To simplify the notation, we make use of the Kronecker product, which can be defined for a $m \times n$ matrix $\mathbf{G}$ with entries g_{ij} and a second matrix $\mathbf{H}$, which does not have to be of the same size as $\mathbf{G}$, as the block matrix

$$\mathbf{G} \otimes \mathbf{H} = \begin{pmatrix} g_{11}\mathbf{H} & g_{12}\mathbf{H} & \dots & g_{1n}\mathbf{H} \\ \vdots & \vdots & \ddots & \vdots \\ g_{m1}\mathbf{H} & g_{m2}\mathbf{H} & \dots & g_{mn}\mathbf{H} \end{pmatrix}.$$

For convenience of the reader, we summarize some elementary properties of the Kronecker product that we use later on. Let $\mathbf{G}$ and $\mathbf{H}$ be defined as above.

1. $(\mathbf{G} \otimes \mathbf{H})^T = \mathbf{G}^T \otimes \mathbf{H}^T$.
2. Let $\mathbf{P}$ and $\mathbf{Q}$ be two further matrices and assume that the number of rows of $\mathbf{P}$ and $\mathbf{Q}$ agrees with the number of columns of $\mathbf{G}$ and $\mathbf{H}$ respectively. Then $(\mathbf{G} \otimes \mathbf{H})(\mathbf{P} \otimes \mathbf{Q}) = (\mathbf{GP}) \otimes (\mathbf{HQ})$.
3. If $\mathbf{G}$ and $\mathbf{H}$ are square and invertible, then $(\mathbf{G} \otimes \mathbf{H})^{-1} = \mathbf{G}^{-1} \otimes \mathbf{H}^{-1}$.

We introduce the $(r+1) \times (r+1)$ matrices $\widehat{\mathbf{A}} := (\alpha_{ij})_{i,j\in\{0,\dots,r\}}$ and $\widehat{\mathbf{B}} := (\beta_{ij})_{i,j\in\{0,\dots,r\}}$ and collect the temporal components $\mathbf{w}_j^l$ and R_j^l of the updates and residuals in the vectors $(\mathbf{w}^l)^T = \begin{pmatrix} (\mathbf{w}_0^l)^T & \cdots & (\mathbf{w}_r^l)^T \end{pmatrix}$ and $(R^l)^T = \begin{pmatrix} (R_0^l)^T & \cdots & (R_r^l)^T \end{pmatrix}$. With these preparations, the Newton update equation (5.6) takes the more compact form

$$\left(\widehat{\mathbf{A}} \otimes \mathbf{M} + k_m \widehat{\mathbf{B}} \otimes \mathbf{A}\right) \mathbf{w}^l = R^l. \tag{5.7}$$

For fixed r we can derive explicit representations for the coefficient matrices $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ as shown in the following lemma.

Lemma 5.2. *If we represent the temporal basis* $\{\psi_j\}$ *as* $\psi_j(t) = \sum_{\mu=0}^{r} c_{\mu j} \left(\frac{t-t_{m-1}}{k_m}\right)^\mu$ *and denote the corresponding coefficient matrix by* $\mathbf{C} := (c_{\mu j})_{\mu,j\in\{0,\dots,r\}}$*, we get the explicit representations*

$$\widehat{\mathbf{A}} = \mathbf{C}^T\mathbf{G}\mathbf{C} \text{ and } \widehat{\mathbf{B}} = \mathbf{C}^T\mathbf{H}\mathbf{C}$$

for the coefficient matrices. Here, $\mathbf{H}$ *denotes the* $(r+1)$*-dimensional Hilbert matrix, that is,* $\mathbf{H}_{\mu\nu} = \frac{1}{\mu+\nu+1}$*, and the entries of* $\mathbf{G}$ *are given by* $\mathbf{G}_{00} = 1$ *and* $\mathbf{G}_{\mu\nu} = \frac{\nu}{\mu+\nu}$ *for the remaining entries.*

$\mathbf{G}$ *can be represented as* $\mathbf{G} = \mathbf{H}\mathbf{D} + \mathbf{E}$ *where* $\mathbf{D}$ *is the representation matrix of the derivative operator on* $\mathcal{P}_r([0,1])$ *with respect to the monomial basis and* $\mathbf{E} = \mathbf{e_1}\mathbf{e_1}^T$*.*

Proof. Transforming the integral $\beta_{ij} = \int_{I_m} \frac{1}{k_m}\psi_j\psi_i \,\mathrm{d}t$ to the unit interval yields

$$\begin{aligned}\beta_{ij} &= \int_0^1 \psi_j(k_m\tau)\psi_i(k_m\tau)\,\mathrm{d}\tau = \sum_{\mu=0}^{r}\sum_{\nu=0}^{r}\int_0^1 c_{\nu j}c_{\mu i}\,\tau^{\mu+\nu}\,\mathrm{d}\tau \\ &= \sum_{\mu=0}^{r} c_{\mu i}\sum_{\nu=0}^{r}\int_0^1 \tau^{\mu+\nu}\,\mathrm{d}\tau\, c_{\nu j} = \sum_{\mu=0}^{r} c_{\mu i}\sum_{\nu=0}^{r}\frac{1}{\mu+\nu+1}\,c_{\nu j}.\end{aligned}$$

Rewriting this identity in terms of matrix products gives the representation formula for $\widehat{\mathbf{B}}$. To derive a representation for $\widehat{\mathbf{A}}$ we start again by transforming the integral to the unit interval and get

$$\begin{aligned}\alpha_{ij} &= \int_0^1 k_m\partial_t\psi_j(k_m\tau)\psi_i(k_m\tau)\,\mathrm{d}\tau + \psi_j(t_{m-1})\psi_i(t_{m-1}) \\ &= \int_0^1 \partial_\tau\psi_j(k_m\tau)\psi_i(k_m\tau)\,\mathrm{d}\tau + c_{0j}c_{0i}.\end{aligned}$$

We proceed as above by plugging in the monomial representation for the temporal basis, evaluating integrals and derivatives of the monomials and rewriting the result in terms of matrix products. □

For order $r = 0$, the resulting scheme is some variant of the well known implicit Euler method, depending on how the temporal integrals in the residual terms are evaluated. This type of scheme is easy to implement on top of existing finite element code for elliptic spatial problems, as long as the matrices $\mathbf{M}$ and $\mathbf{A}$ can be assembled, a linear solver for those matrices is available and elementary vector operations are implemented. If the order r is greater than 0, however, the update equation (5.6) is a coupled system where each block resembles the system matrix of implicit Euler with varying coefficients. Therefore it would be desirable to decouple this block system.

5.1.2. Connection to the Runge-Kutta methods of type Radau-IIA

Multiplying (5.7) from left by $\widehat{\mathbf{A}}^{-1} \otimes \mathbf{M}^{-1}$ gives

$$\left(\mathrm{Id} \otimes \mathrm{Id} + k_m \left(\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}}\right) \otimes \left(\mathbf{M}^{-1}\mathbf{A}\right)\right) \mathbf{w^l} = \widehat{\mathbf{A}}^{-1} \otimes \mathbf{M}^{-1} R^l. \tag{5.8}$$

The identity matrix of the respective dimension is denoted Id. Plugging in the representations for $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{B}}$ from Lemma 5.2 gives for the coefficient matrix of the spatial derivative term

$$\mathcal{A} := \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{B}} = \left(\mathbf{C}^T \left(\mathbf{HD} + \mathbf{E}\right) \mathbf{C}\right)^{-1} \mathbf{C}^T \mathbf{H} \mathbf{C} = \mathbf{C}^{-1} \left(\mathbf{D} + \mathbf{H}^{-1}\mathbf{E}\right)^{-1} \mathbf{C}. \tag{5.9}$$

We denote the inner matrix by $\mathcal{A}_{\mathrm{ref}} = \left(\mathbf{D} + \mathbf{H}^{-1}\mathbf{E}\right)^{-1}$ and after some computations using the inverse of the Hilbert matrix we obtain the explicit representation

$$\mathcal{A}_{\mathrm{ref}} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & a_0 \\ 1 & \ddots & & \vdots & a_1 \\ 0 & 2 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & r & a_r \end{bmatrix} \tag{5.10}$$

with the coefficients $a_k = (-1)^{r+k} \frac{(r+k)!(r+1)!r!}{(k!)^2(r-k+1)!(2r+1)!}$. Together with the notation $\mathbf{L} := k_m \mathbf{M}^{-1}\mathbf{A}$, the Newton update equation can be stated as

$$\left(\mathrm{Id} \otimes \mathrm{Id} + \mathcal{A} \otimes \mathbf{L}\right) \mathbf{w}^l = \widehat{\mathbf{A}}^{-1} \otimes \mathbf{M}^{-1} R^l. \tag{5.11}$$

Lemma 5.3. *When choosing* $\mathbf{C}$ *as the coefficient matrix of the Lagrange basis with the roots of the* r^{th} *right Radau polynomial as nodes, the matrix* $\mathcal{A}$ *is identical to the Butcher tableau of the* $r+1$ *stage Radau-II-A method.*

Remark 5.4. This means that the Newton correction equation (5.11) in this case has the same form as for the corresponding Radau scheme which allows to apply any techniques known from the Radau scheme for its solution. Since for symmetry reason we do not use the Radau nodes for the temporal basis, typically a similarity transform is necessary.

Proof. As at the end of Section 4.1.1, let $\tau_0^{r+1}, \dots \tau_r^{r+1}$ denote the roots of the right Radau polynomial $\mathcal{R}_{r+1}$ of degree $r+1$ on the unit interval $[0,1]$. According to, e. g., [52], the entries a_{ij} of the Butcher scheme $\mathcal{A}_{\mathrm{rad}}$ for the $r+1$ stage Radau-II-A method are determined by the conditions

$$\sum_{j=0}^{r} a_{ij} \tau_j^k = \frac{\tau_i^{k+1}}{k+1} \quad \text{for } i \in 0, \dots, r \text{ and } k \in 0, \dots, r.$$

Rewriting this equation system in terms of matrix products gives

$$\mathcal{A}_{\text{rad}} \cdot \begin{pmatrix} \tau_0^0 & \tau_0^1 & \dots & \tau_0^r \\ \tau_1^0 & \tau_1^1 & \dots & \tau_1^r \\ \vdots & \vdots & \ddots & \vdots \\ \tau_r^0 & \tau_r^1 & \dots & \tau_r^r \end{pmatrix} = \begin{pmatrix} \frac{\tau_0^1}{1} & \frac{\tau_0^2}{2} & \dots & \frac{\tau_0^{r+1}}{r+1} \\ \frac{\tau_1^1}{1} & \frac{\tau_1^2}{2} & \dots & \frac{\tau_1^{r+1}}{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\tau_r^1}{1} & \frac{\tau_r^2}{2} & \dots & \frac{\tau_r^{r+1}}{r+1} \end{pmatrix}.$$

Let $\mathbf{C}$ denote the coefficient matrix of the Lagrange basis with nodes $\tau_0, \dots, \tau_r$. Then the Vandermonde matrix on the left hand side is the inverse $\mathbf{C}^{-1}$. To show our claim, we have to show $\mathcal{A}_{\text{rad}} = \mathbf{C}^{-1}\mathcal{A}_{\text{ref}}\mathbf{C}$. With the above identity this is equivalent to the condition

$$\mathbf{C}^{-1}\mathcal{A}_{\text{ref}} = \mathcal{A}_{\text{rad}}\mathbf{C}^{-1} = \begin{pmatrix} \frac{\tau_0^1}{1} & \frac{\tau_0^2}{2} & \dots & \frac{\tau_0^{r+1}}{r+1} \\ \frac{\tau_1^1}{1} & \frac{\tau_1^2}{2} & \dots & \frac{\tau_1^{r+1}}{r+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\tau_r^1}{1} & \frac{\tau_r^2}{2} & \dots & \frac{\tau_r^{r+1}}{r+1} \end{pmatrix}.$$

Evaluating the matrix product on the left gives

$$\begin{pmatrix} \frac{\tau_0^1}{1} & \frac{\tau_0^2}{2} & \dots & \frac{\tau_0^r}{r} & \sum_{k=0}^r a_k \tau_0^k \\ \frac{\tau_1^1}{1} & \frac{\tau_1^2}{2} & \dots & \frac{\tau_1^r}{r} & \sum_{k=0}^r a_k \tau_1^k \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\tau_r^1}{1} & \frac{\tau_r^2}{2} & \dots & \frac{\tau_r^r}{r} & \sum_{k=0}^r a_k \tau_r^k \end{pmatrix}.$$

Hence the claim is equivalent to the Radau nodes being the zeros of the polynomial

$$\tilde{\mathcal{R}}_{r+1}(t) = \frac{t^{r+1}}{r+1} - \sum_{k=0}^{r} a_k t^k.$$

Using the explicit representation

$$\mathcal{L}_n(t) = (-1)^n \sum_{k=0}^{n} (-1)^k \frac{(n+k)!}{(k!)^2(n-k)!} x^k$$

for the Legendre-polynomials on the interval $[0, 1]$ found, e. g., in [52, Section IV.5],we can verify easily that the polynomial $\tilde{\mathcal{R}}_{r+1}(t)$ is a scaled version of the Radau polynomial $\mathcal{R}_{r+1}(t)$ which shows the claim. □

5.1.3. Spectrum of the Coefficient matrix

The formulation (5.11) of the Newton update suggests performing a similarity transform to turn it into a block diagonal (or block triangular) system with blocks of the form $\mathrm{Id} + \mu_j \mathbf{L}$ on the diagonal where μ_j are the eigenvalues of $\mathcal{A}$. The question of whether this

works without introducing complex coefficients is equivalent to the question whether the eigenvalues of $\mathcal{A}$ are real. We note that since the coefficient matrices for different choices of the temporal basis are similar to each other, the spectrum cannot be influenced by modifying the temporal basis.

We consider the scalar test equation $\partial_t u - \lambda u = 0$ for $u \in H^1(I)$ with $I = (0,1)$ and $u(0) = 1$. It is discretized with the discontinuous Galerkin method of order r on the time grid consisting only of the single interval I. Since the problem is linear without time-dependent coefficients, Newton's method with starting value 0 computes the discrete solution as the first update, that is the nodal representation vector $\mathbf{u} \in \mathbb{R}^{r+1}$ of the discrete solution u_k satisfies

$$(\mathrm{Id} - \lambda\mathcal{A})\,\mathbf{u} = R \tag{5.12}$$

with $R_j = \psi_j(0)$ for $j = 0, \ldots, r$. Let us assume for now that the temporal basis $\{\psi_j\}$ is a Lagrange basis with the last node at the end of the time interval. Then we have $u_k(1) = \mathbf{u}_r$.

Next, we need to establish a link between the dG(r) solution of this problem at final time 1 and a Padé approximation of the exact solution $u(1) = e^\lambda$. The $[r/r+1]$-Padé approximation of e^λ is a rational function in λ with numerator of maximal degree r and denominator of maximal degree $r+1$ that approximates e^λ up to an error of $\mathcal{O}(\lambda^{2r+2})$. If numerator and denominator are required to have no common roots, it is uniquely determined (see, e. g., the survey article by Brezinski and Iseghem [20]). We denote it by $Q_{r,r+1}(\lambda) = \frac{\kappa(\lambda)}{\tau(\lambda)}$. It is shown in [88, § 42] that the Padé table for e^λ is normal and hence the denominator of the $[r/r+1]$-Padé approximation of e^λ has exactly degree $r+1$ for any r and no terms in the Padé approximation cancel out. It is shown by Lesaint and Raviart during the proof of Theorem 2 in [67] that in fact $\mathbf{u}_r = u_k(1) = Q_{r,r+1}(\lambda)$.

An alternative way of representing $\mathbf{u}_r$ is by applying Cramer's rule to (5.12) yielding

$$\mathbf{u}_r = \frac{\det\left((\mathrm{Id} - \lambda\mathcal{A})_{0,\ldots,r,\ 0,\ldots,r-1} \quad R\right)}{\det(\mathrm{Id} - \lambda\mathcal{A})}$$

which is a rational function with respect to λ with degree of numerator and denominator at most r and $r+1$ respectively. Since $\mathbf{u}_r = Q_{r,r+1}(\lambda)$, we conclude that apart from a constant the denominator equals the denominator $\tau(\lambda)$ of the $[r/r+1]$-Padé approximation. Therefore we conclude that the spectrum of $\mathcal{A}$ consists of the reciprocals of the zeros of $\tau(\lambda)$.

Wanner et al. in [113, Theorem 8] show an upper bound for the order of rational approximations of the exponential based on the degree of the numerator and the number of non-real-valued roots of the denominator. Inserting the known approximation order and degree of the numerator for $Q_{r,r+1}$ into this estimate gives immediately that $\tau(\lambda)$ has at most one real-valued root. The following lemma summarizes the above considerations.

Lemma 5.5. *The eigenvalues of the matrix $\mathcal{A}$ are given as $\mu_i = \frac{1}{\lambda_i}$ with $i = 0, \dots, r$, where λ_i are the roots of the denominator $\tau(\lambda)$ of the $[r/r+1]$-Padé approximation of e^λ. In particular at most one eigenvalue is real-valued.*

Remark 5.6. As a consequence of this result, apart from the trivial case $r = 0$, it is not possible to decouple the blocks of the Newton system (5.7) without introducing complex coefficients.

5.2. Approximate decoupling scheme

As shown in the previous section, the matrix $\mathcal{A}$ has apart from at most one exception complex spectrum. Therefore, it is not possible to transform the system (5.8) into block triangular form with real coefficients. Instead, following the idea of Cooper and Butcher [26] for Gauß methods, we propose to approximate $\mathcal{A}$ by a suitable matrix $\mathbf{T}$ with a single real eigenvalue. Convergence analysis for the linear scheme and for the resulting non-linear solver will rely on some assumptions on the stiffness matrix which we discuss in the first part of this section. Next, we introduce the numerical scheme for the approximate Newton update, followed by convergence results, first for linear equations with time-independent coefficients and subsequently for nonlinear problems. Finally, the applicability of the obtained results to a class of semilinear equations is discussed.

5.2.1. Assumptions on the problem

For many standard examples of scalar diffusion-reaction equations, the derivative of the spatial differential operator is self-adjoint and positive. This motivates to optimize the approximate Newton scheme to work well for symmetric positive semidefinite stiffness matrices.

Assumption 5.7. Let $a\colon I \times V \times V \to \mathbb{R}$ be the semilinear form defining the spatial differential operator. Then for all $u \in V$, $t \in I$, and $\varphi, \psi \in V$ we assume

$$\begin{aligned} &a'_u(t,u)(\varphi,\psi) = a'_u(t,u)(\psi,\varphi) \quad \text{and} \\ &a'_u(t,u)(\varphi,\varphi) \geq 0, \end{aligned}$$

i. e., the derivative of a with respect to u is positive and symmetric.

Remark 5.8. 1. In practice it can be sufficient to require the assumption only for a subset of pairs (t,u). It just has to be ensured that all iterates that are encountered in the course of the algorithm are contained in that subset.
2. Obviously, the assumption implies that $\mathbf{A}$ is symmetric positive semi-definite. Therefore, although $\mathbf{L} = k_m\mathbf{M}^{-1}\mathbf{A}$ is not necessarily symmetric, it has non-negative real spectrum. To see that, consider an eigenvalue μ of $\mathbf{L}$ and the corresponding eigenvector $\mathbf{v}$. We have $\mathbf{L}\mathbf{v} = \mu\mathbf{v}$ which is equivalent to $k_m\mathbf{A}\mathbf{v} = \mu\mathbf{M}\mathbf{v}$. Multiplying by the conjugate

transpose $\mathbf{v}^*$ of $\mathbf{v}$ and solving for μ gives $\mu = k_m \frac{\mathbf{v}^*\mathbf{A}\mathbf{v}}{\mathbf{v}^*\mathbf{M}\mathbf{v}}$ and, since the right hand side of this identity is a non-negative real value, the same holds for the eigenvalue μ.

Cases where this assumption clearly is satisfied include spatial differential operators of the form $A(t,u) := -\operatorname{div}(G\nabla u) + g(t,x,u)$ where G is a symmetric $d \times d$ matrix with entries $G_{ij} \in L^\infty(\Omega)$ which is positive semidefinite on Ω. The nonlinearity $g\colon I \times \Omega \times \mathbb{R} \to \mathbb{R}$ is required to be continuously differentiable and non-decreasing with respect to the third argument for almost all $(t,x) \in I \times \Omega$. Obviously the equations for the test problems given in Sections 2.3.1 and 2.3.2 satisfy this characterization and hence Assumption 5.7 is fulfilled by them.

For problems involving transport or systems of equations, this assumption might be too restrictive. However, when accepting less favourable constants, part of the convergence results can be shown for the following more general assumption which is satisfied for a monotone spatial differential operator.

Assumption 5.9. For all $u \in V$, $t \in I$, and $\varphi \in V$ we assume

$$a'_u(t,u)(\varphi,\varphi) \geq 0.$$

Concerning the state equation of the temperature control problem discussed in Section 2.3.3, numerical inspection of the stiffness matrix resulting from a coarse discretization indicates that neither of the above assumptions are satisfied. However, as shown in the numerical tests, the proposed decoupling converges for reasonable size of the time step nevertheless.

To show convergence of the solution process for non-linear problems, besides one of the above assumptions, we need to impose the following condition on the sensitivity of the stiffness matrix to changes of the evaluation points t and u.

Assumption 5.10. For given $u \in V_h$ and $t \in I$ let $\mathbf{A}(t,u)$ denote the corresponding stiffness matrix with the entries $(\mathbf{A}(t,u))_{n,n'} = a'_u(t,u)(\Phi_{n'},\Phi_n)$. We require that for any $(t,u) \in I \times V_h$, there is $\delta > 0$ and $\bar{k} > 0$ such that for any pair of elements (t_1,u_1) and (t_2,u_2) contained in the neighbourhood

$$B_\delta(t,u) = \left\{(\hat{t},\hat{u}) \in I \times V_h \;\middle|\; \|\hat{u}-u\|_H < \delta \wedge |\hat{t}-t| < \delta\right\}$$

and any $k \leq \bar{k}$, the condition

$$\left\|(\mathbf{M} + k\mathbf{A}(t_1,u_1))^{-1}\, k\,(\mathbf{A}(t_2,u_2) - \mathbf{A}(t_1,u_1))\right\|_{\mathbf{M}} \leq \kappa_1|t_1-t_2| + \kappa_2\|u_1-u_2\|_H$$

is satisfied for some positive constants κ_1 and κ_2 depending only on $B_\delta(t,u)$.

Remark 5.11. 1. This resembles an affine-covariant Lipschitz condition for an implicit Euler time stepping equation with time step k. Such conditions are frequently encountered in the corresponding convergence theorems for Newton's method (see for example Deuflhard [28]).

2. For a linear differential operator A without time dependent coefficients, Assumption 5.10 is satisfied trivially. In general, it is difficult to ensure this assumption with constants κ_1 and κ_2 independent of the spatial mesh. A more detailed discussion of the arising issues and possible resolution will be given in Section 5.2.5 for the case of semilinear reaction-diffusion type equations.

5.2.2. Approximation of the coefficient matrix

As pointed out we will replace $\mathcal{A}$ by an approximation $\mathbf{T}$ with a spectrum consisting only of a single $(r+1)$-fold real eigenvalue. Since throughout this section we consider only a single Newton step, for simplicity the iteration index l will be omitted.

Instead of solving (5.11) for the Newton correction $\mathbf{w}$, an approximation $\widetilde{\mathbf{w}}$ satisfying

$$(\mathrm{Id}\otimes\mathrm{Id}+\mathbf{T}\otimes\mathbf{L})\,\widetilde{\mathbf{w}} = \widehat{\mathbf{A}}^{-1}\otimes\mathbf{M}^{-1}R \tag{5.13}$$

is computed. The error between $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ can be written as

$$\mathbf{w}-\widetilde{\mathbf{w}} = \mathbf{w}-(\mathrm{Id}\otimes\mathrm{Id}+\mathbf{T}\otimes\mathbf{L})^{-1}(\mathrm{Id}\otimes\mathrm{Id}+\mathcal{A}\otimes\mathbf{L})\,\mathbf{w} = (\mathrm{Id}\otimes\mathrm{Id}+\mathbf{T}\otimes\mathbf{L})^{-1}(\mathbf{T}-\mathcal{A})\otimes\mathbf{L}\mathbf{w}.$$

For a scalar test equation $\partial_t u+\mu u=f$ this simplifies to

$$\mathbf{w}-\widetilde{\mathbf{w}} = (\mathrm{Id}+\mathbf{T}\lambda)^{-1}(\mathbf{T}-\mathcal{A})\lambda\mathbf{w}, \tag{5.14}$$

with $\lambda = k_m\mu$. We denote the error matrix by $\mathbf{V}(\lambda) = (\mathrm{Id}+\mathbf{T}\lambda)^{-1}(\mathbf{T}-\mathcal{A})\lambda$. Since we are interested in solving equations where $\mathbf{L}$ has positive real spectrum, we choose as a criterion for a good approximation $\mathbf{T}$ that the spectral radius of $\mathbf{V}$ is small for all positive and real λ. It can be verified easily that the spectral radius of $\mathbf{V}$ is invariant under similarity transform of $\mathbf{T}$ and $\mathcal{A}$, therefore we can chose the most convenient basis to construct an approximation $\mathbf{T}$ for $\mathcal{A}$.

We start from the reference representation $\mathcal{A}_{\mathrm{ref}}$ introduced in (5.10) and perform another similarity transform using the diagonal matrix $\mathbf{S}$ with $\mathbf{S}_{jj} = (j)!$ for $j=0,\dots,r$ to obtain

$$\mathcal{A}_{\mathrm{f}} := \mathbf{S}\mathcal{A}_{\mathrm{ref}}\mathbf{S}^{-1} = \begin{bmatrix} 0 & \cdots & \cdots & 0 & b_0 \\ 1 & \ddots & & \vdots & b_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & b_r \end{bmatrix}$$

with $b_k = (-1)^{r+k}\frac{(r+k)!(r+1)!}{k!(r-k+1)!(2r+1)!}$. This matrix has the form of a companion matrix, hence its characteristic polynomial is given explicitly as $\chi(\mu) = \mu^{r+1} - \sum_{k=0}^{r} b_k\mu^k$.

Remark 5.12. Note that the temporal basis corresponding to $\mathcal{A}_{\mathrm{f}}$ is given by $\frac{1}{j!}\left(\frac{t-t_{m-1}}{k_m}\right)^j$ with $j=0,\dots,r$. So the entries of a given vector with respect to this basis can be

interpreted as the coefficients of the Taylor expansion of the corresponding polynomial around 0 after transforming back to the reference interval $[0,1]$. Therefore we subsequently refer to this basis as *Taylor basis*.

Now we want to replace the matrix $\mathcal{A}_\mathrm{f}$ by a suitable approximation $\mathbf{T}_\mathrm{f}$ with spectrum consisting of a single real $(r+1)$-fold eigenvalue, that is, its characteristic polynomial $\tilde{\chi}$ has the form $\tilde{\chi}(\mu) = (\mu-\gamma)^{r+1}$ for a suitable γ. We choose

$$\gamma = \sqrt[r+1]{\frac{r!}{(2r+1)!}}$$

since for this choice the constant coefficients of χ and $\tilde{\chi}$ agree. Denoting the negatives of the coefficients of $\tilde{\chi}$ by $\tilde{b}_k = (-1)^{r+k}\binom{r+1}{k}\gamma^{r+1-k}$ for $k = 0,\dots,r$, we propose the approximation

$$\mathbf{T}_\mathrm{f} := \begin{bmatrix} 0 & \cdots & \cdots & 0 & \tilde{b}_0 \\ 1 & \ddots & & \vdots & \tilde{b}_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & \tilde{b}_r \end{bmatrix}$$

which obviously has the desired characteristic polynomial.

We evaluate the error matrix $\mathbf{V}_\mathrm{f}(\lambda) = (\mathrm{Id} + \mathbf{T}_\mathrm{f}\lambda)^{-1}(\mathbf{T}_\mathrm{f} - \mathcal{A}_\mathrm{f})\lambda$, which gives us

$$\mathbf{V}_\mathrm{f}(\lambda) = \begin{bmatrix} 1 & \cdots & \cdots & 0 & \lambda\tilde{b}_0 \\ \lambda & \ddots & & \vdots & \lambda\tilde{b}_1 \\ 0 & \lambda & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 1 & \lambda\tilde{b}_{r-1} \\ 0 & \cdots & 0 & \lambda & 1+\lambda\tilde{b}_r \end{bmatrix}^{-1} \begin{bmatrix} 0 & \tilde{b} - b \end{bmatrix} \lambda.$$

Obviously only the last column of $\mathbf{V}_\mathrm{f}$ contains non-zero entries, hence its spectrum consists only of zeros and the eigenvalue $\mu_r = (\mathbf{V}_\mathrm{f})_{rr}$. We can compute this value by Gaussian elimination and obtain

$$\mu_r = \frac{\sum_{k=0}^{r}(-1)^k\lambda^{k+1}(\tilde{b}_{r-k} - b_{r-k})}{1+\sum_{k=0}^{r}(-1)^k\lambda^{k+1}\tilde{b}_{r-k}} = 1 - \frac{1+\sum_{k=0}^{r}(-1)^k\lambda^{k+1}b_{r-k}}{(1+\lambda\gamma)^{r+1}}. \tag{5.15}$$

For the polynomial in the numerator of the last term we note that the identity $1+\sum_{k=0}^{r}(-1)^k\lambda^{k+1}b_{r-k} = \det(\mathrm{Id}+\lambda\mathcal{A}_\mathrm{f}) = \det(\mathrm{Id}+\lambda\mathcal{A})$ holds. As noted in Section 5.1.3, $\det(\mathrm{Id}-\lambda\mathcal{A})$ is the denominator of the $[r/r+1]$ Padé approximation of $\exp(\lambda)$. Consequently, $\det(\mathrm{Id}+\lambda\mathcal{A})$ has to be the denominator of the corresponding approximation for $\exp(-\lambda)$. This means μ_{r+1} is identical to the error terms we encountered for the block Gauß elimination in [91]. Therefore for positive and real λ the spectral radius $|\mu_{r+1}|$ of $\mathbf{V}_\mathrm{f}$ is bounded by

$$\rho_r = \sup\left\{|\mu_r| \;\middle|\; \lambda \in \mathbb{R}^+\right\}.$$

Table 5.1.: Characteristic polynomials of $\mathcal{A}$ and $\mathbf{T}$ an bounds on the spectral radius of $\mathbf{V}$ for different values of r.

r	$\chi(\mu)$	$\tilde{\chi}(\mu)$	ρ_r
1	$\mu^2 - \frac{2}{3}\mu + \frac{1}{6}$	$\left(\mu - \sqrt{\frac{1}{6}}\right)^2$	0.092
2	$\mu^3 - \frac{3}{5}\mu^2 + \frac{3}{20}\mu - \frac{1}{60}$	$\left(\mu - \sqrt[3]{\frac{1}{60}}\right)^3$	0.169
3	$\mu^4 - \frac{4}{7}\mu^3 + \frac{1}{7}\mu^2 - \frac{2}{105}\mu + \frac{1}{840}$	$\left(\mu - \sqrt[4]{\frac{1}{840}}\right)^4$	0.238
4	$\mu^5 - \frac{5}{9}\mu^4 + \frac{5}{36}\mu^3 - \frac{5}{252}\mu^2 + \frac{5}{3024}\mu - \frac{1}{15120}$	$\left(\mu - \sqrt[5]{\frac{1}{15120}}\right)^5$	0.301
5	$\mu^6 - \frac{6}{11}\mu^5 + \frac{3}{22}\mu^4 - \frac{2}{99}\mu^3 + \frac{1}{528}\mu^2 - \frac{1}{9240}\mu + \frac{1}{332640}$	$\left(\mu - \sqrt[6]{\frac{1}{332640}}\right)^6$	0.359
6	$\mu^7 - \frac{7}{13}\mu^6 + \frac{7}{52}\mu^5 - \frac{35}{1716}\mu^4 + \frac{7}{3432}\mu^3 - \frac{7}{51480}\mu^2 + \frac{7}{1235520}\mu - \frac{1}{8648640}$	$\left(\mu - \sqrt[7]{\frac{1}{8648640}}\right)^7$	0.412
7	...	...	0.460

This means while superlinear convergence is lost, we can expect fast linear convergence from the approximate Newton update equation

Numerical upper bounds for ρ_r along with the characteristic polynomials χ and $\tilde{\chi}$ are given in Table 5.1.

Remark 5.13. It can be shown by explicit computation for dG(1) that in general, our choice of γ is not optimal in the sense that it minimizes $\sup_{\lambda \geq 0} |\mu_r|$. For relevant orders r ($r \leq 7$) however, our choice still gives a spectral radius less than $\frac{1}{2}$.

For later usage, we also record an explicit formula for the other entries of the matrix $\mathbf{V}_\mathrm{f}$. Let j be in $\{0, \ldots, r-1\}$, then $\mu_j = (\mathbf{V}_\mathrm{f})_{jr}$ is given by

$$\mu_j = \sum_{k=0}^{j} (-1)^{j-k} \lambda^{j-k+1} \left[\tilde{b}_k - b_k - \tilde{b}_k \mu_r \right]. \tag{5.16}$$

Some computations show that μ_j is bounded for $\lambda \to \infty$ since the polynomial degrees of numerator and denominator of the resulting rational function are equal. This is important for the approximation error estimates with respect to some norm discussed in Section 5.2.3, since there we exploit the fact that the functions μ_j stay bounded on the positive half of the complex plane.

To formulate the decoupled approximate Newton update equation, we proceed as shown for example in [87]. We first transform $\mathbf{T}_\mathrm{f}$ back to the basis the Newton equation (5.11) was stated in and obtain

$$\mathbf{T} = \mathbf{C}^{-1}\mathbf{S}^{-1}\mathbf{T}_\mathrm{f}\mathbf{S}\mathbf{C}.$$

For decoupling the approximate system, we want to transform $\mathbf{T}$ to a lower triangular matrix. Since the spectrum of the matrix $\mathcal{T}$ consists only of the $(r+1)$-fold eigenvalue γ, there exists a decomposition of $\mathbf{T}$ of the form $\mathbf{T} = \gamma\mathbf{Q}(\mathrm{Id} - \mathbf{U})^{-1}\mathbf{Q}^{-1}$ with $\mathbf{U}$ being a strictly lower triangular matrix. Note that since $\mathbf{T}$ is similar to $\mathbf{T}_\mathrm{f}$, it is not diagonalizable and therefore we cannot achieve $\mathbf{U} = 0$. To obtain the decomposition of $\mathbf{T}$, we compute a Schur decomposition of the transpose $\mathbf{T}^T$ and take the transpose of the result. Replacing $\mathbf{T}$ in (5.13) by this representation we get

$$\begin{aligned}
&\left(\mathrm{Id}\otimes\mathrm{Id} + \gamma(\mathbf{Q}(\mathrm{Id}-\mathbf{U})^{-1}\mathbf{Q}^{-1})\otimes\mathbf{L}\right)\widetilde{\mathbf{w}} = \widehat{\mathbf{A}}^{-1}\otimes\mathbf{M}^{-1}R \\
\Leftrightarrow\ &\left(\left((\mathrm{Id}-\mathbf{U})\mathbf{Q}^{-1}\right)\otimes\mathrm{Id} + \gamma\mathbf{Q}^{-1}\otimes\mathbf{L}\right)\widetilde{\mathbf{w}} = \left((\mathrm{Id}-\mathbf{U})\mathbf{Q}^{-1}\widehat{\mathbf{A}}^{-1}\right)\otimes\mathbf{M}^{-1}R \\
\Leftrightarrow\ &\mathrm{Id}\otimes(\mathrm{Id}+\gamma\mathbf{L})\left(\mathbf{Q}^{-1}\otimes\mathrm{Id}\right)\widetilde{\mathbf{w}} = \left((\mathrm{Id}-\mathbf{U})\mathbf{Q}^{-1}\widehat{\mathbf{A}}^{-1}\right)\otimes\mathbf{M}^{-1}R \\
&\qquad + \left(\left(\mathbf{U}\mathbf{Q}^{-1}\right)\otimes\mathrm{Id}\right)\widetilde{\mathbf{w}}.
\end{aligned}$$

We substitute $\widetilde{\mathbf{x}} := \left(\mathbf{Q}^{-1}\otimes\mathrm{Id}\right)\widetilde{\mathbf{w}}$, introduce the matrix $\mathbf{F} = (\mathrm{Id}-\mathbf{U})\mathbf{Q}^{-1}\widehat{\mathbf{A}}^{-1}$ and multiply from left with $\mathrm{Id}\otimes\mathbf{M}$ to obtain the update scheme

$$\begin{aligned}
\mathrm{Id}\otimes(\mathbf{M}+\gamma k_m\mathbf{A})\,\widetilde{\mathbf{x}} &= (\mathbf{F}\otimes\mathrm{Id})\,R + (\mathbf{U}\otimes\mathbf{M})\,\widetilde{\mathbf{x}}, \\
\widetilde{\mathbf{w}} &= (\mathbf{Q}\otimes\mathrm{Id})\,\widetilde{\mathbf{x}}.
\end{aligned} \tag{5.17}$$

The first equation decouples into individual equations for the temporal components of $\widetilde{\mathbf{x}}$. For each component one standard θ step type of equation has to be solved. That means disregarding the cost for assembling the stiffness matrix, the computational effort per Newton step is roughly $r+1$ times the computational effort for a Newton step with implicit Euler time stepping. Compared to the schemes we proposed in [91], the computational effort per Newton step is reduced significantly while the speed of convergence is comparable. The complete algorithm for solving the time stepping equation is shown in Algorithm 5.1.

Computing the value γ and the coefficient matrices $\mathbf{F}$, $\mathbf{U}$ and $\mathbf{Q}$ according to the presented approach can be automatized with a computer algebra system. This allows to derive iteration schemes for very high order dG(r) methods with moderate effort. If the transformation from $\mathbf{T}_\mathrm{f}$ to $\mathbf{T}$ is evaluated numerically, sufficient precision of the computation has to be ensured since the involved matrices can become severely ill-conditioned. This is due to the involved transformations to the monomial basis.

Algorithm 5.1 Approximate Newton iteration for solving the dG(r) time stepping equation.

Input: starting value $\mathbf{u}^0$

for $l = 0, 1, 2, \ldots$ **do**
  Compute residual R^l from (5.3)
  for $i = 0, \ldots, r$ **do**
    Compute $\widetilde{\mathbf{x}}_i^l$ as solution of

$$(\mathbf{M} + \gamma k_m \mathbf{A})\, \widetilde{\mathbf{x}}_i^l = \sum_{j=0}^{r} \mathbf{F}_{ij} R_j^l + \sum_{j=0}^{i-1} \mathbf{U}_{ij} \mathbf{M} \widetilde{\mathbf{x}}_j^l.$$

  end for
  Set $\widetilde{\mathbf{w}}^l = (\mathbf{Q} \otimes \mathrm{Id})\, \widetilde{\mathbf{x}}^l$
  Set $\mathbf{u}^{l+1} = \mathbf{u}^l + \widetilde{\mathbf{w}}^l$
  if stopping criterion fulfilled **then**
    break
  end if
end for

5.2.3. Convergence analysis for linear problems with time-independent coefficients

Compared to an exact Newton iteration for the time stepping equation, the proposed solution process involves two sources of error that can affect the convergence behaviour. Averaging the Jacobian of the spatial operator over time introduces an error, another error arises from approximating $\mathcal{A}$ by $\mathbf{T}$. To investigate the latter error independent from the former, we restrict ourselves to linear problems without time-dependent coefficients in this section before discussing the general nonlinear setting in the next section.

For a linear equation, Newton's method computes the exact solution with a single iteration, so the rate of convergence of the approximated scheme can be estimated by

$$\theta^l = \frac{\left\| \widetilde{\mathbf{w}}^l - \mathbf{w}^l \right\|}{\|\mathbf{w}^l\|} \leq \|\mathbf{V}_\mathrm{f}(\mathbf{L})\|$$

where $\mathbf{V}_\mathrm{f}(\mathbf{L})$ denotes the block matrix obtained from replacing all occurrences of λ in $\mathbf{V}_\mathrm{f}$ by $\mathbf{L}$ and all fractions by matrix inverses. Here, $\|\cdot\|$ denotes some vector norm and the corresponding induced matrix norm.

It is desirable that the norm to measure convergence in is independent of the spatial mesh. For a discrete spatial function $v_h \in V_h^s$, a natural choice of a norm with that property is the spatial L^2 norm which in the discrete setting is given by

$$\|v_h\|_{\mathbf{L}^2(\Omega)} = \|\mathbf{v}\|_{\mathbf{M}} = \sqrt{\mathbf{v}^T \mathbf{M} \mathbf{v}}$$

where $\mathbf{v}$ is the nodal representation vector of v_h. We denote the induced matrix norm by $\|\cdot\|_{\mathbf{M}}$ as well. For later use, we need the following technical result.

Lemma 5.14. *Let $\mathbf{X}$ be an $(r+1)\times(r+1)$ matrix, $\mathbf{L} = k_m\mathbf{M}^{-1}\mathbf{A} \in \mathbb{R}^{N\times N}$, and q_k with $k = 0,\dots,r$ rational functions. By $Q(\mathbf{L})$ we denote the block matrix*

$$Q(\mathbf{L}) = \begin{pmatrix} q_0(\mathbf{L}) \\ \vdots \\ q_r(\mathbf{L}) \end{pmatrix}.$$

Correspondingly, for scalar values λ, $Q(\lambda)$ becomes a $\mathbb{R}^N$ vector. If $\mathbf{A}$ is symmetric, then there exists a $\mathbf{M}$-orthonormal basis $\{\mathbf{w}_1,\dots,\mathbf{w}_N\}$ of $\mathbb{R}^N$ such that for any vector $\mathbf{w} = \sum_{j=1}^N \omega_j\mathbf{w}_j$ we have the identity

$$\mathbf{w}^T Q(\mathbf{L})^T(\mathbf{X}\otimes\mathbf{M})Q(\mathbf{L})\mathbf{w} = \sum_{i=0}^{N} \omega_i^2 Q(\lambda_i)^T\mathbf{X}Q(\lambda_i)$$

with λ_i for $i = 1,\dots,N$ denoting the eigenvalues of $\mathbf{L}$. The basis $\{\mathbf{w}_1,\dots,\mathbf{w}_N\}$ does not depend on $\mathbf{X}$ and $Q(\mathbf{L})$.

In particular, the estimate

$$\left|\mathbf{w}^T Q(\mathbf{L})^T(\mathbf{X}\otimes\mathbf{M})Q(\mathbf{L})\mathbf{w}\right| \le \sup\left\{\left|Q(\lambda)^T\mathbf{X}Q(\lambda)\right| \;\middle|\; \lambda\in\sigma(\mathbf{L})\right\}\|w\|_{\mathbf{M}}^2$$

holds true.

Proof. Since the mass matrix $\mathbf{M}$ is symmetric positive definite, it possesses a square root $\mathbf{M}^{\frac{1}{2}}$. Due to symmetry of $\mathbf{A}$, the matrix

$$\mathbf{M}^{\frac{1}{2}}\mathbf{L}\mathbf{M}^{-\frac{1}{2}} = k_m\mathbf{M}^{-\frac{1}{2}}\mathbf{A}\mathbf{M}^{-\frac{1}{2}}$$

is symmetric as well. Therefore there is an orthonormal basis of $\mathbb{R}^N$ consisting of eigenvectors $\{\mathbf{v}_1,\dots,\mathbf{v}_N\}$ of this matrix. The corresponding eigenvalues are denoted by $\bar\lambda_1,\dots,\bar\lambda_N$. Defining the vectors

$$\mathbf{w}_j = \mathbf{M}^{-\frac{1}{2}}\mathbf{v}_j, \quad j = 1,\dots,N,$$

we note that they are the solutions of the generalized eigenvalue problem

$$\mathbf{A}\mathbf{w}_j = \bar\lambda_j\mathbf{M}\mathbf{w}_j$$

and form an orthonormal basis with respect to the inner product corresponding to $\|\cdot\|_{\mathbf{M}}$. From the above identity we see that $\mathbf{w}_j$ are eigenvectors of $\mathbf{L}$ with the eigenvalues $\lambda_j := k_m\bar\lambda_j$.

Let $\mathbf{w} \in \mathbb{R}^N$ with $\mathbf{w} = \sum_{j=1}^N \omega_j \mathbf{w}_j$. Then we get

$$\begin{aligned}\mathbf{w}^T Q(\mathbf{L})^T (\mathbf{X} \otimes \mathbf{M}) Q(\mathbf{L}) \mathbf{w} &= \sum_{i,j=1}^N \omega_i \omega_j \mathbf{w}_i^T Q(\mathbf{L})^T (\mathbf{X} \otimes \mathbf{M}) Q(\mathbf{L}) \mathbf{w}_j \\ &= \sum_{i,j=1}^N \omega_i \omega_j \mathbf{w}_i^T Q(\lambda_i)^T (\mathbf{X} \otimes \mathbf{M}) Q(\lambda_j) \mathbf{w}_j \\ &= \sum_{i,j=1}^N \omega_i \omega_j \mathbf{w}_i^T \mathbf{M} \mathbf{w}_j Q(\lambda_i)^T \mathbf{X} Q(\lambda_j) = \sum_{i=1}^N \omega_i^2 Q(\lambda_i)^T \mathbf{X} Q(\lambda_i).\end{aligned}$$

The inequality follows by taking the supremum over the terms containing λ_i and remembering that $\sum_{i=1}^N \omega_i^2 = \|\mathbf{w}\|_{\mathbf{M}}^2$. □

To measure the convergence of the iterative scheme for the dG time stepping equation, we need a norm on $\mathcal{P}_r(I_m, V_h)$. Using the $L^2(I_m \times \Omega)$ norm seems natural, however the following detailed analysis reveals that we cannot guarantee contraction with respect to this norm. To see that, we restrict the analysis once more to the scalar test problem. For this configuration, the operator norm induced by the L^2 norm can be stated explicitly for the error estimation matrix. Recalling that the Hilbert matrix $\mathbf{H}$ is the mass matrix for the polynomial space $\mathcal{P}_r((0,1))$ with respect to the monomial basis, we see that with the transformation matrix $\mathbf{S}$, $\widehat{\mathbf{B}}_\mathrm{f} = \mathbf{S}^{-1}\mathbf{H}\mathbf{S}^{-1}$ is the temporal mass matrix with respect to the Taylor basis (see Remark 5.12). As before we denote the corresponding norm by $\|\cdot\|_{\widehat{\mathbf{B}}_\mathrm{f}} = \left(\cdot^T \widehat{\mathbf{B}}_\mathrm{f} \cdot\right)^{\frac{1}{2}}$.

Proposition 5.15. *The L^2 operator norm $\|\mathbf{V}_\mathrm{f}\|_{\widehat{\mathbf{B}}_\mathrm{f}}$ of the error matrix for a scalar test equation with $L = \lambda$ is given by*

$$\|\mathbf{V}_\mathrm{f}(\lambda)\|_{\widehat{\mathbf{B}}_\mathrm{f}} = \left(\frac{1}{\beta - \mathbf{b}^T \tilde{\mathbf{B}}^{-1} \mathbf{b}} \sum_{i,j=0}^r (\widehat{\mathbf{B}}_\mathrm{f})_{ij}\ \mu_i(\lambda)\mu_j(\lambda) \right)^{\frac{1}{2}}$$

where $\tilde{\mathbf{B}} \in \mathbb{R}^{r \times r}$, $\mathbf{b} \in \mathbb{R}^r$, and $\beta \in \mathbb{R}$ with

$$\begin{bmatrix} \tilde{\mathbf{B}} & \mathbf{b} \\ \mathbf{b}^T & \beta \end{bmatrix} = \widehat{\mathbf{B}}_\mathrm{f}.$$

Proof. According to definition, we have

$$\|\mathbf{V}_\mathrm{f}(\lambda)\|_{\widehat{\mathbf{B}}_\mathrm{f}} = \sup_{\substack{\mathbf{w} \in \mathbb{R}^{r+1} \\ \mathbf{w} \neq 0}} \frac{\|\mathbf{V}_\mathrm{f}(\lambda)\mathbf{w}\|_{\widehat{\mathbf{B}}_\mathrm{f}}}{\|\mathbf{w}\|_{\widehat{\mathbf{B}}_\mathrm{f}}}. \tag{5.18}$$

The goal is now to compute a $\mathbf{w}$ for which the supremum is attained. Without loss of generality we can assume $\mathbf{w}_r = 1$, since for $\mathbf{w}_r = 0$, the numerator vanishes. Then we get

$$\|\mathbf{V}_\mathrm{f}(\lambda)\mathbf{w}\|_{\widehat{\mathbf{B}}_\mathrm{f}}^2 = \mathbf{w}^T \mathbf{V}_\mathrm{f}(\lambda)^T \widehat{\mathbf{B}}_\mathrm{f} \mathbf{V}_\mathrm{f}(\lambda)\mathbf{w} = \sum_{i,j=0}^r (\widehat{\mathbf{B}}_\mathrm{f})_{ij}\ \mu_i(\lambda)\mu_j(\lambda).$$

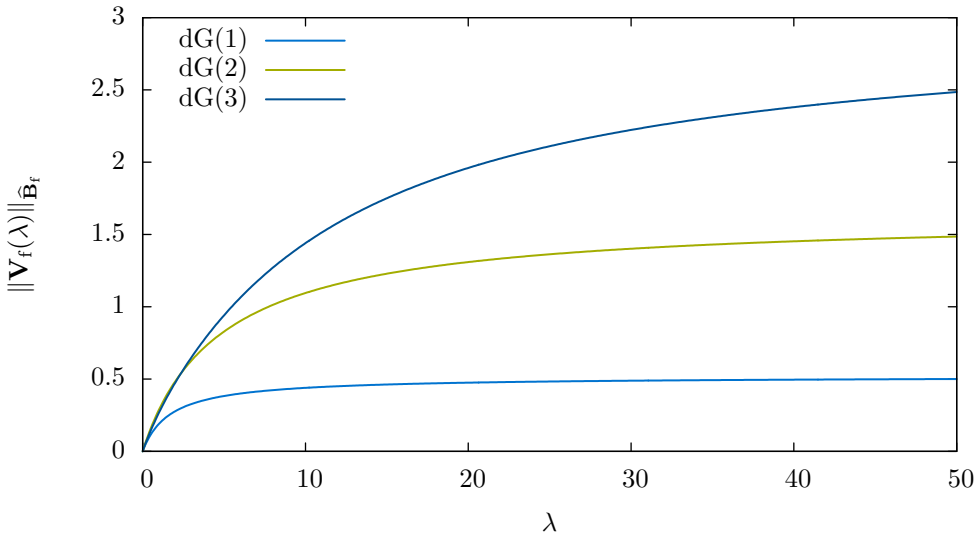


Figure 5.1.: Bound for L^2 accuracy of approximate Newton

In particular, the numerator is independent of the other components of $\mathbf{w}$. Hence, for the supremum to be assumed, we have to choose the remaining components of $\mathbf{w}$ in such a way that the denominator is minimized. So with $\mathbf{w}^T = (\bar{\mathbf{w}}^T \ \mathbf{w}_r)$, we have to minimize

$$\|\mathbf{w}\|^2_{\widehat{\mathbf{B}}_f} = \mathbf{w}^T \widehat{\mathbf{B}}_f \mathbf{w} = \begin{pmatrix} \bar{\mathbf{w}}^T & \mathbf{w}_r \end{pmatrix} \begin{bmatrix} \tilde{\mathbf{B}} & \mathbf{b} \\ \mathbf{b}^T & \beta \end{bmatrix} \begin{pmatrix} \bar{\mathbf{w}} \\ \mathbf{w}_r \end{pmatrix} = \bar{\mathbf{w}}^T \tilde{\mathbf{B}} \bar{\mathbf{w}} + 2\bar{\mathbf{w}}^T \mathbf{b} + \beta.$$

Since $\tilde{\mathbf{B}}$ is positive definite, the sufficient optimality condition for this linear-quadratic problem gives

$$\bar{\mathbf{w}} = -\tilde{\mathbf{B}}^{-1}\mathbf{b}.$$

Evaluating the quotient in (5.18) for $\mathbf{w}^T = \begin{pmatrix} \bar{\mathbf{w}}^T & 1 \end{pmatrix}$ shows the claim. □

In Figure 5.1 we plot the L^2 norm of the error matrix with respect to λ for $r = 1, 2, 3$ and positive real λ. One can see that for $r > 1$, contraction with respect to the L^2 norm can not be guaranteed and also for $r = 1$ the estimate departs significantly from the spectral radius of $\mathbf{V}_f$. Therefore the L^2 norm is of limited use for analyzing convergence.

From the structure of $\mathbf{V}_f(\lambda)$ we see that this matrix is diagonalizable. According to a standard result from linear algebra, there exists a norm such that the induced matrix norm of $\mathbf{V}_f(\lambda)$ equals to its spectral radius. However, we note that this norm depends on λ or $\mathbf{L}$ respectively and therefore in the case of non-linear problems on the current iterate. Hence for convergence analysis, we have to find a norm that admits a reasonable contraction bound independent of λ. An obvious candidate is the Euclidean norm

with respect to the Taylor basis introduced in Remark 5.12. So for a given vector $\mathbf{w}^T = \begin{pmatrix}\mathbf{w}_0^T & \dots & \mathbf{w}_r^T\end{pmatrix}$ we define the norm by

$$\|\mathbf{w}\|_{\mathrm{Id}\otimes\mathbf{M}} = \left(\sum_{j=0}^{r} \mathbf{w}_j^T \mathbf{M}\mathbf{w}_j\right)^{\frac{1}{2}}.$$

The following lemma gives an estimate for the corresponding induced operator norm of the error matrix $\mathbf{V}_\mathrm{f}$.

Lemma 5.16. *If the spatial differential operator determining the stiffness matrix* $\mathbf{A}$ *satisfies Assumption 5.7, the norm* $\|\mathbf{V}_\mathrm{f}(\mathbf{L})\|_{\mathrm{Id}\otimes\mathbf{M}}$ *can be bounded by*

$$\|\mathbf{V}_\mathrm{f}(\mathbf{L})\|_{\mathrm{Id}\otimes\mathbf{M}} \leq \sup_{\lambda\in\sigma(\mathbf{L})} \left(\sum_{j=0}^{r} \mu_j(\lambda)^2\right)^{\frac{1}{2}}$$

where μ_j, $j = 0, \dots, r$ *are the entries in the last column of* $\mathbf{V}_\mathrm{f}$ *given in* (5.15) *and* (5.16).

Under the more general Assumption 5.9 on the differential operator, the estimate reads

$$\|\mathbf{V}_\mathrm{f}(\mathbf{L})\|_{\mathrm{Id}\otimes\mathbf{M}} \leq \sup_{\lambda\in\mathbb{C}^+} \left(\sum_{j=0}^{r} \mu_j(\lambda)^2\right)^{\frac{1}{2}}$$

instead where $\mathbb{C}^+ = \{\lambda \in \mathbb{C} \mid \operatorname{Re}\lambda \geq 0\}$.

Proof. We first consider the case when Assumption 5.7 holds and note that in this case, Lemma 5.14 can be applied. Starting with the definition of the operator norm

$$\|\mathbf{V}_\mathrm{f}(\mathbf{L})\|^2_{\mathrm{Id}\otimes\mathbf{M}} = \sup_{\substack{\mathbf{w}\in\mathbb{R}^{(r+1)\cdot N}\\ \mathbf{w}\neq 0}} \frac{\mathbf{w}^T\mathbf{V}_\mathrm{f}(\mathbf{L})^T\,\mathrm{Id}\otimes\mathbf{M}\mathbf{V}_\mathrm{f}(\mathbf{L})\mathbf{w}}{\mathbf{w}^T\,\mathrm{Id}\otimes\mathbf{M}\mathbf{w}}$$

we can estimate the numerator by Lemma 5.14 with $\mathbf{X} = \mathrm{Id}$ and $q_i = \mu_i$ for $i = 0, \dots, r$

$$\mathbf{w}^T\mathbf{V}_\mathrm{f}(\mathbf{L})^T\,\mathrm{Id}\otimes\mathbf{M}\mathbf{V}_\mathrm{f}(\mathbf{L})\mathbf{w} = \mathbf{w}_r^T \sum_{j=0}^{r} \mu_j(\mathbf{L}^T)\mathbf{M}\mu_j(\mathbf{L})\mathbf{w}_r \leq \sup_{\lambda\in\sigma(\mathbf{L})} \left(\sum_{j=0}^{r} \mu_j(\lambda)^2\right) \|\mathbf{w}_r\|^2_{\mathbf{M}}.$$

For the denominator, we have

$$\mathbf{w}^T\,\mathrm{Id}\otimes\mathbf{M}\mathbf{w} \geq \mathbf{w}_r^T\mathbf{M}\mathbf{w}_r = \|\mathbf{w}_r\|^2_{\mathbf{M}}.$$

This shows the desired bound for the norm.

The more general Assumption 5.9 on the differential operator implies that for any nodal vector $\mathbf{v} \in \mathbb{R}^N$, the stiffness matrix satisfies $\mathbf{v}^T\mathbf{A}\mathbf{v} \geq 0$ and hence $(\mathbf{v}, \mathbf{L}\mathbf{v})_\mathbf{M} \geq 0$. Therefore, for any complex vector $\mathbf{u} \in \mathbb{C}^N$ we have

$$\operatorname{Re}(\mathbf{u}, \mathbf{L}\mathbf{u})_M \geq 0.$$

r	ρ_r	$\tilde{\rho}_r^1$	$\tilde{\rho}_r^2$
1	0.092	0.150	0.170
2	0.169	0.173	0.343
3	0.238	0.244	0.493
4	0.301	0.315	0.618
5	0.359	0.381	0.719
6	0.412	0.442	0.802
7	0.460	0.498	0.868

Table 5.2.: Upper bounds ρ_r, $\tilde{\rho}_r^1$, and $\tilde{\rho}_r^2$ for the spectral radius and the $\|\cdot\|_{\mathrm{Id}\otimes\mathbf{M}}$ norm of the matrix $\mathbf{V}_{\mathrm{f}}(\mathbf{L})$

Hence we can apply the matrix-valued generalization of a theorem going back to von Neumann, which can be found as Corollary 3 in Nevanlinna [86]. This yields immediately the estimate

$$\|\mathbf{V}_{\mathrm{f}}(\mathbf{L})\|_{\mathrm{Id}\otimes\mathbf{M}} \leq \sup_{\lambda\in\mathbb{C}^+} \|\mathbf{V}_{\mathrm{f}}(\lambda)\|_{\mathrm{Id}}$$

and evaluating the matrix norm on the right hand side gives the claim. □

Looking at the structure of μ_j for $j = 0, \ldots, r$, we note that the only singularities occur at $\lambda = -\frac{1}{\gamma}$. Also some computations show that the degree of numerator and denominator are identical for each μ_j. This means in particular that μ_j is bounded for $\operatorname{Re}\lambda \geq 0$ which implies that if the differential operator satisfies one of the Assumptions 5.7 or 5.9, then $\|\mathbf{V}_{\mathrm{f}}(\mathbf{L})\|_{\mathrm{Id}\otimes\mathbf{M}}$ is bounded. Approximations for the upper bounds

$$\tilde{\rho}_r^1 = \sup_{\lambda\in\mathbb{R}^+} \left(\sum_{j=0}^{r} \mu_j(\lambda)^2 \right)^{\frac{1}{2}}$$

and

$$\tilde{\rho}_r^2 = \sup_{\lambda\in\mathbb{C}^+} \left(\sum_{j=0}^{r} \mu_j(\lambda)^2 \right)^{\frac{1}{2}}$$

can be computed numerically and are given in Table 5.2 for dG(1) up to dG(7). As it turns out, while $\tilde{\rho}_r^2$ grows larger than one for $r \geq 10$, for moderate order, all of them are less than 1. If $\mathbf{L}$ has positive real spectrum, the norm is even reasonably close to the maximal spectral radius ρ_r of $\mathbf{V}_{\mathrm{f}}(\mathbf{L})$. Hence the following corollary is an immediate consequence of the above lemma.

Corollary 5.17. *Let the matrix $\mathbf{X}$ be given as $\mathbf{X} = \mathbf{C}^T\mathbf{S}^T\mathbf{S}\mathbf{C}$. We consider the discrete dG(r) time time stepping equation with $r \leq 7$ for a linear problem without time dependent coefficients satisfying either Assumption 5.7 or 5.9. Let $\bar{\mathbf{u}}$ denote the solution of the time*

stepping equation and $\mathbf{u}^0$ *some arbitrary starting value. Then the iterates* $\mathbf{u}^l$ *generated by Algorithm 5.1 satisfy the estimate*

$$\left\|\mathbf{u}^{l+1} - \bar{\mathbf{u}}\right\|_{\mathbf{X}\otimes\mathbf{M}} \leq \rho \left\|\mathbf{u}^{l} - \bar{\mathbf{u}}\right\|_{\mathbf{X}\otimes\mathbf{M}},$$

with contraction rate $\rho = \tilde{\rho}_r^1$ *in the case of Assumption 5.7 and* $\rho = \tilde{\rho}_r^2$ *in the case of Assumption 5.9.*

Although the Taylor norm allows for stating the convergence result Corollary 5.17, it is not the first choice for monitoring convergence in a practical realization of the approximate Newton scheme. For a nodal basis with the Gauß-Legendre points as nodes, we observe that its mass matrix $\mathbf{X}$ becomes severely ill-conditioned for increasing r and already for $r = 7$, the condition number exceeds double precision. Additionally, numerical tests show that when a_u is not constant with respect to time, contraction with respect to this norm is quite sensitive towards approximating the temporal integrals over a_u by a mid point evaluation. Using the L^2 norm in time instead appears to result in more robust convergence although we have shown in Proposition 5.15 that contraction cannot be guaranteed in general. The following result offers a partial explanation of this phenomenon.

Proposition 5.18. *For a linear parabolic equation with time-independent coefficients that fulfills Assumption 5.7, the approximations* $\mathbf{u}^l$ *for the solution* $\bar{\mathbf{u}}$ *of the time stepping equation produced by Algorithm 5.1 satisfy for* $l \geq 1$

$$\left\|\mathbf{u^{l+1}} - \bar{\mathbf{u}}\right\|_{\mathbf{Y}\otimes\mathbf{M}} \leq \rho_r \left\|\mathbf{u^{l}} - \bar{\mathbf{u}}\right\|_{\mathbf{Y}\otimes\mathbf{M}}$$

with the vector norm $\|\cdot\|_{\mathbf{Y}\otimes\mathbf{M}}$ *defined with an arbitrary symmetric positive definite* $(r+1)\times(r+1)$ *matrix* $\mathbf{Y}$*.*

Proof. Without loss of generality we can assume that we operate in the Taylor basis since positive definiteness of $\mathbf{Y}$ is invariant under change of coordinates.

First we observe that for a linear equation with time-independent coefficients the exact Newton update $\mathbf{w}^l$ is given by $\mathbf{w}^l = \bar{\mathbf{u}} - \mathbf{u}^l$ Hence for $l \in \mathbb{N}$ the relationship

$$\bar{\mathbf{u}} - \mathbf{u}^{l+1} = \bar{\mathbf{u}} - \left(\mathbf{u}^l + \widetilde{\mathbf{w}}^l\right) = \mathbf{w}^l - \widetilde{\mathbf{w}}^l = \mathbf{V}_{\mathrm{f}}(\mathbf{L})\mathbf{w}^l = \mathbf{V}_{\mathrm{f}}(\mathbf{L})\left(\bar{\mathbf{u}} - \mathbf{u}^l\right)$$

holds. Taking the structure of $\mathbf{V}_{\mathrm{f}}(\mathbf{L})$ into account, we obtain by induction the explicit representation

$$\bar{\mathbf{u}} - \mathbf{u}^l = \begin{pmatrix} \mu_0(\mathbf{L}) \\ \vdots \\ \mu_r(\mathbf{L}) \end{pmatrix} \left(\mu_r(\mathbf{L})\right)^{l-1} \left(\bar{\mathbf{u}}_r - \mathbf{u}_r^0\right)$$

for the errors if $l > 0$. Setting $\mathbf{X} = \mathbf{Y}$ and $q_j(\lambda) = \mu_j(\lambda)(\mu_r(\lambda))^{l-1}$ in Lemma 5.14, we obtain

$$\left\|\bar{\mathbf{u}} - \mathbf{u}^l\right\|^2_{\mathbf{Y}\otimes\mathbf{M}} = \sum_{i=1}^{N} \omega_i^2 \left(\mu_r(\lambda_i)\right)^{2(l-1)} \begin{pmatrix}\mu_0(\lambda_i) & \cdots & \mu_r(\lambda_i)\end{pmatrix} \mathbf{Y} \begin{pmatrix}\mu_0(\lambda_i) \\ \vdots \\ \mu_r(\lambda_i)\end{pmatrix}$$

where ω_i, $i = 1, \ldots, N$, denote the coefficients of $\bar{\mathbf{u}}_r - \mathbf{u}_r^0$ with respect to the spatial basis defined in Lemma 5.14. Hence, we can estimate

$$\left\|\bar{\mathbf{u}} - \mathbf{u}^{l+1}\right\|^2_{\mathbf{Y}\otimes\mathbf{M}} \leq \sup\left\{\left(\mu_r(\lambda)\right)^2 \;\middle|\; \lambda \in \mathbb{R}_0^+\right\} \left\|\bar{\mathbf{u}} - \mathbf{u}^l\right\|^2_{\mathbf{Y}\otimes\mathbf{M}} = \rho_r^2 \left\|\bar{\mathbf{u}} - \mathbf{u}^l\right\|^2_{\mathbf{Y}\otimes\mathbf{M}}. \quad \square$$

Remark 5.19. This result shows that for linear equations satisfying Assumption 5.7, apart from the initial iteration, contraction with rate ρ_r can be guaranteed for *any* choice of the temporal norm. For nonlinear problems however, the matrix $\mathbf{A}$ depends on the current iterate $\mathbf{u}^l$. Besides, an additional error from averaging the Jacobian enters such that contraction with respect to, e. g. the L^2 norm can not be guaranteed also for later iterates. However, due to the disadvantages of the Taylor norm discussed above, for algorithmic realization we propose a pragmatic approach. We use the L^2 norm of the update for convergence monitoring but do not require contraction in every single iteration. This seems to work well in practice. However, in case it is not sufficient, an alternative would be to use the approximate solution scheme only as an inner linear solver for the Newton update equation. In this case, the linear convergence theory above applies again.

5.2.4. Convergence analysis for nonlinear equations

Our convergence analysis for non-linear equations is based on the corresponding result by Calvo et al. [22] for implicit Runge-Kutta schemes. However, we use a more general Lipschitz condition (Assumption 5.10). Furthermore, some technical difference in the proof arises from the fact that in the dG time stepping formulation, the spatial semilinear form is integrated over time whereas the Runge-Kutta schemes evaluate it only in a few discrete time points. For the convenience of the reader, we give a complete proof for the modified result.

Theorem 5.20. *For a problem satisfying one of Assumptions 5.7 and 5.9 and additionally Assumption 5.10 we consider Algorithm 5.1 for solving the dG(r) time stepping equation on the interval I_m with length k_m. Let $\bar{u}$ denote the exact solution of the discrete time stepping equation. We assume that the initial iterate $u^0 \in \mathcal{P}_r(I_m, V_h)$ satisfies*

$$\left\|u^0 - \bar{u}\right\|_{L^\infty(I_m,H)} \leq C_1 k_m$$

with C_1 independent of k_m. Then there is a time step size $\bar{k}$ such that for any $k_m \leq \bar{k}$, Algorithm 5.1 converges to the solution of the time stepping equation and for the iteration errors $e^l = \bar{u} - u^l$, an estimate of the form

$$\left\|\mathbf{e}^l\right\|_{\mathrm{Id}\otimes M} \leq \left(\rho + C_2 k_m + C_3 \left\|e^0\right\|_{L^\infty(I_m,H)}\right)^l \left\|\mathbf{e}^0\right\|_{\mathrm{Id}\otimes M}$$

for constants C_2 and C_3 independent from k_m holds true where $\mathbf{e}^l$ denotes the nodal representation of the error term e^l with respect to the Taylor basis. The value of ρ is either $\tilde{\rho}_r^1$ or $\tilde{\rho}_r^2$ depending on whether Assumption 5.7 or Assumption 5.9 applies.

Remark 5.21. To obtain the required first order accurate initial solution, usually it is sufficient to use the terminal value from the previous interval, i. e., to set $u^0(t) = u^-_{kh,m-1}$ for $t \in I_m$. If the terminal value is at least first order accurate and the dG(r) solution $\bar{u}$ on the current interval approximates the exact solution with order 1 pointwise in time, then a standard interpolation estimate for u^0 shows the desired accuracy.

Proof. We choose $\hat{\delta} > 0$ such that the initial iterate u^0 is contained in the ball $B_{\hat{\delta}}(\bar{u}) = \left\{u \in \mathcal{P}_r(I_m, V_h) \;\middle|\; \|u - \bar{u}\|_{L^\infty(I_m,V_h)}\right\}$. For $u \in \mathcal{P}_r(I_m, V_h)$, the norms $\|u\|_{L^\infty(I_m,V_h)}$ and $\|\mathbf{u}\|_{\mathrm{Id}\otimes\mathbf{M}}$, where $\mathbf{u}$ is the nodal representation of u with respect to the Taylor norm, are equivalent with constants c_1 and c_2 independent of k_m.

We set $\delta = c_1 c_2 \hat{\delta}$ and consider $B_\delta(\bar{u})$. Let the set N_δ be given by

$$N_\delta = \left\{(t, v) \in I \times V_h \;\middle|\; t \in I_m,\ v = u(\hat{t}) \text{ for some } \hat{t} \in I_m \text{ and } u \in B_\delta(\bar{u})\right\}.$$

For sufficiently small time step, N_δ is contained in some neighbourhood where Assumption 5.10 applies.

Our next aim is to estimate the error of the $(l+1)^{\text{th}}$ iterate $\mathbf{e}^{l+1} = \bar{\mathbf{u}} - \mathbf{u}^{l+1} = \mathbf{e}^l - \widetilde{\mathbf{w}}^l$ in terms of the previous error $\mathbf{e}^l$. For this purpose we assume that, by induction, $u^l \in N_\delta$.

Writing the update as solution of (5.13) and using the fact that the residual vanishes at the exact solution of the time stepping equation, i. e., $R(\bar{\mathbf{u}}) = 0$, we get

$$\mathbf{e}^{l+1} = \mathbf{e}^l + (\mathrm{Id} + \mathbf{T}_\mathrm{f} \otimes \mathbf{L})^{-1} \widehat{\mathbf{A}}^{-1} \otimes \mathbf{M}^{-1} \left(R(\bar{\mathbf{u}}) - R(\mathbf{u}^l)\right) \tag{5.19}$$

The difference of the two residuals can be rewritten by the mean value theorem as

$$R(\bar{\mathbf{u}}) - R(\mathbf{u}^l) = \int_0^1 R'(\mathbf{u}^\mathbf{l} + \tau \mathbf{e}^l)(\mathbf{e}^l)\,\mathrm{d}\tau. \tag{5.20}$$

Using the matrix $\dot{\mathbf{B}}_\mathrm{f}(t)$ with entries $\left(\dot{\mathbf{B}}_\mathrm{f}(t)\right)_{ij} = \psi_j(t)\psi_i(t)$ and the notation from Assumption 5.10 for the linearization of the spatial differential operator, a convenient representation of the derivative of the residual for given nodal solution $\mathbf{u}$ and associated discrete function $u \in \mathcal{P}_r(I_m, V_h)$ reads

$$R'(\mathbf{u}) = -\widehat{\mathbf{A}}_\mathrm{f} \otimes \mathbf{M} - \int_{I_m} \dot{\mathbf{B}}_\mathrm{f}(t) \otimes \mathbf{A}(t, u(t))\,\mathrm{d}t.$$

With this identity, (5.20) transforms to

$$\begin{aligned} R(\bar{\mathbf{u}}) - R(\mathbf{u}^l) &= -\left(\widehat{\mathbf{A}}_\mathrm{f} \otimes \mathbf{M} + \int_0^1 \int_{I_m} \dot{\mathbf{B}}_\mathrm{f}(t) \otimes \mathbf{A}(t, u^l(t) + \tau e^l(t))\,\mathrm{d}t\,\mathrm{d}\tau\right) \mathbf{e}^l \\ &= -\left(\widehat{\mathbf{A}}_\mathrm{f} \otimes \mathbf{M} + k_m \widehat{\mathbf{B}}_\mathrm{f} \otimes \mathbf{A} + \int_0^1 \int_{I_m} \dot{\mathbf{B}}_\mathrm{f}(t) \otimes \Big(\mathbf{A}(t, u^l(t) + \tau e^l(t)) - \mathbf{A}\Big)\,\mathrm{d}t\,\mathrm{d}\tau\right) \mathbf{e}^l. \end{aligned}$$

Inserting this equality into (5.19) and using the definition of the matrix $\mathbf{V}_\mathrm{f}(\mathbf{L})$ results in

(5.21)
$$\begin{aligned} \mathbf{e}^{l+1} &= \mathbf{V}_\mathrm{f}(\mathbf{L})\mathbf{e}^l + (\mathrm{Id} + \mathbf{T}_\mathrm{f} \otimes \mathbf{L})^{-1} \widehat{\mathbf{A}}_\mathrm{f}^{-1} \otimes \mathbf{M}^{-1} \\ &\qquad \cdot \int_0^1 \int_{I_m} \dot{\mathbf{B}}_\mathrm{f}(t) \otimes \Big(\mathbf{A}(t, u^l(t) + \tau e^l(t)) - \mathbf{A}\Big)\,\mathrm{d}t\,\mathrm{d}\tau\ \mathbf{e}^l \\ &= \mathbf{V}_\mathrm{f}(\mathbf{L})\mathbf{e}^l + (\mathrm{Id} + \mathbf{T}_\mathrm{f} \otimes \mathbf{L})^{-1} \widehat{\mathbf{A}}_\mathrm{f}^{-1} \otimes (\mathrm{Id} + \mathbf{L}) \\ &\qquad \cdot \int_0^1 \int_{I_m} \dot{\mathbf{B}}_\mathrm{f}(t) \otimes \Big\{(\mathbf{M} + k_m\mathbf{A})^{-1} \Big(\mathbf{A}(t, u^l(t) + \tau e^l(t)) - \mathbf{A}\Big)\Big\}\,\mathrm{d}t\,\mathrm{d}\tau\ \mathbf{e}^l. \end{aligned}$$

A bound for the matrix norm of $\mathbf{V}_\mathrm{f}$ is provided by Corollary 5.17, so it remains to bound the matrix on the second line. For this purpose, we split it into two factors that we discuss separately. Using once more the matrix valued version of von Neumann's theorem (Corollary 3 in Nevanlinna [86]), we can estimate

(5.22)
$$\begin{aligned} &\left\|(\mathrm{Id} + \mathbf{T}_\mathrm{f} \otimes \mathbf{L})^{-1} \widehat{\mathbf{A}}_\mathrm{f}^{-1} \otimes (\mathrm{Id} + \mathbf{L})\right\|_{\mathrm{Id} \otimes \mathbf{M}} \\ &\qquad\qquad \leq \sup_{\lambda \in \mathbb{C}^+} \left\|(\mathrm{Id} + \lambda \mathbf{T}_\mathrm{f})^{-1} (1 + \lambda) \widehat{\mathbf{A}}_\mathrm{f}^{-1}\right\|_\mathrm{Id} =: \kappa_3. \end{aligned}$$

Since $\mathbf{T}_\mathrm{f}$ has the $(r+1)$-fold eigenvalue γ, which is positive, κ_3 is finite. To estimate the integral terms, we use Assumption 5.10 and note that a short calculation yields $k_m^{-1} \int_{I_m} \left\|\dot{\mathbf{B}}_\mathrm{f}(t)\right\|_\mathrm{Id} \,\mathrm{d}t = \int_0^1 \sum_{j=0}^r \frac{1}{(j!)^2} t^{2j} \,\mathrm{d}t \leq 2$.

(5.23)
$$\begin{aligned} &\left\|\int_0^1 \int_{I_m} \dot{\mathbf{B}}_\mathrm{f}(t) \otimes \Big\{(\mathbf{M} + k_m\mathbf{A})^{-1} \Big(\mathbf{A}(t, u^l(t) + \tau e^l(t)) - \mathbf{A}\Big)\Big\}\,\mathrm{d}t\,\mathrm{d}\tau\right\|_{\mathrm{Id} \otimes M} \\ &\leq \frac{1}{k_m} \int_{I_m} \left\|\dot{\mathbf{B}}_\mathrm{f}(t)\right\|_\mathrm{Id} \,\mathrm{d}t \int_0^1 \sup_{t \in I_m} \left\|(\mathbf{M} + k_m\mathbf{A})^{-1} k_m \Big(\mathbf{A}(t, u^l(t) + \tau e^l(t)) - \mathbf{A}\Big)\right\|_M \,\mathrm{d}\tau \\ &\leq 2\left(\kappa_1 \frac{k_m}{2} + \kappa_2 \left\|e^l\right\|_{L^\infty(I_m, H)}\right). \end{aligned}$$

To bound the error $\mathbf{e}^{l+1}$, we apply the estimates (5.22) and (5.23) to the representation (5.21) and obtain

$$\left\|\mathbf{e}^{l+1}\right\|_{\mathrm{Id} \otimes \mathbf{M}} \leq \left(\|\mathbf{V}_\mathrm{f}(\mathbf{L})\|_{\mathrm{Id} \otimes \mathbf{M}} + 2\kappa_3 \left(\kappa_1 \frac{k_m}{2} + \kappa_2 \left\|e^l\right\|_{L^\infty(I_m, H)}\right)\right) \left\|\mathbf{e}^l\right\|_{\mathrm{Id} \otimes \mathbf{M}}.$$

Under the assumption that all previous iteration steps were contracting with respect to the Taylor norm, we have

$$\left\|e^l\right\|_{L^\infty(I_m,H)} \leq c_1 \left\|\mathbf{e}^l\right\|_{\mathrm{Id}\otimes\mathbf{M}} \leq c_1 \left\|\mathbf{e}^0\right\|_{\mathrm{Id}\otimes\mathbf{M}} \leq c_1 c_2 \left\|e^0\right\|_{L^\infty(I_m,H)} \tag{5.24}$$

and therefore

$$\left\|\mathbf{e}^{l+1}\right\|_{\mathrm{Id}\otimes\mathbf{M}} \leq \left(\rho + C_2 k_m + C_3 \left\|e^0\right\|_{L^\infty(I_m,H)}\right) \left\|\mathbf{e}^l\right\|_{\mathrm{Id}\otimes\mathbf{M}}.$$

with $C_2 = \kappa_1\kappa_3$ and $C_3 = 2c_1c_2\kappa_2\kappa_3$. We see that for sufficiently small time step k_m, the iteration is contracting and due to (5.24), we also have $u^{l+1} \in B_\delta(\bar{u})$ which completes the induction argument. □

Remark 5.22. It deserves mentioning that the constant κ_3 grows very fast with r and hence for higher order schemes, the above convergence result is of limited practical use. However, in numerical tests with a semilinear test equation (see Section 5.4.1), we could not observe extreme growth of the constants C_2 and C_3.

5.2.5. Applicability of the convergence result to a semilinear model problem

As pointed out in Remark 5.11.2, the validity of the assumptions of Theorem 5.20 for concrete problems deserves some discussion. As an example, we consider a semilinear reaction-diffusion equation. For given $u_0 \in V$, find $u \in X \cap L^\infty(I \times \Omega)$ satisfying

$$\begin{aligned} \partial_t u - \Delta u + g(t,x,u) &= f \quad \text{in } I \times \Omega, \\ u(0) &= u_0 \end{aligned}$$

with $u_0 \in L^\infty(\Omega)$, $f \in L^q(I \times \Omega)$, $q > \frac{d}{2} + 1$, and $g\colon I \times \Omega \times \mathbb{R} \to \mathbb{R}$ satisfying

1. g is continuously differentiable with respect to u for almost all $(t,x) \in I \times \Omega$ and there is a $K > 0$ such that

$$\|g(\cdot,\cdot,0)\|_{L^\infty(I\times\Omega)} + \|g'_u(\cdot,\cdot,0)\|_{L^\infty(I\times\Omega)} \leq K.$$

2. The first derivative of g is uniformly Lipschitz-continuous in u on bounded sets, that is, for any $S > 0$ there is $L(S) > 0$ such that

$$\|g'_u(\cdot,\cdot,u_1) - g'_u(\cdot,\cdot,u_2)\|_{L^\infty(I\times\Omega)} \leq L(S)|u_1 - u_2|$$

   for any $u_1, u_2 \in \mathbb{R}$ with $|u_1|, |u_2| \leq S$.

3. For almost all $(t,x) \in I \times \Omega$ and all $u \in \mathbb{R}$, the monotonicity condition

$$g'_u(\cdot,\cdot,u) \geq 0$$

   is fulfilled.

With these assumptions, well-posedness of the equation can be shown as in Neitzel and Vexler [85]. The semilinear form corresponding to the problem is given as

$$a(t,u)(\varphi) = (\nabla u, \nabla\varphi) + \int_\Omega g(t,x,u)\varphi \,\mathrm{d}x - (f,\varphi).$$

Due to the Lipschitz condition on the first derivative of g, the semilinear form is Fréchet differentiable for $u \in V \cap L^\infty(\Omega)$ and the derivative reads

$$a'_u(t,u)(\psi,\varphi) = (\nabla\psi, \nabla\varphi) + \int_\Omega g'_u(t,x,u)\psi\varphi \,\mathrm{d}x.$$

That this expression satisfies Assumption 5.7 follows immediately from the monotonicity of g. To satisfy the Lipschitz condition in Assumption 5.10, a further restriction is required as the proof of the following result shows.

Proposition 5.23. *We consider the semilinear problem with the above assumptions satisfied. Additionally we impose a stronger Lipschitz condition on the derivative of the non-linearity g, that is, there is a constant $L(S)$ for any $S > 0$ such that for any $u_1, u_2 \in \mathbb{R}$ and $t_1, t_2 < S$, we have*

$$\|g'_u(t_1,\cdot,u_1) - g'_u(t_2,\cdot,u_2)\| \le L(S)\left(|u_1 - u_2| + |t_1 - t_2|\right).$$

Besides, we have to assume that the spatial discretization is quasi-uniform, i. e., the quotient of the largest and smallest cell diameter is uniformly bounded independent of the discretization parameter h. Then the discretized semilinear form satisfies Assumption 5.10 with constants κ_1 and κ_2 independent of the discretization parameter h.

Remark 5.24. That the Lipschitz condition on g'_u is enforced globally in u constitutes a significant restriction. Already for the semilinear example problem with $g(t,x,u) = u^3$ given in Section 2.3.2, this assumption does not hold. It is needed because in the proof of the proposition, an estimate of the form $\|g'_u(t_1,\cdot,u_1) - g'_u(t_1,\cdot,u_1)\| \le C\left(\|u_1 - u_2\| + |t_1 - t_2|\right)$ with the constant C independent of the discretization parameter is required. Such an estimate would also hold true without the global Lipschitz condition if it was known that u_1 and u_2 were bounded in $L^\infty(\Omega)$. In this case, the constant C would depend on the L^∞ bound. So if we could show that the iterates produced by the time stepping solver were uniformly bounded in $L^\infty(\Omega)$ (and not only in $L^2(\Omega)$), then a local Lipschitz condition would be sufficient.

Proof. For given $t_1, t_2 \in I_m$ and $u_1, u_2 \in V_h$, let

$$\nu = \left\| \left(\mathbf{M} + k\mathbf{A}(t_1,u_1)\right)^{-1} k\left(\mathbf{A}(t_2,u_2) - \mathbf{A}(t_1,u_1)\right) \right\|_{\mathbf{M}}.$$

We consider the following discrete problem: given $f_h \in V_h$, find $v_h \in V_h$ such that

$$k a'_u(t_1,u_1)(v_h,\varphi) + (v_h,\varphi) = k\left(a'_u(t_2,u_2)(f_h,\varphi) - a'_u(t_2,u_2)(f_h,\varphi)\right)$$

for all $\varphi \in V_h$. It is easy to see that ν is the smallest constant such that

$$\|v_h\| \leq \nu \|f_h\|$$

for any $f_h \in V_h$. For the model problem, the above equation reads

$$k(\nabla v_h, \nabla \varphi) + \int_\Omega \left(1 + k g'_u(t_1, x, u_1)\right) v_h \varphi \,\mathrm{d}x = k \int_\Omega \left(g'_u(t_2, x, u_2) - g'_u(t_1, x, u_1)\right) f_h \varphi \,\mathrm{d}x.$$

To derive the stability estimate, we introduce a discrete dual problem which reads: find $z_h \in V_h$ satisfying

$$k(\nabla \varphi, \nabla z_h) + \int_\Omega \left(1 + k g'_u(t_1, x, u_1)\right) \varphi z_h \,\mathrm{d}x = k(v_h, \varphi) \quad \text{for all } \varphi \in V_h.$$

In the same way as in the proof of Lemma 4.5 in Meidner and Vexler [79], for z_h an a priori estimate of the form

$$\|z_h\|_{L^\infty(\Omega)} \leq C \|v_h\|$$

with C independent of h can be shown. Since the proof involves an inverse estimate, the assumption of quasi-uniformity of the spatial mesh is used. Testing the dual equation with v_h and inserting the primal problem gives

$$\begin{aligned}
k \|v_h\|^2 &= k(\nabla v_h, \nabla z_h) + \int_\Omega \left(1 + k g'_u(t_1, x, u_1)\right) v_h z_h \,\mathrm{d}x \\
&= k \int_\Omega \left(g'_u(t_2, x, u_2) - g'_u(t_1, x, u_1)\right) f_h z_h \,\mathrm{d}x \\
&\leq \left\|g'_u(t_2, \cdot, u_2) - g'_u(t_1, \cdot, u_1)\right\| \|f_h\| \|z_h\|_{L^\infty(\Omega)} \\
&\leq CL(S) \left(\|u_2 - u_1\| + |t_2 - t_1|\right) \|f_h\| \|v_h\|
\end{aligned}$$

Dividing by $\|v_h\|$ shows the desired estimate

$$\nu \leq CL(S) \left(\|u_2 - u_1\| + |t_2 - t_1|\right),$$

which completes the proof. □

5.3. Practical realization

In this section we discuss some aspects of the practical realization of the time stepping solver. First, we briefly elaborate on our choice of the termination criterion and subsequently present a strategy for detecting and handling convergence problems. Consistent with the theoretical results, the iteration should be controlled in the affine covariant setting, i. e., we work in terms of the error of the iterate instead of in terms of the residual. A systematic discussion of convergence control for implicit ODE solvers in this setting is given by Gustafsson and Söderlind [50]. We will adopt their approach for our purposes.

5.3.1. Termination criterion

Since we are interested in approximating the solution of the time stepping equation to a specific accuracy, for the termination criterion, a good estimate of the corresponding error is necessary. From the analysis carried out in the last few sections we know that we can expect at best linear convergence in terms of the error. Assuming the contraction rate can be bounded from above by the value $\theta < 1$ in the considered norm, the error in iteration l, $\mathbf{e}^l$ can be bounded in terms of the update $\widetilde{\mathbf{w}}^l$ by

$$\left\|\mathbf{e}^l\right\| \leq \frac{\theta}{1-\theta}\left\|\widetilde{\mathbf{w}}^l\right\|. \tag{5.25}$$

This can be seen by estimating the sum over all remaining updates.

For linear problems with constant coefficients, rigorous upper bounds for the contraction rate were given in Corrolary 5.17. However, these require the use of the Taylor norm with respect to the temporal discretization, which is not desirable for practical usage for the reasons pointed out in Remark 5.19. Additionally in the general case, the contraction rate can be worse due to the presence of nonlinearities and the averaging error of the stiffness matrix. As a numerical estimate for the contraction rate in the norm, the quotient of the norms of two successive iterates can be used. Since this might be too optimistic, we take the maximum of the contraction rates observed in the last few (e. g., three) steps as estimate for the average contraction rate. So in step l we use $\bar{\theta}^l$ given by

$$\bar{\theta}^l = \max\left\{\|\widetilde{\mathbf{w}}^{l-j}\|/\|\widetilde{\mathbf{w}}^{l-j-1}\| \;\middle|\; j = 0,\ldots,2\right\} \tag{5.26}$$

to estimate the iteration error via (5.25). Apart from some tests with the Taylor norm reported in Section 5.4, we use the norm defined by $\widehat{\mathbf{B}} \otimes \mathbf{M}$ in practice. Since $\widehat{\mathbf{B}}$ is the mass matrix for the reference interval $(0,1)$, this norm is the $L^2(I_m, H)$ norm scaled by $\frac{1}{\sqrt{k_m}}$ and therefore equivalent to the $L^\infty(I_m, H)$ norm.

5.3.2. Controlling the iteration

Theorem 5.20 suggests that convergence failures or slow convergence may occur in the presence of nonlinearities or time-dependent coefficients if the time step was chosen too large. On the other hand, due to the high accuracy of higher order dG schemes, it is desirable to use as large as possible time steps for the discretization in order to minimize the number of degrees of freedom. Ideally, the size of the time steps is determined in such a way that the degradation of the rate of convergence for large time steps is balanced against the extra computational work incurred by using smaller time steps. Such balancing strategies for implicit Runge-Kutta solvers are described for example by Hairer and Wanner [51] and in a more systematic fashion by Gustafsson and Söderlind [50]. In this context where only the solution of the forward equation is required, the cost added through shortening time steps is simply the cost of solving the resulting time stepping equations. When solving optimal control problems, the situation is more complicated.

Algorithm 5.2 Convergence control for the time stepping solver

Input: l, $\left\|\widetilde{\mathbf{w}}^{l}\right\|$, $\left\|\widetilde{\mathbf{w}}^{l-1}\right\|$, $\left\|\widetilde{\mathbf{w}}^{l-2}\right\|$, $\left\|\widetilde{\mathbf{w}}^{l-3}\right\|$, n_{bad}

fix TOL, i_{overhead}, $n_{\text{bad}}^{\max}$
set $\theta^l = \frac{\left\|\widetilde{\mathbf{w}}^{l}\right\|}{\left\|\widetilde{\mathbf{w}}^{l-1}\right\|}$
if $\theta^l >= 1$ **then**
    $n_{\text{bad}} = n_{\text{bad}} + 1$
else
    compute $\bar{\theta}^l$ from (5.26)
    if $\frac{\bar{\theta}^l}{1-\bar{\theta}^l}\left\|\widetilde{\mathbf{w}}^{l}\right\| < \text{TOL}$ **then**
        return "converged"
    end if
    compute i via (5.27) from θ^l
    compute $\theta_{1/2}$, $l_{1/2}$, and $i_{1/2}$
    if $2\left(l_{1/2} + i_{1/2}\right) + i_{\text{overhead}} < l + i$ **then**
        $n_{\text{bad}} = n_{\text{bad}} + 1$
    end if
end if
if $n_{\text{bad}} > n_{\text{bad}}^{\max}$ **then**
    return "shorten time step"
end if

In this setting, the state equation has to be solved repeatedly for different values of the control entering the state equation. This means that once a time step was shortened, the additional cost arises for every subsequent iteration of the optimization loop although the temporal dynamics was potentially changed by the updated control in such a way that the time stepping solver would converge satisfactory also on the larger time step. Another important consideration is that whenever a time step size changes this means modifying the optimal control problem and thereby shifting the optimum. This can severely delay or even inhibit convergence of the optimization. Hence, on the one hand, the initial discretization should be chosen fine enough to ensure convergence of the time stepping solver for the majority of time steps in order to minimize the need for refinements during the optimization algorithm. On the other hand, refinement decisions due to slow convergence of the time stepping solver should be taken conservatively, accounting for the additional overhead they cause to the optimization process.

Compared to the Runge-Kutta implementations, when shortening the time step we do not try to determine an optimal new step size but just half the current discretization interval. This restriction keeps the change to the time discretization localized to the current interval. Therefore extending a time dependent control to the modified discretization is possible without error. To determine whether it pays off to half the time step, we have to model the effect on the computational cost. The number of solution steps i remaining

for given contraction rate θ until a prescribed tolerance TOL is reached can be estimated by the condition

$$\theta^i \left\| \mathbf{e}^l \right\| \leq \mathrm{TOL}.$$

Solving for i and inserting the estimate (5.25) for $\mathbf{e}^l$ gives

$$i \geq \frac{\log \mathrm{TOL} + \log(1-\theta) - \log \widetilde{\mathbf{w}}^l}{\log \theta} - 1. \tag{5.27}$$

Consistent with the estimate in Theorem 5.20, we model the dependence of the contraction rate θ on the size k_m of the time step by

$$\theta \approx \rho + C k_m,$$

where ρ is the systematic error caused by the approximate decoupling of the Newton update and the constant C describes the influence of nonlinearities and averaging the stiffness matrix. In practice, we chose $\rho = \rho_r$. Given an estimate for the current contraction rate θ, the constant C can be determined from this identity. It can be used to predict the contraction rate $\theta_{1/2}$ resulting from halving the time step. We obtain

$$\theta_{1/2} \approx \frac{\theta + \rho}{2}.$$

Inserting this value into (5.27), we estimate the number of solver iterations $i_{1/2}$ required to reach the tolerance for time step size $\frac{k_m}{2}$ when starting with an error of magnitude $\left\| \mathbf{e}^l \right\|$. To obtain an estimate for the total number of solver steps required with time step $\frac{k_m}{2}$, we have to add the number of iterations $l_{1/2}$ it takes to reduce the initial error for the shortened time step to $\left\| \mathbf{e}^l \right\|$. It can be approximated as

$$l_{1/2} = \frac{\log \theta}{\log \theta_{1/2}} l.$$

The decision to shorten the time step is then taken based on whether the relation

$$2\left(l_{1/2} + i_{1/2}\right) + i_{\mathrm{overhead}} < l + i$$

holds true. The constant i_{overhead} can be used to account for any extra cost that arises from increasing the number of time steps by one, e. g., from having to evaluate the cost functional, its gradient and hessian for an additional time step.

Due to the findings discussed in the Propositions 5.15 and 5.18 we concluded in Remark 5.19 that it can make sense to continue the solution process even if a step did not result in the expected contraction. Therefore we make provisions to allow for a limited number of steps with insufficient contraction before shortening the time step. A prototypical realization of the outlined strategy for controlling the solver is shown in Algorithm 5.2. For simplicity we omitted special cases arising in the first few iterations due to insufficient convergence data and some additional heuristics for reusing the Jacobian. In practice, significant computational cost can be saved if the stiffness matrix is not reassembled after every update of the solution. So we reuse the matrix as long as possible and employ a heuristic criterion to detect degrading convergence and to decide about rebuilding it.

5.4. Numerical results

For linear problems with constant coefficients satisfying Assumption 5.7, the solution scheme performs as predicted by the theoretical results. Since we can expect that decoupling schemes based on complex arithmetic perform better in this setting, we do not provide numerical results for it and present results for two nonlinear model problems instead. In the first part, we consider solving the state equation of the semilinear problem introduced in Section 2.3.2. Given that the iterates of the time stepping solver are globally bounded, the presented nonlinear convergence theorem applies to this problem. The second part of this section is concerned with solving the state equation of the combustion example from Section 2.3.3. Since neither of the Assumptions 5.7 and 5.9 apply to this problem, it is not covered by the presented convergence theory. Nevertheless, as we shall see, our solution scheme proves competitive also for this problem.

5.4.1. Semilinear equation

We consider the state equation (2.7b) of the semilinear example presented in Section 2.3.2. The control is fixed as $q = 0$ and the right hand side is chosen as

$$f(t, x_1, x_2) = \sin(\pi x_1)\sin(\pi x_2)\Big\{\pi^2\cos(\pi^2 t) + 10 \\ + \big(\sin(\pi^2 t) + 10t\big)\Big[2\pi^2 + \big(\sin(\pi^2 t) + 10t\big)^2 \sin(\pi x_1)^2 \sin(\pi x_2)^2\Big]\Big\}.$$

Therefore, the exact solution is given by

$$u(t, x_1, x_2) = \big(\sin(\pi^2 t) + 10t\big)\sin(\pi x_1)\sin(\pi x_2).$$

For the time interval, we use $I = (0, 1)$ and the domain is the two-dimensional unit square.

First, we investigate how the choice of the temporal norm for controlling the iteration affects the behaviour of the algorithm. We compare the Taylor norm to the scaled temporal L^2 norm. For this purpose we use a fixed spatial discretization with $N = 4225$ spatial nodes and an initial temporal discretization with a single time step. To monitor convergence and refine the time discretization if necessary, the procedure outlined in Algorithm 5.2 is used with $n_{\text{bad}}^{\max} = 0$ and $i_{\text{overhead}} = 10$. Since the absolute values of the two norms are not expected to be comparable, we use a termination criterion based on a relative tolerance of 10^{-8} instead of an absolute one.

In Table 5.3 the resulting number of time steps and the total number of linear solution steps are compared. We see that especially for larger orders, to ensure convergence in the Taylor norm, significantly shorter time steps are required than for the L^2 norm. This also results in an increase of the total number of linear solution steps required, which can be seen as an indicator for the total computational cost. However, we note that a

Table 5.3.: Number of required time steps and resulting total number of linear iterations when controlling the iteration with either the L^2 or the Taylor norm

	L^2 norm		Taylor norm	
r	M	n_{iter}	M	n_{iter}
1	3	44	3	45
2	6	105	6	94
3	7	121	8	128
4	5	90	13	205
5	5	99	15	280
6	5	119	14	325
7	5	123	10	256

Table 5.4.: Total number of linear iterations needed to solve the discrete semilinear equation for $M = 5$ time steps and N spatial degrees of freedom

$N \setminus r$	1	2	3	4	5	6	7
9	71	103	113	121	125	135	134
25	69	103	104	107	114	121	159
81	70	106	105	121	121	178	144
289	69	104	105	121	118	168	146
1089	69	105	105	122	119	177	146
4225	69	105	105	122	119	156	145
16641	69	105	105	122	121	170	145
66049	69	105	105	122	121	154	145
263169	69	105	105	122	121	186	146
1050625	69	104	105	122	121	174	145

very fast growth of the number of time steps with increasing order r is not observed for the Taylor norm, contrary to what the extreme growth of the constant κ_3 in the proof of Theorem 5.20 would suggest. Nevertheless the cost savings from using the temporal L^2 norm are significant enough to prefer it over the Taylor norm for practical computations.

Next, we want to verify that convergence of the time stepping solver is in fact independent of the fineness h of the spatial mesh for the test example. For this purpose we fix the number of time steps at $M = 5$ and solve on a sequence of uniformly refined grids down to an absolute L^2 tolerance of 10^{-10}. In Table 5.4, the total number of linear solution steps required is listed for orders one up to seven. Clearly, mesh-independent convergence can be observed.

In Figure 5.2 we plot the L^2 error of the computed discrete state relative to the analytic solution u for the finest spatial discretization in the above setting. As expected,

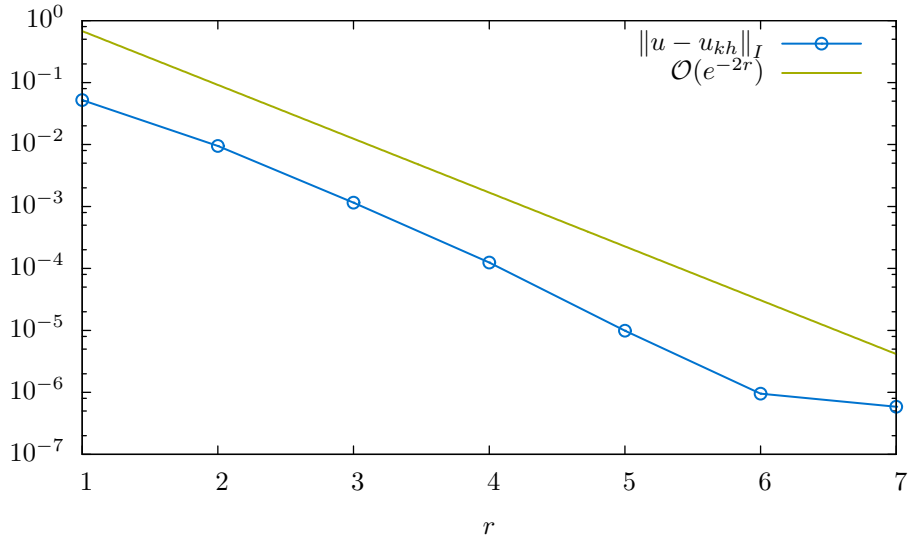


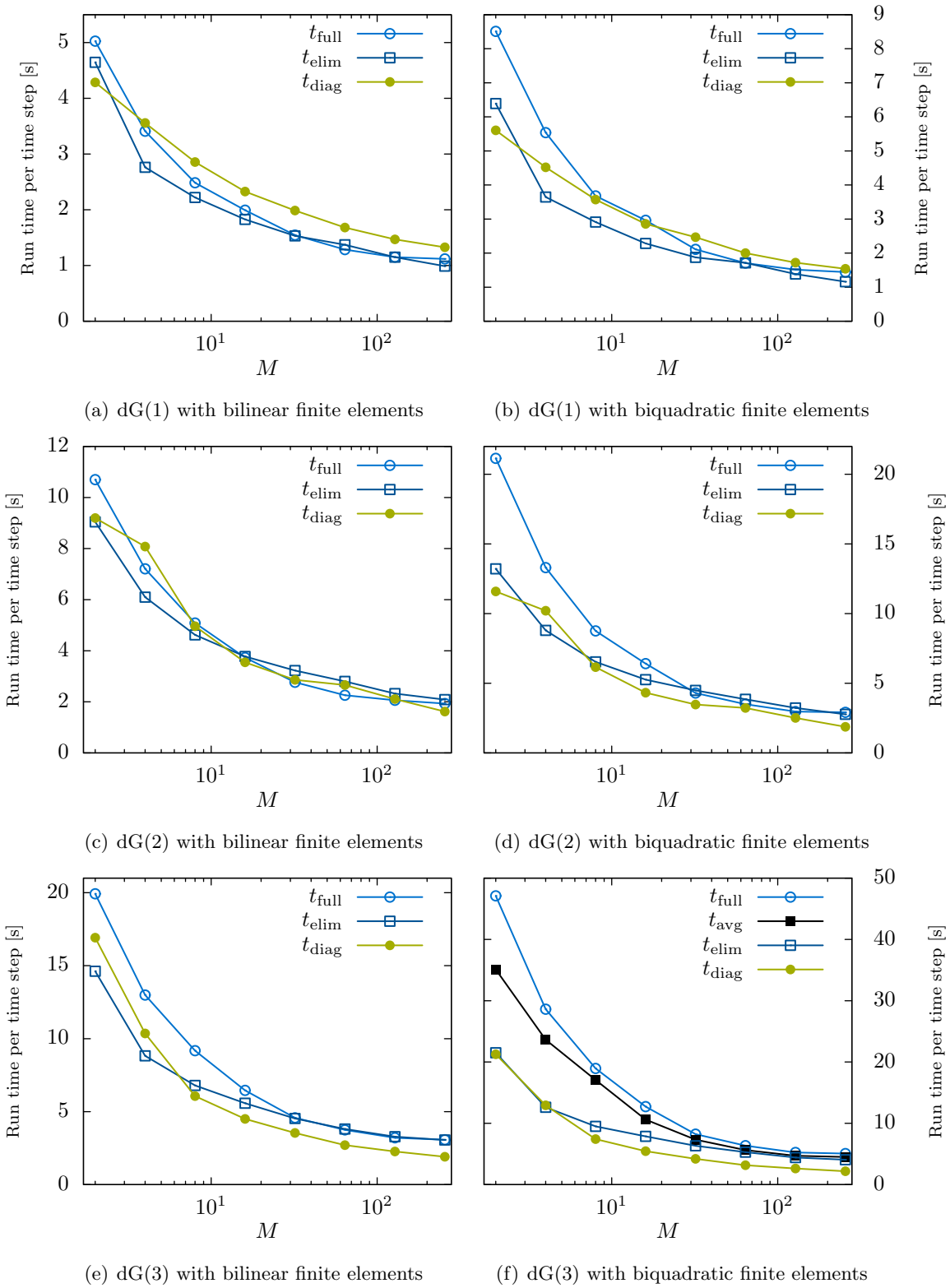
Figure 5.2.: L^2 error of the discrete solution of the semilinear example for $M = 5$ time steps and $N = 1050625$ spatial nodes

scheme \ M	2	4	8	16	32	64	128	256
approx. diagonalization	15.0	12.5	9.5	7.6	6.6	5.7	5.0	4.4
approx. block elimination	10.5	6.8	4.8	3.9	3.2	2.7	2.3	2.0
full system	8.5	6.0	4.4	3.5	2.7	2.2	2.0	2.0

Table 5.5.: Average number of solver iterations per time step for dG(1) with bilinear elements in space

exponential convergence with respect to r is observed, validating the correctness of our implementation.

To assess the computational efficiency of the iterative solver, we compare it to several other approaches. For the comparison, we consider on the one hand solving the full coupled Newton update system (5.4) and on the other hand our previous decoupling approach from [91], which is based on block-wise Gauß elimination with subsequent approximation of the resulting matrix polynomials. While the decoupled variants solve equations of dimension $N \times N$, the coupled system has dimension $(r+1)N \times (r+1)N$. Hence a meaningful comparison based on matrix-vector operation counts is not possible. Instead we compare CPU times for implementations of all variants in the library RoDoBo [93]. For this purpose, we solve the semilinear state equation on a fixed spatial grid with $N = 16641$ nodes and a sequence of uniformly refined temporal grids starting with $M = 2$ time steps. For all approaches, we use the affine covariant termination criterion discussed in Section 5.3.1 and require an absolute tolerance of 10^{-8}.



(a) dG(1) with bilinear finite elements

(b) dG(1) with biquadratic finite elements

(c) dG(2) with bilinear finite elements

(d) dG(2) with biquadratic finite elements

(e) dG(3) with bilinear finite elements

(f) dG(3) with biquadratic finite elements

Figure 5.3.: *Semilinear test problem:* Comparison of average run time per time step for different solution schemes

	fraction of total computational cost		
space discretization	residual assembly	matrix assembly	solution of linear systems
bilinear	56.7%	12.7%	25.7%
biquadratic	42.2%	15.4%	36.8%

Table 5.6.: Distribution of computational cost for dG(1) solver with approximate diagonalization according to Callgrind [114] for 2 and 4 time steps

In Figure 5.3(a), the average computation times per time step are depicted for a dG(1) discretization with bilinear finite elements in space. We denote the computation time per time step for the decoupling based on block diagonalization by t_{diag}, for the decoupling variant from [91] by t_{elim}, and for solving the full block system by t_{full}. As expected, the computational work per time step decreases for all variants as the time steps become shorter. For dG(1), the performance of all three schemes is reasonably similar, however, some disadvantage for the proposed approximate block diagonalization is observed.

For the purpose of understanding how the observed performance difference relates to the performance characteristics of the underlying finite element library, a more detailed investigation is appropriate. First we note that, of all three schemes, the decoupling based on approximate diagonalization requires the highest number of solution steps as seen from Table 5.5. On the other hand, the cost for solving the update equation is lower than for the other two schemes. Per iteration, two systems of dimension N have to be solved whereas the decoupling with block elimination needs to solve three such systems and additionally solve one time for the mass matrix. The full time stepping equation is a system of dimension $2N$. For all three schemes, the cost of evaluating the residual in each iteration is identical. Therefore the decoupled schemes can only be expected to gain an advantage over the full scheme if the total cost per iteration is not dominated by the residual evaluation and therefore the savings in solving the linear systems can compensate for the additional iterations.

To explore how a shift in this cost distribution affects the performance of the three schemes, we compare the result for bilinear finite elements to the result when using biquadratic elements in space. We use the same number of degrees of freedom as in the bilinear case, hence the cost for assembling the residual stays virtually the same. Besides, we observe that the number of iterations necessary remains unchanged for all three schemes. The only major difference from the solver's point of view is that the resulting matrices are more densely populated, leading to an increased workload for the solver in each iteration.

In order to confirm that the workload distribution indeed shifts towards the linear solver when changing the discretization, we examine the decoupling scheme based on approximate diagonalization with the profiling tool Callgrind [114]. The results can be seen in Table 5.6. An increase in the fraction of computing time spent on the linear

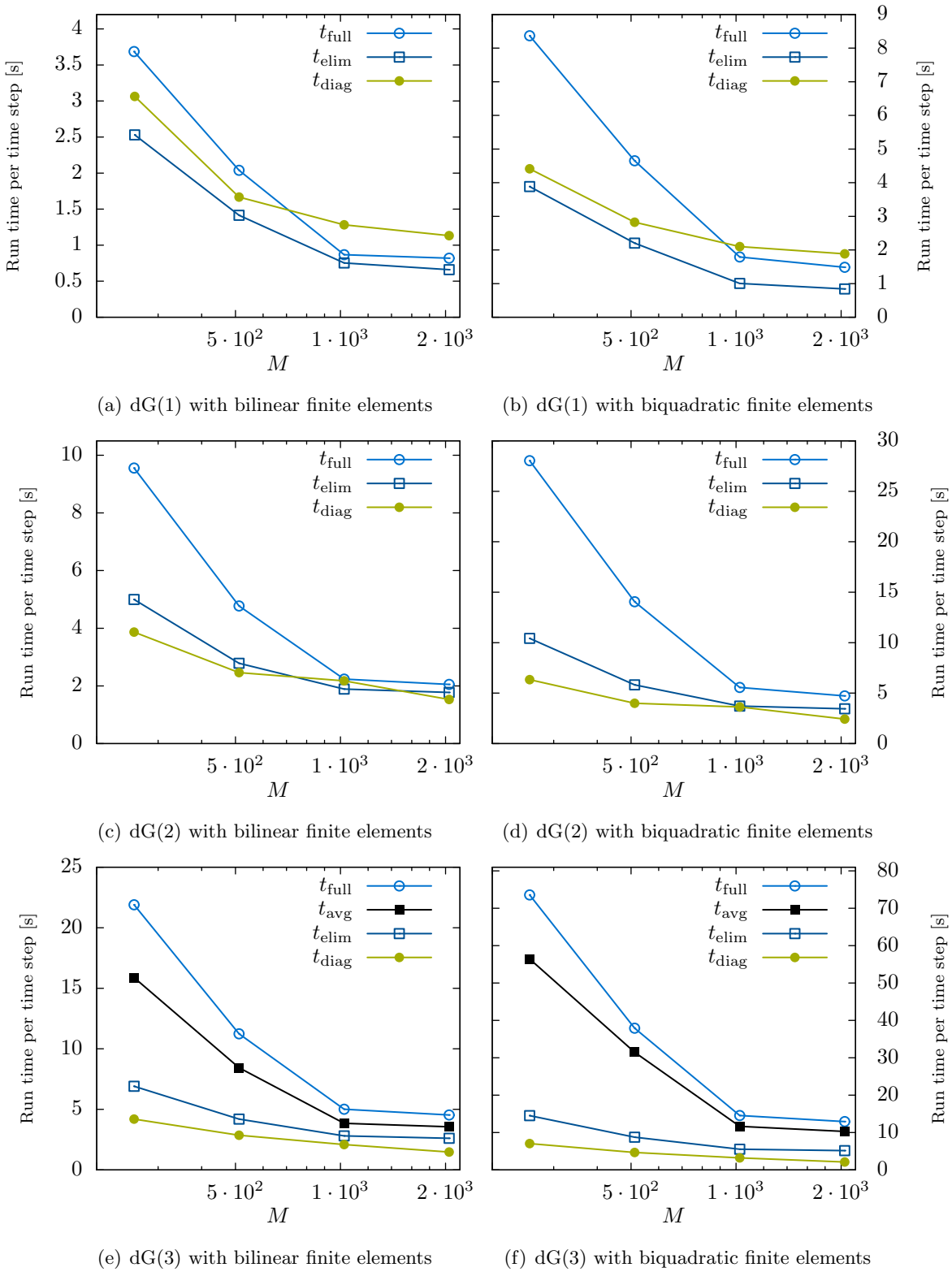
solver is clearly visible as we switch from a bilinear to a biquadratic approximation in space. We also note that, due to the efficient geometric multigrid solver with ILU smoother (see Becker and Braack [11]) we employ for our tests, assembling residuals and matrices constitutes the majority of the computational cost in both configurations.

The computation times per time step for dG(1) time discretization and biquadratic elements in space are reported in Figure 5.3(b). We see that the scheme solving the full system takes a greater performance hit from the change of spatial discretization than the decoupled approaches. However, the savings per time step from using approximate diagonalization for decoupling instead of approximate block elimination are still not sufficient to compensate for the increased number of solution steps. The computing times per time step for dG(2) and dG(3) can be seen from Figures 5.3(c) to 5.3(f). While for dG(2), the two decoupling schemes perform comparably for both spatial discretization variants, for dG(3), a clear advantage of the scheme based on approximate diagonalization is visible for the biquadratic space discretization. For the latter configuration, we additionally report the solution times t_{avg} for a scheme solving the full Newton system with averaged Jacobian (Equation (5.6)). It can be seen that t_{avg} is in between t_{full} and t_{elim}. The same behaviour is observed for the other configurations, therefore we omitted t_{avg} in Figures 5.3(a) to 5.3(e) in order to maintain readability of the graphs.

In conclusion we found that the decoupled schemes proposed here offer comparable performance to previous approaches for solving higher order dG schemes for the semilinear test equation. Since iteration counts tend to be higher than for the alternatives, they perform best when the linear solver contributes a large fraction of the total computational cost. Besides, we saw that with increasing order of the time discretization, greater performance gains can be realized through decoupling.

5.4.2. Combustion problem

As a more realistic problem to test the convergence properties of the decoupling schemes on, we consider the state equation (2.8) of the combustion problem introduced in Section 2.3.3. It should be noted that this problem does not satisfy Assumption 5.9. We simulate the uncontrolled state, that is, we set $q = 0$. As for the semilinear problem, we compare the run time of the decoupling based on diagonalization to the decoupling variant from [91] and to the solution of the exact Newton system. For the spatial discretization, we compare again bilinear to biquadratic finite elements, each with $N = 11041$ spatial degrees of freedom on a uniform mesh. For each scheme, the state equation is solved on a sequence of equidistant temporal grids with $M = 256, 512, 1024$, and 2048 subintervals. The resulting average run times per time step for temporal orders $r = 1$ to $r = 3$ are shown in Figures 5.4(a) to 5.4(f). As before we denote the times for the decoupling through approximate diagonalization by t_{diag}, for the decoupling through approximate block elimination by t_{elim}, for the full Newton update by t_{full}, and for the full Newton update with averaged Jacobian by t_{avg}. For dG(1) and dG(2), t_{avg} is omitted.



(a) dG(1) with bilinear finite elements

(b) dG(1) with biquadratic finite elements

(c) dG(2) with bilinear finite elements

(d) dG(2) with biquadratic finite elements

(e) dG(3) with bilinear finite elements

(f) dG(3) with biquadratic finite elements

Figure 5.4.: *State equation of the combustion problem:* Comparison of average run time per time step for different solution schemes

Particularly for relatively large time steps, solving the full Newton update equation carries a significant performance penalty compared to the decoupled schemes. As for the semilinear test equation, the decoupled schemes suffer less from the increase in cost for linear solutions than the full variants when going from bilinear to biquadratic finite elements in space. While in the case of dG(1), once more the decoupling scheme using diagonalization suffers from a higher number of solver iteration, for orders two and three it is more efficient than the other approaches. We expect this trend to continue when increasing the order further, however due to technical restrictions, we did not perform comparisons with order greater than three.

Overall, the performed tests indicate that the constructed decoupling schemes based on approximate diagonalization result in at least comparable run times to previous approaches. At the same time they allow for a straightforward implementation requiring neither the assembly of large block systems as for the full schemes nor the use of different linear combinations of the mass and stiffness matrix within the same time step as for the block elimination schemes. For this reason, they also have the lowest memory footprint of the considered schemes and allow for convenient change of temporal discretization order over time, which makes them particularly well-suited for *hp* time discretization.

Concerning the fact that the constructed decoupling schemes on average require more iterations to converge than the other decoupling variant, we have to keep in mind that our specific choice of the approximation $\mathbf{T}$ for $\mathcal{A}$ was mainly driven by the requirement of having a simple general construction principle that works the same for all considered orders and that allows for a detailed theoretical analysis of the approximation properties. We expect that a more sophisticated approximation could lead to faster converging numerical schemes.

6. A priori analysis of a third order scheme for time parameter control with constraints

In this chapter, we analyze a discretization scheme for the linear quadratic model problem (2.6). We show that the presented approach yields convergence with almost order three with respect to the temporal discretization parameter k and second order convergence with respect to the spatial discretization parameter h. The materials presented in this chapter have already been published in Springer and Vexler [104].

Due to the box constraints on the control, the solution of the model problem (2.6) has limited regularity which restricts the order of convergence we can achieve for the time discretization. Previous results about time discretization of problems with control constraints include the works by Lasiecka and Malanowski [66], Malanowski [72], and Meidner and Vexler [79] that discuss first order convergent schemes and the article by Rösch [94] that shows convergence with order $\mathcal{O}(k^{\frac{3}{2}})$ for a one-dimensional problem. More recently, Meidner and Vexler [81] presented a time discretization based on a first order continuous Petrov-Galerkin scheme for the state and piecewise constant ansatz functions for the control. After a post-processing step, the resulting approximation for the control converges with second order. The underlying superconvergence property of the Petrov-Galerkin scheme was also shown by Apel and Flaig [5] in the context of parabolic problems without control constraints.

Here, we propose a discontinuous Galerkin time discretization with piecewise linear trial functions for the state variable. The control is treated with the variational discretization concept by Hinze as described in Section 4.2.1. We will show that this results in second order accuracy of the time discretization. Based on this variational solution we perform a post-processing step and show that the resulting improved optimal control $\tilde{q}_{kh}$ (see Equation (6.50) for its definition) converges with order $\mathcal{O}(|\log k|^{\frac{1}{2}}\, k^3)$ with respect to the temporal L^2 norm. The error introduced by the spatial discretization of the state variable can be analyzed independently and decreases for conforming bilinear finite elements with $\mathcal{O}(h^2)$.

The post-processing is based on a slight modification of the higher order continuous reconstruction $\hat{\pi}_k^{\mathrm{D}}$ introduced in Section 4.1. Here, we derive optimal order a priori estimates for this modified reconstruction when applied to the optimal adjoint state.

The remainder of this chapter is organized as follows: In Section 6.1 we recall the problem setting and the optimality conditions for the model problem (2.6). The regularity of the optimal solution is discussed in detail. The second section specifies the discretization of the problem in time and space and collects tools for the a priori analysis. In Section 6.3 we derive estimates for the discretization error when solving the state and adjoint equations for a given fixed control. In particular, we show that, given sufficient regularity, the reconstruction of the semidiscrete adjoint solution converges with order $\mathcal{O}(k^{r+2})$ for a discontinuous Galerkin time discretization of order r. The results for $r = 1$ and for order $s = 1$ of the spatial discretization are used in Section 6.4 to derive the error estimate for the post-processed optimal control. In the final section, we illustrate our results by a numerical example and provide evidence that the variational treatment of the control variable is in fact necessary to achieve the proposed order of convergence after post-processing.

6.1. Optimality conditions and regularity considerations

In this section we recall the problem setting and the resulting optimality conditions and discuss the regularity of the optimal solution in greater detail.

We assume the spatial domain Ω to be polygonal and convex in order to ensure H^2 regularity of the solutions. For a general recipe on how to generalize estimates for time discretization to non-convex domains, we refer to Flaig et al. [41]. As discussed in Section 2.3.1, we set $V := H_0^1(\Omega)$, $H := L^2(\Omega)$, and $Q = L^2(I, \mathbb{R}^{d_Q})$ where d_Q is a positive integer.

A weak form of the state equation (2.6b) can be stated as: Find $u \in X$ such that

$$\begin{aligned} (\partial_t u, \varphi)_I + (\nabla u, \nabla \varphi)_I &= (f + G^q q, \varphi)_I \quad \text{for all } \varphi \in X, \\ u(0) &= u_0 \qquad\qquad\quad \text{in } \Omega. \end{aligned} \tag{6.1}$$

For the data we assume $u_0 \in V$ and $f \in L^2(I, H)$. Then there exists a unique control-to-state mapping $q \mapsto u(q)$, $Q \to X$, where $u = u(q)$ is the solution of (6.1) for the given q. Note that we replaced the duality pairing in the time derivative term by an L^2 inner product. This is justified because, as we will see in Lemma 6.1, under the stated assumptions, the time derivative has values in H. With the usual definitions for the reduced cost functional

$$j(q) = J(q, u(q)) \tag{6.2}$$

and the admissible set

$$Q_{\text{ad}} = \left\{ q \in Q \;\middle|\; q^a \le q(t) \le q^b \quad \text{for almost all } t \in I \right\}, \tag{6.3}$$

we rewrite the optimal control problem in reduced form as

$$\text{Minimize } j(q) \text{ subject to } q \in Q_{\text{ad}}. \tag{6.4}$$

As pointed out in Section 2.3.1, there exists a unique solution $\bar{q}$ with corresponding optimal state $\bar{u}$ for this problem. Due to the linear quadratic structure of the problem and the convexity of Q_{ad}, the first order necessary optimality condition is also sufficient for optimality. It reads

$$j'(\bar{q})(\delta q - \bar{q}) \geq 0 \quad \forall \delta q \in Q_{\text{ad}}. \tag{6.5}$$

For the model problem, the first derivative of the reduced cost functional j can be computed as

$$j'(q)(\delta q - q) = (\alpha q + G^{q*} z, \delta q - q)_Q$$

with the adjoint state z given as solution of the problem: Find $z \in X$ such that

$$\begin{aligned} -(\varphi, \partial_t z)_I + (\nabla \varphi, \nabla z)_I &= (u - u_d, \varphi)_I \quad \forall \varphi \in X, \\ z(T) &= 0. \end{aligned} \tag{6.6}$$

We have seen in Chapter 3 that the first order optimality condition (6.5) can be expressed equivalently as

$$\bar{q} = P_{Q_{\text{ad}}} \left(-\alpha^{-1} G^{q*} \bar{z} \right). \tag{6.7}$$

Lemma 6.1. *The solution of the state equation* (6.1) *has the improved regularity*

$$u \in H^1(I, H) \cap L^2(I, H^2(\Omega) \cap V) \hookrightarrow C(\bar{I}, V)$$

and satisfies the stability estimate

$$\|\partial_t u\|_I + \|\Delta u\|_I + \|\nabla u(T)\| \leq C \left\{ \|f\|_I + \|q\|_Q + \|\nabla u_0\| \right\}.$$

Proof. In Evans [38, Chapter 7, Theorem 2] the statement is shown for a domain Ω with C^2 boundary. Looking at the proof, we note that the smoothness requirement on the boundary of Ω is only needed in step 3 to apply the corresponding elliptic regularity result Theorem 4 in Chapter 6 on the spatial part of the differential operator. However, this result can also be shown for convex polygonal domains, see, e. g.,Grisvard [49, Theorem 4.4.3.7] for a two-dimensional domain and Maz'ya and Rossmann [75, Theorem 4.3.2] for the three-dimensional case. □

In what follows, we need some additional regularity assumptions on the data of the optimal control problem:

Assumption 6.2. We assume the data u_0, f and u_d satisfy the following conditions:

- $u_0 \in V$ with $\Delta u_0 \in V$,
- $f \in H^1(I, H) \cap C(\bar{I}, V)$,
- $u_d \in H^2(I, H) \cap H^1(I, H^2(\Omega) \cap V)$, and $\Delta u_d(T) \in V$.

Lemma 6.3. *Let $(\bar{q}, \bar{u})$ be the solution of the optimal control problem (2.6) and $\bar{z}$ the corresponding adjoint state. If Assumption 6.2 is satisfied, we obtain the improved regularities*

$$\begin{aligned}
&\bar{u} \in H^2(I,H) \cap H^1(I, H^2(\Omega) \cap V) \hookrightarrow C^1(\bar{I}, V),\\
&\bar{z} \in H^3(I,H) \cap H^2(I, H^2(\Omega) \cap V) \hookrightarrow C^2(\bar{I}, V), \text{ and}\\
&\bar{q} \in W^{1,\infty}(I, \mathbb{R}^{d_Q}).
\end{aligned}$$

Moreover we have the stability estimates

$$\begin{aligned}
\|\partial_t \Delta \bar{u}\|_I + \left\|\partial_t^2 \bar{u}\right\|_I \leq C\Big\{ &\|f\|_{H^1(I,H)} + \|\bar{q}\|_{H^1(I,\mathbb{R}^{d_Q})} + \|\nabla f(0)\|\\
&+ \|\nabla \Delta u_0\| \Big\} \text{ and}\\
\left\|\partial_t^3 \bar{z}\right\|_I + \left\|\partial_t^2 \Delta \bar{z}\right\|_I + \left\|\nabla \partial_t^2 \bar{z}(T)\right\| \leq C\Big\{ &\left\|\partial_t^2 u_d\right\|_I + \|\nabla \partial_t u_d(T)\| + \|\nabla \Delta u_d(T)\|\\
&+ \|f\|_{H^1(I,H)} + \|\nabla f(0)\| + \|\nabla f(T)\|\\
&+ \|\bar{q}\|_{H^1(I,\mathbb{R}^{d_Q})} + \|\bar{q}(T)\|_{\mathbb{R}^{d_Q}} + \|\nabla \Delta u_0\| \Big\}.
\end{aligned}$$

Proof. The stated regularity results for $\bar{u}$ and $\bar{q}$ and the stability estimate for $\bar{u}$ are shown by Meidner and Vexler in [81, Proposition 2.3]. Furthermore, the authors prove that the adjoint solution satisfies $\bar{z} \in H^2(I,H) \cap H^1(I, H^2(\Omega) \cap V) \hookrightarrow C^1(\bar{I}, V)$ with the stability estimate

$$\|\partial_t \Delta \bar{z}\|_I + \left\|\partial_t^2 \bar{z}\right\|_I \leq C \left\{ \|u_d\|_{H^1(I,H)} + \|\nabla u_d(T)\| + \|f\|_I + \|\bar{q}\|_Q + \|\nabla u_0\| \right\}.$$

Using the regularity already shown, we verify that $\hat{z} := \partial_t \bar{z}$ satisfies the equation

$$-\partial_t \hat{z} - \Delta \hat{z} = \partial_t (\bar{u} - u_d) \tag{6.8}$$

with the terminal condition

$$\hat{z}(T) = \partial_t \bar{z}(T) = -\Delta \bar{z}(T) - (\bar{u} - u_d)(T) = -(\bar{u} - u_d)(T) \tag{6.9}$$

since $\bar{z}(T) = 0$ and hence $-\Delta \bar{z}(T) = 0$. We differentiate equation (6.8) formally another time with respect to the time variable resulting in

$$-\partial_t \tilde{z} - \Delta \tilde{z} = \partial_t^2 (\bar{u} - u_d). \tag{6.10}$$

for a new variable $\tilde{z}$. For the terminal condition of this equation we set

$$\begin{aligned}
\tilde{z}(T) &= \partial_t \hat{z}(T) = -\Delta \hat{z}(T) - \partial_t (\bar{u} - u_d)(T)\\
&= -\partial_t (\bar{u} - u_d)(T) + \Delta (\bar{u} - u_d)(T) = (\partial_t u_d - \Delta u_d)(T) - (f + G^q \bar{q})(T).
\end{aligned}$$

In the second line, the terminal condition (6.9) for $\hat{z}$ and the state equation (6.1) were plugged in.

We note that with Assumption 6.2 and the shown regularity for $\bar{q}$, the terminal value for $\tilde{z}$ is in V. Hence we can apply the regularity result from Lemma 6.1 to Equation (6.10). This gives $\tilde{z} \in H^1(I,H) \cap L^2(I, H^2(\Omega) \cap V)$. The corresponding stability estimate reads

$$\begin{aligned}\|\partial_t \tilde{z}\|_I + \|\Delta \tilde{z}\|_I + \|\nabla \tilde{z}(T)\| \\ \leq C \left(\left\|\partial_t^2(\bar{u} - u_d)\right\|_I + \|\nabla\left(\partial_t u_d(T) - \Delta u_d(T)\right)\| + \|\nabla f(T)\| + \|\bar{q}(T)\|_{\mathbb{R}^{d_Q}} \right).\end{aligned}$$

The first term on the right hand side can be estimated by the stability estimate for the state equation. We verify in the same way as in the proof of Theorem 27.2 in Wloka [117] that in fact $\tilde{z} = \partial_t^2 \bar{z}$. This completes the proof. □

6.2. Auxiliary results for the semidiscrete and discrete problem

In this section we specify the assumptions needed on the discretization and collect auxiliary results relating to the semidiscrete and discrete problems.

6.2.1. Semidiscrete problem

For the temporal discretization of the state equation, we consider a discontinuous Galerkin discretization as described in Section 4.1.1. However, we do not allow for varying the order of the discretization over time, i. e., the order vector r is constant. Additionally, we require the following regularity condition on the temporal mesh.

Assumption 6.4. We impose a regularity condition on the temporal mesh and require that there is a constant $\kappa \geq 1$ independent of the mesh width k such that

$$\kappa^{-1} \leq \frac{k_m}{k_{m-1}} \leq \kappa \quad \forall m = 2, 3, \ldots, M.$$

Introducing the bilinear form $B : X_k^r \times X_k^r \to \mathbb{R}$ given by

$$B(u_k, \varphi) := \sum_{m=1}^{M} (\partial_t u_k, \varphi)_{I_m} + (\nabla u_k, \nabla \varphi)_I + \sum_{m=1}^{M-1} ([u_k]_m, \varphi_m^+) + (u_0^+, \varphi_0^+) \tag{6.11}$$

the dG(r) semidiscrete state equation reads: For given control q, find $u_k \in X_k^r$ such that

$$B(u_k, \varphi) = (f + G^q q, \varphi)_I + (u_0, \varphi_0^+) \quad \forall \varphi \in X_k^r. \tag{6.12}$$

A semidiscrete adjoint type equation with some given right hand side g and terminal condition $z_k(T) = 0$ takes the form

$$B(\varphi, z_k) = (\varphi, g)_I \quad \forall \varphi \in X_k^r. \tag{6.13}$$

Remark 6.5. As noted for example in [80], the continuous solution u of (6.1) for given control $q \in Q$ fulfills the semidiscrete state equation as well. Hence, although the dG(r) semidiscretization is non-conforming, we get *Galerkin orthogonality*, i. e.,

$$B(u - u_k, \varphi) = 0 \quad \forall \varphi \in X_k^r \tag{6.14}$$

holds.

The semidiscrete optimization problem reads:

$$\text{Minimize } J(q_k, u_k) \text{ subject to (6.12) and } (q_k, u_k) \in Q_{\text{ad}} \times X_k^r. \tag{6.15}$$

The adjoint equation has the form (6.13) with right hand side $g := \bar{u}_k - u_d$. From interval-wise integration by parts with respect to time we obtain a dual representation of the bilinear form B.

$$B(\varphi, \psi) = -\sum_{m=1}^{M} (\varphi, \partial_t \psi)_{I_m} + (\nabla\varphi, \nabla\psi)_I - \sum_{m=1}^{M-1} (\varphi_m^-, [\psi]_m) + (\varphi_M^-, \psi_M^-). \tag{6.16}$$

The first order optimality condition for the semidiscrete problem is given as

$$(\alpha\bar{q}_k + G^{q*}\bar{z}_k, \delta q - \bar{q}_k)_Q \geq 0 \quad \forall \delta q \in Q_{\text{ad}}, \tag{6.17}$$

or, equivalently,

$$\bar{q}_k = P_{Q_{\text{ad}}}\left(-\alpha^{-1} G^{q*}\bar{z}_k\right). \tag{6.18}$$

Later on, we need the following stability estimates for the semidiscrete equations:

Theorem 6.6. *For the solution $u_k \in X_k^r$ of the semidiscrete state equation* (6.12) *with right-hand side $f \in L^2(I, H)$, control $q \in Q_{\text{ad}}$ and initial condition $u_0 \in V$ the stability estimate*

$$\|u_k\|_I^2 + \sum_{m=1}^{M} \|\partial_t u_k\|_{I_m}^2 + \|\Delta u_k\|_I^2 + \sum_{m=1}^{M} k_m^{-1} \|[u_k]_{m-1}\|^2 \leq C\left(\|f + G^q q\|_I^2 + \|u_0\|^2 + \|\nabla u_0\|^2\right)$$

holds when defining the jump $[u_k]_0$ as $u_{k,0}^+ - u_0$. The constant C depends only on the domain Ω, the final time T and the order r of the semidiscretization.

Proof. See Meidner and Vexler [80, Theorems 4.1 and 4.3]. □

Corollary 6.7. *The solution $z_k \in X_k^r$ of the semidiscrete adjoint equation* (6.13) *for any right hand side $g \in L^2(I,H)$ satisfies the stability estimate*

$$\|z_k\|_I^2 + \sum_{m=1}^{M} \|\partial_t z_k\|_{I_m}^2 + \|\Delta z_k\|_I^2 + \sum_{m=1}^{M} k_m^{-1} \|[z_k]_m\|^2 \leq C \|g\|_I^2 .$$

Here, the jump term $[z_k]_M$ at final time is set to be $-z_{k,M}^-$.

Lemma 6.8. *For the solution z_k of the semidiscrete adjoint equation* (6.13) *we have additionally the stability estimate*

$$\|z_k\|_{L^\infty(I,H)} \leq C \|g\|_I$$

with the constant C only depending on the domain Ω, the final time T, and the order r of the discretization.

Remark 6.9. An analogous estimate can be shown for the semidiscrete state solution. It reads

$$\|u_k\|_{L^\infty(I,H)} \leq C \left(\|f + G^q q\|_I + \|u_0\| + \|\nabla u_0\| \right).$$

Proof. We estimate the spatial L^2 norm of z_k at a fixed time $t^* \in I$ with $t^* \neq t_m$ for any m. Therefore let m^* denote the smallest index such that $t_{m^*} > t^*$. Then, the norm of $z_k(t^*)$ can be written as

$$\|z_k(t^*)\| = \left\| -\int_{t^*}^{t_{m^*}} \partial_t z_k \,\mathrm{d}t - \sum_{m=m^*+1}^{M} \int_{I_m} \partial_t z_k \,\mathrm{d}t - \sum_{m=m^*}^{M} [z_k]_m \right\| .$$

We define the function v_k interval-wise by $v_k|_{I_m} = \partial_t z_k$. Together with the triangle inequality we obtain

$$\|z_k(t^*)\| \leq \left\| -\int_{t^*}^{T} v_k \,\mathrm{d}t \right\| + \sum_{m=m^*}^{M} \|[z_k]_m\| .$$

The sum on the right hand side can be estimated by means of Corollary 6.7, giving

$$\sum_{m=m^*}^{M} \|[z_k]_m\| \leq \sum_{m=1}^{M} k_m^{\frac{1}{2}} \cdot k_m^{-\frac{1}{2}} \|[z_k]_m\| \leq \sqrt{T} \left(\sum_{m=1}^{M} \frac{1}{k_m} \|[z_k]_m\|^2 \right)^{\frac{1}{2}} \leq C \|g\|_I .$$

For the integral term, we get with Hölder's inequality and the stability estimate from Corollary 6.7

$$\left\| -\int_{t^*}^{T} v_k \,\mathrm{d}t \right\| \leq \sqrt{T} \|v_k\|_I \leq C \|g\|_I .$$

This shows the claim. □

Additionally we need the following stability estimate for a semidiscrete auxiliary adjoint equation:

Theorem 6.10. *Let $w \in L^2(I, H^2(\Omega) \cap V)$ be given. Then the solution $y_k \in X_k^r$ of the equation*

$$B(\varphi, y_k) = (\varphi, w)_I \quad \forall \varphi \in X_k^r \tag{6.19}$$

satisfies the estimate

$$\left\|\Delta^2 y_k\right\|_I + \left(\sum_{m=1}^{M} \left\|\partial_t \Delta y_k\right\|_{I_m}\right)^{\frac{1}{2}} \leq C \left\|\Delta w\right\|_I .$$

The proof uses a Galerkin approximation in the spatial variable and is given in detail in Appendix A.

As key tools for obtaining and proving almost third order convergence of our time discretization we need several projection and interpolation operators which we collect below. For completeness, we also include the operators π_k^{D} and $\hat{\pi}_k^{\mathrm{D}}$, which were already introduced in Section 4.1.1.

1. The L^2 *projection* Π_k^0 onto the space of piecewise constant functions in time given by
$$\Pi_k^0 \colon L^2(I, V) \to X_k^0, \ \Pi_k^0 v\big|_{I_m} = \frac{1}{k_m} \int_{I_m} v(t) \, \mathrm{d}t.$$

2. A projection $P_k \colon C(\bar{I}, V) \to X_k^r$ that is defined interval-wise by the two conditions
$$(P_k v - v, \varphi)_{I_m} = 0 \quad \forall \varphi \in \mathcal{P}_{r-1}(I_m, V), \tag{6.20a}$$
$$P_k v(t_m)^- = v(t_m)^- \tag{6.20b}$$
for each $m = 1, \ldots, M$. This operator is commonly employed in the error analysis of discontinuous Galerkin methods, see, e. g., Thomée [106, Chapter 12].

3. The interpolation operator $\pi_k^{\mathrm{D}} \colon C(\bar{I}, V) \to X_k^r$ at the left Radau nodes on each interval as introduced in Section 4.1.1.

4. The operator $\hat{\pi}_k^{\mathrm{D}}$, which is also known from Section 4.1.1. Here we will adopt the convention that $\hat{\pi}_k^{\mathrm{D}} v := \hat{\pi}_k^{\mathrm{D}}(v(T)^-, v)$ for a function $v \in X_k^r \cup C(\bar{I}, V)$. For a continuous function, this means that the actual terminal value is used as first parameter, whereas for a piecewise discontinuous function, the operator leaves the function on the final interval I_M unmodified as depicted in Figure 6.1.

5. A modified *reconstruction operator* $\tilde{\pi}_k^{\mathrm{D}} \colon X_k^r \cup C(\bar{I}, V) \to X_k^{r+1} \cap C(\bar{I}, V)$, which treats the terminal interval differently than $\hat{\pi}_k^{\mathrm{D}}$. It is determined by the two conditions
$$\tilde{\pi}_k^{\mathrm{D}} v\big|_{I_{M-1} \cup I_M} \in \mathcal{P}_{r+1}(I_{M-1} \cup I_M, V), \text{ and} \tag{6.21a}$$
$$\tilde{\pi}_k^{\mathrm{D}} v\big|_{I_m} = \hat{\pi}_k^{\mathrm{D}} v \quad \forall m = 1, \ldots, M-1. \tag{6.21b}$$

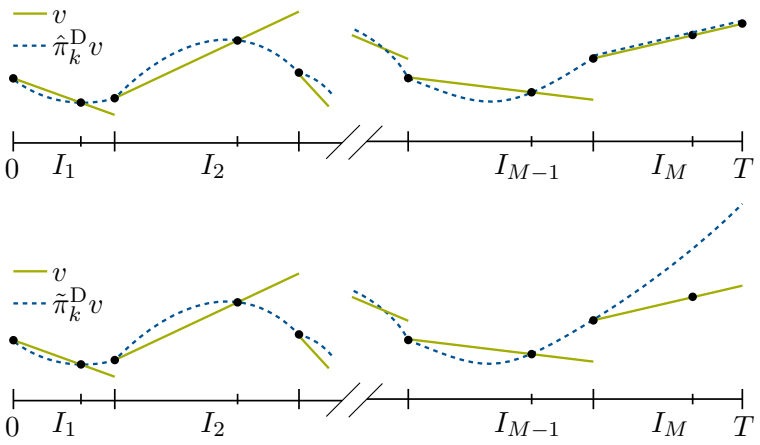


Figure 6.1.: Visualization of the operators $\hat{\pi}_k^{\mathrm{D}}$ and $\tilde{\pi}_k^{\mathrm{D}}$ for $r = 1$.

That means we extend the reconstruction polynomial on the second-last interval onto the last interval. A visualization is given in Figure 6.1. The modification is necessary for Lemma 6.16 to hold.

Since all of the above operators act only on the time variable, we can extend the definitions in the obvious way to control-type variables by replacing the spatial spaces V and H by $\mathbb{R}^{d_Q}$ while requiring the same temporal regularities. We will use the same notations for those operators acting on time-dependent functions with values in $\mathbb{R}^{d_Q}$.

Lemma 6.11. *Let $v \in H^1(I, H)$. For the operators defined above we have the estimates*

$$\left\| v - \Pi_k^0 v \right\|_I \leq Ck \left\| \partial_t v \right\|_I , \tag{6.22}$$

and if we require additionally $v \in C(\bar{I}, V)$

$$\left\| v - P_k v \right\|_I \leq Ck^{r+1} \left\| \partial_t^{r+1} v \right\|_I \qquad \text{for } v \in H^{r+1}(I, H), \tag{6.23}$$

$$\left\| v - \pi_k^{\mathrm{D}} v \right\|_I \leq Ck^{r+1} \left\| \partial_t^{r+1} v \right\|_I \qquad \text{for } v \in H^{r+1}(I, H), \tag{6.24}$$

$$\left\| v - \hat{\pi}_k^{\mathrm{D}} v \right\|_I \leq Ck^{r+2} \left\| \partial_t^{r+2} v \right\|_I \qquad \text{for } v \in H^{r+2}(I, H), \tag{6.25}$$

$$\left\| v - \tilde{\pi}_k^{\mathrm{D}} v \right\|_I \leq Ck^{r+2} \left\| \partial_t^{r+2} v \right\|_I \qquad \text{for } v \in H^{r+2}(I, H). \tag{6.26}$$

Proof. All of the above estimates can be shown in the standard way by transforming each interval to the unit interval, applying the Bramble-Hilbert Lemma and transforming back, for the estimate (6.23) this is done in the proof of Theorem 12.1 in Thomée [106]. For the reconstruction operator $\tilde{\pi}_k^{\mathrm{D}}$, the last two discretization intervals are treated as a single interval. □

6.2.2. Discrete problem

For the spatial discretization of the model problem we limit ourselves to the case of a fixed spatial mesh $\mathcal{T}_h$ for all time steps. Using the corresponding space V_h^s the definition of the fully discrete state space simplifies to

$$X_{k,h}^{r,s} = \left\{ \varphi \in L^2(I,V) \;\middle|\; \varphi|_{I_m} \in \mathcal{P}_r(I_m, V_h^s),\ m = 1, \dots, M \right\} \subseteq X_k^r.$$

Then the discrete state equation reads: for given control $q \in Q$ find $u_{kh} \in X_{k,h}^{r,s}$ such that

$$B(u_{kh}, \varphi) = (f + G^q q, \varphi)_I \quad \forall \varphi \in X_{k,h}^{r,s}. \tag{6.27}$$

For the state-discrete optimal control problem we have

$$\text{Minimize } J(q_{kh}, u_{kh}) \text{ subject to (6.27) and } (q_{kh}, u_{kh}) \in Q_{\text{ad}} \times X_{k,h}^{r,s}, \tag{6.28}$$

and the corresponding adjoint equation is: find $z_{kh} \in X_{k,h}^{r,s}$ such that

$$B(\varphi, z_{kh}) = (\varphi, g)_I \quad \forall \varphi \in X_{k,h}^{r,s} \tag{6.29}$$

with $g := u_{kh} - u_d$. The first order optimality condition for the optimal solution $(\bar{q}_{kh}, \bar{u}_{kh})$ in terms of the corresponding adjoint state $\bar{z}_{kh}$ can be stated as

$$(\alpha \bar{q}_{kh} + G^{q*} \bar{z}_{kh}, \delta q - \bar{q}_{kh})_I \geq 0 \quad \forall \delta q \in Q_{\text{ad}}, \tag{6.30}$$

or, equivalently

$$\bar{q}_{kh} = P_{Q_{\text{ad}}} \left(-\frac{1}{\alpha} G^{q*} \bar{z}_{kh} \right). \tag{6.31}$$

We quote the following stability estimate from Theorem 4.6 and Corollary 4.7 in Meidner and Vexler [80]:

Theorem 6.12. *Let $\Pi_h : V \to V_h^s$ denote the L^2 projection onto the space V_h^s. For the solution $u_{kh} \in X_{k,h}^{r,s}$ of the discrete state equation (6.27) with right hand side $f \in L^2(I,H)$, control $q \in Q_{\text{ad}}$ and initial condition $u_0 \in V$ the stability estimate*

$$\|u_{kh}\|_I^2 + \|\nabla u_{kh}\|_I^2 \leq C \left\{ \|f + G^q q\|_I^2 + \|\nabla \Pi_h u_0\|_I^2 + \|\Pi_h u_0\|_I^2 \right\}$$

holds. The solution z_{kh} of the discrete adjoint equation (6.29) with some right hand side $g \in L^2(I,H)$ satisfies

$$\|z_{kh}\|_I^2 + \|\nabla z_{kh}\|_I^2 \leq C \|g\|_I^2 .$$

We note that the control for the discrete optimal control problem (6.28) is still from the infinite dimensional space Q. As pointed out, we employ the variational approach and hence a discretization of the control is not necessary for an algorithmic solution of the problem.

6.3. Error estimates for the state and adjoint solution with fixed control

In this section we derive error estimates for the semidiscrete and discrete state and adjoint state computed from a given fixed control q. For the state equation we quote results for the temporal discretization error with respect to the $L^2(I,H)$ and the $L^\infty(I,H)$ norms. Subsequently we derive an a priori estimate for the error of the reconstructed adjoint solution which is obtained by applying the reconstruction operator $\tilde{\pi}_k^{\mathrm{D}}$ to the computed semidiscrete adjoint solution. Given sufficient regularity of the solutions, we show that for a dG(r) semidiscretization, the $L^2(I,H)$ error of the reconstruction converges with order $r+2$ with respect to the step size k to the exact adjoint solution, that is, we gain one power of k compared to the plain dG(r) solution. When setting $r=1$, the regularities shown in Lemma 6.3 are sufficient to apply the estimate to the solution of the optimal control problem.

Throughout this section we assume to be given a fixed control $q \in Q$ and denote the corresponding continuous state solution by $u(q)$ and the solution of the semidiscrete problem (6.12) with control q by $u_k(q)$. The solution of the adjoint equation (6.6) with $u(q)$ entering the right hand side is represented by $z(q)$ and the solution of the semidiscrete adjoint equation (6.13) with $u_k(q)$ on the right hand side by $z_k(q)$. The notations $u_{kh}(q)$ and $z_{kh}(q)$ are used analogously.

6.3.1. Estimates for the semidiscrete state solution

The following result for the error with respect to the $L^2(I,H)$ norm can be found for example as Theorem 5.1 in Meidner and Vexler [80].

Theorem 6.13. *The error with respect to the $L^2(I,H)$ norm between the solution $u = u(q)$ of Equation (6.1) and the solution $u_k = u_k(q)$ of its semidiscretization (6.12) can be estimated by*

$$\|u - u_k\|_I \leq Ck^{r+1} \left\|\partial_t^{r+1} u\right\|_I$$

provided that the exact solution u is in $H^{r+1}(I,H)$.

In Thomée [106, Theorem 12.4], an estimate for the $L^\infty(I,H)$ norm error is given:

Theorem 6.14. *For the error between the solution $u = u(q)$ of Equation (6.1) and the solution $u_k = u_k(q)$ of its semidiscretization (6.12) we have the estimate*

$$\|u - u_k\|_{L^\infty(I,H)} \leq C\gamma(k)k^{r+1} \left\|\partial_t^{r+1} u\right\|_{L^\infty(I,H)}$$

with the logarithmic factor $\gamma(k) := |\log k|^{\frac{1}{2}} + 1$.

Note that this estimate requires Assumption 6.4 on the regularity of the temporal mesh.

6.3.2. Superconvergence of the reconstructed semidiscrete adjoint solution

Next we show that the reconstruction $\tilde{\pi}_k^{\mathrm{D}} z_k(q)$ converges with one order more to the exact adjoint solution than $z_k(q)$ w. r. t. the $L^2(I,H)$ norm. We assume $r \geq 1$ since the dG(0) method does not have this superconvergence property.

Theorem 6.15. *For the reconstruction $\tilde{\pi}_k^{\mathrm{D}} z_k$ computed from the solution $z_k = z_k(q)$ of the semidiscrete adjoint equation (6.13) with right hand side $u_k(q) - u_d$ the estimate*

$$\left\| z - \tilde{\pi}_k^{\mathrm{D}} z_k \right\|_I \leq C k^{r+2} \left(\left\| \partial_t^{r+1} \Delta z \right\|_I + \left\| \partial_t^{r+2} z \right\|_I + \left\| \partial_t^{r+1} u \right\|_I \right)$$

holds true where $z = z(q)$ is the solution of the adjoint equation (6.6) with $u = u(q)$ entering the right hand side.

As a preparation for the proof of Theorem 6.15, we need the following Lemma.

Lemma 6.16. *The reconstruction operator $\tilde{\pi}_k^{\mathrm{D}} : X_k^r \to X_k^{r+1}$ is stable with respect to the $L^2(I,H)$ norm, that is, there is a constant C independent of k such that for any $v_k \in X_k^r$*

$$\left\| \tilde{\pi}_k^{\mathrm{D}} v_k \right\|_I \leq C \left\| v_k \right\|_I .$$

Proof. Consider an interval I_m with $m \neq M$. Then

$$\left\| \tilde{\pi}_k^{\mathrm{D}} v_k \right\|_{I_m} \leq \left\| \tilde{\pi}_k^{\mathrm{D}} v_k - v_k \right\|_{I_m} + \left\| v_k \right\|_{I_m}$$

According to Makridakis and Nochetto [71, Lemma 2.2] we have for the first term

$$\left\| \tilde{\pi}_k^{\mathrm{D}} v_k - v_k \right\|_{I_m}^2 = k_m \alpha_2^2 \left\| [v_k]_m \right\|^2 \leq C k_m \left(\left\| v_{k,m}^+ \right\|^2 + \left\| v_{k,m}^- \right\|^2 \right),$$

where the constant α_2 is determined by the order r. The Lobatto quadrature rule with $r+2$ nodes is exact for polynomials of up to degree $2r+1$ and has positive weights (see, e. g., Michels [83]). Let ω_j with $j = 0, \ldots, r+1$ denote the weights of the Lobatto quadrature rule on the unit interval and $c_{m,j}$ the corresponding nodes transformed onto the interval I_m. Note that $c_{m,0} = t_{m-1}$ and $c_{m,r+1} = t_m$. Then we obtain for the left-sided limit $v_{k,m}^-$ at t_m the estimate

$$\begin{aligned} \frac{1}{\omega_{r+1}} \int_{I_m} \left\| v_k \right\|^2 \, \mathrm{d}t &= \frac{k_m}{\omega_{r+1}} \left(\omega_0 \left\| v_{k,m-1}^+ \right\|^2 + \sum_{j=1}^{r} \omega_j \left\| v_k(c_{m,j}) \right\|^2 + \omega_{r+1} \left\| v_{k,m}^- \right\|^2 \right) \\ &\geq k_m \left\| v_{k,m}^- \right\|^2 . \end{aligned}$$

For the right-sided limit $v_{k,m}^+$ we proceed similarly and get

$$k_m \left\| v_{k,m}^+ \right\|^2 \leq \frac{k_m}{k_{m+1}} \frac{1}{\omega_0} \left\| v_k \right\|_{I_{m+1}}^2 .$$

Due to the assumption we made for the temporal mesh, the ratio $\frac{k_m}{k_{m+1}}$ is bounded by κ. So apart from the last subinterval we get

$$\sum_{m=1}^{M-1} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k - v_k\right\|_{I_m}^2 \leq C \sum_{m=1}^{M-1} k_m \left(\|v_k\|_{I_m}^2 + \|v_k\|_{I_{m+1}}^2\right) \leq C \|v_k\|_I^2 . \tag{6.32}$$

The last interval I_M requires a separate treatment. Therefore we will show that there exists a constant $C > 0$ depending only on the mesh regularity parameter κ and the order of discretization r such that

$$\int_{I_M} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k\right\|^2 \mathrm{d}t \leq C \int_{I_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k\right\|^2 \mathrm{d}t \tag{6.33}$$

holds. To see this, we transform the temporal integrals such that the integral over I_{M-1} is transformed to the negative unit interval $(-1, 0)$, which gives

$$\int_{I_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k\right\|^2 \mathrm{d}t = \int_{-1}^{0} \frac{1}{k_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k(t_{M-1} + k_{M-1}\tau)\right\|^2 \mathrm{d}\tau$$

and

$$\int_{I_M} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k\right\|^2 \mathrm{d}t = \int_{0}^{\frac{k_M}{k_{M-1}}} \frac{1}{k_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k(t_{M-1} + k_{M-1}\tau)\right\|^2 \mathrm{d}\tau.$$

We define the polynomial $p\colon \mathbb{R} \to \mathbb{R}$ with maximum degree $2r+2$ by requiring $p(\tau) := \frac{1}{k_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k(t_{M-1} + k_{M-1}\tau)\right\|^2$ for $\tau \in \left(-1, \frac{k_M}{k_{M-1}}\right)$. Since the values of p are non-negative, the second integral can be estimated by

$$\int_{0}^{\frac{k_M}{k_{M-1}}} \frac{1}{k_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k(t_{M-1} + k_{M-1}\tau)\right\|^2 \mathrm{d}\tau = \int_{0}^{\frac{k_M}{k_{M-1}}} p(\tau)\,\mathrm{d}\tau \leq \int_{0}^{\kappa} p(\tau)\,\mathrm{d}\tau$$

When considering both integrals as L^1 norms on the finite dimensional polynomial space $\mathcal{P}_{2r+2}(\mathbb{R})$ we see that those two norms are equivalent and in particular there exists a constant C such that

$$\int_{I_M} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k\right\|^2 \mathrm{d}t \leq \int_{0}^{\kappa} p(\tau)\,\mathrm{d}\tau \leq C \int_{-1}^{0} p(\tau)\,\mathrm{d}\tau = C \int_{I_{M-1}} \left\|\tilde{\pi}_k^{\mathrm{D}} v_k\right\|^2 \mathrm{d}t.$$

The constant C depends only on κ and r. Hence the estimate (6.33) is shown and together with (6.32) we obtain the assertion. □

Proof of Theorem 6.15. The error $\left\|z - \tilde{\pi}_k^{\mathrm{D}} z_k\right\|_I$ is split into two parts. We note that the identity $\tilde{\pi}_k^{\mathrm{D}} \circ \pi_k^{\mathrm{D}} = \tilde{\pi}_k^{\mathrm{D}}$ holds true for arguments in $C(\bar{I}, V)$. Together with Lemma 6.16, we have

$$\left\|z - \tilde{\pi}_k^{\mathrm{D}} z_k\right\|_I \leq \left\|z - \tilde{\pi}_k^{\mathrm{D}} z\right\|_I + \left\|\tilde{\pi}_k^{\mathrm{D}}(z - z_k)\right\|_I \leq \left\|z - \tilde{\pi}_k^{\mathrm{D}} z\right\|_I + C \left\|\pi_k^{\mathrm{D}} z - z_k\right\|_I . \tag{6.34}$$

The first term is bounded by the projection estimate (6.26), which results in

$$\left\|z - \tilde{\pi}_k^{\mathrm{D}} z\right\|_I \leq Ck^{r+2} \left\|\partial_t^{r+2} z\right\|_I .$$

For the second term we pose a discrete dual equation (which is a forward equation again): find $w_k \in X_k^r$ satisfying

$$B(w_k, \varphi) = (\pi_k^{\mathrm{D}} z - z_k, \varphi)_I \quad \forall \varphi \in X_k^r.$$

We choose $\varphi = \pi_k^{\mathrm{D}} z - z_k \in X_k^r$ which gives us

$$\left\|\pi_k^{\mathrm{D}} z - z_k\right\|_I^2 = B(w_k, \pi_k^{\mathrm{D}} z - z_k) = B(w_k, \pi_k^{\mathrm{D}} z - z) + B(w_k, z - z_k). \tag{6.35}$$

To estimate the first term on the right hand side we note that $(\pi_k^{\mathrm{D}} z)_m^+ = z_m^+$ and hence the jump terms in representation (6.11) of the bilinear form B vanish. We get

$$B(w_k, \pi_k^{\mathrm{D}} z - z) = \sum_{m=1}^{M} (\partial_t w_k, \pi_k^{\mathrm{D}} z - z)_{I_m} + (\nabla w_k, \nabla(\pi_k^{\mathrm{D}} z - z))_I. \tag{6.36}$$

For the first term we make use of the fact that the Radau quadrature formula with $r+1$ nodes is exact for polynomials up to degree $2r$ (see [1]). Using the auxiliary operator $\hat{\pi}_k^{\mathrm{D}}$ we get the identity

$$(\partial_t w_k, \pi_k^{\mathrm{D}} z)_{I_m} = (\partial_t w_k, \hat{\pi}_k^{\mathrm{D}} z)_{I_m},$$

since on both sides we have a temporal integral over a polynomial with respect to time with the polynomial on the left having degree $2r-1$ and the polynomial on the right having degree $2r$. Hence, both can be evaluated exactly with Radau's integration formula, which gives the same result in both cases. With the above identity we obtain for the first term of (6.36)

$$\sum_{m=1}^{M} (\partial_t w_k, \pi_k^{\mathrm{D}} z - z)_{I_m} = \sum_{m=1}^{M} (\partial_t w_k, \hat{\pi}_k^{\mathrm{D}} z - z)_{I_m} \leq \left(\sum_{m=1}^{M} \|\partial_t w_k\|_{I_m}^2 \right)^{\frac{1}{2}} \left\|\hat{\pi}_k^{\mathrm{D}} z - z\right\|_I .$$

With the interpolation estimate (6.25) for $\hat{\pi}_k^{\mathrm{D}}$ and the stability estimate for $\partial_t w_k$ from Theorem 6.6, this gives the estimate

$$\sum_{m=1}^{M} (\partial_t w_k, \pi_k^{\mathrm{D}} z - z)_{I_m} \leq Ck^{r+2} \left\|\partial_t^{r+2} z\right\|_I \left\|\pi_k^{\mathrm{D}} z - z_k\right\|_I \tag{6.37}$$

for the time derivative term of (6.36).

To estimate the second term on the right hand side of (6.36), we split it into two parts using the L^2 projection Π_k^0 into the space X_k^0,

$$(\nabla w_k, \nabla(\pi_k^{\mathrm{D}} z - z))_I = (\nabla(w_k - \Pi_k^0 w_k), \nabla(\pi_k^{\mathrm{D}} z - z))_I + (\nabla \Pi_k^0 w_k, \nabla(\pi_k^{\mathrm{D}} z - z))_I. \tag{6.38}$$

To estimate the first term, we integrate by parts with respect to the spatial domain and obtain

$$(\nabla(w_k - \Pi_k^0 w_k), \nabla(\pi_k^{\mathrm{D}} z - z))_I = (w_k - \Pi_k^0 w_k, -\Delta(\pi_k^{\mathrm{D}} z - z))_I$$

The temporal interpolation operator π_k^{D} commutes with the Laplacian and together with the error estimates (6.22) and (6.24) for Π_k^0 and π_k^{D}, we get

$$\begin{aligned}(\nabla(w_k - \Pi_k^0 w_k), \nabla(\pi_k^{\mathrm{D}} z - z))_I &\leq C \left(\sum_{m=1}^{M} k_m^2 \left\| \partial_t w_k \right\|_{I_m}^2 \right)^{\frac{1}{2}} \left\| \pi_k^{\mathrm{D}} \Delta z - \Delta z \right\|_I \\ &\leq C \left(\sum_{m=1}^{M} k_m^2 \left\| \partial_t w_k \right\|_{I_m}^2 \right)^{\frac{1}{2}} k^{r+1} \left\| \partial_t^{r+1} \Delta z \right\|_I .\end{aligned}$$

The stability estimate from Theorem 6.6 gives the desired estimate for the first term of (6.38),

$$(\nabla(w_k - \Pi_k^0 w_k), \nabla(\pi_k^{\mathrm{D}} z - z))_I \leq C k^{r+2} \left\| \pi_k^{\mathrm{D}} z - z_k \right\|_I \left\| \partial_t^{r+1} \Delta z \right\|_I . \tag{6.39}$$

For the second term we have, since the temporal L^2 projection commutes with spatial derivatives,

$$(\nabla \Pi_k^0 w_k, \nabla(\pi_k^{\mathrm{D}} z - z))_I = - \sum_{m=1}^{M} (\Pi_k^0 \Delta w_k, \pi_k^{\mathrm{D}} z - z)_{I_m}.$$

The product $(\Pi_k^0 \Delta w_k, \pi_k^{\mathrm{D}} z)$ is a polynomial of degree $r < 2r$ with respect to time. Thus with the same reasoning as we used when estimating the temporal derivative term, $\pi_k^{\mathrm{D}} z$ can be replaced by $\hat{\pi}_k^{\mathrm{D}} z$ without changing the value of the above expression. Subsequently applying the interpolation estimate (6.25) and taking the continuity of Π_k^0 into consideration gives

$$\begin{aligned}(\nabla \Pi_k^0 w_k, \nabla(\pi_k^{\mathrm{D}} z - z))_I &= - \sum_{m=1}^{M} (\Pi_k^0 \Delta w_k, \hat{\pi}_k^{\mathrm{D}} z - z)_{I_m} \\ &\leq C \sum_{m=1}^{M} k_m^{r+2} \left\| \Delta w_k \right\|_{I_m} \left\| \partial_t^{r+2} z \right\|_{I_m} \\ &\leq C k^{r+2} \left\| \pi_k^{\mathrm{D}} z - z_k \right\|_I \left\| \partial_t^{r+2} z \right\|_I .\end{aligned} \tag{6.40}$$

In the last step we used again Theorem 6.6. Plugging equations (6.37), (6.38), (6.39), and (6.40) into (6.36) we get for the first term of (6.35)

$$B(w_k, \pi_k^{\mathrm{D}} z - z) \leq C k^{r+2} \left\| \pi_k^{\mathrm{D}} z - z_k \right\|_I \left(\left\| \partial_t^{r+2} z \right\|_I + \left\| \partial_t^{r+1} \Delta z \right\|_I \right) . \tag{6.41}$$

We use the projection operator P_k into the semidiscrete space to split the second term on the right hand side of (6.35) into two parts

$$B(w_k, z - z_k) = (w_k, u - u_k)_I = (w_k, u - P_k u)_I + (w_k, P_k u - u_k)_I. \tag{6.42}$$

Due to Condition (6.20a), we have $(\Pi_k^0 w_k, u - P_k u)_I = 0$ and hence the first term can be bounded using the interpolation estimates (6.22) and (6.23) for Π_k^0 and P_k respectively, which gives

$$(w_k, u - P_k u)_I = (w_k - \Pi_k^0 w_k, u - P_k u)_I \leq Ck^{r+2} \left(\sum_{m=1}^{M} \|\partial_t w_k\|_{I_m}^2 \right)^{\frac{1}{2}} \left\| \partial_t^{r+1} u \right\|_I .$$

For the second term we use another duality argument. Let y_k be the solution of the semidiscrete dual equation

$$B(\varphi, y_k) = (\varphi, w_k)_I \quad \forall \varphi \in X_k^r. \tag{6.43}$$

Then testing with $P_k u - u_k$ results in

$$(w_k, P_k u - u_k)_I = B(P_k u - u_k, y_k) = B(P_k u - u, y_k) + B(u - u_k, y_k), \tag{6.44}$$

where the last term vanishes due to Galerkin orthogonality. For the first term we expand the bilinear form in its dual formulation and note that the jump terms vanish due to Condition (6.20b). Since the time derivative $\partial_t y_k|_{I_m}$ is in $\mathcal{P}_{r-1}(I_m, V)$, the time derivative terms vanish as well with Condition (6.20a). This leaves only the spatial operator, that is,

$$B(P_k u - u, y_k) = (\nabla(P_k u - u), \nabla y_k)_I = (P_k u - u, -\Delta y_k + \Pi_k^0 \Delta y_k)_I.$$

In the last step we performed integration by parts and used the orthogonality condition (6.20a). The interpolation estimates (6.22) and (6.23) together with the stability estimates Theorem 6.10 for Equation (6.43) and Theorem 6.6 for the equation for w_k finally result in

$$\begin{aligned} B(P_k u - u, y_k) &\leq Ck^{r+2} \left\| \partial_t^{r+1} u \right\|_I \left(\sum_{m=1}^{M} \|\partial_t \Delta y_k\|_{I_m} \right)^{\frac{1}{2}} \leq Ck^{r+2} \left\| \partial_t^{r+1} u \right\|_I \|\Delta w_k\|_I \\ &\leq Ck^{r+2} \left\| \partial_t^{r+1} u \right\|_I \left\| \pi_k^{\mathrm{D}} z - z_k \right\|_I . \end{aligned}$$

Plugging this estimate into Equation (6.44) gives an estimate for the second term on the right hand side of (6.42). Collecting all estimates shows the claim. □

Remark 6.17. In the same way as in the proof of Theorem 6.15, it can be shown that for the reconstruction $\tilde{\pi}_k^{\mathrm{S}} u_k$ of the semidiscrete state, which is defined analogously to $\tilde{\pi}_k^{\mathrm{D}}$, the estimate

$$\left\| u - \tilde{\pi}_k^{\mathrm{S}} u_k \right\|_I \leq Ck^{r+2} \left(\left\| \partial_t^{r+1} \Delta u \right\|_I + \left\| \partial_t^{r+2} u \right\|_I \right) \tag{6.45}$$

holds if the exact solution u is in $H^{r+2}(I, H) \cap H^{r+1}(I, H^2(\Omega) \cap V)$. For our optimization problem, we cannot apply this estimate, even in the case $r = 1$ since the optimal state $\bar{u}$ will be in $H^2(I, H)$ but in general not in $H^3(I, H)$. However, we will use this kind of reconstruction to approximate the weights for the temporal a posteriori error indicator developed in Section 7.1. The rigorous analysis carried out here provides a partial justification for that. Additionally, the result for the state equation can be used for improving numerical solutions of parabolic PDEs outside the optimization context.

6.3.3. Error analysis for the spatial discretization

We briefly summarize the results we need about the error between semidiscrete and discrete state and adjoint solutions for fixed control.

Theorem 6.18. *For the solution u_{kh} of the discrete state equation* (6.27) *and the semidiscrete solution u_k satisfying* (6.12)*, the a priori error estimate*

$$\|u_k - u_{kh}\|_I \leq Ch^{s+1} \left\|\nabla^{s+1} u_k\right\|_I$$

holds true with the constant C independent of k and h if $u_k \in L^2(I, H^{s+1}(\Omega))$.

If additionally $z_k \in L^2(I, H^{s+1})$, the solutions z_k of the semidiscrete adjoint equation (6.13) *with right hand side $u_k - u_d$ and z_{kh} of the discrete adjoint equation* (6.29) *with right hand side $u_{kh} - u_d$ fulfill the estimate*

$$\|z_k - z_{kh}\|_I \leq Ch^{s+1} \left(\left\|\nabla^{s+1} u_k\right\|_I + \left\|\nabla^{s+1} z_k\right\|_I\right).$$

Proof. Both parts of the claim are shown as Theorem 5.5 and as step in the proof of Lemma 6.2 in Meidner and Vexler [80] respectively. □

Corollary 6.19. *The error $\tilde{\pi}_k^{\mathrm{D}} z_k - \tilde{\pi}_k^{\mathrm{D}} z_{kh}$ between the reconstructed semidiscrete and discrete adjoint solution satisfies the a priori bound*

$$\left\|\tilde{\pi}_k^{\mathrm{D}} z_k - \tilde{\pi}_k^{\mathrm{D}} z_{kh}\right\|_I \leq Ch^{s+1} \left(\left\|\nabla^{s+1} u_k\right\|_I + \left\|\nabla^{s+1} z_k\right\|_I\right).$$

Proof. Applying Lemma 6.16 to the left hand side of the estimate reduces it to the statement of Theorem 6.18. □

6.4. Error analysis for the optimal control problem

Now we turn to the analysis of the discretization error for the control-constrained optimal control problem (6.4). Looking at the temporal and spatial regularity of the optimal state and adjoint state as discussed in Lemma 6.3, we observe that the error estimates in Theorem 6.15 and Corrollary 6.19 can only be applied for the case $r = s = 1$. Therefore, in this section we restrict our considerations to first order elements in both, time and space.

Remark 6.20. Choosing an order $s > 1$ for the spatial discretization can lead to improved convergence with respect to h if we require a domain Ω with smooth boundary and enforce additional compatibility conditions on the data.

6.4.1. Time discretization

As a preparation for our main result we show almost second order convergence of the control with respect to the $L^\infty(I, \mathbb{R}^{d_Q})$ norm. Therefore, we proceed as Hinze [55] by first proving convergence with order $\mathcal{O}(k^2)$ with respect to the $L^2(I, H)$ norm.

Lemma 6.21. *For the error between the solution $\bar{q}$ of the continuous optimal control problem* (6.4) *and the solution $\bar{q}_k$ of the semidiscrete problem* (6.15) *we have the estimate*

$$\|\bar{q} - \bar{q}_k\|_Q \leq Ck^2 \left(\alpha^{-\frac{1}{2}} \left\|\partial_t^2 \bar{u}\right\|_I + \alpha^{-1} \left\|\partial_t^2 \bar{z}\right\|_I \right)$$

and for the corresponding state error we obtain

$$\|\bar{u} - \bar{u}_k\|_I \leq Ck^2 \left(\left\|\partial_t^2 \bar{u}\right\|_I + \alpha^{-\frac{1}{2}} \left\|\partial_t^2 \bar{z}\right\|_I \right).$$

Proof. Testing the optimality condition (6.5) with $\delta q = \bar{q}_k$, its semidiscrete counterpart (6.17) with $\bar{q}$, and adding up the results gives

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_k\|_Q^2 &\leq \left(G^{q*}\left(\bar{z} - \bar{z}_k\right), \bar{q}_k - \bar{q}\right)_Q \\ &= \left(\bar{z} - z_k(\bar{u}), G^q\left(\bar{q}_k - \bar{q}\right)\right)_I + \left(z_k(\bar{u}) - \bar{z}_k, G^q\left(\bar{q}_k - \bar{q}\right)\right)_I. \end{aligned} \tag{6.46}$$

Here, $z_k(\bar{u})$ denotes the solution of the semidiscrete adjoint with $\bar{u}$ entering the right hand side. For the second term on the right, we apply the semidiscrete state equation followed by Galerkin orthogonality and the semidiscrete adjoint and obtain

$$\begin{aligned} &\left(z_k(\bar{u}) - \bar{z}_k, G^q\left(\bar{q}_k - \bar{q}\right)\right)_I = B(\bar{u}_k - \bar{u}, z_k(\bar{u}) - \bar{z}_k) \\ &= B(\bar{u}_k - u_k(\bar{q}), z_k(\bar{u}) - \bar{z}_k) = (\bar{u}_k - u_k(\bar{q}), \bar{u} - \bar{u}_k)_I. \end{aligned}$$

We plug this result into (6.46) and add on both sides $\|\bar{u} - \bar{u}_k\|_I^2$. With the scaled version of Young's inequality, this yields

$$\begin{aligned} \alpha \|\bar{q} - \bar{q}_k\|_Q^2 + \|\bar{u} - \bar{u}_k\|_I^2 &\leq \left(\bar{z} - z_k(\bar{u}), G^q\left(\bar{q}_k - \bar{q}\right)\right)_I \\ &\quad + (\bar{u}_k - u_k(\bar{q}), \bar{u} - \bar{u}_k)_I + (\bar{u} - \bar{u}_k, \bar{u} - \bar{u}_k)_I \\ &\leq \|G^q\| \left\|\bar{z} - z_k(\bar{u})\right\|_I \|\bar{q} - \bar{q}_k\|_Q + \|\bar{u} - u_k(\bar{q})\|_I \|\bar{u} - \bar{u}_k\|_I \\ &\leq \frac{1}{2\alpha} \|G^q\|^2 \|\bar{z} - z_k(\bar{u})\|_I^2 + \frac{\alpha}{2} \|\bar{q} - \bar{q}_k\|_Q^2 \\ &\quad + \frac{1}{2} \|\bar{u} - u_k(\bar{q})\|_I^2 + \frac{1}{2} \|\bar{u} - \bar{u}_k\|_I^2. \end{aligned}$$

Therefore we have with the a priori estimate Theorem 6.13, applied once to the state equation and once to the adjoint equation

$$\begin{aligned} \frac{\alpha}{2} \|\bar{q} - \bar{q}_k\|_Q^2 + \frac{1}{2} \|\bar{u} - \bar{u}_k\|_I^2 &\leq \frac{1}{2\alpha} \|G^q\|^2 \|\bar{z} - z_k(\bar{u})\|_I^2 + \frac{1}{2} \|\bar{u} - u_k(\bar{q})\|_I^2 \\ &\leq Ck^4 \left(\left\|\partial_t^2 \bar{u}\right\|_I^2 + \alpha^{-1} \left\|\partial_t^2 \bar{z}\right\|_I^2 \right) \end{aligned}$$

which results in the desired estimates

$$\|\bar{q} - \bar{q}_k\|_Q \leq Ck^2 \left(\alpha^{-\frac{1}{2}} \left\| \partial_t^2 \bar{u} \right\|_I + \alpha^{-1} \left\| \partial_t^2 \bar{z} \right\|_I \right)$$

and

$$\|\bar{u} - \bar{u}_k\|_I \leq Ck^2 \left(\left\| \partial_t^2 \bar{u} \right\|_I + \alpha^{-\frac{1}{2}} \left\| \partial_t^2 \bar{z} \right\|_I \right). \qquad \square$$

Lemma 6.22. *For the error between the solution $\bar{q}$ of the continuous optimal control problem* (6.4) *and the solution $\bar{q}_k$ of the semidiscrete problem* (6.15) *with respect to the $L^\infty(I, \mathbb{R}^{d_Q})$ norm, the estimate*

$$\|\bar{q} - \bar{q}_k\|_{L^\infty(I,\mathbb{R}^{d_Q})} \leq \alpha^{-1} k^2 \left\{ \gamma(k) \left\| \partial_t^2 \bar{z} \right\|_{L^\infty(I,H)} + \left\| \partial_t^2 \bar{u} \right\|_I + \alpha^{-\frac{1}{2}} \left\| \partial_t^2 \bar{z} \right\|_I \right\}$$

holds true with the logarithmic factor $\gamma(k) = |\log k|^{\frac{1}{2}} + 1$ as introduced in Theorem 6.14.

Proof. Taking into account that $\|P_{Q_{\mathrm{ad}}}(f) - P_{Q_{\mathrm{ad}}}(g)\|_{L^\infty(I,\mathbb{R}^{d_Q})} \leq \|f - g\|_{L^\infty(I,\mathbb{R}^{d_Q})}$, the optimality conditions (6.7) and (6.18) yield

$$\|\bar{q} - \bar{q}_k\|_{L^\infty(I,\mathbb{R}^{d_Q})} \leq \alpha^{-1} \|G^q\| \, \|\bar{z} - \bar{z}_k\|_{L^\infty(I,H)}. \tag{6.47}$$

We introduce an auxiliary adjoint solution $\tilde{z}_k \in X_k^r$ satisfying

$$B(\varphi, \tilde{z}_k) = (\varphi, \bar{u} - u_d)_I \quad \forall \varphi \in X_k^r$$

and split the adjoint error into

$$\|\bar{z} - \bar{z}_k\|_{L^\infty(I,H)} \leq \|\bar{z} - \tilde{z}_k\|_{L^\infty(I,H)} + \|\tilde{z}_k - \bar{z}_k\|_{L^\infty(I,H)}.$$

The first term can be estimated by the supremum norm estimate from Theorem 6.14 applied backward in time, which gives

$$\|\bar{z} - \tilde{z}_k\|_{L^\infty(I,H)} \leq C\gamma(k) k^2 \left\| \partial_t^2 \bar{z} \right\|_{L^\infty(I,H)}. \tag{6.48}$$

For the second term we apply the stability estimate from Lemma 6.8 and the state error bound from Lemma 6.21 to obtain

$$\|\tilde{z}_k - \bar{z}_k\|_{L^\infty(I,H)} \leq C \|\bar{u} - \bar{u}_k\|_I \leq Ck^2 \left(\left\| \partial_t^2 \bar{u} \right\|_I + \alpha^{-\frac{1}{2}} \left\| \partial_t^2 \bar{z} \right\|_I \right). \tag{6.49}$$

Plugging the inequalities (6.48) and (6.49) into (6.47) shows the claim. $\square$

Definition 6.23. Let $\mathcal{M} := \{1, 2, \ldots, M\}$ denote the set of all time indices. For a given $z \in X \cup X_k^r$ we define the sets of active and inactive indices for each of the d_Q components of the resulting control by

$$\mathcal{A}_i(z) := \left\{ m \in \mathcal{M} \;\middle|\; \exists t \in I_m : \left(-\alpha^{-1} G^{q*} z(t)\right)_i > q_i^b \vee \left(-\alpha^{-1} G^{q*} z(t)\right)_i < q_i^a \right\} \text{ and}$$
$$\mathcal{I}_i(z) := \left\{ m \in \mathcal{M} \;\middle|\; \exists t \in I_m : \left(-\alpha^{-1} G^{q*} z(t)\right)_i \in (q_i^a, q_i^b) \right\}$$

respectively where $i \in \{1, \dots, d_Q\}$. For convenience we also define the sets

$$\mathcal{E}_i(z) := \left\{ m \in \mathcal{M} \;\middle|\; \forall t \in I_m : \left(-\alpha^{-1} G^{q*} z(t)\right)_i = q_i^b \vee \left(-\alpha^{-1} G^{q*} z(t)\right)_i = q_i^a \right\}.$$

Note that $\mathcal{A}_i(z) \cup \mathcal{I}_i(z) \cup \mathcal{E}_i(z) = \mathcal{M}$ for any $i \in \{1, \dots, d_Q\}$ and any $z \in X \cup X_k^r$.

With this notation we can introduce the set $\mathcal{K}$ of critical indices collecting all intervals where at least one component is both active and inactive for either of the functions $\bar{z}$ and $\pi_k^{\mathrm{D}} \bar{z}$. The remaining indices are collected in the set $\mathcal{R}$.

$$\begin{aligned} \mathcal{K} &:= \bigcup_{i=1}^{d_Q} \left\{ [\mathcal{A}_i(\bar{z}) \cap \mathcal{I}_i(\bar{z})] \cup \left[\mathcal{A}_i(\pi_k^{\mathrm{D}} \bar{z}) \cap \mathcal{I}_i(\pi_k^{\mathrm{D}} \bar{z})\right] \right\}, \\ \mathcal{R} &:= \mathcal{M} \setminus \mathcal{K}. \end{aligned}$$

Assumption 6.24. We assume that the set $\mathcal{K}$ satisfies

$$\sum_{m \in \mathcal{K}} |I_m| \leq Ck$$

for a constant C independent from k.

Remark 6.25. Similar assumptions are used frequently in the context of error estimates for higher order schemes and post-processing of optimal control problems. As examples, we mention Meyer and Rösch [82] and Vexler and coworkers [14, 79, 81, 96]. The assumption is satisfied if the boundary of the active set for the continuous problem consists of finitely many points and additionally the time derivative of $\left(-\alpha^{-1} G^{q*} \bar{z}\right)_i$ in those points has non-zero value.

With this assumption, we get the following estimate for the reconstructed semidiscrete adjoint solution.

Theorem 6.26. *Let the Assumptions 6.2, 6.4, and 6.24 be fulfilled. Then the error between the optimal adjoint $\bar{z}$ of the continuous problem* (6.4) *and the piecewise quadratic reconstruction $\tilde{\pi}_k^{\mathrm{D}} \bar{z}_k$ of the adjoint for the semidiscrete problem* (6.15) *satisfies*

$$\left\| \bar{z} - \tilde{\pi}_k^{\mathrm{D}} \bar{z}_k \right\|_I \leq C(\alpha) k^3 \left\{ \gamma(k) \left\| \partial_t^2 \bar{z} \right\|_{L^\infty(I,H)} + \left\| \partial_t^2 \bar{u} \right\|_I + \left\| \partial_t^2 \bar{z} \right\|_I + \left\| \partial_t^3 \bar{z} \right\|_I + \left\| \partial_t^2 \Delta \bar{z} \right\|_I \right\}$$

where the constant $C(\alpha)$ can be estimated by $C(\alpha) \leq C\left(1 + \alpha^{-\frac{5}{2}}\right)$ and $\gamma(k)$ is given by $\gamma(k) = |\log k|^{\frac{1}{2}} + 1$.

With Lipschitz continuity of $P_{Q_{\mathrm{ad}}}$ we immediately get the main result of this section.

Corollary 6.27. *Let $(\bar{q}_k, \bar{u}_k, \bar{z}_k)$ be the solutions of the semidiscrete optimization problem. Then for the control solution $\tilde{q}_k$ obtained by the post-processing step*

$$\tilde{q}_k = P_{Q_{\mathrm{ad}}}\left(-\alpha^{-1} G^{q*} \tilde{\pi}_k^{\mathrm{D}} \bar{z}_k\right) \tag{6.50}$$

the estimate

$$\|\bar{q} - \tilde{q}_k\|_Q \leq C(\alpha)k^3 \left\{ \gamma(k) \left\|\partial_t^2 \bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^2 \bar{u}\right\|_I + \left\|\partial_t^2 \bar{z}\right\|_I + \left\|\partial_t^3 \bar{z}\right\|_I + \left\|\partial_t^2 \Delta \bar{z}\right\|_I \right\}$$

holds true with $C(\alpha) \leq C\left(\alpha^{-1} + \alpha^{-\frac{7}{2}}\right)$.

A key ingredient for the proof of Theorem 6.26 are the following two auxiliary controls which are constructed in a similar fashion as in Rösch and Simon [95].

Definition 6.28. The function $p_k \in L^2(I, \mathbb{R}^{d_Q})$ is given piecewise as

$$p_k|_{I_m} = \begin{cases} \bar{q}_k, & \text{if } m \in \mathcal{K}, \\ \pi_k^{\mathrm{D}} \bar{q}, & \text{if } m \in \mathcal{R}, \end{cases}$$

that is, it is identical to the semidiscrete solution on the critical set and interpolates the exact solution on the remaining intervals. In particular, this function is a linear polynomial on each subinterval. Additionally, we will use the function $\hat{p}_k \in L^2(I, \mathbb{R}^{d_Q})$ given by

$$\hat{p}_k|_{I_m} = \begin{cases} \bar{q}_k, & \text{if } m \in \mathcal{K}, \\ \bar{q}, & \text{if } m \in \mathcal{R}. \end{cases}$$

We emphasize that p_k and $\hat{p}_k$ are not related to the unprojected control we introduced in Chapter 3.

Lemma 6.29. *For the difference between the semidiscrete adjoint states computed from the exact control* $\bar{q}$ *and the auxiliary control* p_k*, the estimate*

$$\|z_k(\bar{q}) - z_k(p_k)\|_I \leq C(\alpha)k^3 \left(\gamma(k) \left\|\partial_t^2 \bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^2 \bar{u}\right\|_I + \left\|\partial_t^2 \bar{z}\right\|_I + \left\|\partial_t^3 \bar{z}\right\|_I \right)$$

holds true with $C(\alpha) \leq C\alpha^{-1}\left(1 + \alpha^{-\frac{1}{2}}\right)$ *and* $\gamma(k) = |\log k|^{\frac{1}{2}} + 1$.

Proof. We start with an estimate for the difference between the corresponding semidiscrete states $u_k(\bar{q})$ and $u_k(p_k)$. With the semidiscrete adjoint (6.13) and the semidiscrete state equation (6.12) we get

$$\begin{aligned} \|u_k(\bar{q}) - u_k(p_k)\|_I^2 &= B(u_k(\bar{q}) - u_k(p_k), z_k(\bar{q}) - z_k(p_k)) \\ &= (G^q(\bar{q} - p_k), z_k(\bar{q}) - z_k(p_k))_I. \end{aligned} \tag{6.51}$$

For convenience, we introduce the abbreviation $v_k := z_k(\bar{q}) - z_k(p_k)$. Then, using the auxiliary control $\hat{p}_k$ and the L^2 projection Π_k^0 onto the space of piecewise constant functions we split the right hand side into

$$\begin{aligned} (G^q(\bar{q} - p_k), v_k)_I = (G^q(\bar{q} - \hat{p}_k), v_k)_I + (G^q(\hat{p}_k - p_k), v_k - \Pi_k^0 v_k)_I \\ + (G^q(\hat{p}_k - p_k), \Pi_k^0 v_k)_I. \end{aligned} \tag{6.52}$$

Each of the three terms on the right hand side can be estimated separately. For the first term, we obtain since $\bar{q}$ and $\hat{p}_k$ agree on I_m for $m \in \mathcal{R}$

$$\begin{aligned}(G^q\bar{q} - G^q\hat{p}_k, v_k)_I &= \sum_{m\in\mathcal{K}} \int_{I_m} (G^q\bar{q} - G^q\hat{p}_k, v_k)\,\mathrm{d}t \\ &\leq \sum_{m\in\mathcal{K}} |I_m| \, \|G^q\| \, \|\bar{q} - \hat{p}_k\|_{L^\infty(I_m,\mathbb{R}^{d_Q})} \, \|v_k\|_{L^\infty(I_m,H)} \\ &\leq C \sum_{m\in\mathcal{K}} |I_m| \, \|\bar{q} - \bar{q}_k\|_{L^\infty(I,\mathbb{R}^{d_Q})} \, \|v_k\|_{L^\infty(I,H)} \,.\end{aligned}$$

Plugging the estimate from Lemma 6.22, the maximum norm stability estimate from Lemma 6.8, and Assumption 6.24 into this inequality yields

$$\begin{aligned}&(G^q(\bar{q} - \hat{p}_k), v_k)_I \\ &\quad\leq C\alpha^{-1}k^3 \left(\gamma(k) \left\|\partial_t^2 \bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^2 \bar{u}\right\|_I + \alpha^{-\frac{1}{2}} \left\|\partial_t^2 \bar{z}\right\|_I\right) \|u_k(\bar{q}) - u_k(p_k)\|_I \,.\end{aligned} \tag{6.53}$$

The second term of (6.52) vanishes on all intervals in the critical set $\mathcal{K}$ since $p_k = \hat{p}_k$ there. Hence we have

$$\begin{aligned}&(G^q(\hat{p}_k - p_k), v_k - \Pi_k^0 v_k)_I \leq \|G^q\| \sum_{m\in\mathcal{R}} \left\|\bar{q} - \pi_k^{\mathrm{D}}\bar{q}\right\|_{\mathbb{R}^{d_Q}} \left\|v_k - \Pi_k^0 v_k\right\|_{I_m} \\ &\qquad\leq Ck \left(\sum_{m\in\mathcal{R}} \sum_{i=1}^{d_Q} \left\|\bar{q}_i - \pi_k^{\mathrm{D}}\bar{q}_i\right\|_{L^2(I_m)}^2\right)^{\frac{1}{2}} \left(\sum_{m=1}^{M} \|\partial_t v_k\|_{I_m}^2\right)^{\frac{1}{2}}.\end{aligned} \tag{6.54}$$

To estimate the factor involving the control, we note that for each component $\bar{q}_i$ of the optimal control, all interval indices in $\mathcal{R}$ are contained either in the active set $\mathcal{A}_i(\bar{z})\setminus\mathcal{I}_i(\bar{z})$, the inactive set $\mathcal{I}_i(\bar{z}) \setminus \mathcal{A}_i(\bar{z})$, or the rest set $\mathcal{E}_i(\bar{z})$ belonging to the index i. For the indices in the active set and the rest set, component $\bar{q}_i$ is constant on the corresponding intervals. Hence the difference $(\bar{q} - \pi_k^{\mathrm{D}}\bar{q})_i$ vanishes on those intervals. On the inactive intervals, the component is in $H^3(I,\mathbb{R})$ since the relationship $\bar{q}_i = -\alpha^{-1}(G^{q*}\bar{z})_i$ holds and $\bar{z}$ is in $H^3(I,H)$. Hence we can apply an interpolation estimate for π_k^{D} and obtain

$$\begin{aligned}\Big(\sum_{m\in\mathcal{R}} \sum_{i=1}^{d_Q} \left\|\bar{q}_i - \pi_k^{\mathrm{D}}\bar{q}_i\right\|_{L^2(I_m)}^2\Big)^{\frac{1}{2}} &= \Big(\sum_{i=1}^{d_Q} \sum_{m\in\mathcal{R}\cap\mathcal{I}_i(\bar{z})} \left\|\bar{q}_i - \pi_k^{\mathrm{D}}\bar{q}_i\right\|_{L^2(I_m)}^2\Big)^{\frac{1}{2}} \\ &\leq Ck^2\Big(\sum_{i=1}^{d_Q} \sum_{m\in\mathcal{R}\cap\mathcal{I}_i(\bar{z})} \left\|\partial_t^2\bar{q}_i\right\|_{L^2(I_m)}^2\Big)^{\frac{1}{2}} \\ &\leq C\alpha^{-1}k^2\Big(\sum_{m=1}^{M} \left\|\partial_t^2\bar{z}\right\|_{I_m}^2\Big)^{\frac{1}{2}}.\end{aligned}$$

Plugging this estimate into Equation (6.54) yields for the second term of (6.52) the bound

$$(G^q(\hat{p}_k - p_k), v_k - \Pi_k^0 v_k)_I \leq C\alpha^{-1}k^3 \left\|\partial_t^2 z\right\|_I \left(\sum_{m=1}^{M} \|\partial_t v_k\|_{I_m}^2\right)^{\frac{1}{2}} \leq C\alpha^{-1}k^3 \left\|\partial_t^2 z\right\|_I \|u_k(\bar{q}) - u_k(p_k)\|_I . \tag{6.55}$$

In the last step the stability estimate from Corollary 6.7 was used.

For estimating the third term on the right hand side of (6.52), we introduce a third auxiliary control $\tilde{p}_k$ given by

$$\tilde{p}_k|_{I_m} = \begin{cases} \bar{q}_k, & \text{if } m \in \mathcal{K}, \\ \hat{\pi}_k^{\mathrm{D}} \bar{q}, & \text{if } m \in \mathcal{R}. \end{cases}$$

We observe that $(G^q p_k, \Pi_k^0 v_k)$ is a polynomial of degree one with respect to time on each interval I_m with $m \in \mathcal{R}$. Since the two point Radau quadrature formula integrates polynomials up to degree two exactly, we have the identity $(G^q p_k, \Pi_k^0 v_k)_{I_m} = (G^q \tilde{p}_k, \Pi_k^0 v_k)_{I_m}$ and hence

$$(G^q(\hat{p}_k - p_k), \Pi_k^0 v_k)_I = \sum_{m \in \mathcal{R}} (G^q(\hat{p}_k - \tilde{p}_k), \Pi_k^0 v_k)_{I_m} \leq C\alpha^{-1}k^3 \left\|\partial_t^3 \bar{z}\right\|_I \|v_k\|_I .$$

The last step involves the interpolation estimate (6.25) which can be applied since for intervals with index in $\mathcal{R}$ every component of the control $\bar{q}$ is either constant or in $H^3(I_m, \mathbb{R})$. Using the stability estimate from Corollary 6.7 for the semidiscrete adjoint, we obtain as estimate for the third term

$$(G^q(\hat{p}_k - p_k), \Pi_k^0 v_k)_I \leq C\alpha^{-1}k^3 \left\|\partial_t^3 \bar{z}\right\|_I \|u_k(\bar{q}) - u_k(p_k)\|_I . \tag{6.56}$$

Putting Equations (6.51), (6.52), (6.53), (6.55), and (6.56) together and dividing everything by $\|u_k(\bar{q}) - u_k(p_k)\|_I$ gives

$$\|u_k(\bar{q}) - u_k(p_k)\|_I \leq C\alpha^{-1}k^3 \left\{\gamma(k) \left\|\partial_t^2 \bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^2 \bar{u}\right\|_I + (1 + \alpha^{-\frac{1}{2}}) \left\|\partial_t^2 \bar{z}\right\|_I + \left\|\partial_t^3 \bar{z}\right\|_I\right\} .$$

The desired estimate for $\|z_k(\bar{q}) - z_k(p_k)\|_I$ is obtained from this inequality by means of the stability estimate Corollary 6.7 for the semidiscrete adjoint. □

With these preparations we can prove the main result of this section:

Proof of Theorem 6.26. We start by splitting the error into the two parts

$$\left\|\bar{z} - \tilde{\pi}_k^{\mathrm{D}} \bar{z}_k\right\|_I \leq \left\|\bar{z} - \tilde{\pi}_k^{\mathrm{D}} z_k(\bar{q})\right\|_I + \left\|\tilde{\pi}_k^{\mathrm{D}}(z_k(\bar{q}) - \bar{z}_k)\right\|_I .$$

For the first term, we use the estimate in Theorem 6.15. To bound the second term, we exploit the L^2 stability of the reconstruction operator $\tilde{\pi}_k^{\mathrm{D}}$ and split further

$$\begin{aligned}\left\|\tilde{\pi}_k^{\mathrm{D}}(z_k(\bar{q}) - \bar{z}_k)\right\|_I &\leq C \left\|z_k(\bar{q}) - \bar{z}_k\right\|_I \\ &\leq C \left(\left\|z_k(\bar{q}) - z_k(p_k)\right\|_I + \left\|z_k(p_k) - \bar{z}_k\right\|_I\right).\end{aligned}$$

The first term on the right hand side can be estimated with Lemma 6.29 and for the second term the stability estimates from Corollary 6.7 and Theorem 6.6 give

$$\left\|z_k(p_k) - \bar{z}_k\right\|_I \leq C \left\|p_k - \bar{q}_k\right\|_Q. \tag{6.57}$$

To estimate the term on the right hand side, we first observe that the inequality

$$(G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k, \bar{q}_k - p_k)_Q \geq 0 \tag{6.58}$$

holds true. This can be seen as follows: we write

$$\begin{aligned}(G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k, \bar{q}_k - p_k)_Q = &\sum_{m\in\mathcal{K}} (G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k, \bar{q}_k - p_k)_{\mathbb{R}^{d_Q}} \\ &+ \sum_{m\in\mathcal{R}} \sum_{i=1}^{d_Q} ((G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k)_i, (\bar{q}_k - p_k)_i)_{L^2(I_m)}\end{aligned}$$

and show that each of the addends is non-negative. For the critical set $\mathcal{K}$ the factor $\bar{q}_k - p_k$ vanishes on the corresponding intervals due to the definition of p_k. On the remaining intervals we have to distinguish whether constraints are active or not, that is, for each component $\bar{q}_i$ of the control we have to consider the cases $m \in \mathcal{I}_i(\bar{z}) \setminus \mathcal{A}_i(\bar{z})$, $m \in \mathcal{A}_i(\bar{z}) \setminus \mathcal{I}_i(\bar{z})$, and $m \in \mathcal{E}_i(\bar{z})$. If $m \in \mathcal{I}_i(\bar{z}) \setminus \mathcal{A}_i(\bar{z})$, then we have pointwise $(G^{q*}\bar{z} + \alpha\bar{q})_i = 0$ on the interval I_m. Therefore, on I_m the interpolant $(G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k)_i = \pi_k^{\mathrm{D}}(G^{q*}\bar{z} + \alpha\bar{q})_i$ vanishes.

In the other two cases, one of the constraints is active, that is, the component $\bar{q}_i$ has either the constant value q_i^a or the constant value q_i^b on I_m. So in particular we have $p_k = \bar{q}$. Since $\pi_k^{\mathrm{D}}\bar{z}$ interpolates $\bar{z}$ in two points on I_m, we know that m is also in $\mathcal{A}_i(\pi_k^{\mathrm{D}}\bar{z})$ or $\mathcal{E}_i(\pi_k^{\mathrm{D}}\bar{z})$ respectively and therefore according to the definition of the set $\mathcal{R}$ not in $\mathcal{I}_i(\pi_k^{\mathrm{D}}\bar{z})$. From the optimality condition (6.7), the value of $-\alpha^{-1}(G^{q*}\bar{z})_i$ is either less or equal q_i^a or greater or equal q_i^b on I_m and hence we get for the projection

$$-\alpha^{-1}(G^{q*}\pi_k^{\mathrm{D}}\bar{z})_i \begin{cases} \leq q_i^a, & \text{if } \bar{q}_i = q_i^a, \\ \geq q_i^b, & \text{if } \bar{q}_i = q_i^b. \end{cases}$$

Therefore on the interval I_m, we have pointwise

$$(G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k)_i \begin{cases} \geq 0, & \text{if } \bar{q}_i = q_i^a, \\ \leq 0, & \text{if } \bar{q}_i = q_i^b \end{cases}$$

and, since $\bar{q}_k$ is in the admissible set and $p_k = \bar{q}$ for $m \in \mathcal{R}$,

$$(\bar{q}_k - p_k)_i \begin{cases} \geq 0, & \text{if } \bar{q}_i = q_i^a, \\ \leq 0, & \text{if } \bar{q}_i = q_i^b. \end{cases}$$

So in total we have $((G^{q*}\pi_k^{\mathrm{D}}\bar{z} + \alpha p_k)_i, (\bar{q}_k - p_k)_i)_{L^2(I_m)} \geq 0$.

Testing the semidiscrete optimality condition (6.17) with $\delta q = p_k$ gives

$$(G^{q*}\bar{z}_k + \alpha\bar{q}_k, p_k - \bar{q}_k)_Q \geq 0. \tag{6.59}$$

By adding up the relations (6.58) and (6.59) we get

$$(G^{q*}(\pi_k^{\mathrm{D}}\bar{z} - \bar{z}_k), \bar{q}_k - p_k)_Q - \alpha \left\|\bar{q}_k - p_k\right\|_Q^2 \geq 0.$$

We split the first term and obtain the estimate

$$\begin{aligned} \alpha \left\|\bar{q}_k - p_k\right\|_Q^2 \leq\ & (G^{q*}(\pi_k^{\mathrm{D}}\bar{z} - z_k(\bar{q})), \bar{q}_k - p_k)_Q \\ & + (G^{q*}(z_k(\bar{q}) - z_k(p_k)), \bar{q}_k - p_k)_Q + (G^{q*}(z_k(p_k) - \bar{z}_k), \bar{q}_k - p_k)_Q. \end{aligned} \tag{6.60}$$

The first term on the right hand side is estimated as in the proof of Theorem 6.15 giving

$$\begin{aligned} (G^{q*}(\pi_k^{\mathrm{D}}\bar{z} - z_k(\bar{q})), \bar{q}_k - p_k)_Q &\leq C \left\|\pi_k^{\mathrm{D}}\bar{z} - z_k(\bar{q})\right\|_I \left\|\bar{q}_k - p_k\right\|_Q \\ &\leq Ck^3\big(\left\|\partial_t^2\Delta z\right\|_I + \left\|\partial_t^3 z\right\|_I + \left\|\partial_t^2 u\right\|_I \big) \left\|\bar{q}_k - p_k\right\|_Q . \end{aligned} \tag{6.61}$$

To bound the second term we use Lemma 6.29 to get

$$\begin{aligned} (G^{q*}(z_k(\bar{q}) - z_k(p_k)), \bar{q}_k - p_k)_Q &\leq C \left\|z_k(\bar{q}) - z_k(p_k)\right\|_I \left\|\bar{q}_k - p_k\right\|_Q \\ &\leq C(\alpha)k^3\Big[\gamma(k) \left\|\partial_t^2\bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^3\bar{z}\right\|_I \\ &\qquad + \left\|\partial_t^2\bar{u}\right\|_I + \left\|\partial_t^2\bar{z}\right\|_I\Big] \left\|\bar{q}_k - p_k\right\|_Q \end{aligned} \tag{6.62}$$

with $C(\alpha) \leq C\alpha^{-1}\left(1 + \alpha^{-\frac{1}{2}}\right)$.

Finally the third term on the right hand side of (6.60) is estimated by applying the semidiscrete state equation followed by the semidiscrete adjoint equation, which results in

$$\begin{aligned} (G^{q*}(z_k(p_k) - \bar{z}_k), \bar{q}_k - p_k)_Q &= -B(u_k(p_k) - \bar{u}_k, z_k(p_k) - \bar{z}_k) \\ &= -\left\|u_k(p_k) - \bar{u}_k\right\|_I^2 \leq 0. \end{aligned} \tag{6.63}$$

Plugging the estimates (6.61), (6.62) and (6.63) into (6.60), dividing by $\alpha \left\|\bar{q}_k - p_k\right\|_Q$ and using Young's inequality to obtain $1 + \alpha^{-1} + \alpha^{-\frac{3}{2}} \leq C\left(1 + \alpha^{-\frac{3}{2}}\right)$ gives

$$\begin{aligned} \left\|p_k - \bar{q}_k\right\|_Q \leq C(\alpha)k^3\Big\{&\gamma(k) \left\|\partial_t^2\bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^2\bar{u}\right\|_I \\ &+ \left\|\partial_t^2\bar{z}\right\|_I + \left\|\partial_t^3\bar{z}\right\|_I + \left\|\partial_t^2\Delta\bar{z}\right\|_I \Big\} \end{aligned} \tag{6.64}$$

with $C(\alpha) \leq C\left(\alpha^{-1} + \alpha^{-\frac{5}{2}}\right)$. Plugging this estimate into (6.57), collecting all the resulting terms and estimating the factors involving α where appropriate with Young's inequality yields the claim. □

6.4.2. Spatial discretization

The error that the spatial discretization causes on the post-processed solution can be assessed independently. We show the following estimate.

Theorem 6.30. *For the error $\tilde{\pi}_k^{\mathrm{D}} \bar{z}_k - \tilde{\pi}_k^{\mathrm{D}} \bar{z}_{kh}$ between the reconstruction of the semidiscrete optimal adjoint and the reconstruction of the adjoint $\bar{z}_{kh}$ belonging to the discrete optimal control problem* (6.28) *the estimate*

$$\left\|\tilde{\pi}_k^{\mathrm{D}} \left(\bar{z}_k - \bar{z}_{kh}\right)\right\|_I \leq C\left(1 + \alpha^{-1}\right) h^2 \left\{\left\|\nabla^2 \bar{u}_k\right\|_I + \left\|\nabla^2 \bar{z}_k\right\|_I\right\}$$

holds true.

Proof. To show the claim, we split

$$\left\|\tilde{\pi}_k^{\mathrm{D}} \left(\bar{z}_k - \bar{z}_{kh}\right)\right\|_I \leq \left\|\tilde{\pi}_k^{\mathrm{D}} \bar{z}_k - \tilde{\pi}_k^{\mathrm{D}} z_{kh}(\bar{q}_k)\right\|_I + \left\|\tilde{\pi}_k^{\mathrm{D}} \left(z_{kh}(\bar{q}_k) - \bar{z}_{kh}\right)\right\|_I . \tag{6.65}$$

Corollary 6.19 estimates the first term on the right hand side. For the second term we use first the stability estimate in Lemma 6.16 and subsequently the stability estimates in Theorem 6.12 for the fully discrete state and adjoint equations which give

$$\begin{aligned} \left\|\tilde{\pi}_k^{\mathrm{D}} \left(z_{kh}(\bar{q}_k) - \bar{z}_{kh}\right)\right\|_I &\leq C \left\|z_{kh}(\bar{q}_k) - \bar{z}_{kh}\right\|_I \leq C \left\|u_{kh}(\bar{q}_k) - \bar{u}_{kh}\right\|_I \\ &\leq C \left\|\bar{q}_k - \bar{q}_{kh}\right\|_Q . \end{aligned} \tag{6.66}$$

To estimate the term $\bar{q}_k - \bar{q}_{kh}$ we test the optimality conditions (6.17) and (6.30) with $\delta q = \bar{q}_{kh}$ and $\delta q = \bar{q}_k$ respectively and add up the results. This gives

$$\left(\alpha \bar{q}_k + G^{q*} \bar{z}_k - \left(\alpha \bar{q}_{kh} + G^{q*} \bar{z}_{kh}\right), \bar{q}_{kh} - \bar{q}_k\right)_Q \geq 0.$$

Hence,

$$\begin{aligned} \alpha \left\|\bar{q}_k - \bar{q}_{kh}\right\|_Q^2 &\leq \left(\bar{z}_k - \bar{z}_{kh}, G^q \left(\bar{q}_{kh} - \bar{q}_k\right)\right)_I \\ &= \left(\bar{z}_k - z_{kh}(\bar{q}_k), G^q \left(\bar{q}_{kh} - \bar{q}_k\right)\right)_I + \left(z_{kh}(\bar{q}_k) - \bar{z}_{kh}, G^q \left(\bar{q}_{kh} - \bar{q}_k\right)\right)_I . \end{aligned} \tag{6.67}$$

For the second term, using the discrete state equation (6.27) followed by the discrete adjoint (6.29) we obtain

$$\begin{aligned} \left(z_{kh}(\bar{q}_k) - \bar{z}_{kh}, G^q \left(\bar{q}_{kh} - \bar{q}_k\right)\right)_I &= B\left(z_{kh}(\bar{q}_k) - \bar{z}_{kh}, \bar{u}_{kh} - u_{kh}(\bar{q}_k)\right) \\ &= - \left\|\bar{u}_{kh} - u_{kh}(\bar{q}_k)\right\|_I^2 \leq 0. \end{aligned}$$

Therefore, Equation (6.67) gives the estimate

$$\left\|\bar{q}_k - \bar{q}_{kh}\right\|_Q \leq C\alpha^{-1} \left\|\bar{z}_k - z_{kh}(\bar{q}_k)\right\|_I . \tag{6.68}$$

Plugging this result into (6.66), applying Theorem 6.18 and collecting the resulting terms on the right hand side of Equation (6.65) shows the assertion. □

Corollary 6.31. *For the reconstructed fully discrete solution* $\tilde{q}_{kh} = P_{Q_{\mathrm{ad}}}\left(-\alpha^{-1}\tilde{\pi}_k^{\mathrm{D}}\bar{z}_{kh}\right)$, *we have the estimate*

$$\begin{aligned}\|\tilde{q}_{kh} - \bar{q}\|_Q \le C_1(\alpha)k^3 &\left\{\gamma(k)\left\|\partial_t^2\bar{z}\right\|_{L^\infty(I,H)} + \left\|\partial_t^2\bar{u}\right\|_I + \left\|\partial_t^2\bar{z}\right\|_I + \left\|\partial_t^3\bar{z}\right\|_I + \left\|\partial_t^2\Delta\bar{z}\right\|_I\right\} \\ &+ C_2(\alpha)h^2\left\{\left\|\nabla^2\bar{u}_k\right\|_I + \left\|\nabla^2\bar{z}_k\right\|_I\right\}\end{aligned}$$

with $C_1(\alpha) \le C\left(\alpha^{-1} + \alpha^{-\frac{7}{2}}\right)$ *and* $C_2(\alpha) \le C\left(\alpha^{-1} + \alpha^{-2}\right)$.

Proof. The error is split into

$$\|\tilde{q}_{kh} - \bar{q}\|_Q \le \|\tilde{q}_{kh} - \tilde{q}_k\|_Q + \|\tilde{q}_k - \bar{q}\|_Q\,,$$

we estimate the first term using Lipschitz continuity of $P_{Q_{\mathrm{ad}}}$ and Theorem 6.30, and the second term by Corollary 6.27. □

6.5. Numerical validation

For the numerical tests, we consider the case $d_Q = 1$, that is, a control consisting of one time dependent parameter. As spatial domain we use the unit square $\Omega = (0,1)^2$. The data and exact solutions of the test problems are stated in terms of the eigenfunctions

$$w_k(x) = 2\sin(k\pi x_1)\sin(k\pi x_2)$$

of the Laplacian on the unit square. We denote the corresponding eigenvalues by $\lambda_k = 2k^2\pi^2$. The operator G^q is defined through $(G^q q)(t,x) := q(t)w_k(x)$ and for the right hand side of the state equation we set $f = 0$.

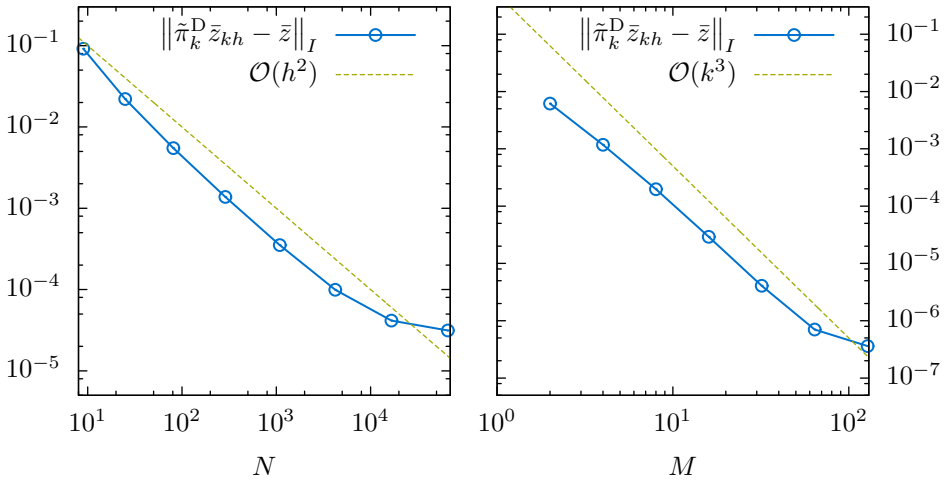
For solving the discrete optimal control problem, we use the semismooth Newton method discussed in Section 3.2. All computations were done using the software package RoDoBo [93].

Our test example is constructed such that the third derivative of the adjoint state with respect to time has a jump at the point where the control constraint becomes inactive. We consider the time interval $I = (\frac{1}{2}, 1)$ and the control constraints $q^a = -\frac{\sqrt{3}}{2\alpha}$ and $q^b = \frac{\sqrt{3}}{2\alpha}$. The remaining data are chosen as

$$\begin{aligned}u_d(t,x) &= \left(\pi\cos(\pi t) - \lambda_k\sin(\pi t)\right)\left(1 + \frac{1}{\alpha(\lambda_k^2 + \pi^2)}\right)w_k(x), \\ u\left(\frac{1}{2}, x\right) &= \left(\frac{\pi e^{\frac{1}{6}\lambda_k}\left(\pi\sqrt{3} - \lambda_k\right)}{2\lambda_k\alpha(\lambda_k^2 + \pi^2)} - \frac{\sqrt{3}}{2\lambda_k\alpha}\right)w_k(x),\end{aligned}$$

$k = 1$, and $\alpha = 0.1$. For this choice of data, the optimal control is given by

$$\bar{q}(t) = \begin{cases} -\frac{\sqrt{3}}{2\alpha}, & \text{if } t \le \frac{2}{3}, \\ -\frac{\sin(\pi(1-t))}{\alpha}, & \text{otherwise.}\end{cases}$$



(a) Refinement of the spatial grid for $M = 16$ time steps

(b) Refinement of the time steps for triangulation with $N = 1050625$ nodes

Figure 6.2.: Discretization error $\left\|\tilde{\pi}_k^{\mathrm{D}} \bar{z}_{kh} - \bar{z}\right\|_I$ for spatial and temporal refinement

and for the optimal adjoint we obtain

$$\bar{z}(t,x) = \begin{cases} z_1(t) w_k(x), & \text{if } t \leq \frac{2}{3}, \\ \sin(\pi(1-t)) w_k(x), & \text{otherwise,} \end{cases}$$

where

$$z_1(t) = \frac{\pi^2\sqrt{3}\cosh\left(\lambda_k\left(t-\frac{2}{3}\right)\right) + \pi\lambda_k \sinh\left(\lambda_k\left(t-\frac{2}{3}\right)\right)}{2\alpha\lambda_k^2\left(\lambda_k^2+\pi^2\right)} + \sin(\pi t)\left(1 + \frac{1}{\alpha\left(\lambda_k^2+\pi^2\right)}\right) - \frac{\sqrt{3}}{2\alpha\lambda_k^2}.$$

We assess the $L^2(I,H)$ errors of the reconstructed adjoint for spatial and temporal discretization separately: to investigate the error of the spatial discretization, we fix the number of time steps at $M = 16$; for the temporal error we consider a fixed uniform spatial triangulation with $N = 1050625$ nodes. In Figure 6.2(a), the error $\left\|\tilde{\pi}_k^{\mathrm{D}} \bar{z}_{kh} - \bar{z}\right\|_I$ for a sequence of uniform refinements of the spatial grid is shown. Second order convergence with respect to the mesh width $h = \sqrt{\frac{1}{N}}$ is observed down to where the error contribution of the time discretization dominates.

Figure 6.2(b) shows the development of the error when refining the width $k = \frac{1}{M}$ of the time steps. The highest numerical order of convergence we observe is about 2.85, considerably less than predicted by Theorem 6.26. However, we note a slight increase of the observed order of convergence as the time steps decrease, up to the point where the

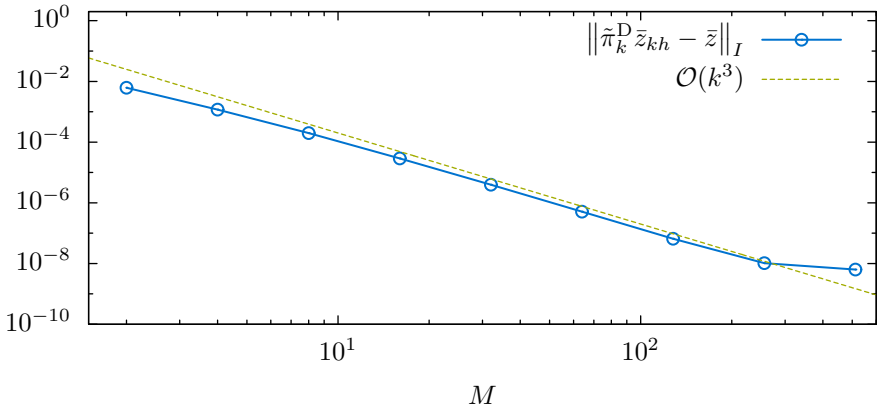


Figure 6.3.: Discretization error $\left\|\tilde{\pi}_k^{\mathrm{D}} \bar{z}_{kh} - \bar{z}\right\|_I$ for biquadratic space discretization ($s = 2$) with $N = 263169$ nodes and uniform refinement of the time steps.

spatial discretization error becomes dominant. Hence, it is reasonable to assume that the third order convergence will become apparent for smaller time steps and therefore can only be observed for an even finer spatial discretization. We substantiate this assumption by a computation using biquadratic finite elements in space, i. e., $s = 2$, which makes sense here since the solutions of our test problem are smooth with respect to the spatial variable. The result when refining the time step for a fixed spatial discretization with $N = 263169$ nodes is plotted in Figure 6.3. We observe a maximal estimated order of convergence of 2.97.

As evidence that the variational treatment of the control is in fact necessary to guarantee the almost third order error estimate for the post-processed solution, we consider an ODE version of the model problem (2.6). This avoids the spatial discretization error and therefore makes it easier to observe the influence of the control treatment on the time discretization. We compare our solution approach to two variants of control discretization with piecewise linear discontinuous Ansatz functions. On the one hand we consider the discretization with constraints only enforced at the Gauß nodes discussed in Section 4.2.2. On the other hand we look at the more conventional approach where the control constraints are enforced globally for the discrete control. Since application of the semismooth Newton method is problematic in the latter case, we use a primal-dual active set strategy instead, see, e. g., Kunisch and co-workers [15, 63]. The scalar ODE problem reads: Minimize

$$J(q,u) = \frac{1}{2}\int_0^1 (u(t) - u_d(t))^2 \,\mathrm{d}t + \frac{\alpha}{2}\int_0^1 q(t)^2 \,\mathrm{d}t$$

subject to

$$\begin{aligned} \partial_t u + u &= q, \quad u(0) = 0, \quad \text{and} \\ 1 \le q(t) &\le 2 \quad \text{a. e.} \end{aligned}$$

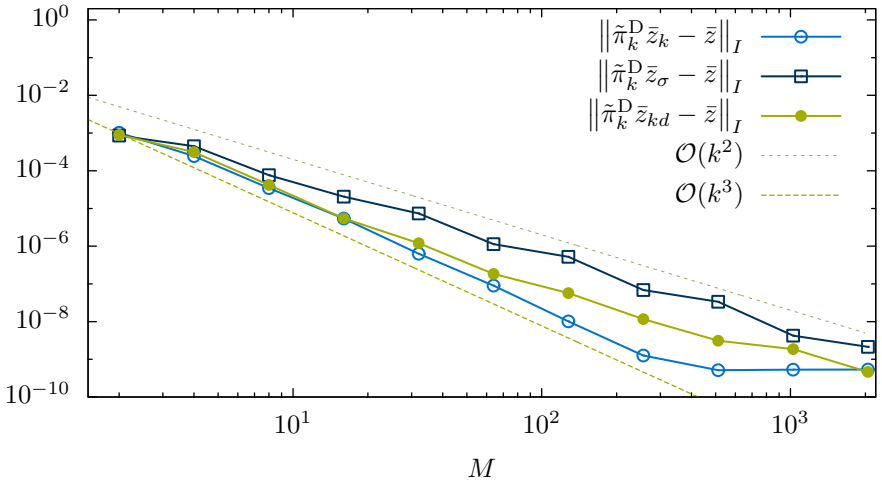


Figure 6.4.: Discretization errors $\left\|\tilde{\pi}_k^{\mathrm{D}}\bar{z}_k - \bar{z}\right\|_I$, $\left\|\tilde{\pi}_k^{\mathrm{D}}\bar{z}_\sigma - \bar{z}\right\|_I$, and $\left\|\tilde{\pi}_k^{\mathrm{D}}\bar{z}_{kd} - \bar{z}\right\|_I$ for the ODE example

with $q \in L^2((0,1))$ and $u \in H^1((0,1))$. We specify the data as

$$u_d(t) = \frac{3}{100}(2-t) + \begin{cases} 2 - 2e^{-t}, & t < \frac{1}{3}, \\ 6 - 3t - \left(3e^{\frac{1}{3}} + 2\right)e^{-t}, & \frac{1}{3} \le t \le \frac{2}{3}, \\ 1 + \left(3e^{\frac{2}{3}} - 3e^{\frac{1}{3}} - 2\right)e^{-t}, & t > \frac{2}{3}, \end{cases}$$

and $\alpha = 0.01$. Our regularity assumptions are obviously satisfied for the given data. The resulting optimal solutions are given by

$$\bar{q}(t) = \begin{cases} 2, & t < \frac{1}{3}, \\ 3 - 3t, & \frac{1}{3} \le t \le \frac{2}{3}, \\ 1, & t > \frac{2}{3}, \end{cases}$$

$$\bar{u}(t) = u_d(t) + \frac{3}{100}(t-2), \quad \text{and}$$

$$\bar{z}(t) = \frac{3}{100}(t-1).$$

We denote the discrete optimal control, state, and adjoint resulting from the discretized control with discretized constraints by $\bar{q}_{kd}$, $\bar{u}_{kd}$, and $\bar{z}_{kd}$ respectively and for the solution with discrete control but exact enforcement of the constraints by $\bar{q}_\sigma$, $\bar{u}_\sigma$, and $\bar{z}_\sigma$.

We examine the $L^2(I)$ error of the reconstructed adjoint for all three treatments of the control on a sequence of uniformely refined temporal grids. In Figure 6.4 the development of the errors $\left\|\tilde{\pi}_k^{\mathrm{D}}\bar{z}_k - \bar{z}\right\|_{L^2(I)}$, $\left\|\tilde{\pi}_k^{\mathrm{D}}\bar{z}_\sigma - \bar{z}\right\|_{L^2(I)}$ and $\left\|\tilde{\pi}_k^{\mathrm{D}}\bar{z}_{kd} - \bar{z}\right\|_{L^2(I)}$ with respect to the number of time steps M is shown. It can be seen that while the reconstruction obtained from variational control treatment converges with third order, the error of $\tilde{\pi}_k^{\mathrm{D}}\bar{z}_\sigma$ decreases only with second order. The error with respect to $\tilde{\pi}_k^{\mathrm{D}}\bar{z}_{khd}$ appears to converge slightly better, however, its order is still clearly less than three.

Remark 6.32. In the example the control is piecewise linear, so the main contribution to the observed error stems from the two discretization intervals containing kinks. For more general examples, it can be difficult to observe the different behaviour of the two discretizations because frequently the error caused by the kinks is dominated by the error contribution from approximating the control on the rest of the interval. Since the latter seems to be of third order for both, the error for the discretized control will be dominated by the second order component originating from the kinks only for very small time steps.

7. A posteriori error control and hp adaptivity

The a priori analysis presented in Chapter 6 approaches the goal of constructing a rapidly converging higher order dG time discretization for a control-constrained parabolic problem by carefully tailoring the discretization scheme to the regularities of the optimization variables. In this chapter, we explore a different approach. For many problems, low regularity occurs on the boundary between the active and inactive sets resulting from the control constraints. Since this boundary in the case of space-independent control frequently consists of only a few isolated kinks, typically we expect that non-smoothness is only encountered locally in time.

In this chapter we aim at resolving this localized non-smoothness by an adaptive mesh refinement procedure based on appropriate *a posteriori* error indicators. On the parts of the time domain where the solution is smooth, we would like to profit from the good approximation properties of high order dG schemes. However, on the parts of the time interval where the solution is not smooth, using high order schemes implies unnecessary overhead. Therefore, we use an hp adaptive algorithm, i. e., on time steps that are marked for refinement by the adaptive algorithm, either the temporal grid can be refined (h refinement) or the order of discretization can be increased (p refinement).

For an hp adaptive algorithm, two main ingredients are required. After solving the optimization problem for a given discretization, a procedure is needed to determine which regions of the temporal and spatial meshes need refinement in order to reach the desired accuracy. Suitable a posteriori error indicators form the basis of this step. In order to perform hp refinement, as opposed to pure h adaptivity, additionally we have to decide where to increase the order and where to refine the mesh. For this purpose, an *hp refinement strategy* has to be defined.

Due to the high computational cost involved in solving optimization problems with PDE constraints, h adaptive discretization schemes for such problems have received considerable attention in recent years. In many cases, error estimates with respect to the energy norm of the underlying PDE are used. As an example, we mention the early contributions [68, 70] by Liu and coworkers. For optimal control problems however, it is frequently reasonable to base the refinement procedure on the approximation error of the value of the cost functional. This has the desirable effect that primarily those parts of the optimization variables are resolved well which have most influence on the cost functional.

Such goal oriented error estimates are usually formulated in the context of the *dual weighted residual* (DWR) error estimation framework. For a general introduction to the method, we refer to the survey article [13] by Becker and Rannacher. For parabolic optimal control problems with dG time discretization and continuous finite elements in space, error estimates of DWR type with respect to the cost functional are derived by Meidner and Vexler [78]. We adapt their error estimates to our context and augment them by estimates for the errors due to control constraints and numerical quadrature. To approximate the exact state and adjoint solution occurring in the weights of the temporal error estimators, we exploit the superconvergent reconstruction that was introduced in Section 4.1.1. This allows for an interval-wise evaluation of the error estimator, implying no restrictions for the choice of the temporal *hp* grid.

To our knowledge, the only works on *hp* discretization in the context of optimal control problems are the articles [16,110,111] by Wachsmuth, Wurst, and co-authors that consider spatial *hp* refinement for linear quadratic elliptic problems. Their approach for handling problems with control constraints is based on the energy error. For the solution of elliptic PDEs without optimal control, the combination of goal oriented error estimation and *hp* refinement was considered for example by Heuveline and Rannacher [54] and Šolín and Demkowicz [103]. Schötzau and co-authors [100–102] discuss *hp* adaptivity for dG time discretization driven by a priori knowledge on the solution and energy-based error indicators.

For the decision whether to perform h or p refinement, many different strategies have been proposed; for an overview we refer to Mitchell and McClain [84], where the numerical performance of thirteen *hp* refinement strategies is compared. The strategies producing the optimal discretizations in this comparison rely on computing a reference solution on a mesh that is uniformly refined in both h and p. Šolín and Demkowicz [103] employ such a strategy. However, they also use the reference solution for evaluating the weights in their DWR error estimator such that the computational effort for computing the reference solution can be justified.

Since the reconstruction procedure from Chapter 6 allows us to evaluate the error estimator without additional computations on finer discretizations, we would like to avoid them also for the *hp* refinement strategy. Under this restriction, the strategy performing best in Mitchell and McClain [84] is due to Mavriplis [74]. It uses the decay rate of the Legendre coefficients of the local discrete approximation to estimate the smoothness of the solution. The strategy we will employ here is a simple heuristics proposed more recently by Wihler [116] and is based on a similar idea: the local smoothness of the exact solution is estimated by monitoring certain Sobolev embedding constants for the local discrete solution. Its main advantages are its simple structure and the fact that it is applicable also to low order schemes.

We organize this chapter as follows: Sections 7.1 and 7.2 respectively discuss the two main ingredients of the *hp* adaptive procedure, the a posteriori error indicators for the various error contributions, and the smoothness indicator driving the *hp* refinement strategy. In

Section 7.3 we show how they can be combined to form the complete adaptive algorithm. Numerical results are reported in Section 7.4.

7.1. DWR error estimators for the error with respect to the cost functional

In this section, we derive a posteriori error estimators for the discretization error with respect to the cost functional. Our derivation is similar to the one of Meidner and Vexler [78]. However, due to the control constraints, additional terms arise. For treating them, we proceed similar as Vexler and Wollner [109] in the case of elliptic problems with control constraints. A further extension becomes necessary due to the use of high order time discretization: as noted for example by Schmich and Vexler [99] and confirmed by some numerical experiments, the quadrature error from integrating nonlinearities and, in particular problem data, can contribute significantly to the overall discretization error when using high order schemes. Although we use quadrature formulas of sufficiently high order, taking the quadrature error into account improves the accuracy of the error prediction on coarse meshes. In [77], Meidner and Richter analyze a time stepping scheme that can be interpreted as a Galerkin scheme with numerical quadrature. We adopt their approach for estimating the quadrature error numerically.

As in Chapter 4, we will discuss only the case of a time dependent parameter control, i. e., $Q = L^2(I, \mathbb{R}^{d_Q})$. However, an extension of the presented error estimates to distributed control in space is straightforward. Since we use identical time discretization for control and state, we estimate the errors resulting from both time discretizations combined. The error due to the spatial discretization of the state however has to be accounted for separately in order to decide whether refinement with respect to time or space has to be performed.

7.1.1. Derivation of the error estimators

In order to fix notation, we introduce the semidiscrete state equation with numerical quadrature for the temporal integrals as: given q_τ, find $u_\tau \in X_k^r$ satisfying

$$\begin{aligned}(7.1)\quad \sum_{m=1}^{M} \left[(\partial_t u_\tau, \varphi)_{I_m} + \mathcal{Q}_m\left(a(q_\tau, u_\tau)(\varphi)\right)\right] + \sum_{m=1}^{M-1} \left([u_\tau]_m, \varphi_m^+\right) \\ + \left(u_{\tau,0}^+, \varphi_0^+\right) = \left(u_0, \varphi_0^+\right) \quad \text{for any } \varphi \in X_k^r,\end{aligned}$$

where $\mathcal{Q}_m \colon C(\bar{I}_m) \to \mathbb{R}$ is the interpolatory quadrature rule used on interval I_m for evaluation of the temporal integrals. We assume that the quadrature rule has sufficiently high order that bilinear terms involving only discrete quantities are integrated exactly.

Here, we will typically use the $(r_m + 1)$-point Gauß formula, which clearly has this property. The time discrete optimization problem with quadrature reads

$$\begin{aligned} &\text{Minimize } J^\tau(q_\tau, u_\tau) = J_1^\tau(u_\tau) + J_2(u_{\tau,M}^-) + \frac{\alpha}{2}\|q_\tau\|_Q^2 \\ &\text{subject to } \begin{cases} (q_\tau, u_\tau) \in Q_d \times X_k^r \text{ satisfying (7.1)}, \\ q_\tau \in Q_{d,\mathrm{ad}}. \end{cases} \end{aligned} \tag{7.2}$$

The notation J_1^τ indicates that any temporal integrals occurring in J_1 are replaced by the chosen quadrature rule. As usual we denote the optimal solution of this problem along with the corresponding adjoint state by $(\bar{q}_\tau, \bar{u}_\tau, \bar{z}_\tau)$ and the associated Lagrangian by $\hat{\mathcal{L}}^\tau$.

Let $\mathcal{Q}_m^h$ denote the spatial quadrature formula employed on the mesh $\mathcal{T}_h^m$ for the m^{th} time interval. As for the time discretization we assume the quadrature to have sufficiently high order that all bilinear terms in the state equation are integrated exactly. We also ensure that the jump terms are integrated exactly. In practice this is accomplished by working on the common refinement of two subsequent spatial discretizations. We state the fully discrete problem with numerical quadrature as:

$$\text{Minimize } J^\sigma(q_\sigma, u_\sigma) = J_1^\sigma(u_\sigma) + J_2^\sigma(u_{\sigma,M}^-) + \frac{\alpha}{2}\|q_\sigma\|_Q^2 \tag{7.3}$$

subject to $(q_\sigma, u_\sigma) \in Q_{d,\mathrm{ad}} \times X_{k,h}^{r,s}$ satisfying

$$\begin{aligned} &\sum_{m=1}^{M} (\partial_t u_\sigma, \varphi)_{I_m} + \sum_{m=1}^{M} \mathcal{Q}_m\left(\mathcal{Q}_m^h\left(a(q_\sigma, u_\sigma)(\varphi)\right)\right) + \sum_{m=1}^{M-1} ([u_\sigma]_m, \varphi_m^+) \\ &\qquad + (u_{\sigma,0}^+, \varphi_0^+) = \mathcal{Q}_1^h\left(u_0 \varphi_0^+\right) \quad \text{for any } \varphi \in X_{k,h}^{r,s}. \end{aligned} \tag{7.4}$$

Again, the superscript σ for the two components of the cost functional indicates that all integrals are replaced by quadrature rules. The optimal solutions are denoted by $(\bar{q}_\sigma, \bar{u}_\sigma, \bar{z}_\sigma)$ and the Lagrangian by $\hat{\mathcal{L}}^\sigma$.

To separate influences of the time and space discretization, we split the functional error by

$$J(\bar{q}, \bar{u}) - J^\sigma(\bar{q}_\sigma, \bar{u}_\sigma) = J(\bar{q}, \bar{u}) - J^\tau(\bar{q}_\tau, \bar{u}_\tau) + J^\tau(\bar{q}_\tau, \bar{u}_\tau) - J^\sigma(\bar{q}_\sigma, \bar{u}_\sigma)$$

into the two parts

$$\eta_\tau \approx J(\bar{q}, \bar{u}) - J^\tau(\bar{q}_\tau, \bar{u}_\tau) \qquad \text{and} \qquad \eta_\sigma \approx J^\tau(\bar{q}_\tau, \bar{u}_\tau) - J^\sigma(\bar{q}_\sigma, \bar{u}_\sigma)$$

for which we derive separate error indicators. For the convenient formulation of the error indicators, we introduce some abbreviations. The triple consisting of optimal control, state, and adjoint state is denoted by $\xi = (\bar{q}, \bar{u}, \bar{z})$. Correspondingly we define

$\xi_\tau = (\bar{q}_\tau, \bar{u}_\tau, \bar{z}_\tau)$ and $\xi_\sigma = (\bar{q}_\sigma, \bar{u}_\sigma, \bar{z}_\sigma)$. For given $\zeta_k = (q_k, u_k, z_k) \in Q \times X_k^r \times X_k^r$, the residuals of the optimality system without quadrature are defined as

$$\begin{aligned}
\rho_q(\zeta_k)(\delta q) &= \hat{\mathcal{L}}_q'(\zeta_k)(\delta q) = \alpha(q_k, \delta q)_Q - \int_I a_q'(q_k, u_k)(\delta q, z_k)\,\mathrm{d}t,\\
\rho_u(\zeta_k)(\varphi) &= \hat{\mathcal{L}}_z'(\zeta_k)(\varphi) = (u_0 - u_{k,0}^-, \varphi_0^+) - \sum_{m=1}^{M} (\partial_t u_k, \varphi)_{I_m}\\
&\quad - \sum_{m=1}^{M-1} ([u_k]_m, \varphi_m^+) - \int_I a(q_k, u_k)(\varphi)\,\mathrm{d}t,\\
\rho_z(\zeta_k)(\varphi) &= \hat{\mathcal{L}}_u'(\zeta_k)(\varphi) = J_1'(u_k)(\varphi) + J_2'(u_{k,M}^-)(\varphi_M^-) - (z_{k,M}^-, \varphi_M^-)\\
&\quad + \sum_{m=1}^{M} (\partial_t z_k, \varphi)_{I_m} + \sum_{m=1}^{M-1} ([z_k]_m, \varphi_m^-) - \int_I a_u'(q_k, u_k)(z_k, \varphi)\,\mathrm{d}t,
\end{aligned}$$

with $\delta q \in Q$ and $\varphi \in X + X_k^r$.

Assumption 7.1. We assume that the continuous, time-discrete, and discrete optimal control problems admit optimal solutions ξ, ξ_τ, and ξ_σ. Furthermore we require the functionals J_1, J_2, and the semilinear form a to be three times Gâteaux differentiable with respect to q and u.

Proposition 7.2. *Under Assumption 7.1, the temporal discretization error with respect to the cost functional admits the representation*

$$\begin{aligned}
J(\bar{q}, \bar{u}) - J^\tau(\bar{q}_\tau, \bar{u}_\tau) = \frac{1}{2}\big\{&\rho_u(\xi_\tau)(\bar{z} - \pi_k^{\mathrm{D}}\bar{z}) + \rho_z(\xi_\tau)(\bar{u} - \pi_k^{\mathrm{S}}\bar{u})\\
&+\rho_q(\xi_\tau)(\bar{q} - \bar{q}_\tau) + \rho_q(\xi)(\bar{q} - \bar{q}_\tau)\big\} + \hat{\mathcal{L}}(\xi_\tau) - \hat{\mathcal{L}}^\tau(\xi_\tau) + \mathcal{R}^\tau
\end{aligned}$$

with the remainder term

$$\mathcal{R}^\tau = \frac{1}{2}\left\{\rho_u(\xi_\tau)(\pi_k^{\mathrm{D}}\bar{z} - \bar{z}_\tau) + \rho_z(\xi_\tau)(\pi_k^{\mathrm{S}}\bar{u} - \bar{u}_\tau) + \int_0^1 \hat{\mathcal{L}}'''(\xi_\tau + s e_\tau)(e_\tau, e_\tau, e_\tau) s(s-1)\,ds\right\}$$

where π_k^{S} and π_k^{D} denote the temporal interpolation operators at the Radau nodes as introduced in Section 4.1.1 and $e_\tau = \xi - \xi_\tau$ the difference between the exact and the time discrete optimal triple.

Proof. For the time discretization error, we have the identity

$$J(\bar{q}, \bar{u}) - J^\tau(\bar{q}_\tau, \bar{u}_\tau) = \underbrace{\hat{\mathcal{L}}(\xi) - \hat{\mathcal{L}}(\xi_\tau)}_{\textbf{(I)}} + \underbrace{\hat{\mathcal{L}}(\xi_\tau) - \hat{\mathcal{L}}^\tau(\xi_\tau)}_{\textbf{(II)}}.$$

The term **(II)** represents the quadrature error for evaluating the Lagrangian and can be estimated numerically. For the first term, we proceed as in the proof of Proposition 2.1

in Becker and Rannacher [13] by rewriting the difference as an integral and applying the trapezoidal rule. Setting $e_\tau = \xi - \xi_\tau$, we get

$$\textbf{(I)} = \int_0^1 \hat{\mathcal{L}}'(\xi_\tau + se_\tau)(e_\tau)\,\mathrm{d}s = \underbrace{\frac{1}{2}\hat{\mathcal{L}}'(\xi_\tau)(e_\tau)}_{\textbf{(III)}} + \underbrace{\frac{1}{2}\hat{\mathcal{L}}'(\xi)(e_\tau)}_{\textbf{(IV)}} + \mathcal{R}_1^\tau$$

with the remainder term

$$\mathcal{R}_1^\tau = \frac{1}{2}\int_0^1 \hat{\mathcal{L}}'''(\xi_\tau + se_\tau)(e_\tau, e_\tau, e_\tau)s(s-1)\,\mathrm{d}s.$$

In the term **(III)**, the optimal triple ξ_τ does not match the Lagrangian since the latter is defined without numerical quadrature. For the state and adjoint residual, we proceed as in Meidner and Richter [77] by inserting the desired interpolants $\pi_k^{\mathrm{S}}\bar{u}$ and $\pi_k^{\mathrm{D}}\bar{z}$. The gradient residual is left unmodified yielding

$$\textbf{(III)} = \frac{1}{2}\left\{\rho_q(\xi_\tau)(\bar{q} - \bar{q}_\tau) + \rho_u(\xi_\tau)(\bar{z} - \pi_k^{\mathrm{D}}\bar{z}) + \rho_z(\xi_\tau)(\bar{u} - \pi_k^{\mathrm{S}}\bar{u})\right\} + \mathcal{R}_2^\tau$$

with the remainder term

$$\mathcal{R}_2^\tau = \frac{1}{2}\left\{\rho_u(\xi_\tau)(\pi_k^{\mathrm{D}}\bar{z} - \bar{z}_\tau) + \rho_z(\xi_\tau)(\pi_k^{\mathrm{S}}\bar{u} - \bar{u}_\tau)\right\}.$$

Since, as noted in Remark 6.5, the continuous state and adjoint satisfy the semidiscrete equations, the corresponding derivatives of the Lagrangian at ξ vanish and we have

$$\textbf{(IV)} = \frac{1}{2}\rho_q(\xi)(\bar{q} - \bar{q}_\tau)$$

Setting $\mathcal{R}^\tau = \mathcal{R}_1^\tau + \mathcal{R}_2^\tau$ and collecting all terms shows the desired error representation for the temporal discretization error. □

Remark 7.3. 1. Assuming at least first order convergence of the discrete scheme with quadrature, we expect that in many configurations, the remainder term $\mathcal{R}_2^\tau$ is of higher order than the quadrature error term **(II)**. For a rigorous discussion in the case of a fractional-step-θ scheme and a semilinear model problem, we refer to Meidner and Richter [77]. The remainder term $\mathcal{R}_1^\tau$ is the usual term encountered in DWR error indicators and is typically expected to be of higher order as well.
2. If no control constraints are present, the gradient condition is satisfied with equality. Hence, the two control residual terms vanish, and—apart from the modifications due to quadrature error—we get the usual error representation for a problem without inequality constraints as given in Meidner and Vexler [78].
3. On time intervals where both, $\bar{q}$ and $\bar{q}_\tau$ are inactive, the control residuals vanish. Where both controls are active over the whole discretization interval, they cancel each other out. Therefore, the control residuals can be considered as a measure for the error caused by mismatch between the active sets of the continuous and semidiscrete solutions.

Additionally, they account for errors due to the difference between the admissible sets Q_{ad} and $Q_{d,\mathrm{ad}}$.
4. Vexler and Wollner [109] derive their error representation from an extended Lagrangian that contains separate multipliers for the inequality constraints. However it can be verified that the resulting terms containing the residuals with respect to control and multiplier are equivalent to the sum of the control residuals at the exact and at the discrete solution, weighted by the difference of the solutions. Therefore our representation for the discretization error is actually very similar to theirs.

For the spatial error indicator, we have the following representation. The proof is virtually identical to the time discretization error.

Proposition 7.4. *Under Assumption 7.1, the spatial discretization error with respect to the cost functional can be represented as*

$$J(\bar{q}_\tau, \bar{u}_\tau) - J^\tau(\bar{q}_\sigma, \bar{u}_\sigma) = \frac{1}{2}\{\rho_u^\tau(\xi_\sigma)(\bar{z}_\tau - i_h\bar{z}_\tau) + \rho_z^\tau(\xi_\sigma)(\bar{u}_\tau - i_h\bar{u}_\tau) + \rho_q^\tau(\xi_\sigma)(\bar{q}_\tau - \bar{q}_\sigma) + \rho_q(\xi_\tau)(\bar{q}_\tau - \bar{q}_\sigma)\} + \hat{\mathcal{L}}^\tau(\xi_\sigma) - \hat{\mathcal{L}}^\sigma(\xi_\sigma) + \mathcal{R}^\sigma$$

with the remainder term

$$\mathcal{R}^\sigma = \frac{1}{2}\left\{\rho_u^\tau(\xi_\sigma)(i_h\bar{z}_\tau - \bar{z}_\sigma) + \rho_z^\tau(\xi_\sigma)(i_h\bar{u}_\tau - \bar{u}_\sigma) + \int_0^1 (\hat{\mathcal{L}}^\tau)'''(\xi_\sigma + se_\sigma)(e_\sigma, e_\sigma, e_\sigma)s(s-1)\,ds\right\}$$

where $i_h\colon X_k^r \to X_{k,h}^{r,s}$ is defined interval-wise as the standard nodal interpolation operator into the space $V_{h,m}$ applied pointwise in time and $e_\sigma = \xi_\tau - \xi_\sigma$ denotes the difference between the time discrete and the fully discrete solution triple. The residuals ρ_u^τ, ρ_z^τ, and ρ_q^τ are obtained from ρ_u, ρ_z, and ρ_q by replacing all temporal integrals with the corresponding quadrature rule.

Remark 7.5. Although the control discretization does not change when moving from the time discrete to the fully discrete problem, the two control residual terms do not vanish when inequality constraints are present. This is due to the fact that the space discretization of the state will in general affect the active and inactive sets for the control constraints.

7.1.2. Practical realization

In this section we show how the abstract error representations given in Propositions 7.2 and 7.4 can be used to construct computable a posteriori error indicators η_τ and η_σ, and how the error contributions can be localized.

Temporal estimator

We recall the representation for the time discretization error from Proposition 7.2

$$J(\bar{q},\bar{u}) - J^\tau(\bar{q}_\tau,\bar{u}_\tau) = \underbrace{\frac{1}{2}\rho_u(\xi_\tau)(\bar{z}-\pi_k^{\mathrm{D}}\bar{z}) + \frac{1}{2}\rho_z(\xi_\tau)(\bar{u}-\pi_k^{\mathrm{S}}\bar{u})}_{\textbf{(I)}} + \underbrace{\frac{1}{2}\rho_q(\xi_\tau)(\bar{q}-\bar{q}_\tau) + \frac{1}{2}\rho_q(\xi)(\bar{q}-\bar{q}_\tau)}_{\textbf{(II)}} + \underbrace{\hat{\mathcal{L}}(\xi_\tau) - \hat{\mathcal{L}}^\tau(\xi_\tau)}_{\textbf{(III)}} + \mathcal{R}^\tau. \tag{7.5}$$

This error representation still depends on several unknown quantities, which we have to approximate by suitable means. For this purpose, we discuss each of the terms **(I)** to **(III)** separately. The remainder term $\mathcal{R}^\tau$ is dropped for numerical evaluation.

For the state and adjoint residuals collected in **(I)**, the weights $\bar{z}-\pi_k^{\mathrm{D}}\bar{z}$ and $\bar{u}-\pi_k^{\mathrm{S}}\bar{u}$ have to be approximated. The superconvergence results shown in Chapter 6 serve as a motivation to replace $\bar{u}$ and $\bar{z}$ by the higher order reconstructions $\hat{\pi}_k^{\mathrm{S}}(u_0,\bar{u}_\tau)$ and $\hat{\pi}_k^{\mathrm{D}}(\bar{z}_{\tau,M}^+,\bar{z}_\tau)$ of the semidiscrete solutions where the terminal value $\bar{z}_{\tau,M}^+$ is given by the identity $(\bar{z}_{\tau,M}^+,\varphi) = J_2'(\bar{u}_{\tau,M}^-)(\varphi)$ for any $\varphi\in V$. We note that true superconvergence of the reconstruction is not required here, as long as the *local* interpolation error is captured well enough. Therefore, although we did not discuss whether a suitable generalization of the superconvergence result of Theorem 6.15 also holds for a variable order discretization and although regularity requirements might be violated for higher orders, it is still reasonable to expect the approximation of the weight via reconstruction to be sufficient.

If the order is $r=0$, the reconstruction operator $\hat{\pi}_k^{\mathrm{S}}$ interpolates the values of the solution at the end of each discretization interval with a piecewise linear function. This means it is identical to the usual reconstruction employed in this case (see, e. g., Meidner and Vexler [78]).

After approximation of the weights, the error representation still depends on the time-discrete solution ξ_τ. However, only the fully discrete solution ξ_σ is accessible numerically. Therefore, we replace all time-discrete quantities by their fully discrete counterparts. For convenience of notation, we write the resulting reconstructions as

$$\tilde{u} := \hat{\pi}_k^{\mathrm{S}}(\bar{u}_{\sigma,0}^-,\bar{u}_\sigma) \text{ and } \tilde{z} := \hat{\pi}_k^{\mathrm{D}}(\bar{z}_{\sigma,M}^+,\bar{z}_\sigma)$$

where $\bar{u}_{\sigma,0}^-$ is the L^2 projection of the initial value onto the space $V_{h,1}$ and $\bar{z}_{\sigma,M}^+$ denotes the terminal value for the optimal adjoint of the fully discrete problem. Taking note of the identities $\pi_k^{\mathrm{S}}\circ\hat{\pi}_k^{\mathrm{S}} = \mathrm{Id}$ and $\pi_k^{\mathrm{D}}\circ\hat{\pi}_k^{\mathrm{D}} = \mathrm{Id}$ for functions in X_k^r, we obtain the computable approximation

$$\textbf{(I)} \approx \frac{1}{2}\rho_u(\xi_\sigma)\,(\tilde{z}-\bar{z}_\sigma) + \frac{1}{2}\rho_z(\xi_\sigma)\,(\tilde{u}-\bar{u}_\sigma) =: \eta_\tau^1. \tag{7.6}$$

The evaluation of the state and adjoint residual terms can be simplified by a structural observation about the weight functions which is found as part of Lemma 2.2 in Makridakis

and Nochetto [71]. On each discretization interval I_m, the weight $\tilde{u} - \bar{u}_\sigma$ is a polynomial of degree $r_m + 1$ that vanishes at the $r_m + 1$ Radau points $\theta^S_{m,j}$ with $j = 0, \dots, r_m$ and takes the value $[\bar{u}_\sigma]_{m-1}$ at the left end point t_{m-1} of the interval. Let ℓ^S_m denote the Lagrange polynomial of degree $r_m + 1$ that is one at t_{m-1} and zero at all $\theta^S_{m,j}$. Then obviously, the weight satisfies

$$\tilde{u} - \bar{u}_\sigma\big|_{I_m} = -[\bar{u}_\sigma]_{m-1}\, \ell^S_m$$

for $m = 1, \dots, M$. Recalling the definition of the adjoint residual ρ_z, we see that the jump terms vanish since the left limit of the weight function at each temporal node is zero. Additionally we note that the term involving the time derivative of the adjoint is an intervalwise polynomial of degree $2r_m$. The $(r_m + 1)$-point Radau quadrature integrates such a polynomial exactly and since the weight function is zero at the quadrature nodes, also the time derivative terms vanish. Together, the adjoint residual has the simplified representation

$$\begin{aligned}(7.7)\quad &\rho_z(\xi_\sigma)\,(\tilde{u} - \bar{u}_\sigma)\\ &\qquad = \sum_{m=1}^{M} \left\{ J_1'(\bar{u}_\sigma)(-[\bar{u}_\sigma]_{m-1}\, \chi_{I_m}\, \ell^S_m) - \int_{I_m} a_u'(\bar{q}_\sigma, \bar{u}_\sigma)(\bar{z}_\sigma, -[\bar{u}_\sigma]_{m-1}\, \ell^S_m)\, \mathrm{d}t \right\}.\end{aligned}$$

In the same way, using the Lagrange polynomial $\ell^D_m \in \mathcal{P}_{r_m+1}(I_m)$ that is one at the right interval boundary and zero at the reversed Radau nodes $\theta^D_{m,j}$ with $j = 0, \dots, r_m$, we obtain the identity

$$\tilde{z} - \bar{z}_\sigma\big|_{I_m} = [\bar{z}_\sigma]_m\, \ell^D_m$$

for $m = 1, \dots, M$. The resulting representation for the state residual reads

$$(7.8)\qquad \rho_u(\xi_\sigma)\,(\tilde{z} - \bar{z}_\sigma) = -\sum_{m=1}^{M} \int_{I_m} a(\bar{q}_\sigma, \bar{u}_\sigma)([\bar{z}_\sigma]_m\, \ell^D_m)\, \mathrm{d}t.$$

To evaluate the control residuals collected in the term **(II)** of Equation (7.5), an approximation for the exact control $\bar{q}$ is required. Since the smoothness of $\bar{q}$ is limited by the control constraints, we cannot expect an interpolation of $\bar{q}_\tau$ into a space with higher polynomial order to yield an improved approximation. Therefore we follow the approach of Vexler and Wollner [109] and reconstruct an improved control $\tilde{q}$ by a post-processing step. For this purpose, we evaluate the projection condition (3.11c*) using again the higher order reconstructions for state and adjoint. Due to the true time-discrete solutions being inaccessible for numerical computation, we once again substitute the fully discrete values. The reconstructed control reads

$$\tilde{q} = P_{Q_{\mathrm{ad}}}\left(-\frac{1}{\alpha} G_{\mathrm{impl}}\,(\bar{q}_\sigma, \tilde{u}, \tilde{z}) \right).$$

Then the control residual for the time-discrete solution can be evaluated after replacing semidiscrete by discrete quantities. In place of the exact solution quantities in the second control residual term we use the approximation $\tilde{\xi} = (\tilde{q}, \tilde{u}, \tilde{z})$ such that we get

$$\textbf{(II)} \approx \frac{1}{2}\rho_q(\xi_\sigma)(\tilde{q} - \bar{q}_\sigma) + \frac{1}{2}\rho_q(\tilde{\xi})(\tilde{q} - \bar{q}_\sigma) =: \eta_\tau^2. \tag{7.9}$$

When evaluating those two terms in practice, it has to be taken into account that the control reconstruction $\tilde{q}$ can be non-smooth also in the interior of a discretization interval. Therefore care has to be taken to ensure sufficient accuracy when integrating the residuals. For a modest number d_Q of control parameters it is feasible to use an adaptive quadrature algorithm such as the one given by Gander and Gautschi [43] for this purpose. This ensures a sufficiently accurate resolution of possible kinks in the reconstructed control $\tilde{q}$.

Finally, in the third term in (7.5) stemming from the quadrature error, we replace the semidiscrete by the discrete solution yielding

$$\textbf{(III)} \approx \hat{\mathcal{L}}(\xi_\sigma) - \hat{\mathcal{L}}^\tau(\xi_\sigma) =: \eta_\tau^3. \tag{7.10}$$

Collecting the terms from Equations (7.6), (7.9), and (7.10), we get the temporal error indicator

$$\eta_\tau = \eta_\tau^1 + \eta_\tau^2 + \eta_\tau^3. \tag{7.11}$$

To make evaluation of this error estimator feasible in practice, we approximate the occurring exact integrals by some suitable quadrature procedure that is more accurate than the one used in the computation of the discrete solution. We opt for subdividing the discretization intervals into a number of microintervals and applying the $(r_m + 1)$-point Gauß rule on each of them.

Since the goal of an adaptive procedure is not only an accurate estimation of the overall discretization error but also the targeted refinement of those parts of the discretization that contribute most to the global error, localized error information for each time interval is needed besides the global error indicator η_τ. Localizing the contributions η_τ^1 and η_τ^3 is straightforward since both record interval-wise information: the weights in η_τ^1 are approximations of local interpolation errors, and quadrature errors occur independently on each discretization interval.

The control residual terms in η_τ^2 contain an approximation of the difference between exact and discrete control as weight. So an error measured on one interval can not necessarily be reduced by refining that particular interval because the weight is a global quantity. This is a difficulty encountered generally when applying a dual weighted residual type error estimator to a problem with control constraints. It also arises for the error estimators given by Wollner and Vexler. Therefore, although there is no rigorous justification, we localize also η_τ^2 by splitting the residuals into interval-wise contributions. In practice, we observed that η_2 was usually considerably smaller than η_τ^1 and that it decreased with approximately the same rate.

Spatial estimator

The weights in the representation for the spatial error given in Proposition 7.4 depend on the unknown semidiscrete solution variables $\bar{q}_\tau$, $\bar{u}_\tau$, and $\bar{z}_\tau$. To derive a computable error indicator, we proceed in the same way as for the temporal error estimator and approximate them by an improved reconstruction $\tilde{\xi}_\tau = (\tilde{q}_\tau, \tilde{u}_\tau, \tilde{z}_\tau)$ computed from the discrete solutions $\bar{q}_\sigma$, $\bar{u}_\sigma$, and $\bar{z}_\sigma$. In the case of bilinear elements ($s = 1$), we use the usual patch-wise biquadratic interpolation operator (see, e. g., Meidner and Vexler [78]) for the reconstruction of state and adjoint, which we denote $i_{2h}^{(2)} \colon X_{k,h}^{r,1} \to X_{k,2h}^{r,2}$. We set $\tilde{u}_\tau = i_{2h}^{(2)} \bar{u}_\sigma$ and $\tilde{z}_\tau = i_{2h}^{(2)} \bar{z}_\sigma$. A post-processing step gives a corresponding control reconstruction

$$\tilde{q}_\tau = P_{Q_{d,\mathrm{ad}}} \left(-\frac{1}{\alpha} G_{\mathrm{impl}} \left(\bar{q}_\sigma, \tilde{u}_\tau, \tilde{z}_\tau \right) \right).$$

With the identity $i_h \circ i_{2h}^{(2)} = \mathrm{Id}$, the resulting error indicator reads

(7.12)
$$\eta_\sigma = \frac{1}{2} \left\{ \rho_u(\xi_\sigma)(\tilde{z}_\tau - \bar{z}_\sigma) + \rho_z(\xi_\sigma)(\tilde{u}_\tau - \bar{u}_\sigma) + \rho_q(\xi_\sigma)(\tilde{q}_\tau - \bar{q}_\sigma) + \rho_q(\tilde{\xi}_\tau)(\tilde{q}_\tau - \bar{q}_\sigma) \right\} + \hat{\mathcal{L}}^\tau(\xi_\sigma) - \hat{\mathcal{L}}^\sigma(\xi_\sigma)$$

Same as for the time discretization error, exact spatial integrals have to be approximated by some more accurate quadrature formula. For biquadratic elements ($s = 2$), we use an interpolation operator $i_{2h}^{(4)} \colon X_{k,h}^{r,2} \to X_{k,2h}^{r,4}$ into the space of patch-wise bi-quartic elements instead of the operator $i_{2h}^{(2)}$. Subsequently, we will not consider spatial discretizations with even higher order.

Concerning the localization of the spatial error indicators, we follow the approach of Braack and Ern [18] for treating the state and adjoint residual terms, which allows to filter out oscillatory behaviour. For a detailed description in the setting of parabolic optimal control problems, we refer to Meidner [76]. The quadrature error terms are already defined in terms of local contributions. More problematic is the question of a suitable localization for the control terms. For time dependent parameter control they do not contain information on the spatial distribution of the origin of the error. Therefore we propose to localize them only in time and distribute the resulting error terms uniformly to all cells of the spatial grid corresponding to each time step.

7.2. Smoothness indicator based on continuous Sobolev embeddings

To decide whether to perform h or p refinement on a given time interval marked for refinement, we use a heuristic proposed by Wihler [116] to estimate the smoothness

properties of the solution. It is based on the observation that for the one-dimensional embedding

$$H^1(I) \hookrightarrow L^\infty(I),$$

the embedding constant given by

$$\sup_{v \in H^1(I), v \neq 0} \frac{\|v\|_{L^\infty(I)}}{\|v\|_{H^1(I)}}$$

is attained for a very smooth v whereas if v approaches a non-smooth function, i. e., a function in L^∞ which is not in H^1, the quotient of the norms approaches zero. This observation is exploited by monitoring this quotient for appropriate derivatives of the discrete approximation. If the quotient grows too small, we assume that the function being approximated by the discretization is not contained in the considered Sobolev space.

Here, we will adopt an improved formulation of this smoothness indicator given by Fankhauser, Wihler, and Wirz in [39]. For each discretization interval I_m, we consider the functional $\mathcal{S}_m \colon H^1(I_m, H) \to \mathbb{R}$ given by

$$\mathcal{S}_m(v) = \begin{cases} \|v\|_{L^\infty(I_m,H)} \left(k_m^{-\frac{1}{2}} \|v\|_{I_m} + \sqrt{\frac{k_m}{2}} \|\partial_t v\|_{I_m} \right)^{-1}, & \text{if } v \neq 0, \\ 1, & \text{if } v = 0. \end{cases} \tag{7.13}$$

Proposition 7.6. *The smoothness indicators $\mathcal{S}_m$ for $m = 1, \dots, M$ can be bounded by*

$$0 \leq \mathcal{S}_m(v) \leq 1$$

for any $v \in H^1(I_m, H)$. If v is constant in time, they take the value 1.

Proof. Both parts of the claim are shown for scalar functions in the proof of Proposition 1 in [39]. However, the proofs given there apply without changes to H-valued functions. □

From Theorem 6.13 we know that a dG(r) discretization converges with optimal order $r+1$ with respect to the $L^2(I, H)$ norm if the solution is in $H^{r+1}(I, H)$. Therefore when deciding for a given interval I_m whether to increase the order r_m in the next iteration to r_m+1, it would be ideal to check whether the exact solution is in $H^{r_m+2}(I_m, H)$. However, since the discrete solution is a polynomial of degree r_m on I_m, all time derivatives of order greater than r_m vanish. Therefore the smoothness indicator $\mathcal{S}_m$ yields no useful information when applied to the r_m^{th} or any higher time derivative of the discrete solution.

Wihler [116] proposes to base p-refinement decisions on the value of $\mathcal{S}_m(\partial_t^{r_m-1} u_k)$ where u_k is the discrete solution. While this is possible for any piecewise polynomial discretization of order one or greater, specifically for dG time discretization, there is a further possibility. Since the reconstructions $\hat{\pi}_k^{\mathrm{S}}$ and $\hat{\pi}_k^{\mathrm{D}}$ offer a—typically even more accurate—approximation of the exact solution which has one polynomial order more

than the discrete solution, we can also evaluate the smoothness indicators for the r_m^{th} time derivative of the reconstructed solution. An added benefit is that smoothness information can also be obtained on intervals with order $r_m = 0$.

For our optimal control problem, we propose to monitor the smoothness of state and adjoint variable. We will not consider a smoothness indicator for the control variable since there is no meaningful higher order reconstruction for the control available in the presence of constraints. However, limited temporal regularity of the control will typically also limit the temporal regularity of the state solution, therefore, to a certain extent, it should be captured by the smoothness indicator for the state solution.

This leads to the following simple refinement strategy: on each interval I_m that is marked for refinement, we compute the values

$$\mathcal{S}_m^S = \mathcal{S}_m\left(\partial_t^{r_m}\left[\hat{\pi}_k^{\mathrm{S}} u_\sigma\big|_{I_m}\right]\right) \text{ and } \mathcal{S}_m^D = \mathcal{S}_m\left(\partial_t^{r_m}\left[\hat{\pi}_k^{\mathrm{D}} z_\sigma\big|_{I_m}\right]\right).$$

These values are compared to a threshold value $\hat{\tau}$, which is fixed a priori. If one of the smoothness indicators is less than $\hat{\tau}$, the time interval is refined by bisection; if both indicators exceed the threshold, indicating a sufficiently smooth local solution, the order r_m is increased instead.

Concerning the choice of the threshold value, we note that the r_m^{th} time derivative of the higher order reconstruction on the interval I_m is a linear polynomial with respect to time. In [39], it was shown that for scalar linear polynomials, the smoothness indicator is bounded from below by $\frac{\sqrt{3}}{\sqrt{6}+1} \approx 0.502$. Therefore, a sensible value for $\hat{\tau}$ should be larger than this bound.

7.3. Adaptive algorithm

The adaptive algorithm follows the usual pattern with a main loop consisting of the four steps

$$\textbf{solve} \rightarrow \textbf{estimate} \rightarrow \textbf{mark} \rightarrow \textbf{refine}.$$

We give a brief description of each of them.

The step **solve** involves solving the optimal control problem on the current discretization with the proposed semismooth Newton trust region algorithm, see Algorithm 3.1. As a result, we get the optimal triple $(\bar{q}_\sigma, \bar{u}_\sigma, \bar{z}_\sigma)$. Since the reconstruction of the control used in the a posteriori error indicators relies on the assumption that the computed solution satisfies the discrete optimality system, an accurate error estimation can only be expected if the optimization algorithm has converged to a reasonably low tolerance.

Within the **estimate** operation, the error indicators η_τ and η_σ are evaluated as described in Section 7.1. Additionally, the corresponding localized error contributions have to be stored. For the temporal error estimator this results in one local error indicator η_τ^m for

Algorithm 7.1 Adaptive algorithm with *hp* adaptivity in time and *h* adaptivity in space.

chose an initial temporal mesh given by the node vector k and the order vector r and spatial meshes $\mathcal{T}_m$ for each time step
loop
    solve the discrete optimal control problem (7.3) with Algorithm 3.1
    compute the error indicators η_τ and η_σ by (7.11) and (7.12) and store local contributions
    if desired accuracy reached **then**
        break
    end if
    if $c_e\,|\eta_\sigma| > |\eta_\tau|$ **then**
        determine spatial cells to be refined by Dörfler marking
        compute new spatial grids by refining the marked cells
    end if
    if $c_e\,|\eta_\tau| > |\eta_\sigma|$ **then**
        determine time intervals to be refined by Dörfler marking
        for each marked interval I_m **do**
            evaluate $\mathcal{S}_m^S$ and $\mathcal{S}_m^D$
            if $\min(\mathcal{S}_m^S, \mathcal{S}_m^D) \geq \hat{\tau}$ **then**
                set discretization order on I_m to $r_m + 1$
            else
                refine I_m by inserting a new temporal node
            end if
        end for
    end if
end loop

each time interval I_m and for the spatial estimator, we get localized indicators $\eta_\sigma^{m,n}$ with $n = 1, \ldots, N_m$ and $m = 1, \ldots, M$ for each spatial node on each time interval.

Based on the localized error indicators, the **mark** step decides which time discretization intervals and which spatial grid cells require refinement. For an efficient space-time discretization, it is desirable to balance the spatial and temporal discretization errors. This can be done by an equilibration strategy as discussed for example in Meidner and Vexler [78]. We fix an equilibration constant $c_e > 1$ of moderate size. The spatial discretization is considered for refinement whenever $c_e\,|\eta_\sigma| > |\eta_\tau|$ and conversely the time discretization is refined when $c_e\,|\eta_\tau| > |\eta_\sigma|$. Consequently, if both errors differ by a factor larger than c_e, we do not mark any cells of the discretization with the smaller error for refinement.

For the marking itself, many different criteria have been proposed. Since the models involved in mesh optimization approaches like the one described in Richter [90] become rather complex when working on *hp* grids, we opt for a simple marking strategy which is

a variant of the approach originally proposed by Dörfler [30]. We apply it separately to the space and time discretizations. Given a set of local error indicators $\{\eta^j \mid j \in \mathcal{M}\}$ with some index set $\mathcal{M}$, a minimal subset $\mathcal{R} \subseteq \mathcal{M}$ is determined such that

$$\sum_{j \in \mathcal{R}} \left|\eta^j\right| \geq \theta_D \sum_{j \in \mathcal{M}} \left|\eta^j\right|$$

with a chosen parameter $\theta_D \in (0, 1)$. In practice this is accomplished by sorting the error indicators by absolute value. Subsequently, the discretization cells corresponding to the index set $\mathcal{R}$ are marked for refinement. The parameter θ_D controls how aggressively the mesh is being refined. While $\theta_D = 1$ would correspond to global refinement, smaller values cause less cells to be refined in each iteration. Whereas a careful refinement tends to lead to a more optimized final mesh, this has to be traded off against the cost of the higher number of mesh adaption iterations required to reach a prescribed error tolerance. For practical purposes, we have found values around $\theta_D = 0.5$ to yield a reasonable balance.

After having marked subsets of spatial cells and time discretization intervals, the final step **refine** performs the actual refinement. For the spatial discretization, which in our case has fixed polynomial order s, we can rely on well-tested strategies, which also ensure that the new meshes still satisfy the mesh regularity assumptions made in Section 4.1.2 with regard to hanging nodes and patch structure. This is accomplished by refining further cells where necessary.

When refining the marked time intervals, a decision has to be made for each interval whether to perform h or p refinement. For this purpose, the smoothness indicators $\mathcal{S}_m^S$ and $\mathcal{S}_m^D$ are evaluated. If $\min(\mathcal{S}_m^S, \mathcal{S}_m^D) \geq \hat{\tau}$, then the discretization order r_m on the current interval is increased by one, otherwise the interval is split into two subintervals and each of them is assigned the order r_m.

An outline of the complete adaptive algorithm is given in Algorithm 7.1.

7.4. Numerical tests

Subsequently we present test results for the adaptive algorithm on the three test examples introduced in Section 2.3. For the linear quadratic model problem and the semilinear problem, we increase the difficulty through the choice of the problem data.

7.4.1. Linear quadratic model problem

In this section we revisit the linear quadratic model problem introduced in Section 2.3.1. To test how well the adaptive solver handles rapidly changing temporal dynamics and non-smoothness in time, we select on the one hand a small value for the regularization parameter α, leading to an optimal control which approaches bang-bang structure on

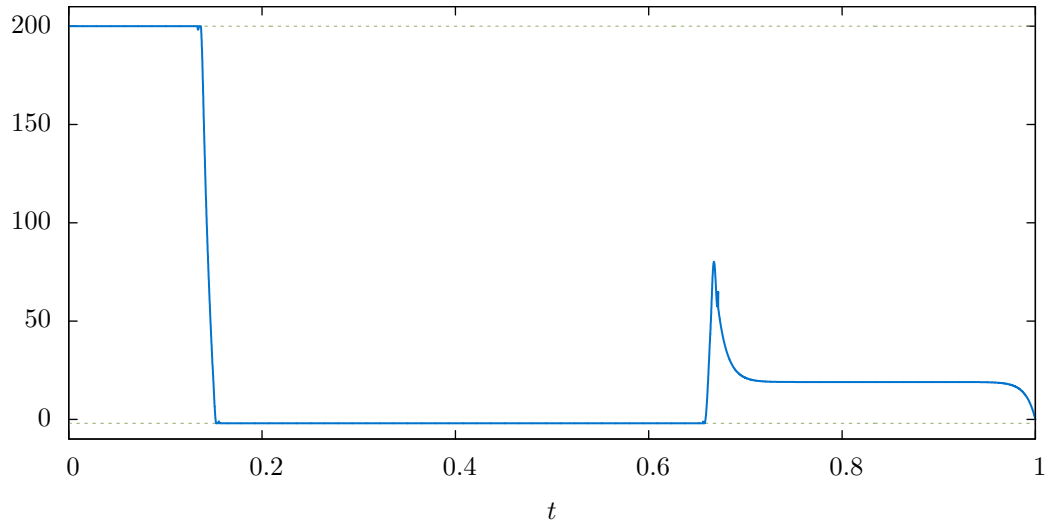


Figure 7.1.: Optimal control $\bar{q}$ for the model problem

part of the time interval. On the other hand, we chose a desired state with a discontinuity with respect to time. Additionally, the initial state is specified far from the desired state such that a large control activity at the beginning of the discretization interval needs to be resolved.

For the time interval and the spatial domain, we consider $I = (0, 1)$ and $\Omega = (0, 1)^2$. As in Section 6.5, the problem data are specified in terms of the first eigenfunction of the Laplacian given by

$$w_1(x) = 2\sin(\pi x_1)\sin(\pi x_2). \tag{7.14}$$

and we consider a one-dimensional time parameter control (i. e., $d_Q = 1$) with the control operator defined by $G^q(q)(t) = q(t)w_1$. The remaining problem data are given by

$$f = 0, \quad u_0 = -200w_1, \quad u_d(t, x) = \begin{cases} -w_1(x), & \text{where } t < \frac{2}{3}, \\ w_1(x), & \text{where } t \geq \frac{2}{3} \end{cases},$$

$q^a = -2$, $q^b = 200$, and $\alpha = 10^{-4}$. The resulting optimal control $\bar{q}$ can be seen in Figure 7.1. We note that there is very large control activity up to around $t = 0.14$ in order to steer the state from the initial value closer to the desired state. Up to about $t = \frac{2}{3}$ the control approaches a bang bang structure, and after that, it transitions into a free arc with an additional discontinuity in the first derivative caused by the jump in the desired state. From a computation on a fine uniform discretization with 1024 dG(4) steps in time and biquadratic finite elements in space, we obtain a reference value $\bar{J}$ for the cost functional at the optimum $(\bar{q}, \bar{u})$.

To investigate whether the proposed smoothness indicators are suited to detect the local temporal irregularities of the optimal solution $(\bar{q}, \bar{u}, \bar{z})$, we solve the problem for various

Table 7.1.: Effectivity index I_{eff}^{τ} for uniform time discretization with M time steps and fixed order r on a fixed spatial discretization

M \ r	0	1	2	3	4	5	6	7
2	12.66	4.96	4.02	4.81	-1.74	1.55	0.81	-0.04
4	4.27	3.40	4.06	2.09	-0.90	-0.65	0.06	0.53
8	2.37	2.08	1.62	1.22	2.49	0.37	-1.08	0.89
16	1.65	1.50	1.35	0.39	0.54	-0.70	0.46	1.04
32	1.32	1.23	1.28	-1.28	0.67	0.52	1.48	0.75
64	1.16	1.11	-0.16	0.71	0.91	0.89	0.57	0.47
128	1.08	1.04	0.90	1.74	1.19	0.91	0.75	0.72
256	1.04	1.02	1.32	1.18	1.23	1.06	0.96	0.63

orders of the time discretization on an equidistant time grid with $M = 32$ intervals, and a fixed uniform spatial discretization with biquadratic elements and $N = 1089$ degrees of freedom.

In Figures 7.2(a) to 7.2(d), we plot the resulting values of the smoothness indicators $\mathcal{S}_m^S$ and $\mathcal{S}_m^D$ exemplary for orders $r = 0, 2, 4, 6$. For reference, we also include the corresponding discrete optimal control $\bar{q}_\sigma$. Clearly, for all displayed orders, the smoothness indicators identify the transitions between active and inactive sets as regions with reduced regularity of the state. For the adjoint state, the lack of regularity at the jump of the desired state is indicated. The results for odd time discretization orders are similar.

To assess the quality of the *hp* adaptive time refinement independent from the spatial discretization at first, we consider once again a fixed biquadratic spatial discretization, this time with $N = 289$ nodes. A reference value $\bar{J}_h$ of the functional for this space discretization is obtained by solving with constant order 4 on a fine uniform time grid with $M = 8192$ time steps. To assess the quality of the temporal error estimator, we define the effectivity index

$$I_{\text{eff}}^{\tau} = \frac{\bar{J}_h - J^{\sigma}(\bar{q}_\sigma, \bar{u}_\sigma)}{\eta_\tau}.$$

In Table 7.1, the values of this effectivity index for uniform temporal grids and fixed order r are listed. For orders 0 and 1 we observe relatively accurate error quantification as soon as the temporal grid is fine enough to resolve the basic features of the solution. For the higher orders however, we see some outliers which are presumably related to the fact that higher order polynomials can fail to approximate the non-smooth features of the solution even qualitatively.

For the *hp*-adaptive algorithm, we set the parameter controlling the *hp*-refinement strategy to $\hat{\tau} = 0.6$ and start the computation with a dG(1) discretization consisting of 4 intervals. In Table 7.2 , we list the resulting number of time steps M, the total number of temporal degrees of freedom $M_{\text{tot}} = \sum_{m=1}^{M}(r_m + 1)$, and the effectivity index for each iteration. Despite the low number of temporal degrees of freedom and the varying

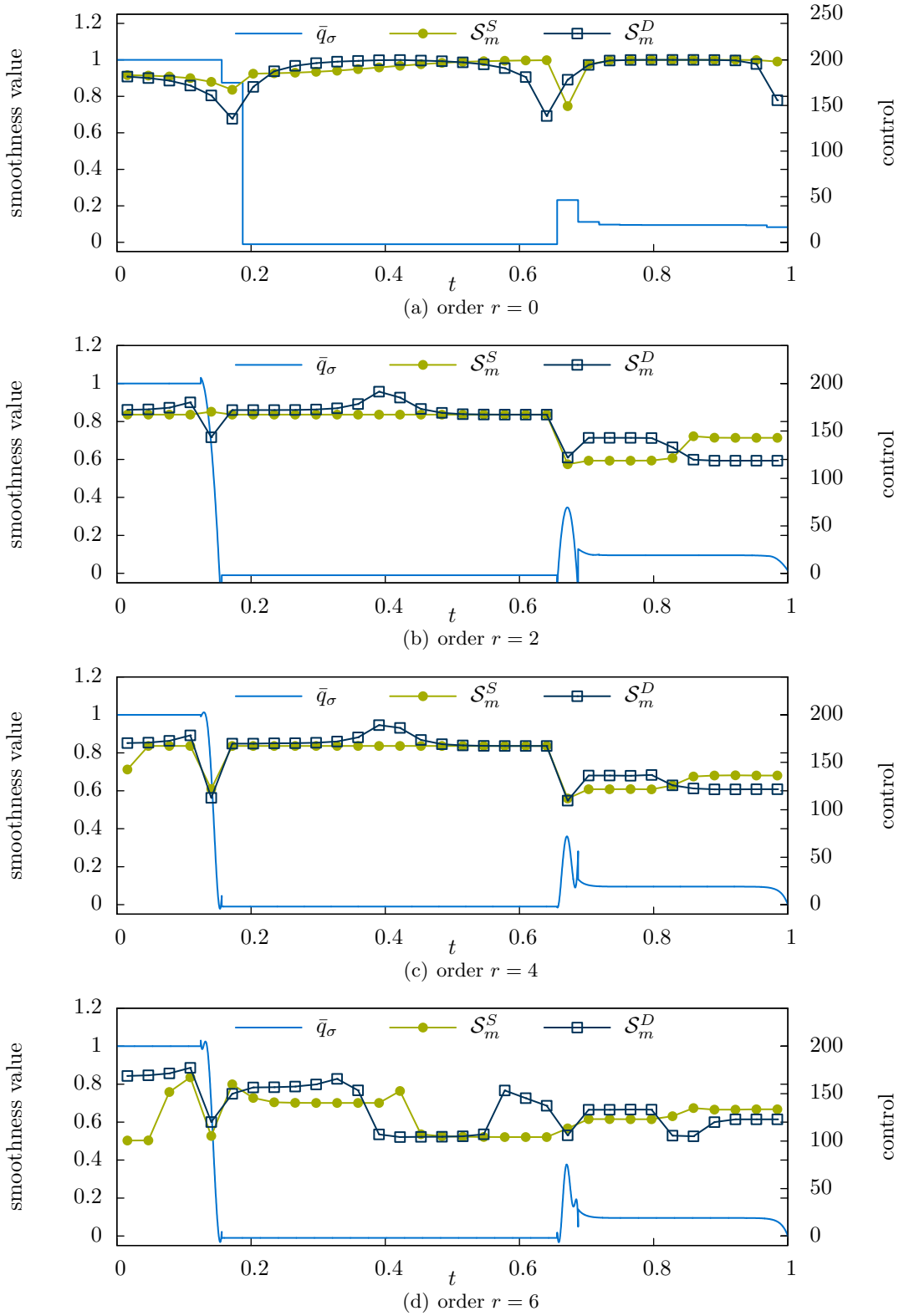
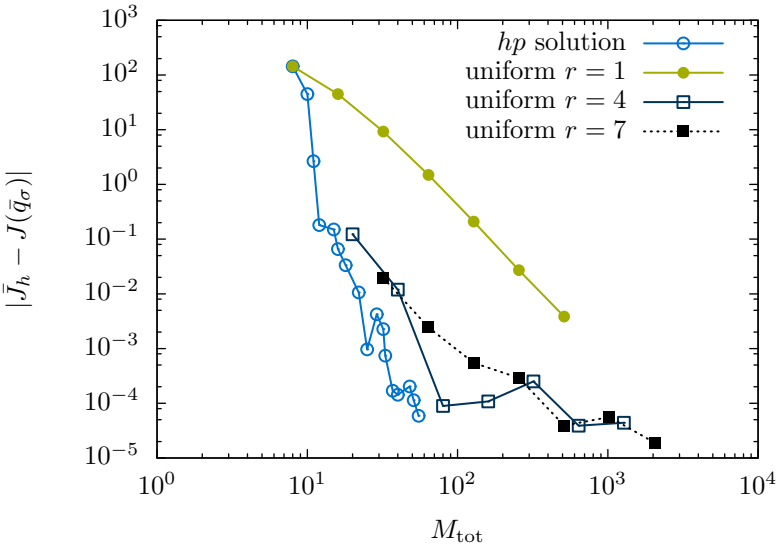


Figure 7.2.: Values of the smoothness indicators $\mathcal{S}_m^S$ and $\mathcal{S}_m^D$ for discrete solutions on $M = 32$ time intervals and $N = 1089$ spatial degrees of freedom



M	M_{tot}	I^τ_{eff}
4	8	3.40
5	10	2.08
5	11	1.59
5	12	1.39
6	15	1.82
6	16	2.55
7	18	2.34
8	22	0.87
9	25	0.43
10	29	1.05
11	32	1.18
11	33	0.96

Figure 7.3 & Table 7.2: Convergence plot and effectivity indices for *hp*-adaptive time discretization of the linear model problem on fixed spatial grid

order of discretization, the accuracy of the error estimation appears to be at least as good as for the uniform discretizations. A convergence plot, comparing the convergence speed with respect to the cost functional to uniform time discretizations of orders 1, 4, and 7, is given in Figure 7.3. We observe that the adaptive procedure requires a considerably lower number of degrees of freedom to approximate the cost functional to a given accuracy than any of the uniform discretizations.

To test the error equilibration between time and space discretization, we run the full Algorithm 7.1 on the model problem. The initial discretization consists of 4 time intervals with dG(1) and a biquadratic spatial discretization with $N = 81$ nodes. Due to the simple structure of the solutions in space, we keep the spatial mesh fixed over the whole time domain. The equilibration factor c_e is set to 5. From the results reported in Table 7.3, we see that the equilibration leads to fast convergence of the overall error. Besides, it can be observed that the value of the spatial error indicator is practically independent from the temporal discretization. The effectivity index for the full discretization is given by

$$I_{\text{eff}} = \frac{\bar{J} - J^\sigma(\bar{q}_\sigma, \bar{u}_\sigma)}{\eta}.$$

The temporal mesh produced in the final iteration of the *hp* adaptive procedure is visualized in Figure 7.4. For reference, we also visualize the resulting discrete control. Whereas the large control activity at the beginning of the discretization interval is resolved by *p*-refinement, the two critical areas are resolved by *h*-refinement at order 2.

Table 7.3.: Space-time adaptivity with error equilibration and variable temporal order for the linear model problem

M	M_{tot}	N	η_τ	η_σ	$\bar{J} - J^\sigma(\bar{q}, \bar{u})$	I_{eff}
4	8	81	$4.234 \cdot 10^{1}$	$2.347 \cdot 10^{-1}$	$1.444 \cdot 10^{2}$	3.39
5	10	81	$2.155 \cdot 10^{1}$	$2.481 \cdot 10^{-1}$	$4.500 \cdot 10^{1}$	2.06
5	11	81	$1.675 \cdot 10^{0}$	$2.291 \cdot 10^{-1}$	$2.888 \cdot 10^{0}$	1.52
5	12	81	$1.304 \cdot 10^{-1}$	$2.258 \cdot 10^{-1}$	$3.991 \cdot 10^{-1}$	1.12
6	15	289	$8.234 \cdot 10^{-2}$	$1.386 \cdot 10^{-2}$	$1.638 \cdot 10^{-1}$	1.70
6	16	289	$2.583 \cdot 10^{-2}$	$1.384 \cdot 10^{-2}$	$7.950 \cdot 10^{-2}$	2.00
7	18	1089	$1.427 \cdot 10^{-2}$	$8.636 \cdot 10^{-4}$	$3.428 \cdot 10^{-2}$	2.26
8	22	1089	$1.217 \cdot 10^{-2}$	$8.638 \cdot 10^{-4}$	$1.145 \cdot 10^{-2}$	0.88
9	25	1089	$2.236 \cdot 10^{-3}$	$8.638 \cdot 10^{-4}$	$1.829 \cdot 10^{-3}$	0.59
10	29	1089	$4.039 \cdot 10^{-3}$	$8.636 \cdot 10^{-4}$	$5.092 \cdot 10^{-3}$	1.04
11	32	4049	$1.924 \cdot 10^{-3}$	$6.279 \cdot 10^{-5}$	$2.331 \cdot 10^{-3}$	1.17
11	33	4049	$7.769 \cdot 10^{-4}$	$6.279 \cdot 10^{-5}$	$8.112 \cdot 10^{-4}$	0.97
12	37	4049	$3.231 \cdot 10^{-4}$	$6.279 \cdot 10^{-5}$	$2.370 \cdot 10^{-4}$	0.61

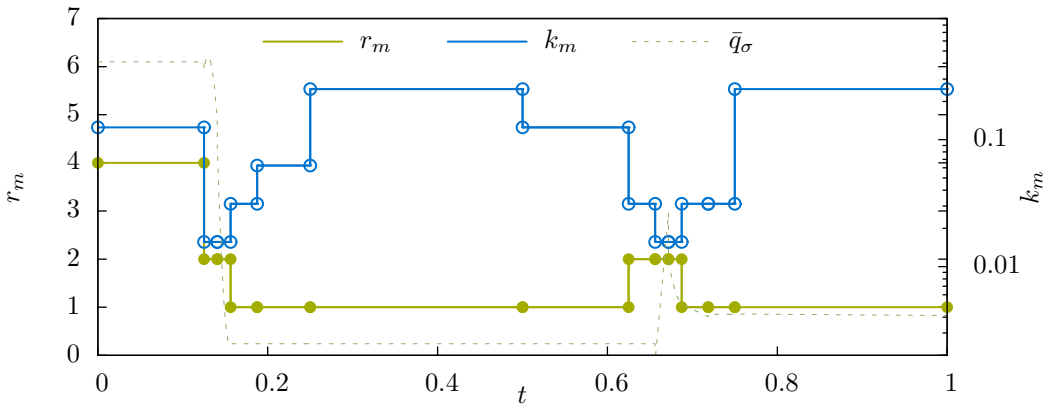


Figure 7.4.: Temporal mesh produced by Algorithm 7.1 for the linear quadratic model problem

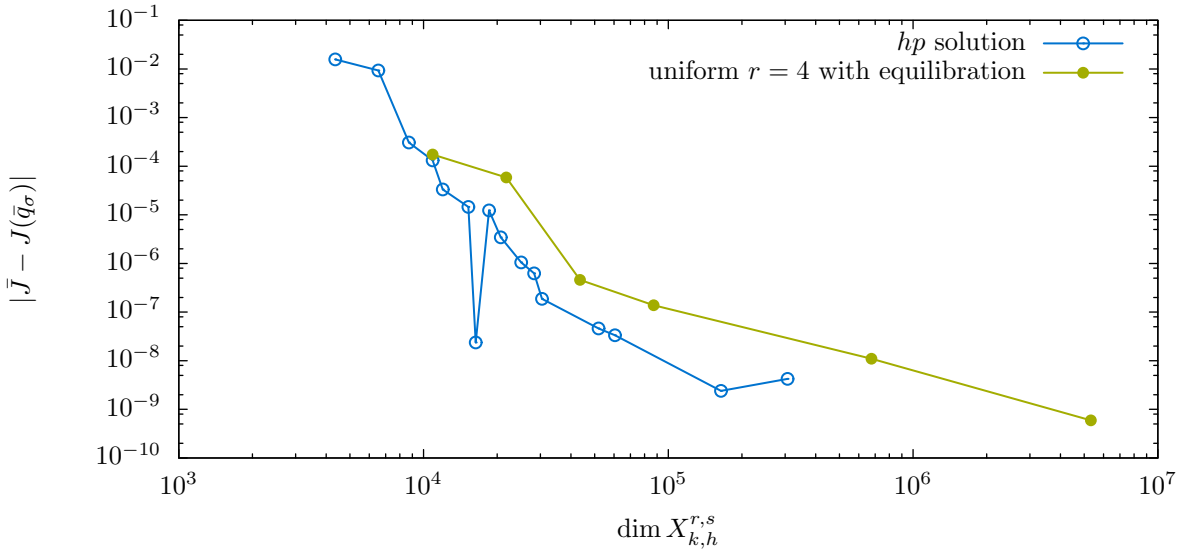


Figure 7.5.: Convergence of the functional value for *hp*-adaptive discretization of the semilinear problem

7.4.2. Semilinear problem with incompatible terminal observation

We consider the semilinear problem given in Section 2.3.2 on the unit square $\Omega = (0,1)^2$ and the unit time interval $(0,1)$. As problem data, we specify $u_0 = 0$, $f = 0$, $u_d = 1$, $q^a = -1$, $q^b = 1$, and $\alpha = 10^{-2}$. The control consists of two time-dependent functions (i. e., $d_Q = 2$) and the control-to-right-hand-side operator G^q is given by $G^q(q_1, q_2) = q_1 w_1(x) + q_2$, where w_1 again denotes the first eigenfunction of the Laplacian as defined in (7.14).

The particular challenge in the problem setup lies in the choice of the desired state $u_d = 1$ which is not contained in the space V. This results in an incompatible terminal condition—and therefore a startup singularity—for the adjoint equation. For the space discretization of the problem, we use $\mathcal{Q}_2$ elements and a reference solution is obtained by a computation with dG(4) on a fine temporal and spatial mesh. Since we do not expect the solution to have features that travel through the spatial domain, we do not consider moving meshes, but rather use the same spatial discretization for the whole time domain.

A convergence plot for the full *hp* adaptive algorithm is given in Figure 7.5 For reference, we included the results of a uniform discretization in space and time with error equilibration and dG(4) elements. We remark that in the last iterate of the adaptive computation, the time discretization uses as little as 12 time intervals with 41 temporal degrees of freedom. The corresponding *hp* mesh can be seen in Figure 7.4. Strong h refinement can be seen at the end of the time interval, where the startup singularity of the adjoint has to be resolved. Moving away from the singularity, gradually both

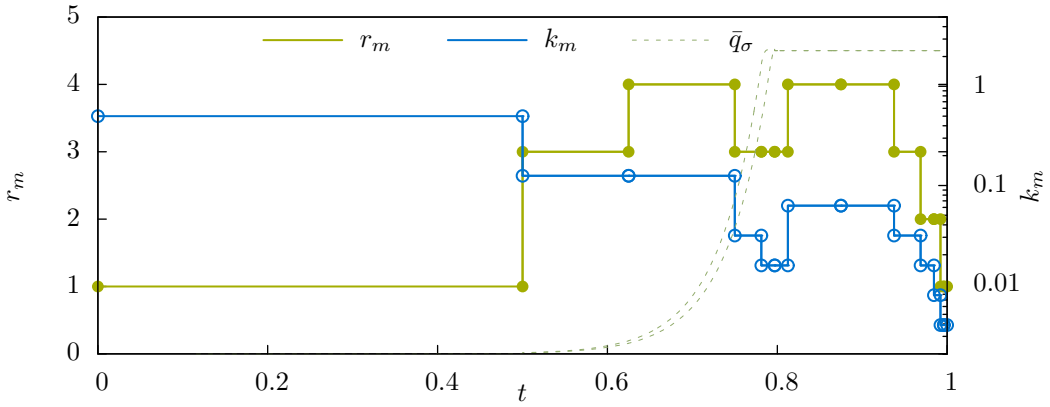


Figure 7.6.: Temporal mesh produced by Algorithm 7.1 for the semilinear problem

Table 7.4.: Space-time adaptivity for the combustion problem with moving meshes and variable temporal order

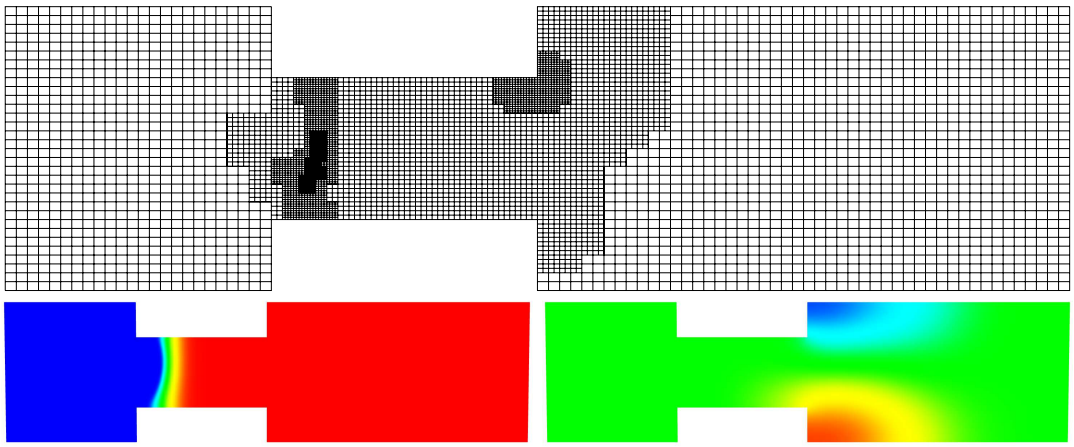
M	M_{tot}	$N_{\max}$	$\dim X_{k,h}^{r,s}$	η_τ	η_σ	$\bar{J} - J^\sigma(\bar{q}, \bar{u})$	I_{eff}
140	280	2833	793240	$1.708 \cdot 10^{-2}$	$-8.793 \cdot 10^{-2}$	$2.730 \cdot 10^{-1}$	-3.85
163	335	3599	1052583	$-3.223 \cdot 10^{-4}$	$1.248 \cdot 10^{-1}$	$1.978 \cdot 10^{-1}$	1.59
163	335	5575	1344053	$-2.506 \cdot 10^{-3}$	$7.024 \cdot 10^{-2}$	$8.319 \cdot 10^{-2}$	1.23
164	337	10059	2020319	$-2.778 \cdot 10^{-3}$	$2.673 \cdot 10^{-2}$	$3.088 \cdot 10^{-2}$	1.29
167	374	16917	3736426	$-5.279 \cdot 10^{-4}$	$1.378 \cdot 10^{-2}$	$8.711 \cdot 10^{-3}$	0.66
171	382	25855	5554526	$-1.116 \cdot 10^{-4}$	$9.732 \cdot 10^{-3}$	$6.561 \cdot 10^{-3}$	0.68

the time step size and the order are increased up to the irregularity due to the control becoming inactive, which is resolved with h refinement at order 3.

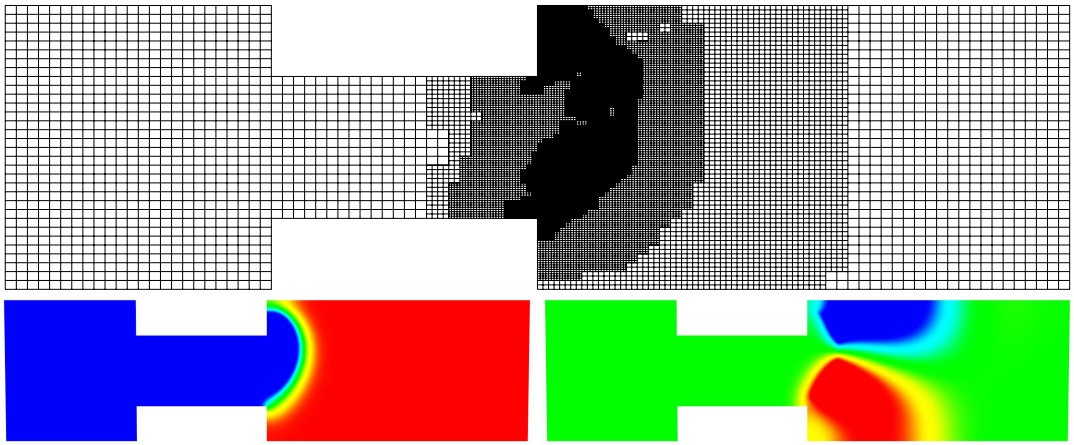
7.4.3. Combustion problem

In this section we revisit the combustion control problem introduced in Section 2.3.3.

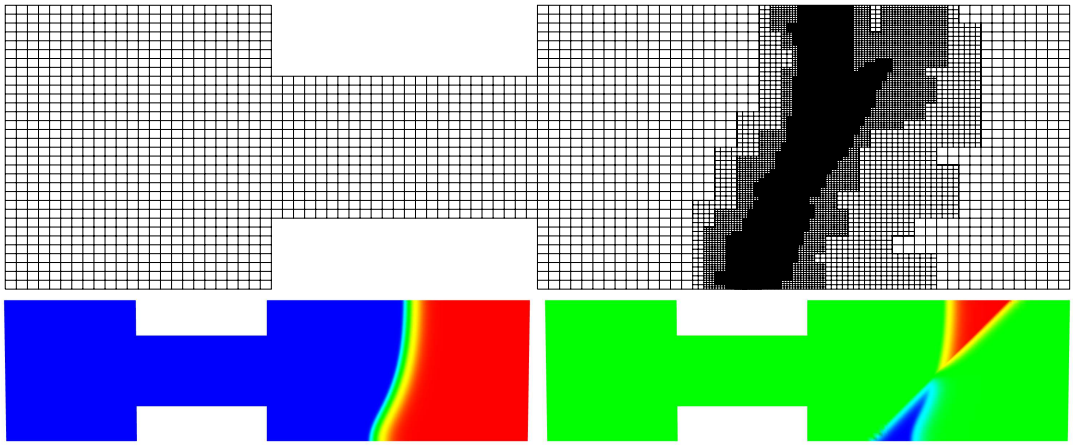
To test adaptivity in space and time for the combustion problem, we start with an initial discretization consisting of $M = 80$ equidistant dG(1) time steps and bilinear finite elements in space on a uniform grid consisting of $N = 2833$ nodes. Presumably due to reentrant corners in the domain, in our experiments, the increased accuracy of $\mathcal{Q}_2$ elements did not justify their higher computational cost. For an accurate solution of the problem, the travelling flame front has to be resolved in space. Hence, for an efficient solution, it is crucial to allow the spatial mesh to vary over time. In Figures 7.7(a) to 7.7(c), we plot the spatial grids, the fluid concentration Y, and the associated component of the adjoint state for a few selected time steps to give an impression of the solution and the effects of adaptivity in space. It can be seen that the spatial meshes resolve the features of both solutions comparatively well.



(a) $t = 10$



(b) $t = 27$



(c) $t = T = 40$

Figure 7.7.: Adapted spatial mesh, concentration component Y of the state solution and corresponding component of the adjoint for selected time steps

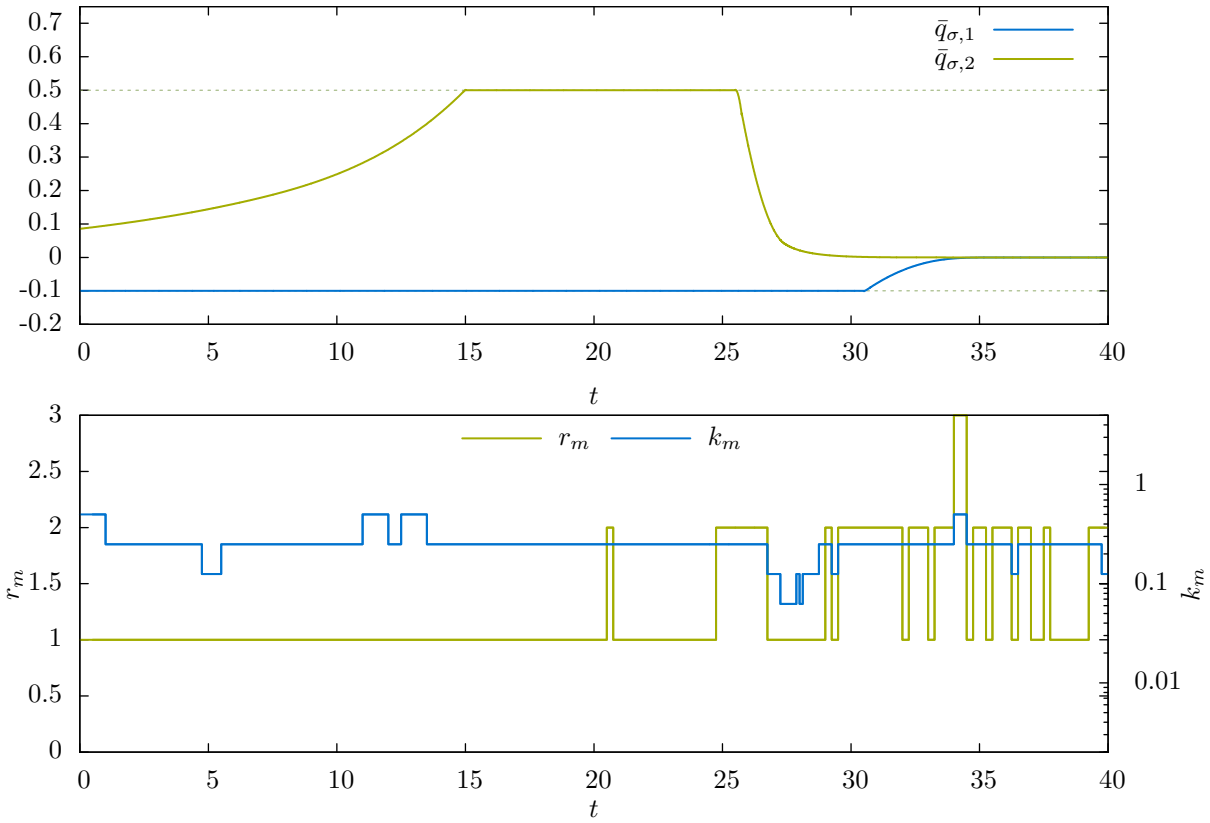


Figure 7.8.: Control and temporal grid for the last iteration of the adaptive algorithm for the combustion problem

In Table 7.4, the results produced by the adaptive algorithm are listed. The reference value $\bar{J} \approx 18.66172$ for the cost functional was obtained from additional iterations of the adaptive algorithm. This is due to the fact that a computation on a uniformly refined discretization with $\dim X_{k,h}^{r,s} = 414\,259\,800$, which took more than 8 CPU months to complete, was only able to confirm the first three significant digits.

The increase in the number of temporal grid cells observed in Table 7.4 is for the most part not caused by adaptive refinement but rather results from convergence failures of the time stepping solver. We also see that the temporal error seems to decrease considerably faster than the spatial error. This highlights on the one hand the superior accuracy of higher order dG methods over more conventional schemes like implicit Euler. On the other hand it shows that for solving this type of problem with high accuracy requirements, the *hp* time discretization should be complemented by a space discretization of comparable accuracy. Since an appropriate discretization has to resolve features like reentrant corners in the domain and steep gradients around the flame front, we would expect *hp* adaptivity on the space domain to be the most promising approach for this purpose.

In Figure 7.8, we depict the temporal mesh used during the last iteration of the adaptive algorithm, along with the corresponding discrete optimal control. Looking at the distribution of the few temporal p-refinements that take place, we see that as we would expect for terminal observation, mainly the end of the time interval needs to be resolved with higher accuracy. As already pointed out, practically all h-refinements visible where not triggered by the adaptive algorithm but by convergence failures of the time stepping solver.

8. Conclusion and outlook

In this work we demonstrated the viability of higher order discontinuous Galerkin methods as an efficient means of time discretization for optimal control problems governed by parabolic PDEs. Two key questions we identified in this context were how to efficiently solve the resulting large equation systems for the time step equations and how to obtain fast convergence in spite of relatively low regularity of the solutions of optimal control problems when additional inequality constraints are present.

To address the first question, we proposed an approximation of the Newton update matrix for the time stepping equation based on ideas originally developed by Cooper and Butcher [26] for Runge-Kutta methods. This approximation results in a simplified update system where the temporal components of the solution decouple and therefore in a significant reduction of the computational cost per solver iteration. The implications of this approximate decoupling on the convergence of the time stepping solver were analyzed in detail. For linear problems with constant coefficients, we established fast linear convergence of the resulting scheme independent of the size of the time step. Numerical results showed that the proposed scheme compared favourably to other approaches in terms of computational costs.

For the second question of dealing with the regularity restrictions caused by control constraints, we presented two approaches. The first one was exploiting the improved regularity of the adjoint state over control and state to construct a discretization that achieves globally almost third order convergence with respect to the time step. It is based on a piecewise linear dG time discretization combined with variational treatment of the control variable and a post processing step using superconvergence properties of the discontinuous Galerkin method. For this scheme, we carried out a rigorous a priori error analysis for a linear-quadratic model problem resulting in error estimates of optimal order.

Our second approach to resolving irregularities caused by inequality constraints was to resolve them by adaptive *hp* refinement of the time discretization. We developed an *hp*-adaptive procedure for the time discretization, coupled with *h*-adaptivity for the spatial discretization and tested it on three example problems. We observed that the algorithm successfully detected non-smoothness caused by control constraints, but also non-smoothness due to irregular data and incompatible initial conditions. However, a more realistic test example also showed that in the presence of strong nonlinearities which limit the length of the time steps, gains from *hp* adaptivity can only be realized for very strict accuracy requirements. Nevertheless we argue that also in this case, the

use of dG schemes with order 1 or 2 can help to produce more accurate results than common time stepping schemes like implicit Euler or Crank-Nicholson.

Overall we demonstrated that a proper use of higher order discontinuous Galerkin time discretization schemes can help to obtain highly accurate solutions of optimization problems with parabolic PDEs while minimizing the number of degrees of freedom spent on the time discretization. Based on our results, we identify some promising questions left for further investigations.

Concerning the time stepping equation solver, a convergence analysis with respect to stronger norms would be desirable since this would help in showing mesh-independence of the maximal size of the time step for more general problems. A classical application domain for Radau methods, which, as we have seen are closely related to the discontinuous Galerkin approach, are differential algebraic problems. Therefore it might be promising to investigate under which conditions the proposed decoupling scheme can also be applied to partial differential algebraic equations (PDAEs). Furthermore, an application of the proposed approximation techniques to higher order continuous Galerkin methods should allow the derivation of efficient schemes also for this class of discretizations.

Possible generalizations of the a priori analysis include terminal observation and semilinear problems. Concerning *hp* adaptivity, as already pointed out in Section 7.4, it would be desirable to consider *hp* adaptivity in both, time and space such that the accuracy of the spatial discretization can match the temporal precision while minimizing the number of degrees of freedom also in space.

Acknowledgements

First of all, I would like to thank my supervisor Boris Vexler, not only for giving me the opportunity to work on this interesting topic and offering support whenever it was needed. He also provided me with ample possibilities to take part in the scientific community and to gain teaching experience.

I gratefully acknowledge the financial support provided by the Munich Center of Advanced Computing and the International Graduate School of Science and Engineering at the Technische Universität München during the first few years of my work and their support for conference participations, training courses, and international exchange.

Furthermore, I thank Thomas Wihler at the University of Bern for accommodating me at his institute and sharing his experience on *hp*-adaptivity and smoothness indicators. A particular thanks goes to Konstantin Pieper for countless fruitful discussions, highly productive collaboration on software development, proofreading, and for providing encouragement whenever it was needed. Besides, I would like to thank Dominik Meidner for always being ready to solve my computer problems or to help out with any other question. Last but by no means least I thank the colleagues at the chairs M1 and M17 for creating a positive and encouraging working atmosphere that made the last four years such a positive experience, and my wife and my family for their continued support.

A. Proof of Theorem 6.10

The proof is similar to Theorem 5 in [38, Chapter 7.1] where higher order stability estimates for a continuous parabolic problem are shown. Just as there, a Galerkin approximation with respect to the spatial variable is used. Let $\{v_n\}_{n\in\mathbb{N}}$ be an orthonormal basis of H consisting of eigenfunctions of $-\Delta$ defined on V. Then we define the spaces V_N and X_{kN}^r by

$$\begin{aligned} V_N &:= \operatorname{span}\{v_n \mid n \le N\}, \\ X_{kN}^r &:= \left\{v \in L^2(I, V_N) \;\middle|\; v|_{I_m} \in \mathcal{P}_r(I_m, V_N),\ m = 1, \dots, M\right\}. \end{aligned}$$

Replacing the test and trial spaces in Equation (6.19) by X_{kN}^r leads to a sequence of Galerkin approximations y_{kN} of the semidiscrete solution y_k. In a first step we have to show that for those approximations the stated stability estimates hold.

Lemma A.1. *For the Galerkin approximations y_{kN} as defined above we have the stability estimate*

$$\left\|\Delta^2 y_{kN}\right\|_I + \left(\sum_{m=1}^{M} \left\|\partial_t \Delta y_{kN}\right\|_{I_m}^2\right)^{\frac{1}{2}} \le C \left\|\Delta w\right\|_I$$

with a constant C independent of N.

Proof. To get the estimate for the first term we test with $\varphi = \Delta^3 y_{kN}$, which exists since y_{kN} is a linear combination of eigenvectors of Δ, resulting in

$$\begin{aligned} -\sum_{m=1}^{M} (\Delta^3 y_{kN}, \partial_t y_{kN})_{I_m} - (\Delta^3 y_{kN}, \Delta y_{kN})_I \\ - \sum_{m=1}^{M-1} (\Delta^3 y_{kN,m}^-, [y_{kN}]_m) - (\Delta^3 y_{kN,M}^-, y_{kN,M}^-) = (\Delta^3 y_{kN}, w)_I. \end{aligned}$$

We apply Green's formula to each term and get

$$\begin{aligned} \sum_{m=1}^{M} (\nabla\Delta y_{kN}, \partial_t \nabla\Delta y_{kN})_{I_m} - \left\|\Delta^2 y_{kN}\right\|_I^2 + \sum_{m=1}^{M-1} (\nabla\Delta y_{kN,m}^-, [\nabla\Delta y_{kN}]_m) + \left\|\nabla\Delta y_{kN,M}^-\right\|^2 \\ = (\Delta^2 y_{kN}, \Delta w)_I. \end{aligned}$$

With the two identities

$$(\nabla\Delta y_{kN}, \partial_t\nabla\Delta y_{kN})_{I_m} = \frac{1}{2}\left\|\nabla\Delta y^-_{kN,m}\right\|^2 - \frac{1}{2}\left\|\nabla\Delta y^+_{kN,m-1}\right\|^2$$

and

$$(\nabla\Delta y^-_{kN,m}, [\nabla\Delta y_{kN}]_m) = \frac{1}{2}\left\|\nabla\Delta y^+_{kN,m}\right\|^2 - \frac{1}{2}\left\|\nabla\Delta y^-_{kN,m}\right\|^2 - \frac{1}{2}\left\|[\nabla\Delta y_{kN}]_m\right\|^2$$

we obtain

$$-\left\|\Delta^2 y_{kN}\right\|_I^2 - \sum_{m=1}^{M-1}\frac{1}{2}\left\|[\nabla\Delta y_{kN}]_m\right\|^2 - \frac{1}{2}\left\|\nabla\Delta y^-_{kN,M}\right\|^2 = (\Delta^2 y_{kN}, \Delta w)_I,$$

which immediately gives the estimate for the first term

$$\left\|\Delta^2 y_{kN}\right\|_I \leq \left\|\Delta w\right\|_I.$$

In order to obtain the second estimate, we test with the interval-wise defined function φ where $\varphi\big|_{I_m} = (t-t_m)\partial_t\Delta^2 y_{kN}$ for a fixed index m and $\varphi = 0$ otherwise. Using the dual formulation (6.16) of the bilinear form, we note that the jump terms vanish and we get

$$-((t-t_m)\partial_t\Delta^2 y_{kN}, \partial_t y_{kN})_{I_m} - ((t-t_m)\partial_t\Delta^2 y_{kN}, \Delta y_{kN})_{I_m} = ((t-t_m)\partial_t\Delta^2 y_{kN}, w)_{I_m}.$$

We apply Green's formula with respect to the spatial variable on each of the three terms and obtain after reordering

$$\begin{aligned}\int_{I_m}(t_m-t)\left\|\partial_t\Delta y_{kN}\right\|^2\,\mathrm{d}t &= \int_{I_m}(t_m-t)(\partial_t\Delta y_{kN}, -\Delta w - \Delta^2 y_{kN})\,\mathrm{d}t \\ &\leq \left(\int_{I_m}(t_m-t)\left\|\partial_t\Delta y_{kN}\right\|^2\,\mathrm{d}t\right)^{\frac{1}{2}}\left(\int_{I_m}(t_m-t)\left\|-\Delta w - \Delta^2 y_{kN}\right\|^2\,\mathrm{d}t\right)^{\frac{1}{2}}.\end{aligned}$$

Together with the inverse estimate (4.5) from [80], which reads in our case

$$\left\|y_{kN}\right\|_{I_m}^2 \leq Ck_m^{-1}\int_{I_m}(t_m-t)\left\|y_{kN}\right\|^2\,\mathrm{d}t$$

with C independent of N, we obtain the estimate

$$\begin{aligned}\left\|\partial_t\Delta y_{kN}\right\|_{I_m}^2 &\leq Ck_m^{-1}\int_{I_m}(t_m-t)\left\|\partial_t\Delta y_{kN}\right\|^2\,\mathrm{d}t \\ &\leq Ck_m^{-1}\int_{I_m}(t_m-t)\left\|-\Delta w - \Delta^2 y_{kN}\right\|^2\,\mathrm{d}t \\ &\leq C\left\|-\Delta w - \Delta^2 y_{kN}\right\|_{I_m}^2 \leq C\left(\left\|\Delta w\right\|_{I_m}^2 + \left\|\Delta^2 y_{kN}\right\|_{I_m}^2\right).\end{aligned}$$

Summing over all time intervals yields

$$\sum_{m=1}^{M}\left\|\partial_t\Delta y_{kN}\right\|_{I_m}^2 \leq C\left(\left\|\Delta w\right\|_I^2 + \left\|\Delta^2 y_{kN}\right\|_I^2\right)$$

which shows the second estimate. □

Proof of Theorem 6.10. From Lemma A.1 we have

$$\left\|\Delta^2 y_{kN}\right\|_I + \left(\sum_{m=1}^{M} \|\partial_t \Delta y_{kN}\|_{I_m}^2\right)^{\frac{1}{2}} \leq C \|\Delta w\|_I$$

with C independent of N. Therefore the sequence $\{y_{kN}\}_{n\in\mathbb{N}}$ is bounded with respect to the norm $\|\cdot\|_Y$ given by

$$\|y_k\|_Y^2 = \|y_k\|_I^2 + \left\|\Delta^2 y_k\right\|_I^2 + \sum_{m=1}^{M} \|\partial_t \Delta y_k\|_{I_m}^2$$

and there exists a sub-sequence $\left(y_{kN_j}\right)_{j\in\mathbb{N}}$ that converges weakly with respect to the Y norm to a limit $\tilde{y}_k$ which satisfies the estimate

$$\left\|\Delta^2 \tilde{y}_k\right\|_I + \|\partial_t \Delta \tilde{y}_k\|_I \leq C \|\Delta w\|_I .$$

To complete the proof, we need to show that $\tilde{y}_k$ is in fact the solution y_k of the semidiscrete problem (6.19). Therefore we note that the stability estimate in Corollary 6.7 also works for the Galerkin approximations y_{kN} with the constant C independent of N. We fix $\bar{N}$, then for any $\varphi \in X_{k\bar{N}}^r$ and for any $N_j \geq \bar{N}$ the identity

$$-\sum_{m=1}^{M} (\varphi, \partial_t y_{kN_j})_{I_m} - (\varphi, \Delta y_{kN_j})_I - \sum_{m=1}^{M} (\varphi^-, [y_{kN_j}]_m) = (\varphi, w)_I \tag{A.1}$$

holds true. Since $\sum_{m=1}^{M} \left\|\partial_t y_{kN_j}\right\|_I^2$, $\left\|\Delta y_{kN_j}\right\|_I$ and $\sum_{m=1}^{M} \left\|[y_{kN_j}]_m\right\|^2$ are bounded by the stability estimate we can extract a subsequence such that (A.1) holds for the weak limit which has to be $\tilde{y}_k$ again. Passing to the limit $\bar{N} \to \infty$ shows that in fact $\tilde{y}_k = y_k$. □

List of Tables

List of Figures

List of Algorithms

Bibliography

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover Publications Inc., New York, 1992.

[2] S. Adjerid, K. D. Devine, J. E. Flaherty, and L. Krivodonova. A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 191(11-12):pp. 1097–1112, 2002.

[3] G. Akrivis, C. Makridakis, and R. H. Nochetto. Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numer. Math.*, 118(3):pp. 429–456, 2011.

[4] T. Apel. *Anisotropic finite elements: Local estimates and applications.* Advances in Numerical Mathematics. Teubner, 1999.

[5] T. Apel and T. G. Flaig. Crank–Nicolson schemes for optimal control problems with evolution equations. *SIAM J. Numer. Anal.*, 50(3):pp. 1484–1512, 2012.

[6] J. Appell and P. P. Zabrejko. *Nonlinear superposition operators.* Cambridge University Press, 1990.

[7] O. Axelsson. A class of A-stable methods. *Nordisk Tidskr. Informationsbehandling (BIT)*, 9:pp. 185–199, 1969.

[8] O. Axelsson. On the efficiency of a class of A-stable methods. *Nordisk Tidskr. Informationsbehandling (BIT)*, 14:pp. 279–287, 1974.

[9] O. Axelsson. High-order methods for parabolic problems. *J. Comput. Appl. Math.*, 1:pp. 5–16, 1975.

[10] S. Basting and S. Weller. Efficient preconditioning of variational time discretization methods for parabolic partial differential equations. *ESAIM Math. Model. Numer. Anal.*, 49(2):pp. 311–347, 2015.

[11] R. Becker and M. Braack. Multigrid techniques for finite elements on locally refined meshes. *Numer. Linear Algebra Appl.*, 7(6):pp. 363–379, 2000.

[12] R. Becker, D. Meidner, and B. Vexler. Efficient numerical solution of parabolic optimization problems by finite element methods. *Optim. Methods Softw.*, 22(5):pp. 813–833, 2007.

[13] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numer.*, 10:pp. 1–102, 2001.

[14] R. Becker and B. Vexler. Optimal control of the convection-diffusion equation using stabilized finite element methods. *Numer. Math.*, 106(3):pp. 349–367, 2007.

[15] M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. *SIAM J. Control Optim.*, 37(4):pp. 1176–1194, 1999.

[16] S. Beuchler, K. Hofer, D. Wachsmuth, and J.-E. Wurst. Boundary concentrated finite elements for optimal control problems with distributed observation. *Comput. Optim. Appl.* doi:10.1007/s10589-015-9737-5, 2015.

[17] T. A. Bickart. An efficient solution process for implicit Runge-Kutta methods. *SIAM J. Numer. Anal.*, 14:pp. 1022–1027, 1977.

[18] M. Braack and A. Ern. A posteriori control of modeling errors and discretization errors. *Multiscale Model. Simul.*, 1(2):pp. 221–238, 2003.

[19] D. Braess. *Finite Elemente*. Springer, Berlin, fourth edition, 2007.

[20] C. Brezinski and J. Van Iseghem. A taste of Padé approximation. *Acta Numer.*, 4:pp. 53–103, 1995.

[21] J. Butcher. On the implementation of implicit Runge-Kutta methods. *Nordisk Tidskr. Informationsbehandling (BIT)*, 16:pp. 237–240, 1976.

[22] M. Calvo, S. González-Pinto, and J. Montijano. On the iterative solution of the algebraic equations in fully implicit Runge–Kutta methods. *Numer. Algorithms*, 23(1):pp. 97–113, 2000.

[23] M. Chipot. *Elements of nonlinear analysis.* Birkhäuser, Basel, 2000.

[24] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 2002.

[25] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. The development of discontinuous Galerkin methods. In *Discontinuous Galerkin methods. Theory, computation and applications. 1st international symposium on DGM, Newport, RI, USA, May 24–26, 1999*, pp. 3–50. Springer, Berlin, 2000.

[26] G. Cooper and J. Butcher. An iteration scheme for implicit Runge-Kutta methods. *IMA J. Numer. Anal.*, 3:pp. 127–140, 1983.

[27] R. Dautray and J.-L. Lions. *Evolution Problems I*, volume 5 of *Mathematical Analysis and Numerical Methods for Science and Technology*. Springer-Verlag, Berlin, 1992.

[28] P. Deuflhard. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*. Springer series in computational mathematics. Springer Verlag, Berlin, 2004.

[29] J. Dieudonné. *Foundations of modern analysis*, volume 10-I of *Pure and Applied Mathematics*. Academic Press, New York-London, 1969.

[30] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):pp. 1106–1124, 1996.

[31] B. L. Ehle. On Páde approximations to the exponential function and A-stable methods for the numerical solution of initial value problems. *Technical report*, University of Waterloo, 1969.

[32] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. I. A linear model problem. *SIAM J. Numer. Anal.*, 28(1):pp. 43–77, 1991.

[33] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. II. Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$. *SIAM J. Numer. Anal.*, 32(3):pp. 706–740, 1995.

[34] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. IV. Nonlinear problems. *SIAM J. Numer. Anal.*, 32(6):pp. 1729–1749, 1995.

[35] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems. V. Long-time integration. *SIAM J. Numer. Anal.*, 32(6):pp. 1750–1763, 1995.

[36] K. Eriksson, C. Johnson, and S. Larsson. Adaptive finite element methods for parabolic problems. VI: Analytic semigroups. *SIAM J. Numer. Anal.*, 35(4):pp. 1315–1325, 1998.

[37] K. Eriksson, C. Johnson, and V. Thomée. Time discretization of parabolic problems by the discontinuous Galerkin method. *RAIRO Modél. Math. Anal. Numér.*, 19(4):pp. 611–643, 1985.

[38] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Series in Mathematics*. AMS, Providence, second edition, 2010.

[39] T. Fankhauser, T. P. Wihler, and M. Wirz. The *hp*-adaptive FEM based on continuous Sobolev embeddings: isotropic refinements. *Comput. Math. Appl.*, 67(4):pp. 854–868, 2014.

[40] K. J. Fidkowski. Output error estimation strategies for discontinuous Galerkin discretizations of unsteady convection-dominated flows. *Internat. J. Numer. Methods Engrg.*, 88(12):pp. 1297–1322, 2011.

[41] T. Flaig, D. Meidner, and B. Vexler. Petrov-Galerkin Crank-Nicolson Scheme for Parabolic Optimal Control Problems on Nonsmooth Domains. In G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich, editors, *Trends in PDE Constrained Optimization*,

volume 165 of *ISNM, Int. Ser. Numer. Math.*, pp. 421–435. Springer International Publishing, 2014.

[42] P. D. Frank and G. R. Shubin. A comparison of optimization-based approaches for a model computational aerodynamics design problem. *J. Comput. Phys.*, 98(1):pp. 74–89, 1992.

[43] W. Gander and W. Gautschi. Adaptive quadrature – Revisited. *BIT*, 40(1):pp. 84–101, 2000.

[44] Gascoigne. The finite element toolkit. URL `http://www.gascoigne.de`.

[45] S. González-Pinto, C. González-Concepción, and J. Montijano. Iterative schemes for Gauss methods. *Comput. Math. Appl.*, 27(7):pp. 67–81, 1994.

[46] S. González-Pinto, J. Montijano, and L. Rández. Iterative schemes for three-stage implicit Runge-Kutta methods. *Appl. Numer. Math.*, 17(4):pp. 363–382, 1995.

[47] S. González-Pinto, S. Pérez-Rodríguez, and J. Montijano. Implementation of high-order implicit Runge-Kutta methods. *Comput. Math. Appl.*, 41(7-8):pp. 1009–1024, 2001.

[48] C. Gräser and R. Kornhuber. Nonsmooth Newton methods for set-valued saddle point problems. *SIAM J. Numer. Anal.*, 47(2):pp. 1251–1273, 2009.

[49] P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[50] K. Gustafsson and G. Söderlind. Control strategies for the iterative solution of nonlinear equations in ODE solvers. *SIAM J. Sci. Comput.*, 18(1):pp. 23–40, 1997.

[51] E. Hairer and G. Wanner. Stiff differential equations solved by Radau methods. *J. Comput. Appl. Math.*, 111:pp. 93–111, 1999.

[52] E. Hairer and G. Wanner. *Solving ordinary differential equations II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2010.

[53] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.*, 49:pp. 409–436, 1952.

[54] V. Heuveline and R. Rannacher. Duality-based adaptivity in the *hp*-finite element method. *J. Numer. Math.*, 11(2):pp. 95–113, 2003.

[55] M. Hinze. A variational discretization concept in control constrained optimization: The linear-quadratic case. *Comput. Optim. Appl.*, 30(1):pp. 45–61, 2005.

[56] M. Hinze and K. Kunisch. Second order methods for optimal control of time-dependent fluid flow. *SIAM J. Control Optim.*, 40(3):pp. 925–946, 2001.

[57] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints.* Dordrecht: Springer, 2009.

[58] M. Hinze and M. Vierling. The semi-smooth Newton method for variationally discretized control constrained elliptic optimal control problems; implementation, convergence and globalization. *Optim. Methods Softw.*, 27(6):pp. 933–950, 2012.

[59] K. Ito and K. Kunisch. *Lagrange multiplier approach to variational problems and applications.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[60] P. Jamet. Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain. *SIAM J. Numer. Anal.*, 15:pp. 912–928, 1978.

[61] B. Janssen and T. P. Wihler. Existence Results for the Continuous and Discontinuous Galerkin Time Stepping Methods for Nonlinear Initial Value Problems, 2014. Submitted.

[62] K. Kunisch, K. Pieper, and A. Rund. Time-optimal control for the monodomain equations – a monolithic approach, 2014. Submitted.

[63] K. Kunisch and A. Rösch. Primal-dual active set strategy for a general class of constrained optimal control problems. *SIAM J. Optim.*, 13(2):pp. 321–334, 2002.

[64] K. Kunisch and A. Rund. Time optimal control of the monodomain model in cardiac electrophysiology, 2014. Accepted, IMA J. Appl. Math.

[65] J. Lang. *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems*, volume 16 of *Lecture Notes in Computational Science and Engineering.* Springer-Verlag, 2001.

[66] I. Lasiecka and K. Malanowski. On discrete-time Ritz-Galerkin approximation of control constrained optimal control problems for parabolic systems. *Control Cybernet.*, 7(1):pp. 21–36, 1978.

[67] P. Lesaint and P.-A. Raviart. On a finite element method for solving the neutron transport equation. In *Mathematical aspects of finite elements in partial differential equations*, pp. 89–123. Publication No. 33. Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974.

[68] R. Li, W. Liu, H. Ma, and T. Tang. Adaptive finite element approximation for distributed elliptic optimal control problems. *SIAM J. Control Optim.*, 41(5):pp. 1321–1349, 2002.

[69] J. L. Lions. *Optimal Control Of Systems Governed By Partial Differential Equations.* Springer Verlag, 1971.

[70] W. Liu and N. Yan. A posteriori error estimates for distributed convex optimal control problems. *Adv. Comput. Math.*, 15(1-4):pp. 285–309, 2001.

[71] C. Makridakis and R. H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numer. Math.*, 104:pp. 489–514, 2006.

[72] K. Malanowski. Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems. *Appl. Math. Optimization*, 8:pp. 69–95, 1982.

[73] G. Matthies and F. Schieweck. Higher order variational time discretizations for nonlinear systems of ordinary differential equations, 2011. Preprint.

[74] C. Mavriplis. Adaptive mesh strategies for the spectral element method. *Comput. Methods Appl. Mech. Engrg.*, 116(1-4):pp. 77–86, 1994.

[75] V. Maz'ya and J. Rossmann. *Elliptic Equations in Polyhedral Domains.*, volume 162 of *Math. Surv. Monogr.* AMS, 2010.

[76] D. Meidner. *Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems.* Ph.D. thesis, Universität Heidelberg, 2008.

[77] D. Meidner and T. Richter. Goal-oriented error estimation for the fractional step theta scheme. *Comput. Methods Appl. Math.*, 14(2):pp. 203–230, 2014.

[78] D. Meidner and B. Vexler. Adaptive space-time finite element methods for parabolic optimization problems. *SIAM J. Control Optim.*, 46(1):pp. 116–142, 2007.

[79] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems. II: Problems with control constraints. *SIAM J. Control Optim.*, 47(3):pp. 1301–1329, 2008.

[80] D. Meidner and B. Vexler. A priori error estimates for space-time finite element discretization of parabolic optimal control problems part I: problems without control constraints. *SIAM J. Control Optim.*, 47(3):pp. 1150–1177, 2008.

[81] D. Meidner and B. Vexler. A Priori Error Analysis of the Petrov–Galerkin Crank–Nicolson Scheme for Parabolic Optimal Control Problems. *SIAM J. Control Optim.*, 49(5):pp. 2183–2211, 2011.

[82] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems. *SIAM J. Control Optim.*, 43(3):pp. 970–985, 2004.

[83] H. H. Michels. Abscissas and weight coefficients for Lobatto quadrature. *Math. Comp.*, 17:pp. 237–244, 1963.

[84] W. F. Mitchell and M. A. McClain. A Comparison of hp-Adaptive Strategies for Elliptic Partial Differential Equations. *ACM Trans. Math. Softw.*, 41(1):pp. 2:1–2:39, October 2014.

[85] I. Neitzel and B. Vexler. A priori error estimates for space-time finite element discretization of semilinear parabolic optimal control problems. *Numer. Math.*, 120(2):pp. 345–386, 2012.

[86] O. Nevanlinna. Matrix valued versions of a result of von Neumann with an application to time discretization. *J. Comput. Appl. Math.*, 12/13:pp. 475–489, 1985.

[87] S. Perez-Rodriguez, S. Gonzalez-Pinto, and B. Sommeijer. An iterated Radau method for time-dependent PDEs. *J. Comput. Appl. Math.*, 231(1):pp. 49–66, 2009.

[88] O. Perron. *Die Lehre von den Kettenbrüchen. Band II. Analytisch-funktionentheoretische Kettenbrüche.* Teubner, Stuttgart, third edition, 1957.

[89] K. Pieper. *Finite element discretization for elliptic and parabolic sparse control problems.* Ph.D. thesis, Technische Universität München, 2015.

[90] T. Richter. *Parallel Multigrid Method for Adaptive Finite Elements with Application to 3D Flow Problems.* Ph.D. thesis, Universität Heidelberg, 2005.

[91] T. Richter, A. Springer, and B. Vexler. Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems. *Numer. Math.*, 124(1):pp. 151–182, 2013.

[92] S. M. Robinson. Normal maps induced by linear transformations. *Math. Oper. Res.*, 17(3):pp. 691–714, 1992.

[93] RoDoBo. A C++ library for optimization with stationary and nonstationary PDEs with interface to Gascoigne [44]. URL `http://www.rodobo.org`.

[94] A. Rösch. Error estimates for parabolic optimal control problems with control constraints. *Z. Anal. Anwend.*, 23(2):pp. 353–376, 2004.

[95] A. Rösch and R. Simon. Superconvergence properties for optimal control problems discretized by piecewise linear and discontinuous functions. *Numer. Funct. Anal. Optim.*, 28(3-4):pp. 425–443, 2007.

[96] A. Rösch and B. Vexler. Optimal control of the Stokes equations: a priori error analysis for finite element discretization with postprocessing. *SIAM J. Numer. Anal.*, 44(5):pp. 1903–1920, 2006.

[97] A. Schiela. A simplified approach to semismooth Newton methods in function space. *SIAM J. Optim.*, 19(3):pp. 1417–1432, 2008.

[98] M. Schmich. *Adaptive Finite Element Methods for Computing Nonstationary Incompressible Flows.* Ph.D. thesis, Ruprecht-Karls-Universität Heidelberg, 2009.

[99] M. Schmich and B. Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM J. Sci. Comput.*, 30(1):pp. 369–393, 2008.

[100] D. Schötzau. *hp-DGFEM for Parabolic Evolution Problems.* Ph.D. thesis, Swiss Federal Institute of Technology, Zürich, 1999.

[101] D. Schötzau and C. Schwab. Time discretization of parabolic problems by the *hp*-version of the discontinuous Galerkin finite element method. *SIAM J. Numer. Anal.*, 38(3):pp. 837–875, 2000.

[102] D. Schötzau and T. Wihler. A posteriori error estimation for hp-version time-stepping methods for parabolic partial differential equations. *Numer. Math.*, 115:pp. 475–509, 2010.

[103] P. Šolín and L. Demkowicz. Goal-oriented *hp*-adaptivity for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 193(6-8):pp. 449–468, 2004.

[104] A. Springer and B. Vexler. Third order convergent time discretization for parabolic optimal control problems with control constraints. *Comput. Optim. Appl.*, 57(1):pp. 205–240, 2014.

[105] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numer. Anal.*, 20:pp. 626–637, 1983.

[106] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems.* Springer series in computational mathematics. Springer-Verlag, Berlin, second edition, 2006.

[107] F. Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen.* Vieweg+Teubner, Wiesbaden, second edition, 2009.

[108] M. Ulbrich. Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces. Habilitation thesis, 2002. Technische Universtät München.

[109] B. Vexler and W. Wollner. Adaptive finite elements for elliptic optimization problems with control constraints. *SIAM J. Control Optim.*, 47(1):pp. 509–534, 2008.

[110] D. Wachsmuth and J.-E. Wurst. A short step method designed for solving linear quadratic optimal control problems with *hp* finite elements, 2014. Preprint 323, Institut für Mathematik, Universität Würzburg.

[111] D. Wachsmuth and J.-E. Wurst. Optimal control of interface problems with hp-finite elements, 2014. Preprint 326, Institut für Mathematik, Universität Würzburg.

[112] G. Wachsmuth. Differentiability of implicit functions: beyond the implicit function theorem. *J. Math. Anal. Appl.*, 414(1):pp. 259–272, 2014.

[113] G. Wanner, E. Hairer, and S. P. Nørsett. Order stars and stability theorems. *Nordisk Tidskr. Informationsbehandling (BIT)*, 18:pp. 475–489, 1978.

[114] J. Weidendorfer, M. Kowarschik, and C. Trinitis. A Tool Suite for Simulation Based Analysis of Memory Access Behavior. In M. Bubak, G. van

Albada, P. Sloot, and J. Dongarra, editors, *Computational Science - ICCS 2004*, volume 3038 of *Lecture Notes in Computer Science*, pp. 440–447. Springer Berlin Heidelberg, 2004.

[115] T. Werder, K. Gerdes, D. Schötzau, and C. Schwab. *hp*-discontinuous Galerkin time stepping for parabolic problems. *Comput. Methods Appl. Mech. Engrg.*, 190(49-50):pp. 6685–6708, 2001.

[116] T. P. Wihler. An *hp*-adaptive strategy based on continuous Sobolev embeddings. *J. Comput. Appl. Math.*, 235(8):pp. 2731–2739, 2011.

[117] J. Wloka. *Partial differential equations*. Cambridge University Press, Cambridge, 1987.